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]

**Collaborateurs extérieurs**

François Barthélemy [MC, CNAM]

**Doctorants**

Vitor Rocio [Thèse en co-tutelle avec l'Université Nouvelle de Lisbonne]

François Role [Fonctionnaire au DISTNB-MESR, Université d'Orléans]

**Stagiaires**

Linda Kaouane [mars–septembre 2000, Université d'Orléans]

## 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**, dans ses aspects syntaxiques, voire sémantiques. Ce domaine de recherche est en effet 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 (petite) 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 a considérablement crû au cours de cette année.

# 3 Fondements scientifiques

## 3.1 Formalismes grammaticaux

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

**Participants** : Pierre Boullier, Éric Villemonte de la Clergerie.

**Résumé :** *Ce thème concerne l'analyse syntaxique de 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*
**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 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) [MK96] et *Head-Driven Phrasal Structure Grammars* (HPSG) [PS94].

**Grammaires stochastiques** Pratiquement, toute grammaire peut être décorée avec des probabilités ou des pondérations, afin de mieux coïncider avec l'usage de la langue rencontrée sur un corpus de textes. Ces probabilités peuvent être vues comme des décorations prises dans un demi-anneau, dont les propriétés algébriques sont exploitables au cours de l'analyse syntaxique [Ten97].

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 5.2).

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.

[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.

[MK96] J. T. Maxwell, R. M. Kaplan, «An efficient parser for LFG», *in: Proc. of 1st LFG Conference*, Grenoble, 1996.

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

[Ten97] F. Tendeau, *Analyse syntaxique et sémantique avec évaluation d'attributs dans un demi-anneau – Application à la linguistique calculatoire*, thèse de doctorat, Université d'Orléans, 1997.

### 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 à la fois 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.

### 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 [4]. 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 [3].

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, telles les transformations Magic-Set [Ram88].

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

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

[Ram88] R. RAMAKRISHNAN, « Magic Templates: A Spellbinding Approach to Logic Programs », *in : Proc. of the 5th Int. Conf. and Symp. on Logic Programming*, p. 140–159, 1988.

### 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 possibles ou l'ensemble des dérivations possibles pour une phrase et sont en général assimilables à des grammaires ou à des programmes logiques [2]. 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 5.3), constituent de surcroît un point de départ pour des traitements linguistiques ultérieurs (prise en compte de contraintes syntaxiques ou sémantiques complémentaires, traduction, ...).

## 3.2 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 Logiciels

## 4.1 Logiciel SYNTAX

**Participants** : Pierre Boullier, Philippe Deschamp.

La version 3.9 de SYNTAX a été réalisée. 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[1]. Ce constructeur est utilisé dans le module de LECL qui traite les mots-clés et pour la fabrication de dictionnaires. Nous avons par exemple construit pour l'anglais un dictionnaire d'environ 320 000 entrées qui associe à chaque forme fléchie l'ensemble des arbres élémentaires d'une TAG dans lesquels elle peut apparaître ;
- un prototype pour le traitement des RCG (voir la section 5.2).

En outre, suite à la présentation[2] effectuée par Robert Sedgewick au Colloquium de Rocquencourt sous l'égide du projet ALGO, nous avons décidé d'évaluer les procédures de tri utilisées par les constructeurs de SYNTAX. Il en a résulté une implémentation du tri *Multi-Key QuickSort* qui conserve les propriétés d'efficacité théorique (complexité) du modèle et qui est en pratique plus efficace que la version antérieure et que les versions système des diverses plates-formes que nous avons testées.

Cette version 3.9 existe pour les environnements Linux, SunOs, Solaris, DOS et Windows. Pour ce dernier, il est à noter que dorénavant SYNTAX utilise directement les bibliothèques dynamiques standard du système d'exploitation.

À sa demande, nous avons fourni à une grande entreprise internationale, leader mondial sur le marché du traitement automatique de la langue, une version spécifique de SYNTAX aux fins d'évaluation. L'analyseur produit est non-déterministe mais retourne au plus un seul arbre d'analyse en cas d'ambiguïté. Des comparaisons de vitesse ont été faites avec ses propres outils, développés en interne ou acquis par croissance externe, à partir d'une grammaire d'interrogation d'un sous-domaine en langue naturelle. Le résultat a été une victoire écrasante pour SYNTAX — un à plusieurs ordres de grandeur. Des problèmes de stratégie et de politique internes de cette entreprise ont toutefois fait abandonner les collaborations envisagées !

## 4.2 Logiciel DYALOG

**Participant** : Éric Villemonte de la Clergerie.

Le logiciel DYALOG est un compilateur de grammaires et de programmes logiques produisant des exécutables tabulaires. Il est plus spécifiquement dédié à la construction d'analyseurs syntaxiques pour le traitement de la langue naturelle mais est aussi utile pour remplacer des

1. Ces techniques sont appliquées pour représenter sous forme compacte les automates à états finis utilisés. On peut noter que cette optimisation en place ne s'effectue pas au détriment de l'efficacité : on assure qu'un mot de $n$ caractères est reconnu (ou rejeté) en au plus $n$ comparaisons.
2. Voir `http://www.inria.fr/actualites/colloques/COLLOQUIUM200516-fra.html`

systèmes PROLOG traditionnels dans le cadre d'applications très ambiguës avec potentiellement du partage de calculs.

Les sources ou une distribution RPM de la version courante de DyALog (1.6) sont disponibles pour les plates-formes Linux (Pentium) et SunOS (Sparc) sous FTP à `http://atoll.inria.fr/~clerger`.

La version actuelle permet le traitement des DCG (*Definite Clause Grammars*) et des FTAG (*Feature Tree Adjoining Grammars*). Elle offre la possibilité d'utiliser des structures typées de traits et des domaines finis pour des écritures plus compactes des grammaires. Il est également possible d'interfacer DyALog avec du code C.

# 5 Résultats nouveaux

## 5.1 Atelier TAG

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

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

**Glossaire :**
**TAG** *Tree Adjoining Grammars*
**XML** *eXtensible Markup Language*
**DTD** *Document Type Definition*

**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 travail théorique sur l'analyse syntaxique des Grammaires d'Arbres Adjoints (TAG) s'est traduit par le développement d'analyseurs syntaxiques et plus généralement par le développement d'un environnement de travail pour les TAG.

Le premier problème rencontré est celui de la représentation de grammaires TAG. Le format XTAG [DEH+94] `http://www.cis.upenn.edu/~xtag/` est un standard de facto dans la communauté TAG mais a le désavantage de ne pas être très lisible et de ne pas être clairement défini. Nous avons ainsi plusieurs variantes d'encodage des TAG en format XTAG. Suivant d'autres travaux, nous avons opté pour une représentation en format XML qui suit l'architecture XTAG.

En relation avec le groupe de travail TAGML, nous avons ainsi défini une DTD `http://atoll.inria.fr/~clerger/tag.dtd,xml` pour la représentation des grammaires TAG.

F. Barthélemy a réalisé le travail de conversion des grammaires au format XTAG vers le format XML. À ce jour, nous disposons en format XML, d'une petite grammaire du français (50 schémas d'arbres) et d'une grammaire moyenne de l'anglais (500 schémas d'arbres). Des

[DEH+94] C. Doran, D. Egedi, B. A. Hockey, B. Srinivas, M. Zaidel, « XTAG System — A Wide Coverage Grammar for English», *in : Proc. of the 15th International Conference on Computational Linguistics (COLING'94)*, p. 922–928, Kyoto, Japan, August 1994.

travaux sont en cours pour deux autres grammaires, dont une grammaire à grande couverture pour le français (5000 schémas d'arbres) développée par le groupe TALANA (Paris 7).

D'autre part, nous avons également développé un ensemble de modules Perl dits de «maintenance», s'appuyant fortement sur la DTD. Ces modules servent pour des tâches de conversion (vers le format RCG, vers le format «programmation logique» pour DyALog, vers un format SQL de base de données, ...), pour des tâches de vérification des grammaires, ....

Pour la petite grammaire du français, nous disposons actuellement de nombreux analyseurs syntaxiques qui nous permettent de tester et comparer différentes approches, telles les RCG (Section 5.2), tel DyALog en tabulation forte ou faible et stratégie descendante ou hybride, et tel DyALog guidé par les RCG (Section 5.3). Pour faciliter l'accès à ces différents analyseurs, nous avons mis en place un serveur d'analyseurs syntaxiques[3]. Ce serveur, après sélection d'un analyseur et envoi d'une phrase, retourne la forêt (partagée) des dérivations possibles. Cette forêt peut être retournée dans divers formats, dont un format HTML et un format XML s'appuyant sur une DTD `http://atoll.inria.fr/~clerger/forest.dtd,xml`.

Le format HTML pour les forêts de dérivations est exploité par une interface WEB d'accès au serveur d'analyseurs `http://medoc.inria.fr/pub-cgi-bin/parser.cgi`. Le format XML est exploité dans le cadre d'une interface en Java, développée par L. Kaouane, qui appelle elle aussi le serveur d'analyseurs. Cette interface permet de naviguer dans les dérivations (représentées sous forme arborescente), ce qui se révèle utile pour vérifier et comprendre les analyses d'une phrase. Nous pensons également que cette interface est un outil pédagogique intéressant pour découvrir et comprendre les TAG. L'interface en Java permet aussi de naviguer dans les divers composants d'une grammaire et en particulier d'en afficher les arbres.

Un document préliminaire [21], non encore publié mais présenté au groupe de travail TAGML, décrit l'ensemble de ces activités et complète le mémoire de DEA de L. Kaouane [22].

## 5.2 Analyse contextuelle

**Participant** : Pierre Boullier.

**Mots clés** : formalismes grammaticaux contextuels, forêts partagées, 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 leurs 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

3. accessible par `telnet medoc.inria.fr 8999`

puissantes, elles englobent les grammaires non-contextuelles (CFG) et les formalismes faiblement dépendant du contexte (*mildly context-sensitive*—MCS). Elles permettent même de décrire des phénomènes linguistiques dont certains nécessitaient auparavant des grammaires indexées[4] ainsi que 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, 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 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 la quantité d'information contenue dans un arbre élémentaire est majoritairement bien plus grande que celle contenue dans une production non-contextuelle, mais également que les 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 $\mathcal{O}(n^6)$, 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 commencé à réaliser des analyseurs RCG pour des grammaires à large couverture de l'anglais et du français. Le but de cette réalisation est non seulement de montrer que cela est possible avec la technologie RCG, mais aussi de montrer que les analyseurs obtenus ont des performances égales, voire supérieures, aux analyseurs dédiés originaux.

### 5.2.1 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 trois publications internationales.

Une première publication, présentée à IWPT [18], définit les RCG, expose quelques propriétés fondamentales et présente le principe d'un algorithme d'analyse. On y trouve également les justifications théoriques qui font des RCG un (le seul ?) formalisme modulaire dont les phrases s'analysent en temps polynomial. D'après notre définition, la modularité nécessite non seulement des notions de réutilisabilité mais également des propriétés de fermeture. Par exemple, les RCG sont modulaires vis-à-vis de l'intersection car les RCG sont closes par intersection : si $G_1$ et $G_2$ sont des RCG définissant les langages $L_1$ et $L_2$, le langage $L = L_1 \cap L_2$ peut être défini par une RCG $G$ qui peut se construire (de façon très simple) à partir de $G_1$ et $G_2$ sans

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

toucher ni à $G_1$ ni à $G_2$. Ces modules (grammaires) peuvent donc être rassemblés dans des librairies mises à la disposition des linguistes.

Une deuxième publication dans la revue *Grammars* [5], qui se fonde sur les travaux exposés l'an dernier au cours de la $6^{\text{ème}}$ conférence sur les mathématiques de la langue (*Sixth Meeting on Mathematics of Language (MOL6)*), étudie plus particulièrement la sous-classe des RCG qui ne contient que des prédicats unaires, les 1-RCG. Cette classe, déjà très puissante (elle peut définir les langages non-contextuels, leurs intersections et leur complémentaire) a la particularité de pouvoir s'analyser, comme les CFG, en temps cubique. Nous envisageons d'ailleurs d'exploiter cette propriété pour améliorer les analyseurs RCG en les guidant par des analyseurs 1-RCG, dans l'esprit de ce que nous réalisé pour les TAG avec les RCG et DyALog (Section 5.3) [16].

Il est bien connu que les formalismes grammaticaux ne savent pas très bien compter ! Par exemple les CFG ne savent compter que jusqu'à deux : elles savent reconnaître les parenthèses ouvrantes et fermantes de structures bien parenthésées. Nous avons étudié quelques-unes des possibilités des RCG dans ce domaine et nous avons présenté ces résultats, sur invitation, au $2^{nd}$ *AMAST*[5] Workshop on Language Processing [17].

Nous y montrons que, dans ce domaine aussi, les RCG ont un intérêt et peuvent décrire des propriétés que même les grammaires indexées sont incapables de définir. Par exemple, le langage qui décrit l'ensemble des chaînes de $a$ dont les longueurs sont les nombres premiers (tous les nombres premiers et uniquement ceux-là) est un RCL et l'analyseur correspondant permet, en temps $\mathcal{O}(n \log n)$, que $n$ soit premier ou non, de décider de l'appartenance de la chaîne $a^n$ à ce langage.

### 5.2.2 Utilisation des RCG pour l'anglais et le français

En 1998, nous avions montré qu'on pouvait traduire une TAG en une RCG équivalente et l'an dernier nous avons généralisé cet algorithme pour lui faire accepter en entrée des TAG et des contraintes d'adjonction quelconques tout en assurant que les phrases reconnues par la RCG produite s'analysent au pire en temps $\mathcal{O}(n^6)$.

D'autre part, depuis une bonne dizaine d'années, des équipes prestigieuses développent des grammaires à large couverture pour l'anglais (projet XTAG, à l'Université de Pennsylvanie à Philadelphie [DEH+94]) et le français (projet FTAG, au TALANA à l'Université Paris 7). Ces deux descriptions sont réalisées en utilisant le formalisme des TAG. Puisque l'accès à ces grammaires était possible[6], nous avons décidé qu'elles serviraient de base pour nos expérimentations en vraie grandeur dans le cadre du laboratoire d'expérimentation de grammaires et d'analyseurs que nous sommes en train de concevoir et de réaliser (voir la section 5.1). Les premiers résultats pratiques sur la pertinence de l'utilisation des RCG pour analyser des TAG ont été présentés dans le cadre plus large de [16], où une mini-grammaire (50 arbres) du français à servi de support à nos expérimentations. Après ces résultats très encourageants, nous

5. Algebraic Methodology and Software Technology.

6. La grammaire de l'anglais est accessible sur le web et celle du français nous a été concédée par licence.

[DEH+94] C. Doran, D. Egedi, B. A. Hockey, B. Srinivas, M. Zaidel, « XTAG System — A Wide Coverage Grammar for English», *in : Proc. of the 15th International Conference on Computational Linguistics (COLING'94)*, p. 922–928, Kyoto, Japan, August 1994.

allons poursuivre l'expérimentation en trois étapes avec une première grammaire de l'anglais (environ 500 arbres), puis la grammaire complète de l'anglais (plus de 1000 arbres) et enfin la grammaire du français (plus de 5000 arbres).

### 5.3 DyALog: Automates à piles et Programmation dynamique

**Participant** : Éric Villemonte de la Clergerie.

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

**Glossaire :**
**TAG** *Tree Adjoining Grammars*
**Feature TAG** TAG avec attributs
**2SA** *2-stack automata*
**LPDA** *Logical Push-Down 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, permettant l'analyse tabulaire des grammaires d'arbres adjoints [TAG]. Dans le cadre du travail de thèse de M. Alonso Pardo [Alo00], nous avons depuis étudié les liens existants avec d'autres formalismes d'automates comme les automates à piles enchâssées et les automates linéaires indexés. Nous avons constaté une forte équivalence entre ces formalismes et les interprétations en programmation dynamique que l'on peut en dériver [15, 14, 13, 9].

D'autre part, en calquant l'extension des PDA Automates Logiques à Piles (LPDA) pour passer des CFG aux DCG, nous avons rendu possible l'emploi des 2SA pour les Feature TAG en autorisant l'emploi de termes logiques dans les piles et l'emploi de l'unification pour appliquer les transitions. A cette occasion, nous avons pu réutiliser de nombreux résultats obtenus dans le cadre des LPDA, touchant en particulier les problèmes de propagation de l'information.

L'extension aux Feature TAG s'est faite en conjonction avec une implantation dans le système DyALog, à partir d'une première ébauche réalisée par Djamé Seddah en 1999. Cette implantation a été validée sur une petite grammaire du français et décrite dans [12].

Néanmoins, cette première implantation s'est révélée relativement complexe à mettre en œuvre, sans donner des résultats satisfaisants. L'analyse de ces problèmes nous a conduit à penser que l'interprétation originelle en programmation dynamique des 2SA, trop générique, ne prend pas assez en compte les propriétés des TAG. L'interprétation originelle est également complexe à mettre en œuvre car elle cherche à assurer des complexités optimales dans les pires cas, à savoir $O(n^6)$ en temps et $O(n^5)$ en place où $n$ est la longueur de la chaîne analysée. Or, en pratique, les cas conduisant à ces complexités ne semblent pas fréquents dans le cadre

[Alo00] M. Alonso Pardo, *Interpretación tabular de autómatas para lenguajes de adjunctión de árboles*, thèse de doctorat, Universidade da Coruña (Spain), sept 2000.

du traitement de la langue. Ces remarques nous ont incités à développer une nouvelle interprétation plus spécifique des TAG et plus simple à mettre en œuvre. Elle est appelée «faible» par opposition à l'interprétation originelle dite «forte», car la tabulation y est affaiblie. En conséquence, les complexités des pires cas ne sont plus optimales, mais restent néanmoins polynomiales. Les premiers résultats ont confirmé les meilleures performances de l'interprétation «faible». Ce travail est décrit dans un article en cours de soumission à NAACL'01[24].

La mise au point de l'interprétation «faible» a été facilitée par une réflexion menée sur les relations existant entre les notions de tabulation et de continuation. En effet, la tabulation mène à des modèles opérationnels où les calculs ne se déroulent pas de manière séquentielle : ils peuvent être suspendus et réactivés. D'autre part, le partage de calcul, objectif important des approches tabulaires, est obtenu en supprimant l'information non pertinente dans un calcul. Une reprise de calcul revient à suivre une continuation, qui peut être stockée dans une transition (de l'automate sous-jacent), dans l'item (représentant le calcul suspendu) et dans une pile de contrôle (comme cela se fait normalement en Prolog). Ces diverses possibilités sont associées à différentes contraintes de suspension et à différents niveaux de partage des calculs. En particulier, stocker les continuations dans les items affaiblit la tabulation (en diminuant le partage de calcul et en augmentant les complexités des pires cas). La nouvelle interprétation «faible» des TAG exploite cette gestion des continuations. Plus généralement, le système DyALog offre la possibilité de gérer ces différentes formes de continuations au niveau des prédicats, au travers de directives de compilation. Un document sur ce sujet, intitulé «Tabulation and Continuations», est en cours de rédaction.

Enfin, une autre piste de réflexion concerne la notion de forêt partagée. Un analyseur syntaxique comme DyALog retourne par défaut l'ensemble des analyses possibles pour une phrase sous la forme d'une forêt d'analyse très compacte. Il est donc important de mieux comprendre ce que sont exactement ces forêts et si elles peuvent servir d'entrée pour d'autres traitements linguistiques. Ces questions ont motivé l'organisation d'une Journée ATALA sur le sujet, où nous avons également présenté un article [20].

Par ailleurs, ce thème s'est retrouvé dans le cadre des travaux de l'équipe sur les TAG. Il s'est révélé que, pour les TAG comme pour d'autres formalismes, la notion importante n'est pas celle de forêt des arbres d'analyse mais celle de forêt des arbres de dérivation. Les dérivations TAG s'expriment en effet à l'aide de CFG et peuvent être représentées par des arbres. Cette meilleure compréhension des forêts de dérivation pour les TAG nous a permis de les utiliser comme *guide* dans une expérience d'analyse multi-passe : dans le cadre d'une petite grammaire du français, une première passe avec un analyseur RCG (Section 5.2) produit une forêt de dérivation, sous forme CFG, qui est ensuite utilisée pour guider un analyseur DyALog (vérifiant les attributs et les contraintes de co-ancrages). Cette expérience est relatée dans [16].

**Développement du système DyALog** Le travail théorique autour de la notion de tabulation s'est poursuivi en parallèle avec le développement du système DyALog.

En particulier, nous avons étendu DyALog pour traiter les grammaires FTAG organisées selon une architecture XTAG [DEH+94], une grammaire comprenant des schémas d'arbres avec

[DEH+94] C. Doran, D. Egedi, B. A. Hockey, B. Srinivas, M. Zaidel, « XTAG System — A Wide Coverage Grammar for English», *in : Proc. of the 15th International Conference on Computational*

des ancres et des co-ancres, des entrées morphologiques et des entrées syntaxiques. Nous disposons actuellement dans DyALog d'un format de représentation de ces grammaires et d'un compilateur. Différents tests ont été réalisés sur une petite grammaire du français et sont relatés dans [12, 16, 24].

Les réflexions sur la notion de continuation nous ont aidés à simplifier la machine abstraite de DyALog et à proposer une nouvelle classe de prédicats, extrêmement proche des prédicats de PROLOG. Ces prédicats permettent d'améliorer le traitement des calculs de ceux qui n'ont pas besoin d'être tabulés.

Pour faciliter le développement futur de DyALog, nous avons formalisé et simplifié les interfaces entre C et DyALog. Il est maintenant très simple d'intégrer une API C dans DyALog. À titre d'expérience, nous avons intégré une API vers le système de base de données POSTGRESQL. L'appel de prédicats DyALog à partir de C est possible mais il reste encore un effort à effectuer pour en faciliter l'emploi.

Pour faciliter l'écriture de programmes et surtout de grammaires, nous continuons à ajouter différentes fonctionnalités à DyALog. En particulier, il est maintenant possible d'utiliser des variables logiques à valeur dans un domaine fini de constantes.

Ces différentes améliorations ont apporté des gains importants dans le compilateur de DyALog, permettant entre autres de diviser par 10 le temps de compilation de la grammaire du français (passant de 500s à 50s). Ceci a été en particulier rendu possible par l'utilisation de l'interface avec C, permettant la migration en C de certaines parties cruciales du compilateur.

Enfin, nous avons consacré plus de temps cette année au développement d'exemples avec DyALog, ce qui nous confirme sa stabilité et sa maturité. Outre le compilateur («boostrapé» en DyALog) et le traitement de grammaires TAG, nous avons également adapté un analyseur morphologique des verbes en Akkadien originellement développé par F. Barthélemy, adaptation grandement facilitée par la présence des variables à domaine fini dans DyALog. Les bonnes performances de DyALog ont également été confirmées dans le cadre des travaux de thèse de V. Rocio sur une grammaire du portugais [10].

cette «maturation» de DyALog nous incite à faire un effort accru de promotion et de diffusion pour l'année à venir, effort déjà lancé cette année avec par exemple une présentation lors d'une journée ATALA [23]. En vue de cet objectif, il est envisagé de consacrer du temps à la rédaction de la documentation.

### 5.4 Bibliothèques électroniques

**Participant** : François Role.

**Mots clés** : métadonnée, bibliothèque électronique.

**Glossaire :**
**TEI** *Text Encoding Initiative*
**DOM** *Document Object Model*

Dans le cadre de son travail de thèse, F. Role continue d'évaluer 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. Il est intervenu sur ce thème en tant que conférencier à l'école d'été IST'2000 organisé par l'INRIA en septembre 2000. Cette intervention a été accompagnée par la publication d'une contribution dans un ouvrage consacré aux bibliothèques numériques [11].

Parallèlement à la rédaction de sa thèse, F. Role a développé, en coopération avec l'Université de Paris X, un système de consultation de documents XML basé sur des vues dynamiques. La présentation d'un prototype de ce système a donné lieu à la publication d'un article dans les actes de la 3ème conférence internationale sur les documents électroniques (CIDE 2000) qui s'est tenue à Lyon en juin 2000 [19].

## 5.5 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.[7] 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 trois ans.

Les logiciels libres, et plus généralement les ressources libres, sont des ressources immatérielles qui, ayant été produites, sont mises à la disposition du public avec tous les moyens techniques et autorisations juridiques de les utiliser, de les faire évoluer, et de les rediffuser. Il s'agit donc d'un modèle de création très similaire à celui de la recherche scientifique traditionnelle, mais s'appliquant aussi à des ressources pouvant être directement utilisables, par des spécialistes, par des entreprises, ou par le grand public.

Nos travaux dans ce domaine ne portent pas spécifiquement sur les aspects linguistiques, mais plus généralement sur une analyse de l'intérêt et de l'impact des logiciels libres sur l'économie, sur le fonctionnement de la recherche, et sur la stratégie des entreprises. Il s'agit d'un travail de défrichage d'un modèle nouveau de production et qui comporte notamment des volets économiques et juridiques [8, 7]. Notre travail a demandé une forte composante d'activité de terrain, permettant notamment une bonne communication avec les entreprises.

Notre réflexion [8] sur l'intérêt des logiciels libres pour la maîtrise des standards (et donc de certains marchés), ainsi que sur leur intérêt compétitif dans le cas des systèmes embarqués est en voie de confirmation par les décisions stratégiques de plusieurs grandes entreprises françaises, ainsi que des grandes administrations [26].

7. Notre attention fut initialement attirée sur ce sujet par la mise en oeuvre 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.

# 6 Contrats industriels (nationaux, européens et internationaux)

## 6.1 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 doit démarrer en 2001.

# 7 Actions régionales, nationales et internationales

## 7.1 Actions nationales

En raison de son changement d'activité, Ph. Deschamp a quitté fin 1998 l'Academic Advisory Council de la société Sun Corp.

Ph. Deschamp est membre de la Commission spécialisée de terminologie de l'informatique et des composants électroniques, et diffuse sur la toile le glossaire[8] résultant de ses travaux (téléchargé 55 000 fois cette année).

Ph. Deschamp a également fait partie de la Commission de normalisation de l'AFNOR CGTI/CN 1 « Vocabulaire » jusqu'à sa clôture le 27 avril 1999 pour des raisons financières.

B. Lang est secrétaire de l'AFUL (`http://www.aful.org`), Association Francophone des Utilisateurs de Linux et des Logiciels Libre, et membre du conseil d'administration de l'ISoc-France (`http://www.isoc.asso.fr`), chapitre français de l'Internet Society.

### 7.1.1 Logiciels Libres

B. Lang a présenté les logiciels libres dans des séminaires, tables-rondes et conférences organisés par plusieurs entreprises, collectivités locales et administrations, dont la communauté des communes de la région de Soissons, Thomson CSF, Linagora, GEMPLUS, EFE (formation), FSU (Fédération Syndicale Unitaire), Alcove, MatraDatavision, Mandrake, Alcatel, CIRAD (Montpellier), ESA, et a eu une activité de conseil chez Bull.

## 7.2 Réseaux et groupes de travail internationaux

### 7.2.1 Logiciels Libres

B. Lang a été invité à plusieurs reprises à s'exprimer sur les logiciels libres dans les pays francophones (Belgique, Suisse, Tunisie) ainsi qu'aux Pays-Bas à l'Agence Spatiale Européenne et par téléphone à une réunion du PITAC sur les logiciels libres pour le calcul à hautes performances à San Diego (USA).

8. Voir `http://www-Rocq.INRIA.Fr/qui/Philippe.Deschamp/CMTI/index.html`

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/`) qui a produit un rapport sur les logiciel libres. [26]

A l'invitation du Secrétariat d'État à l'Informatique de Tunisie, B. Lang a participé à l'organisation d'un Workshop sur les logiciels libres.

#### 7.2.2 Réseau franco-portugais de formation par la recherche

G. Pereira Lopes et V. Rocio du groupe CENTRIA de l'Université Nouvelle de Lisbonne et É. de la Clergerie ont poursuivi leur collaboration portant sur l'emploi de DyALog pour la réalisation d'un analyseur syntaxique robuste pour le portugais. Cet analyseur comprend plusieurs couches, dont deux exploitent DyALog, et s'appuie en particulier sur le formalisme des *grammaires à mouvements restreints* (BMG). Un article «Tabulation for multi-purpose partial parsing» [10] relate cette expérience. De plus, une demande de coopération ICTII entre le groupe ATOLL, le groupe CENTRIA et l'université d'Orléans a été déposée pour renforcer les échanges entre nos équipes.

## 8 Diffusion de résultats

### 8.1 Encadrement

É. de la Clergerie a encadré le stage de DEA de L. Kaouane [22].

É. de la Clergerie a suivi le travail de thèse de M. Alonso Pardo de l'Université de la Corogne (Espagne) sur le thème «Analyse Tabulaire pour les TAG et les LIGs» et celui d'Isabelle Debourges de l'Université d'Orléans sur le thème «Fouille de textes».

B. Lang suit le travail de thèse de F. Role.

### 8.2 Jury

É. de la Clergerie a participé à la Commission Mixte d'Évaluation en Informatique de l'Université d'Orléans.

É. de la Clergerie (en tant que codirecteur de thèse) et P. Boullier ont participé en Septembre au jury de thèse de M. Alonso Pardo à l'Université de la Corogne (Espagne). La thèse est intitulée «Interprétation tabulaire d'automates pour les langages d'adjonction d'arbres» [Alo00].

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

### 8.3 Enseignement

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

B. Lang a contribué à une formation des moniteurs du CIES de Versailles en mai.

[Alo00] M. Alonso Pardo, *Interpretación tabular de autómatas para lenguajes de adjunctión de árboles*, thèse de doctorat, Universidade da Coruña (Spain), sept 2000.

## 8.4 Comités de programme

Pierre Boullier a été membre du Comité de Programme TAG+5, au cours duquel il a présidé une session. Il a en outre jugé des articles proposés à IWPT'00 (Sixth International Workshop on Parsing Technologies), TAG+5 (Fifth International Workshop on Tree Adjoining Grammars and Related Frameworks), ACL'00 (38th Annual Meeting of the Association for Computational Linguistics, commission analyse syntaxique) et CC01 (Compiler Compilers 2001).

É. de la Clergerie a été membre des comités de programme de TAG+5 et de TAPD'00. Il a également présidé une session à IWPT'00 et à TAPD'00.

B. Lang est membre du comité éditorial de la revue FUTUR(e)S.

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

- Membre du comité de programme, Linux Expo / Linux World, et animateur de la Keynote Session, de la session «Linux en entreprise» et de la table ronde «les enjeux éducatifs des logiciels libres» (février, Paris)
- Membre du comité de programme, Linux Expo / Linux World (février 2001, Paris)
- Salon et conférence Intranet 2000, organisateur et animateur de la session sur les logiciels libres, (mars, Paris).
- Salon et conférence Net 2001, (mars 2001, Paris)
- Membre du comité d'organisation du «Workshop sur le Logiciel Libre», organisé par le Secrétariat d'État à l'Informatique de Tunisie (octobre, Tunis).
- Membre du jury du concours Electrophées des administrations, organisé par la MTIC et le ministère de la fonction publique (décembre, Paris)

F. Role est membre du comité d'organisation de la conference ISKO 2001.

## 8.5 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 :

- Exposé «Les logiciels libres, une réponse aux enjeux socio-économiques de la société de l'information» à la conférence «Les logiciels libres: vers une informatique citoyenne» (février, Bruxelles, Belgique).
- Animateur de la conférence «Le logiciel libre» au MICAD 2000 mars, Paris).
- Animateur de la table ronde «Les logiciels libres : une orientation du marché ? pour quelle cible ? Pour quels usages ?», Aristote, Collège de Polytechnique (mars, Palaiseau).
- Exposé à la soirée débat «Les pièges de l'informatique propriétaire» (avril, Louvain-La-Neuve, Belgique).

- Exposé «Les publications scientifiques libres» aux IIIèmes Journées du Libre de Strasbourg (mai, Strasbourg-Illkirch).
- Exposé «Les logiciels libres - Enjeux économiques et opportunités pour l'entreprise», journée professionnelle RUBIS, Réunion des Utilisateurs de la Base Ingres en Suisse, (mai, Lausanne).
- Séminaire Logiciel Libre organisé par la DIT et la délégation Scientifique MIA au CIRAD (Montpellier, juin).
- Journées logiciels libres organisées par ULiCe (juin, Bourges).
- Exposé «Coopération sur les logiciels libres» à la deuxième université d'été francophone «Développement durable et information pour la prise de décision», Ecole des Mines de Saint-Etienne (juin, Saint-Etienne).
- Exposé invité à la table ronde inaugurale «Open Source Movement» de la Conférence INET 2000, 10th Annual Internet Society Conference (juillet, Yokohama, Japon).
- Table ronde «Les logiciels libres dans l'éducation nationale», organisée par le CNDP à la 21e Université d'été de la Communication (août, Hourtin).
- Exposé «User Perspective» au séminaire «The Role of Open-Source Software in the Space Business» organisé par l'ESA (octobre, ESTEC, Noordwijk, The Netherlands).
- Exposé «Enjeux économiques et perspectives» à la 11ème Journée de Rencontre de l'Observatoire Technologique de Genève : «Les Logiciels Libres» (octobre, Genève).
- Exposé «Logiciels libres, licence et stratégie d'entreprise» au Workshop sur le Logiciel Libre du Secrétariat d'État à l'Informatique de Tunisie (octobre, Tunis).
- Rencontres de l'ORME, participation à la table ronde «Innovation, coopération et propriété intellectuelle : l'approche libre» (octobre, Marseille).
- Conférence «Enjeux de la société de l'information, éthique, politique, liberté, éducation», exposé «Aspects politiques et économiques», Collectif d'associations pour les logiciels libres (Octobre, Bruxelles).
- Conférence IST 2000 (invité, sans participation orale), Commission Européenne, DG Information Society (octobre, Nice).
- Salon «Cyber 2000», 2 exposés «Utiliser les logiciels libres dans l'entreprise» et «Développer un logiciel dans le cadre de travail libre» (novembre, Saint Denis, La Réunion).
- Tables rondes «Les logiciels libres : état des lieux et enjeux dans l'éducation nationale» (participant) et «Linux et les logiciels libres pour les établissements scolaires : dotations informatiques, fonctionnement des parcs» (animateur), Salon Educatec (Paris, novembre).
- Exposé «Logiciel libre et Entreprise» au séminaire «Support Linux et logiciel libres», AFUU (novembre, Boulogne).

- Exposé aus journées «Ada et le Logiciel Libre», Ada-France et AFUL (novembre, ENST, Paris).

- Table ronde de synthèse au workshop du Projet RNTL «Nouvelle économie du logiciel» (décembre, Rocquencourt).

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

- Présentation «Intellectual Property Rights» au workshop «Free Software / Open Source : Information Society Opportunities for Europe ?», organisé par la DG Information Society de la Commission Européenne (mars, Bruxelles).

- Exposé «Gestion des ressources sur l'Internet et diffusion électronique de l'information : questions et perspectives», Journées Multimédia, organisées par le CUTO à la Bibliothèque Nationale de France (mars, Paris).

- Exposé «Les réponses techniques» à la journée d'étude «Les données personelles, la loi et l'internet» organisée par le Groupement Français de l'Industrie de l'Information et le Geste (mai, Paris).

- Exposé «Logiciel libre et logiciel breveté», 18-ème Forum APP «Liberté et droits des auteurs numériques», APP, CEIPI et Université Robert Schuman (décembre, Strasbourg).

- Table ronde «Renforcement ou affaiblissement des monopoles» à la conférence débat «Que sera la propriété intellectuelle dans 10 ans» organisée par le CUERPI, Faculté de Droit, Université Pierre Mendes France, Grenoble 2 (décembre)

- Participation à la table ronde «Les nouvelles technologies nécessitent-elles un droit nouveau ?» du colloque «L'internet à l'épreuve du droit», ANAAFA (décembre, Cergy-Pontoise).

B. Lang a assisté aux Rencontres d'Autrans de l'ISOC-France (janvier, Autrans).

B. Lang a présenté un exposé invité sur «Enoncé mal formés et traitement de l'ambiguité», à la Journée ATALA «Représentation et traitement de l'ambiguité pour l'analyse syntaxique» (Paris, 29 Janvier 2000).

É. de la Clergerie a co-organisé la Journée ATALA avec actes «Représentation et traitement de l'ambiguité pour l'analyse syntaxique» (Paris, 29 Janvier 2000) `http://www.biomath.jussieu.fr/ATALA/je/je-000129.html`.

É. de la Clergerie, P. Boullier, Ph. Deschamp et F. Barthélemy ont participé aux réunions du groupe de travail **TAGML** `http://www.loria.fr/~lopez/TAG_XML/` concernant l'utilisation des Grammaires d'Arbres Adjoints en général et l'emploi de représentations XML de ces grammaires en particulier. É. de la Clergerie et P. Boullier y ont présenté à plusieurs reprises les travaux du groupe ATOLL dans le domaine des TAG.

É. de la Clergerie a participé aux réunions du groupe de travail **A3CTE** `http://www-lipn.univ-paris13.fr/groupes-de-travail/A3CTE/` qui cherche à faire émerger des applications combinant Apprentissage et Traitement Automatique des Langues Naturelles.

Participation de P. Boullier à IWPT'00 (Trente, Italie) avec présentation [18]. Il a été conférencier invité à AMILP'00 (Iowa, USA) [17].

Pierre Boullier a été invité par l'équipe Langue et Dialogue du LORIA à présenter ses travaux sur les RCG. À la suite de ce séminaire, Bertrand Gaiffe, à fait implanter par un stagiaire un analyseur RCG fondé sur l'algorithme de principe décrit en [18].

Pierre Boullier a été invité à l'Université de Marne-la-Vallée, Institut Gaspard Monge, à un séminaire au cours duquel il a présenté ses travaux sur les RCG.

Participation de É. de la Clergerie à TAG+5 (Paris) (Présentation [12] et Démonstration) et à TAPD'00 (Vigo, Espagne) [16].

É. de la Clergerie a été conférencier invité à SEPLN'00 (Vigo, Espagne, Septembre 2000) [25].

É. de la Clergerie a présenté le système DyALog lors de la Journée ATALA «Outils pour le Traitement automatique des langue» (Paris, Novembre 1999) [23]. Il a également été invité par l'équipe Langue et Dialogue du LORIA à présenter ses travaux sur les TAG.

# 9 Bibliographie

## Ouvrages et articles de référence de l'équipe

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

[2] 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.

[3] Éric 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`.

[4] Éric 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.

## Articles et chapitres de livre

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

[6] B. Lang, «Internet libère les logiciels», *La Recherche*, février 2000, `http://pauillac.inria.fr/~~lang/ecrits/larecherche`.

[7] B. Lang, «Le nouveau protectionnisme est intellectuel», *in : Libres enfants du savoir numérique*, O. Blondeau et F. Latrive (éditeurs), Éditions de l'éclat, Perreux, mars 2000, ISBN 2-84162-043-3, `http://pauillac.inria.fr/~~lang/ecrits/latrive`.

[8] B. Lang, «Logiciels libres et entreprises», *Terminal*, 80/81, 2000, `http://pauillac.inria.fr/~~lang/ecrits/monaco`.

[9] M.-J. Nederhof, M. Alonso Pardo, E. Villemonte de la Clergerie, «Tabulation of Automata for Tree-Adjoining Languages», *Grammars 3*, 23, 2000, To be published.

[10] V. J. Rocio, G. P. Lopes, E. Villemonte de la Clergerie, «Tabulation for multi-purpose parsing», *Grammars*, 2000, à paraître.

[11] F. Role, *Bibliothèques numériques - Cours INRIA*, INRIA et éditions de l'ADBS, 2000, ch. Métadonnées et structuration des bibliothèques numériques, p. 143–170.

## Communications à des congrès, colloques, etc.

[12] M. Alonso Pardo, D. Seddah, E. Villemonte de la Clergerie, «Practical aspects in compiling tabular TAG parsers», *in : Proceedings of the 5th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+5)*, p. 27–32, Université Paris 7, Jussieu, Paris, France, mai 2000, `ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/tag5b.ps.gz`.

[13] M. Alonso Pardo, E. Villemonte de la Clergerie, J. Graña, M. Vilares, « New tabular algorithms for LIG Parsing », *in : Proceedings of the 6th International W orkshop on Parsing Technologies (IWPT2000)*, p. 29–40, Trento, Italy, février 2000.

[14] M. Alonso Pardo, E. Villemonte de la Clergerie, M. Vilares, « A formal definition of Bottom-Up Embedded Push-Down Automata and their tabulation techniques », *in : Proc. of 2nd workshop on Tabulation in Parsing and Deduction (TAPD'00)*, p. 101–112, Vigo, Spain, sept 2000.

[15] M. Alonso Pardo, E. Villemonte de la Clergerie, M. Vilares, « A redefinition of Embedded Push-Down Automata », *in : Proceedings of the 5th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+5)*, p. 19–26, Université Paris 7, Jussieu, Paris, France, mai 2000.

[16] F. Barthelémy, P. Boullier, P. Deschamp, E. Villemonte de la Clergerie, «Shared Forests can guide Parsing», *in : Proc. of 2nd workshop on Tabulation in Parsing and Deduction (TAPD'00)*, p. 165–174, Vigo, Spain, sept 2000, `ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/tapd00.ps.gz`.

[17] P. Boullier, «'Counting' with Range Concatenation Grammars», *in : Proceedings of the second AMAST workshop on Algebraic Methods in Language Processing (AMILP 2000)*, Iowa City, Iowa, USA, mai 2000.

[18] P. Boullier, « RANGE CONCATENATION GRAMMARS », *in : Proceedings of the Sixth International Workshop on Parsing Technologies (IWPT2000)*, p. 53–64, Trento, Italy, février 2000.

[19] I. M. S. Chaudiron, F. Role, « Codex: un système pour la définition de vues multiples guidée par les usages», *in : Actes du 3ème Colloque International sur le Document Electronique : CIDE'2000*, M. G. et E. trupin (éditeur), p. 71–81, juillet 2000.

[20] E. Villemonte de la Clergerie, « Créer, extraire et manipuler des forêts partagées avec DyALog », *in : Actes de la Journée ATALA "Représentation et traitement de l'ambiguité pour l'analyse syntaxique"*, P. Blache, E. Villemonte de la Clergerie (éditeurs), ATALA, INRIA, janvier 2000.

## Divers

[21] F. Barthélemy, P. Boullier, P. Deschamp, L. Kaouane, E. Villemonte de la Clergerie, « Tools and Resources for Tree Adjoining Grammars », Présenté au groupe de travail TAGML, octobre 2000.

[22] L. Kaouane, *Adaptation et utilisation d'un environnement graphique pour les TAG au dessus du système DyALog*, Mémoire, Université d'Orléans, sept 2000, Mémoire de DEA.

[23] E. Villemonte de la Clergerie, « Construire des analyseurs syntaxiques tabulaires avec le système DyALog (slides)», slides in French presented at the workshop "Outils pour le traitement automatique des langues" organized by ATALA and GdR I3, novembre 1999.

[24] E. Villemonte de la Clergerie, « Refining Tabular Parsers for TAGs », Submitted to NAACL'01, novembre 2000.

[25] E. Villemonte de la Clergerie, « Tabulation for Natural Language Processing(slides) », slides presented at SEPLN'00, sept 2000, `ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/sepln00-slides.ps.gz`.

[26] Working group on Libre Software, Information Society Directorate General of the European Commission, «Free Software / Open Source: Information Society Opportunities for Europe?», Version 1.2, avril 2000, `http://eu.conecta.it/paper.pdf`.