Avant-Projet ATOLL

Atelier d'outils logiciels pour le langage naturel

Localisation : *Rocquencourt*

Mots-clés : analyse syntaxique, Lambda-Prolog, langage naturel, linguistique, programmation dynamique, programmation logique, programmation par contraintes, interface homme-machine, hypertexte, document électronique, agent intelligent, CD-ROM, World Wide Web.

1 Composition de l'équipe

Responsable scientifique

Bernard Lang, directeur de recherche, INRIA

Responsable permanent

Pierre Boullier, directeur de recherche, INRIA

Secrétaire

Josy Baron, en commun avec LOCO et GENOME

Personnel INRIA

Éric Villemonte de la Clergerie, chargé de recherche

Ingénieur expert

Roland Dachelet

Chercheurs doctorants

Alain Hui Bon Hoa, AMN CNAM
Frédéric Tendeau, boursier INRIA, université d'Orléans
Miguel Alonso Pardo, Université de la Coruña, Espagne, Janvier-Décembre 1996
François Role, Fonctionnaire au DISTNB-MESR, université d'Orléans, à partir de Décembre 1996

Stagiaire longue durée

Cédric Billaud, Stagiaire Ingénieur CNAM, collaboration avec Consortium W3, d'octobre 1995 à octobre 1996

Collaborateurs Extérieurs

François Barthélemy, maître de conférence, CNAM
Jean-Marie Larchevèque, maître de conférence, IUT de Vélizy

2 Présentation du projet

Notre équipe s'est initialement constituée autour d'une compétence des trois membres permanents, établie dans les techniques d'analyse syntaxique et d'évaluation tabulaire des programmes logiques. Nos motivations étaient à l'origine issues des problèmes de la compilation des langages de programmation, mais elles se sont ensuite orientées vers l'analyse syntaxique, voire sémantique, du langage naturel, ce domaine ayant pris une importance considérable tant du point de vue des développements scientifiques que de celui des applications industrielles.

Cependant, si l'on considère les problèmes du traitement de la langue dans toute leur étendue, notre équipe ne couvre en fait qu'un champ très restreint de problèmes, l'algorithmique de l'analyse du langage, même si nous avons acquis quelques connaissances sur des aspects plus spécifiquement linguistiques. Réaliser une chaîne complète de traitement d'un problème lingistique, que ce soit l'analyse de documents ou la traduction automatique (pour prendre deux exemples importants), dépasse à la fois les moyens et les compétences actuels de notre équipe.

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. Le développement de l'usage des documents structurés, dû largement mais pas uniquement au développement de la "toile" WWW (le "World Wide Web") nous parait une opportunité à exploiter, notamment en raison de notre expérience concernant les environnements de programmation. En conséquence, nous sommes en train de diversifier notre domaine de compétence 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 implémentent.

Le développement de nos activités présente donc actuellement deux aspects, que nous faisons converger :

1. Poursuite de nos travaux sur les techniques fondamentales en analyse syntaxique et évaluation tabulaire de programmes logiques.

2. Recherche, traitement et gestion des documents électroniques, et mise en œuvre de leurs principaux supports : la "toile" WWW et le CDROM.

3 Actions de recherche

3.1 Formalismes grammaticaux

Traditionnellement, pour un langage de programmation, on appelle sémantique statique tout ce qui ne peut pas (ou ne peut pas facilement) s'exprimer dans le formalisme choisi pour décrire la syntaxe. Cette partie non syntaxique est de l'information contextuelle qui est décrite par un autre formalisme (les grammaires attribuées par exemple) qui coexiste avec le formalisme syntaxique. Cette partition, qui correspond à une vue et à un traitement non unifiés, est essentiellement justifiée par des questions de modularité, d'expressivité, voire d'efficacité. Bien entendu, cette dualité existe également dans les formalismes décrivant les langues naturelles, mais ici la difficulté est d'un autre ordre de grandeur. Alors que la syntaxe des langages de programmation se définit 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 (TAG et variantes, LIG, GPSG, HPSG, CG, LFG, DCG, etc...). Et de fait, l'apparition de nouveaux formalismes est très fréquente.

Vis-à-vis de ce foisonnement, notre approche est double :

1. avoir un modèle d'exécution très général (automate à pile logique avec tabulation) dans lequel il est possible d'exprimer et de comparer les formalismes linguistiques. Bien entendu, la contrepartie de cette généralité peut être une certaine inefficacité à l'exécution, qui demande en contrepartie le développement de techniques d'optimisation;

2. l'autre approche, qualifiée de pragmatique, est de proposer des implantations efficaces, pour des formalismes linguistiques possédant une base structurelle très forte.

Ces deux approches ne s'opposent pas. Au contraire, chacune doit enrichir l'autre. L'examen de particularités mis en évidence au niveau pragmatique permet des avancées théoriques ; réciproquement, des concepts théoriques bien compris vont se traduire par un élargissement du champs d'action pragmatique.

Dans tous les cas, les différents formalismes grammaticaux entrant dans notre champ d'application se situent dans ce qu'on pourrait appeler le "continuum de Horn", c'est-à-dire un ensemble de formalismes de complexité croissante, allant des Clauses de Horn propositionelles aux Clauses de Horn du premier ordre (grosso-modo Prolog), et même au-delà.

3.2 Évaluation tabulaire: DyALog

Participants : Eric Villemonte de la Clergerie, Miguel Alonso Pardo

Mots-clés : tabulation, linguistique, programmation en logique, programmation dynamique, automates logiques à piles.

Eric de la Clergerie étudie sur le plan théorique et pratique la mise au point d'évaluateurs tabulaires pour des calculs pouvant être décrits à l'aide d'Automates à Piles (plus précisément de SPDA). De telles évaluations qui reposent sur la mémorisation de traces compactes des calculs (appelées *items*) permettent le partage de sous-calculs (et dans une certaine mesure, la détection des boucles). Ceci est utile dans le cadre d'applications fortement non-déterministes, telles l'analyse syntaxique, la programmation logique ou (au confluent des deux premières) en Linguistique.

Ces algorithmes tabulaires reposent sur l'utilisation d'interprétations en Programmation Dynamique des automates qui précisent comment "casser" et "recoller" les calculs. Plusieurs de ces interprétations ont été proposées et utilisées ces dernières années. Eric de la Clergerie a cherché à unifier ces différentes interprétations en précisant les concepts élémentaires sous-jacents [65].

Ce travail de compréhension des principes permettant le développement d'algorithmes tabulaires se retrouve également dans le sujet de thèse de Miguel Alonso Pardo, étudiant de l'Université de la Corogne qu'encadre Eric de la Clergerie. Miguel Alonso Pardo examine comment construire des interprétations en Programmation Dynamique pour des extensions des automates à piles (telles les automates à deux piles) qui permettraient d'expliquer et d'étendre différents algorithmes tabulaires ayant été proposés pour les grammaires d'arbres adjoints (TAG) et linaires indexées (LIG).

D'autre part, en collaboration avec François Barthélemy, Eric de la Clergerie a affiné la notion de flux d'information, qui permet de formaliser proprement la quantité d'information devant être présente dans un item, ce qui est primordial pour assurer un meilleur partage de calculs [53].

L'évaluateur tabulaire de programme logiques DyALog a été réécrit en Scheme puis en C dans l'optique d'une migration de DyALog vers un compilateur de grammaires linguistiques logiques. Le modèle sous-jacent identifie les items et transitions de l'automate à des objets encapsulant leur code d'application. Dans le contexte d'un compilateur, ceci permet de compiler du code d'application beaucoup plus spécialisé (et donc efficace) que le mécanisme uniforme d'application actuel des transitions sur les items.

L'efficacité d'algorithmes tabulaires repose en grande partie sur l'utilisation de systèmes performants d'indexation des items en vue de recherches rapides dans la table des items. Dans le cadre de programmes logiques, cette indexation doit se faire sur des termes logiques. Eric de la Clergerie a développé et testé un tel algorithme d'indexation de termes logiques, qui, bien que performant du point de

vue indexation, se révèle assez lent et coûteux en place. L'analyse des résultats permet de voir différentes amélioration à apporter. D'autre part, d'autres algorithmes d'indexation pour les termes logiques ont enfin commencé à apparaître, d'où la nécessité de conduire différents tests.

3.3 Architecture des analyseurs

Participants : François Barthélemy

Le système APOC-II a été écrit en 1993-94 par F. Barthélemy pour implémenter rapidement diverses variantes des algorithmes d'analyse syntaxique en utilisant une architecture modulaire, ou chaque module peut avoir plusieurs implémentations. Ces variantes peuvent affecter, entre autres, la stratégie d'analyse (ascendante ou descendante), la classe de grammaires utilisée (variation de forme des attributs et des unifications subséquentes), la gestion du non-déterminisme (retour arrière ou tabulation) et ainsi de suite.

En 1996, F. Barthélemy a entrepris le portage du système en OCAML. Le but de ce portage est d'utiliser APOC-II pour une expérimentation: une comparaison empirique du comportement de différents algorithmes d'analyse syntaxique pour une grammaire donnée.

Parallèlement au portage en cours, F. Barthélemy a commencé à rechercher une grammaire décrivant une langue naturelle (anglais ou français) qui puisse servir de banc d'essai, à la fois conséquent en taille, tout en restant lisible en complexité. Cette recherche porte principalement sur les grammaires d'adjonction d'arbres (TAG).

3.4 Analyse contextuelle

Participant : Pierre Boullier

3.4.1 Les grammaires indexées linéaires

Parmi les formalismes décrivant les fondements syntaxiques des langues naturelles en vue de leur traitement automatique, nous nous intéressons à ceux qui reposent explicitement ou implicitement sur un squelette non-contextuel (CF) et dans lesquels la partie contextuelle peut être vue comme un système de contraintes dans le sens où un arbre donné de la forêt partagée du squelette CF est un arbre d'analyse du formalisme considéré si et seulement si la résolution du système appliqué à cet arbre produit au moins une solution. Un traitement se décompose donc en deux phases : une analyse CF produisant en sortie une forêt partagée sur laquelle est appliquée, dans une seconde phase, le système de contraintes. Le résultat final est une structure compatible à la fois avec la partie CF et la partie contextuelle du formalisme initial. De plus, il est possible d'appliquer sur cette nouvelle structure d'autres contraintes non exprimées par le formalisme syntaxique initial.

Dans un premier temps, nous considérons la classe des langages modérément contextuels (Mildly Context-Sensitive). Ces langages peuvent se décrire par un certain nombre de formalismes grammaticaux équivalents dont les grammaires d'arbres adjoints (TAG) et les grammaires indexées linéaires (LIG). D'un point de vue purement linguistique, il semble que les phénomènes contextuels se décrivent plus aisément sous forme de TAG. Cependant, d'un point de vue calculatoire, les LIG sont mieux adaptées et leurs étude et traitement jouent un rôle primordial, sachant que leur équivalence formelle avec les TAG permet de passer de l'une à l'autre.

Nous avons donc conçu un nouvel algorithme de reconnaissance pour les LIG qui se démarque des autres reconnaisseurs sur deux points :

1. il utilise en entrée la forêt partagée, résultat de l'analyse CF de la chaîne source ;
2. il ne calcule explicitement aucune pile de symboles mais utilise des relations binaires qui, par compositions, permettent de déterminer les "dorsales" (spines) valides.

Le point 1 ajoute à la modularité de l'algorithme puisqu'il ne dépend pas de la méthode d'analyse utilisée, quant au point 2, il a permis une généralisation de ce reconnaisseur en un analyseur.

Il est clair que pour une grammaire CF (CFG) déterministe la sortie d'un analyseur est un arbre et que pour une CFG générale la sortie est une structure appelée forêt partagée qui, d'un point de vue formel, peut être considérée comme la CFG résultat de l'intersection de l'automate fini représentant la chaîne source et de la CFG générale initiale. En revanche, pour les formalismes contextuels, il faut trouver une structure de sortie compatible à la fois avec les parties non-contextuelle et contextuelle. Remarquons que cette structure ne peut-être une forêt partagée puisque l'intersection d'un automate fini et d'une grammaire contextuelle n'est pas, dans le cas général, une CFG.

Vijay-Shanker et Weir ont montré que les dérivations dans les TAG (c'est-à-dire le processus de composition des arbres élémentaires) est un phénomène CF. Nous avons appliqué ce résultat au formalisme équivalent des LIG pour définir un analyseur dont la structure de sortie est une CFG qui décrit, non pas la forêt partagée, mais le langage des dérivations valides. Cette structure appelée grammaire des dérivations définit un langage dont les chaînes terminales sont les dérivations valides (et uniquement celles-là) de la LIG qui conduisent de l'axiome à la chaîne source.

Si n est la longueur du source, cette grammaire de taille $\mathcal{O}(n^6)$ peut être construite en temps $\mathcal{O}(n^6)$. Il semble que les taille et temps obtenus en pratique sont bien meilleurs que ces bornes maximales. De plus, les arbres valides peuvent être extraits de ces grammaires en un temps qui est une fonction linéaire de leur taille.

3.5 Analyse contextuelle

Participants : Pierre Boullier, Bernard Lang, Frédéric Tendeau

Nous avons approfondi le travail sur l'utilisation des séries de puissance formelles pour faire de l'évaluation d'attribut dans un demi-anneau. Nous avions vu l'année précédente que les calculs de cœfficient pour les séries algébriques se prétaient bien à l'utilisation de la programmation dynamique mais quelques difficultés persistaient. Le calcul formel de cœfficient n'était pas explicité ; les problèmes de limite posés dans le cas de grammaires cycliques n'étaient pas clairement résolus ; et le formalisme de décoration dans un demi-anneau ne convenait pas au demi-anneau des forêts d'analyse. Enfin, nous avons commencé à développer des applications linguistiques de ces travaux abstraits.

3.5.1 Calcul formel et limites

Nous avons montré que, sous certaines hypothèses, il est possible de calculer formellement le cœfficient σ d'une chaîne $x \in \Sigma^*$ par rapport à une grammaire G et une décoration dans un demi-anneau abstrait A.

Ces hypothèses sont la donnée

1. des opérations de A,
2. d'une application de décoration dans A,
3. d'une application qui donne le cœfficient du mot vide pour la série associée à chaque non-terminal (ceci parce que le mot vide donne lieu à des équations non linéraires qu'il n'est pas possible de résoudre en toute généralité),
4. d'une fonction de calcul de limite dans A correspondant à la résolution d'une équation affine de la forme $X = aXb + c$ dans A. Cette dernière se ramène à une fermeture de Kleene si A est commutatif ($aXb = abX$ donc $X = (ab)^*c$).

Cette preuve a la forme d'un algorithme de construction du cœfficient σ par une méthode ascendante, de type Cocke-Younger-Kasami. Tout autre algorithme conviendrait, par exemple nos algorithmes stochastiques Earley et Left-Corner peuvent être utilisés (moyennant une abstraction des opérations et des calculs de limite).

3.5.2 Décoration

Une A-décoration de G est une application de l'ensemble de ses productions dans A. Si l'on considère qu'elle joue le rôle du point 2 ci-dessus, alors on ne peut pas trouver une interprétation du demi-anneau des forêts qui fasse de σ la forêt partagée de x par rapport à G.

En fait, au cours du calcul de σ, il est possible de savoir pour chaque production dans quel contexte elle est utilisée. On peut donc enrichir l'application de décoration : en faire une application depuis l'ensemble des instances de règles de G dans A.

Ainsi nous avons montré qu'en prenant l'identité comme application de décoration enrichie, et dans l'interprétation des forêts d'analyse, le calcul de σ produit la forêt d'analyse de x par rapport à G.

3.5.3 Applications

Il existe plusieurs formalismes grammaticaux pour les langues naturelles, mais on retrouve toujours un squelette indépendant du contexte, sur lequel se greffe une information contextuelle. Nous avons présenté les Definite Clause Grammars sous la forme d'un demi-anneau de façon à instancier nos algorithmes génériques. Dans le cas où la grammaire non-contextuelle sous jacente n'est pas cyclique, il est même possible de représenter le domaine des termes comme un demi-anneau et de proposer une application de décoration de la forêt partagées dans le demi-anneau des termes. Dans un tel domaine, les calculs ne sont plus cubiques comme avec les probabilités, ils ne sont même plus polynomiaux.

3.5.4 Analyse syntaxique par formes arborescentes

Participant : Jean-Marie Larchevèque

L'objectif est d'analyser des énoncés en langue naturelle de façon à permettre une analyse sémantique incrémentale, de tolérer des variations dans l'ordre des mots, ainsi que des erreurs ne gênant pas la compréhension.

Le principe général consiste à partir d'une grammaire non contextuelle, et à générer pour chaque catégorie syntaxique terminale (nom, adjectif, etc.) et pour certaines catégories non terminales (comme le syntagme nominal) des formes arborescentes (tree patterns). Ces formes ont des valences dont la fonction est la même que celle des liens dans les "link grammars" (Sleator et Temperley, 1995).

L'analyse se fait de gauche à droite sans lecture anticipée (lookahead). Lorsqu'un mot est rencontré, les formes arborescentes associées sont intégrées à la liste courante de formes arborescentes, soit par simple concaténation à la liste (initialement vide), soit par combinaison avec des formes déjà présentes. La combinaison joue le rôle de la réduction en analyse LR.

3.6 Le Poste de Travail Informationnel

La nécessité de trouver des débouchés applicatifs à nos travaux, ainsi qu'un intérêt de l'équipe pour les nouveaux média (principalement CDROM et Internet) dont le rôle économique, social et culturel va croissant, nous a naturellement amenés à nous impliquer dans diverses actions dont nous espérons à terme des synergies avec nos compétences en analyse syntaxique et déduction décrites plus haut, ainsi qu'avec nos compétences plus anciennes en génie logiciel et traitement de documents structurés.

Ce travail plus applicatif présente deux volets complémentaires, qui concernent d'une part la conception et le développement d'outils pour maîtriser un support matériel des documents qui est 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 efforcant 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. Nous commençons par décrire une étude de cas, le logiciel Astrolabe, qui illustre la variété des problèmes à traiter, dont beaucoup se retrouvent dans d'autres applications.

3.7 Cartographie de site Web

Participants : Cédric Billaud, Bernard Lang

Dans le cadre d'une coopération d'un an avec le Consortium W3, nous avons développé un prototype d'outil de cartographie de site W3, appelé Astrolabe [61]. Cet outil explore et ramène localement l'ensemble des pages accessibles depuis une page origine, et satisfaisant une contrainte simple, fixée par l'utilisateur, pour limiter l'espace de recherche (par exemple, les pages d'un même site).

Les pages ramenées et leurs liens sont analysées automatiquement de façon à produire un document HTML qui est une description, de l'ensemble des pages explorées. Cette table des matières permet à un utilisateur de se rendre directement sur les pages du site qui lui semblent pertinentes sans être obligé de naviguer, ou bien d'effectuer une exploration systématique, ce qui est en général difficile avec les hyperdocuments. Cet outil peut également servir à la maintenance des sites en permettant d'en documenter automatiquement la structure.

Ce premier projet lié au Web, et plus généralement à l'Internet, car nous acceptons les documents correspondant à d'autres protocoles comme FTP ou Gopher, nous a permis d'identifier divers problèmes qui nous paraissent essentiels pour le développement d'outils de recherche et de gestion de l'information :

- recherche systématique des pages accessibles à partir d'une adresse initiale ;
- *analyse de similarité* de documents ;
- *analyse des liens hypertextuels* d'un ensemble de pages ;
- *extraction d'information*, par analyse structurelle ou linguistique ;
- *uniformisation des interfaces homme-machine* par l'utilisation des standards du World Wide Web.

3.8 Technologie de l'Internet

Participants : Bernard Lang, François Rouaix du projet Cristal

Le développement d'applications d'analyse et de gestion de l'information électronique nous impose bien entendu de maîtriser l'accès au support de cette information, en particulier sur l'Internet, que ce soit pour accéder aux exemples d'informations à traiter, ou parce que les outils à développer devront en particulier être mis en œuvre dans le cadre des réseaux Internet ou intranet.

Ces travaux sont menés en collaboration avec François Rouaix du projet Cristal, qui est pour le moment l'auteur de l'essentiel des réalisations.

3.8.1 Architecture des clients internet

Plusieurs expériences nous ont montré l'intérêt d'utiliser les outils d'affichage du Web, à ce jour fondés principalement sur le langage HTML, pour réaliser des interfaces simples, satisfaisant pour une bonne part à des paradigmes d'interaction largement diffusés, et facilement portables ou utilisables à distance. Il s'avère cependant que les brouteurs ("browsers") commerciaux les plus répandus ne permettent cela qu'avec quelque difficulté car ils n'autorisent pas simplement l'activation de programmes locaux sans passer par toute la machinerie des serveurs. Dans le cadre du développement du brouteur MMM, une telle possibilité a été ajoutée qui nous a permis d'expérimenter dans de bonnes conditions le développement et l'usage d'interfaces en HTML.

Ceci nous a conduit à repenser l'architecture des clients web, en ce sens qu'il nous est apparu que ces clients comportent essentiellement deux aspects relativement indépendants : un outil sophistiqué d'affichage et d'interaction (les formulaires HTML) d'une part, et un outil de gestion de l'information (communication avec l'Internet, gestion d'historique, cache, ...).

Faute de pouvoir développer la composante graphique d'une architecture ainsi modularisée, nous avons réalisé le proxy personnel *V6* [56] qui peut être vu comme un mécanisme destiné à découpler l'interaction utilisateur, gérée par un brouteur quelconque (commercial ou autre), des fonctionnalités d'accès, de traitement et de gestion des documents électroniques. Alternativement, on peut considérer que V6 est un moyen uniforme de greffer sur un brouteur quelconque une extension de ses fonctionnalités.

Le principe de fonctionnement de V6 est en fait très simple : il filtre et éventuellement modifie les requêtes émises vers le réseau, ainsi que les réponses correspondantes, et au besoin les détourne pour un traitement local indépendant du réseau. Ce mécanisme permet toutes sortes d'applications : caches locaux programmables et indexables, caches diffusables sur CDROM, contrôle de fonctions locales (dont la gestion et la configuration de V6 lui-même), analyse automatique des documents en transit pour les indexer, les modifier, etc., travail coopératif, systèmes PICS, amélioration des communications, contrôles de sécurité, extensions de protocoles existants, ...

Notre intention est de faire de V6 le support technologique pour la composition et la coopération des outils que nous comptons développer et/ou mettre en œuvre pour réaliser un poste de travail informationnel.

3.9 Technologie des CDROMs

Participant : Bernard Lang

Les deux principaux supports de diffusion de l'information électronique sont l'Internet et le CDROM. Ils sont de notre point de vue complémentaires, le premier étant particulièrement adapté aux informations limitées en volume ou en durée, tandis que le second se prête mieux à la diffusion de masse de gros volumes et d'informations pérennes, ou tout simplement à la sauvegarde de la base d'information d'une personne ou d'un groupe.

Nous avons tout d'abord expérimenté les techniques de production de CD-ROM sous Unix à des fins strictement utilitaires (sauvegardes). Ceci nous a conduit à examiner les problèmes de structure des systèmes de fichiers sur CD-ROM (normes ISO 9660 et Rockridge), et à mieux connaître les logiciels de prématriçage (création sur disque de l'image binaire du futur CD-ROM).

Cette expérience, dont nous espérons qu'elle pourra être utile à l'institut pour ses propres productions, s'est prolongée dans la réalisation de *CD-Punch*, un outil (prototype) simple d'exploration des systèmes de fichiers ISO-9660. L'intérêt de cet outil réside en fait surtout dans sa réalisation, car ce fut notre première expérience, très concluante, de production d'un outil logiciel dont l'interface est entièrement réalisée par la technologie Web, et offrant un très haut niveau d'intégration de la mise en œuvre de l'outil et de l'accès à l'aide en ligne et à la documentation, pour un coût de développement dérisoire. Nous pensons que cette approche peut se révéler fructueuse, non seulement pour réaliser des outils plus

conviviaux, mais aussi, dans bien des cas, pour assurer la formation à diverses techniques informatiques (voire à d'autre domaines).

3.10 Bibliothèques électroniques.

Participant : Roland Dachelet

ATOLL participe au projet Aquarelle dont l'objet est de permettre un accès distribué aux ressources d'information concernant le patrimoine muséologique et architectural détenues par les organisations culturelles européennes. Notre participation concerne cet aspect de l'interopérabilité sémantique du système qui a trait aux ressources terminologiques multilingues à mettre en œuvre ou à produire.

Nous procédons dans un premier temps à une étude des besoins des utilisateurs concernant aussi bien l'utilisation de ces ressources lors de l'interrogation que la plateforme requise pour la production coopérative de celles-ci.

Pour ce faire, nous élaborons une typologie de ces ressources multilingues. D'autre part, nous esquissons une description de la méthodologie suivie pour leur élaboration.

4 Actions industrielles

4.1 Avec le CNET : projet VADA

Participants : Pierre Boullier, Bernard Lang, Frédéric Tendeau, Éric Villemonte de la Clergerie

Une collaboration a été lancée avec le CNET dans le cadre de ses consultations thématiques (thème : "Traitement Automatique de la Langue Naturelle"). Cette convention concerne le projet intitulé "VADA : Analyse Syntaxique avec **V**aluation d'**A**ttributs dans un **D**emi-**A**nneau".

Le travail est effectué principalement dans le cadre de la thèse de Frédéric Tendeau, mais concerne également les travaux de Pierre Boullier (utilisation du système SYNTAX comme support expérimental) et de Éric Villemonte de la Clergerie (analyse des flots d'informations dans les automates à pile).

Sur le plan scientifique, notre objectif est de déterminer un cadre théorique unifié et simple pour la prise en compte uniforme de divers systèmes d'attributs dans les formalismes syntaxiques et dans les outils qui en dérivent, dans le cas assez courant où ces attributs sont valués dans un domaine possédant une structure de demi-anneau. L'un des objectifs pratiques est d'utiliser les résultats obtenus pour déterminer une architecture d'analyseur modulaire, permettant de prendre en compte efficacement divers systèmes d'attributs. La réalisation d'un prototype est prévue.

4.2 Astrolabe : Outil d'aide à la navigation sur WWW

Participants : Cédric Billaud, Bernard Lang

Mots-clés : World Wide Web, navigation, interface homme-machine.

Dans le cadre d'une coopération avec le Consortium W3, nous avons réalisé le prototype d'Astrolabe, un outil de cartographie de site WWW. L'objet de ce travail est de produire automatiquement un outil qui analyse les pages d'un site de façon à produire un document HTML qui en sera la description, la table des matières, et permet à un utilisateur de se rendre directement sur les pages du site qui lui semblent pertinentes à son besoin, sans être obligé de naviguer à travers les pages intermédiaires. Cet outil peut aussi servir à la maintenance des sites en permettant d'en documenter automatiquement la structure.

4.3 Acquisition de données en ligne

Participants : Roland Dachelet, Bernard Lang

Nous sommes en train d'établir avec la société Air France une collaboration qui portera sur l'acquisition de données en ligne. Cette collaboration comportera deux parties : l'une méthodologique à partir d'une analyse de cas pour définir le besoin, et l'autre technique par identification, implantation puis validation de solutions logicielles.

4.4 Gestion de mémoire de masse sur CDROM

Participant : Bernard Lang

Cette collaboration, en cours d'établissement avec la société ASSTEC-2, porte sur une activité de conseil et de suivi technique pour la réalisation d'un logiciel de gestion de mémoire de masse sur jukebox de CDROM muni d'un graveur. Cette collaboration devrait nous permettre de mieux maîtriser le support CDROM, ainsi que d'avoir une meilleure connaissance des techniques de stockage de masse nécessaires pour la réalisation de robots explorateurs, voire de bibliothèques personnelles. Ce projet a en outre l'intérêt de devoir être réalisé entièrement à base de logiciels libres (en raison de contraintes de réalisation souhaitées par la société ASSTEC-2) qui sont précisément les logiciels que nous utilisons dans le projet.

4.5 Collaboration avec Xerox Grenoble

Participants : Pierre Boullier, Éric Villemonte de la Clergerie, Roland Dachelet, Bernard Lang, Frédéric Tendeau

Nous avons commencé à établir une collaboration avec Marc Dymetman, qui porte sur le développement de l'utilisation des techniques d'analyse tabulaires pour la traduction dans des sous-langages techniques. L'intérêt de Xerox pour ces techniques est qu'elles permettent de préserver de façon compacte les ambiguïtés (grâce au concept de forêt partagée), et de mieux formaliser le processus de désambiguïsation.

5 Actions nationales et internationales

5.1 Collaborations internes à l'Inria

Comme le montrent les sections précédentes, nous avons une importante collaboration avec le projet CRISTAL, principalement avec François Rouaix pour ce qui concerne le développement d'outils d'interaction avec l'Internet (MMM, V6, robot d'exploration). Nous collaborons également avec Pierre Weis en ce qui concerne la réalisation de CDROMs, et plus particulièrement la reproduction de sites Web sur CDROM.

Nous avons également des relations avec l'action de développement Dyade sur ces sujets.

Nous (principalement Roland Dachelet) travaillons avec l'action de développement Mediaculture dans le cadre du projet Européen Télématics AQUARELLE.

Nous entretenons des relations avec le projet CALLIGRAMME (INRIA-Lorraine), en particulier pour comparer nos approches respectives de l'expression des problèmes syntaxiques.

5.2 Actions nationales

Nous avons une convention de formation doctorale, avec l'Université d'Orléans où plusieurs membres du projet ont déjà enseigné. La thèse de F. Role se fera dans le cadre de cette convention.

Nous préparons avec C. Sedogbo de l'ENST Bretagne, et ultérieurement d'autres partenaires, un projet de collaboration pour la réalisation d'outils d'aide à la maintenance de sites Web multilingues.

Roland Dachelet travaille avec Jacques Virbel de l'IRIT (Toulouse) et Elsa Pascual sur les questions de métalangage concernant la structure des documents, ainsi que sur les bibliothèques électroniques.

Bernard Lang participe aux travaux de la Commission Beaussant pour la définition d'un code de l'Internet.

Éric Villemonte de la Clergerie a fait partie du comité de programmation pour la conférence LACL'96 (Logical Aspects of Computationnal Linguistics) qui s'est tenue fin Septembre 96 à Nancy. Bernard Lang à participé au comité d'organisation de cette conférence.

5.3 Actions internationales

Nous participons au projet Européen Télématics AQUARELLE qui concerne la mise en réseau des ressources muséologiques d'un certain nombre de musées européens. Notre rôle consiste à définir un modèle d'harmonisation des schémas d'indexation utilisé par chacun des participants et pour les divers media considérés. Nous envisageons ce travail en relation avec notre intérêt pour les techniques de classification fondées sur les métadata. En effet, le projet Aquarelle devrait fournir des exemple intéressant et originaux d'utilisation de ces approches.

Notre participation aux projets ERCIM de bibliothèque électronique (projets SAMOS et DELOS) s'est pour le moment limitée à une participation aux réunions du projet et à la maintenance d'une page d'information WWW. Cependant nous envisageons une participation plus active dont les modalités restent à déterminer, vraissemblablement sur des thèmes voisins de ceux considérés dans le projet Aquarelle.

Nous collaborons depuis plusieurs années sur les techniques d'évaluation tabulaires avec Manuel Vilares Ferro du groupe LFCIA (dirigé par J.L. Freire Nistal) à l'Université de la Corogne (Espagne). Miguel Alonso Pardo est un thésard de cette université venu passer un an dans notre équipe. Pour formaliser cette collaboration, nous avons déposé conjointement une proposition intitulée CATALINA dans le cadre des Actions Intégrées franco-espagnoles PICASSO 1997, qui a été acceptée.

Nous avons établi des contacts avec le projet Gate (dirigé par Yorick Wilks) de l'université de Sheffield. Hamish Cunningham, membre de ce projet, est venu nous le présenter, et nous comptons utiliser l'infrastructure logicielle modulaire qu'ils ont développée pour tester nos techniques d'analyse syntaxique.

Enfin, nous avons participé au contrat INRIA-NSF impliquant le projet INRIA LOCO et le professeur Dale Miller de l'Université de Pennsylvanie. Notre participation concernait les travaux d'Alain Hui Bon Hoa sur le langage λProlog.

R. Dachelet a été membre du comité de programme de NLDB'96 (Natural Language and Databases) qui s'est tenue à Amsterdam du 26 au 28 Juin 1996.

6 Diffusion des résultats

6.1 Actions d'enseignement

Eric de la Clergerie a encadré Miguel Alonso Pardo, étudiant en première année de thèse de l'université de la Corogne (Espagne). Miguel travaille sur l'interprétation et l'extension de différents algorithmes tabulaires proposés pour les grammaires d'arbres adjoints (TAG) et linéaires indexées (LIG).

Eric de la Clergerie a assuré des vacations à l'Ecole Polytechnique pour des travaux dirigés dans la mineure d'informatique "Programmation Logique par Contraintes" de François Fages.

Bernard Lang a été rapporteur pour la thèse de PhD de Mohamed Younis (New Jersey Institute of Technology, Juin), la thèse de Doctorat de Laurent Angeli (Université de Nice, Décembre), et le Mémoire d'Habilitation de Christian Jacquemin (Université de Nantes, Janvier 1997).

Bernard Lang a encadré le mémoire d'Ingénieur CNAM de Cédric Billaud.

Frédéric Tendeau a été chargé des cours et travaux dirigés deux unités de valeur en MIAGE à l'Université de Paris XII Val-de-Marne : *Techniques de compilation*, et *Programmation en Ada*.

6.2 Participation à des colloques

Roland Dachelet a participé, sur financement ERCIM, au séminaire OCLC-UKOLN (Warwick, Avril) concernant les metadata dans le contexte des bibliothèques électroniques. Il a également participé au titre du projet Aquarelle au séminaire organisé par la Museum Documentation Association (Oxford, Septembre) concernant les ressources terminologiques dans le domaine de l'information culturelle.

On se reportera à la bibliographie pour avoir la liste de nos autres participations.

6.3 Conférences invitées, tutoriels, cours, etc.

Nous avons présenté en janvier, au CNET-Lannion, nos travaux sur l'analyse tabulaire et stochastique.

Éric Villemonte de la Clergerie a présenté ses travaux lors de rencontres Franco-Portugaises qui se sont tenues à l'INRIA en Mai, et à l'université d'Orléans en Novembre.

Roland Dachelet a été invité à l'atelier « Texte et Communication » sur les textes de type consigne organisé par PRESCOT (Programme de Recherches en Sciences Cognitives de Toulouse) à Mons (France) du 30 Septembre au 3 Octobre 1996. Il a également présenté une communication sur la mise en œuvre et la production des thesauri multilingues dans le projet Aquarelle à Rome le 25 septembre 1996.

7 Publications

Articles et chapitres de livre

[53] F. BARTHÉLEMY, E. VILLEMONTE DE LA CLERGERIE, « Information flow in Tabular Interpretations for generalized Push-Down Automata », *Theoretical Computer Science*, 1997, à paraître.

[54] F. TENDEAU, « Computing abstract decorations of parse-forests using dynamic programming and algebraic power series », *Theoretical Computer Science*, 1997, à paraître.

Communications à des congrès, colloques, etc.

[55] P. BOULLIER, « Another Facet of LIG parsing », *in : Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL'96)*, p. 87–94, 1996. also INRIA Research Report 2858 (http://www.inria.fr/RRRT/RR-2858.html).

[56] B. LANG, F. ROUAIX, « The V6 Engine », *in : WWW5 Workshop: Programming the Web - a search for APIs*, March 1996. http://pauillac.inria.fr/ lang/Papers/v6/.

[57] M. VILARES FERRO, M. A. ALONSO PARDO, D. CABRERO SOUTO, « Autómatas Lógicos y Lenguaje Natural », *in : Proc. of IBERAMIA'96, V Iberoamerican Conference on Artificial Intelligence*, Cholula, Puebla, Mexico, Novembre 1996.

[58] M. VILARES FERRO, M. A. ALONSO PARDO, D. CABRERO SOUTO, « An Experience on Natural Language Parsing », *in : Lenguajes Naturales y Lenguajes Formales*, C. Martín Vide (réd.), *XII*, PPU, p. 555–562, Barcelone, Espagne, Septembre 1996.

[59] M. VILARES FERRO, M. A. ALONSO PARDO, « An LALR Extension for DCGs in Dynamic Programming », *in : Proc. of APPIA-GULP-PRODE'96 Joint Conference on Declarative Programming*, P. Lucio, M. Martell, M. Navarro (réd.), p. 79–88, San Sebastián, Espagne, Juillet 1996.

[60] M. VILARES FERRO, M. A. ALONSO PARDO, «Towards Analyzers Based on Efficient Logical Frames», *in : II International Conference on Mathematical Linguistics (ICML'96): abstracts*, *Report 7/96, Grup de Recerca en Lingüística Matemática i Enginyeria del Llenguatge, Universitat Rovira i Virgili*, p. 97–98, Tarragona, Espagne, June 1996.

Rapports de recherche et publications internes

[61] C. BILLAUD, *Outils d'aide à la navigation sur le World Wide Web, rapatriement et analyse d'un site*, Mémoire, Conservatoire National des Arts et Métiers, Paris, Octobre 1996, Mémoire d'Ingénieur CNAM.

[62] P. BOULLIER, «Another Facet of LIG parsing (extended version)», *Rapport de recherche n°2858*, INRIA, Avril 1996, http://www.inria.fr/RRRT/RR-2858.html.

[63] R. DACHELET, «Multilingual thesauri : typology and production practices», *Final technical report*, Décembre 1996, Task WP7.1, INRIA-Aquarelle.

[64] B. LANG, P. BOULLIER, «Projet VADA», *Notes de séminaire*, INRIA, Novembre 1996, Convention CNET 95-1B 030.

[65] E. VILLEMONTE DE LA CLERGERIE, «The canonical Dynamic Programming interpretation of stack-like computations», *rapport de recherche*, May 1996, Publication interne.

Divers

[66] B. LANG, «Code de Bonne Conduite sur Internet - contribution au debat», Juillet 1996, http://trappiste.hmg.inpg.fr/is/Documents/debat/IsocFR.DB001.html.

Plusieurs références citées sont disponible électroniquement à l'URL suivante :

http://pauillac.inria.fr/atoll/publioscope.html

8 Abstract

The Atoll research group was created around a competence in syntax analysis and tabular evaluation for logical formalisms. This work, both theorical and practical, attempts to explore uniformly a continuum of formalisms ranging from context-free languages to Horn clauses and even beyond to higher-order logic and constraint logic. It also includes many formalisms relevant to natural language processing. Our main objective is to develop algorithms that are efficient both in execution time and in the representation of their results, with a particular emphasis on shared parsed forests and other disjonctive representations allowing a factorized processing of ambiguity.

On the applicative side, we aim at developing an *information workstation*, i.e. an environment to assist information workers in discovering, acquiring, filtering, memorizing, organizing and processing information, and especially linguisic information. Our design is based on the use of a personnal Web proxy as the support for integrating the various information tools, either developed by us, or taken among available existing tools. We intend to cover the major media, notably the Internet and CDROMs.