INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Atoll

Atelier d'Outils Logiciels pour le Langage naturel

Rocquencourt

THEME SYM

Activity Report

2004

Table of contents

1. Team

Head of project team
Éric Villemonte de la Clergerie [CR]

Vice-head of project team
Pierre Boullier [DR]

Administrative assistant
Nadia Mesrar [AJT]
Emmanuelle Grousset [CDD - until June 30]

Staff members Inria
Bernard Lang [DR]
Philippe Deschamp [CR]
François Thomasset [DR]

External members
François Barthélemy [Maître de conférences, CNAM]
Areski Nait Abdallah [Professeur, Univ. of Brest]

Visiting scientist
Francisco Riberra [June and December 2004, University of La Coruña]

Ph. D. student
Benoît Sagot [Détachement du corps des Télécoms]

Technical staff
Lionel Clément [until July 31st]
Guillaume Rousse

Student intern
Tatiana Samoussina [Engineer Internship, École Polytechnique, Summer 2004]
Mehdi Ben Hmida [DEA, University of Paris Dauphine, Spring/Summer 2004]
Alexandra Mounier [CNAN Engineer Internship, starting September 1st]

2. Overall Objectives

2.1. Tools for Natural Language Processing

Project-team ATOLL was formed by people with strong competences in Parsing, essentially acquired in the context of Programming Language Compilation. This competence is now applied to *Natural Language Processing* (NLP), mainly in its parsing aspects but evolving toward more semantic aspects. Besides promising industrial applications, this domain of research also offers many scientific problems that may benefit from a strong formal and algorithmic approach.

In our exploration of fundamental parsing techniques, we focus on the use of tabular techniques, almost mandatory to efficiently handle the ambiguities inherent in any human language. The genericity of our techniques is also an asset because of the large diversity of grammatical formalisms. We also explore more recent and important issues related to robustness. We validate these techniques through the development of two prototype environments (SYNTAX and DyALog) that may be used for building and running parsers.

However, a parser is only one component of a linguistic processing chain that requires other tools and also linguistic resources like lexicons. Besides interesting software engineering issues, designing and running such a chain raises questions about the availability and reusability of linguistic resources. These observations motivate our interest about the normalization, distribution and exploitation of linguistic resources. In particular, we explore how the production cost of some linguistic resources could be reduced by using automatic or semi-automatic acquisition methods, possibly based on parsing corpora with our parsers. Obviously, such an approach is also an opportunity to test ATOLL's tools on a larger scale. We also believe that the use of well-designed tools for linguists can speed up the hand-crafting of linguistic resources as we try to promote with MetaGrammars, a level of abstraction above grammars allowing easier linguistic descriptions.

From a wider point of view, the acquisition of linguistic resources share some common aspects with the extraction of information from corpora or documents, a rapidly growing domain of research and applications. Indeed, the huge development of the World Wide Web and the recent emergence of the notion of Semantic WEB plead for accessing information rather than simply accessing raw documents. As a consequence, tools are needed for extracting information from documents.

The diversity of the tools and resources needed to process natural language overcomes the capacities of project-team ATOLL. Therefore, we favor partnerships for reusing existing tools and resources or for developing new ones in common. An important issue, related to these cooperations and also very present in the NLP community, concerns the standardization and reusability of these tools and resources.

While marginal within ATOLL but nevertheless related to better accessing linguistic resources and tools, a reflexion is led by Bernard Lang on the issues of free access to scientific and technical resources, issues whose scientific, economical, and political interest becomes more and more visible.

3. Scientific Foundations

3.1. Grammatical formalisms

Keywords: *NLP*, *Parsing*, *computational linguistics*, *dynamic programming*, *logic programming*.

Participants: Pierre Boullier, Éric Villemonte de la Clergerie.

CFG *Context-Free Grammars*

DCG *Definite Clause Grammars*

TAG *Tree Adjoining Grammars*

TIG *Tree Insertion Grammars*

LIG *Linear Indexed Grammars*

LFG *Lexical Functional Grammars*

HPSG *Head-driven Phrasal Structure Grammars*

RCG *Range Concatenation Grammars*

MCG *Mildly Context-sensitive Grammars*

LPDA *Logical Push-Down Automata*

2SA *2-Stack Automata*

TA *Thread Automata*

Dynamic Programming Algorithmic method based on dividing a problem into elementary sub-problems whose solutions are tabulated to be reused whenever possible

This theme explores the use of generic parsing techniques covering a large continuum of NLP grammatical formalisms, focusing especially on efficient handling of ambiguities.

3.1.1. From programming languages to linguistic grammars

The shift from programming language grammars to NLP grammars seriously increases complexity and requires ways to handle the ambiguities inherent in every human language. It is well known that these ambiguities are the sources of many badly handled combinatorial explosions.

Furthermore, while most programming languages are expressed by (subclasses) of well-understood context-free grammars (CFGs), no grammatical formalism has yet been accepted by the linguistic community for the description of human languages. On the contrary, new formalisms (or variants of older ones) appear constantly. Many of them may be classified into the two following large families:

Mildly context-sensitive formalisms : They manipulate possibly complex elementary structures with enough restrictions to ensure the possibility of parsing with polynomial time complexities. They include, for instance, Tree Adjoining Grammars (TAGs) with trees as elementary structures, Linear Indexed Grammars (LIGs), and Range Concatenation Grammars (RCGs).

Unification-based formalisms : They combine a context-free backbone with logic arguments as decoration on non-terminals. Most famous representatives are Definite Clause Grammars (DCGs) where PROLOG powerful unification is used to compute and propagate these logic arguments. More recent formalisms, like Lexical Functional Grammars (LFGs) [28] and Head-Driven Phrasal Structure Grammars (HPSGs) [30] rely on more expressive Typed Feature Structures (TFS) [26] or constraints.

The above-mentioned characteristics may be combined, for instance by adding logic arguments or constraints to non-terminals in TAGs. We should also mention that we also concur to this large diversity of formalisms with the introduction of RCGs (Section 6.1).

However, despite this diversity, most formalisms take place in a so-called **Horn continuum**, i.e. a set of formalisms with increasing complexities, ranging from Propositional Horn Clauses to first-order Horn Clauses (roughly speaking equivalent to PROLOG), and even beyond.

This observation motivates our exploration of generic parsing techniques covering this continuum, through two complementary approaches. Both of them use dynamic programming ideas to reduce the combinatorial explosions resulting from ambiguities :

Multi-pass approach : Parsing is broken into a sequence (or cascade) of parsing passes, of (practical or theoretical) increasing complexities, each phase guiding the next one ;

Global Approach : It is mainly based on the use of Push-Down Automata [PDA] to describe parsing strategies for complex formalisms.

These two approaches enrich each other: studying some specificities observed for the multi-pass approach has triggered theoretical advances; conversely, well-understood and identified theoretical concepts have suggested a widening of the scope of the multi-pass approach.

3.1.2. Multi-pass approach

Programming languages processing is usually broken into several successive phases of increasing complexity : lexical analysis, parsing, static semantics,... The decomposition is motivated by theoretical and practical reasons. The finite state automata (FSA) that model lexical analysis are very efficient but do not have enough expressive power to describe the syntax, which requires, at least, Context-Free Grammars. Similarly, CFGs are not powerful enough to describe some contextual phenomena needed in static semantics. Beside a better efficiency (each phase being handled with the best level of complexity), decomposing increases modularity.

The multi-pass approach for NLP results from similar observations. We try to identify and capture, within adequate grammatical formalisms, subparts of grammars which can guide the remaining processing. For instance, we observe that most formalisms found in the Horn continuum are structured by a non-contextual backbone. This backbone may be first parsed with a very efficient and generic non-contextual parser, namely

SYNTAX (cf. 5.1). More formalism-specific treatment can then be applied to check additional constraints, as done this year for LFG decorations (cf. 6.1).

3.1.3. Global approach

The multi-pass approach is less easy to implement when there is no obvious decomposition, for instance when the CF backbone of a formalism cannot be extracted (as in PROLOG) or when the possible phases would be mutually dependent (for instance, when some constraints have a strong impact on the processing of the CF backbone). A more global approach is then needed where constraints and parsing are handled simultaneously.

This very general approach relies on abstract Push-Down Automata formalisms that may be used to describe parsing strategies for various unification-based formalisms [8]. The notion of stack allows us to apply dynamic programming techniques to share elementary sub-computations between several contexts : the intuitive idea relies upon temporarily forget information found in stack bottoms. Elementary sub-computations are represented in a compact way by *items*. The introduction of 2-Stack Automata [2SA] allowed us to handle formalisms such as TAGs and LIGs [9]. More recently, *Thread Automata* (TA) [7] have been introduced to cover mildly-context sensitive formalisms such as Multi-Component TAGs (MC-TAGs).

This global approach may be related to *chart parsing* [27] or *parsing as deduction* [29] and generalizes several approaches found in Parsing but also in Logic Programming. The DYALOG system (cf. 5.2) implements this approach for Logic Programming and several grammatical formalisms.

3.1.4. Shared parse and derivation forests

Both previously presented approaches share several characteristics, for instance the use of dynamic programming ideas and also the notion of *shared forest*. A shared forest groups in a compact way the whole set of possible parses or derivations for a given sentence. Formally, a shared forest may be seen as a grammar or a logic program [6]. For instance, parsing with a CFG may lead to an exponential (or unbounded) number of parse trees for a given sentence, but the parse forest remains cubic in the length of the sentence and is itself equivalent to a CFG (as an instantiation of the original CFG by intersection with the parsed sentence). Moreover, these shared forests are natural intermediary structures to be exchanged from one pass to the next one in the multi-pass approach. They are also promising candidates for further linguistic processing (semantic processing, translation, ...). One can also relatively easily extract dependency information between words from these forests, as done this year in the context of the parsing evaluation campaign EASY. Disambiguation algorithms can also be applied on such shared structures (cf. 6.1 and 6.2).

3.2. Linguistic Infrastructure and Normalization

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

We are interested in the many issues related to the installation of a whole linguistic processing chain, in particular for accessing and representing the needed linguistic resources (cf. 6.5).

To facilitate the installation of such linguistic chains, we develop two systems to build parsers, namely SYNTAX (cf. 5.1) and DYALOG (cf. 5.2). We also develop and distribute several linguistic components (cf. 5.4).

Because we realized that diffusing or reusing tools and resources is not really possible without some standardization, ATOLL is involved in on-going national and international efforts to normalize linguistic resources, using XML-based representations (cf. 7.1). This decision follows preliminary experimentations we have conducted to normalize TAGs and shared forests.

3.3. Resource acquisition and crafting

Participants: Éric Villemonte de la Clergerie, Benoît Sagot, Lionel Clément.

MG *MetaGrammars*

Linguistic resources are scarce and expensive, because they are difficult to build, especially when hand-crafted. This observation motivates us to investigate methods to automatically or semi-automatically acquire linguistic resources. In particular, we would like to reach some bootstrap level where parsing corpora may be used to enrich lexica that may themselves be used for better parsing.

Preliminary experiments have been conducted during the now ended ARC (Action de Recherche Concertée) RLT « Linguistic resources for TAGs » and we are currently working on processing botanical corpora (cf. 6.8).

For hand-crafted resources, we try to design adequate tools and adequate levels of representation for linguists. For instance, we are currently involved in developing grammars through a more abstract notion of *MetaGrammar* (MG) (cf. 6.3). Introduced by [25], a MetaGrammar allows the linguist to focus on a modular description of the linguistic aspects of a grammar, rather than focusing on the specific aspects of a given grammatical formalism. Translation from MGs to grammatical formalisms such as TAG or LFG may be automatically handled (cf. 5.3). Graphical environments can be used to design MGs and their modularity provides a promising way for sharing the description of common linguistic phenomena across human languages [4].

The constitution of resources such as lexica or grammars raises the issues of the evaluation of these resources to assess their quality and coverage. For this reason, ATOLL has been deeply involved during 2004 in the Parsing Evaluation campaign EASY (cf. 6.7). We have also started investigating different kinds of feedback mechanisms to detect problems when using resources (unknown words, error mining, ...).

4. Application Domains

4.1. Applications

Computational Linguistics offers a wide range of potential applications, especially with the emerging of information systems. More specifically for ATOLL, one can (non exhaustively) list the following application domains:

Grammatical checking Parsing is used to detect grammatical errors and to suggest corrections. Tabulation-based parsing techniques present a great potential for grammatical checking because they allow the exploration of many alternatives (for correcting errors) without combinatorial explosions.

Knowledge acquisition Linguistic (and statistical) techniques may be used to extract knowledge from corpora, ranging from a simple terminological list of words to more complex semantic networks with concepts and relations. In this continuum, we also find lexicons, thesaurus, and ontologies. We strongly believe that this domain can benefit from more sophisticated parsing-based techniques.

Text mining and Questions/Answers Parsing and possibly semantic or pragmatic processing may be used to extract precise information from a document, for instance to feed a (knowledge) database or to answer questions formulated by users.

Among these various application domains, ATOLL focuses its efforts on knowledge acquisition and text mining, in particular through the action BIOTIM for processing botanical corpora (cf. 7.2).

5. Software

5.1. System Syntax

Participants: Pierre Boullier [maintainer], Philippe Deschamp.

The (not yet released) version 6.0 of the SYNTAX system has been extended and now includes SXSPELL, a spelling error corrector and SXLFG a Lexical Functional Grammar processor which is divided in two main parts (Section 6.1) : the constructor part which compiles the LFG specifications and the parser part which processes a source text w.r.t. these compiled specifications.

This version of SYNTAX runs on various 32bit platforms such as Linux, Solaris, HP/UX and Windows. A first 64-bit port has been made for HP/UX. Optimized ports for 32-bit compatible 64-bit architectures are currently in progress, including 64-bit x86 running Linux and IBM G5 running Mac OS X.

Release 3.9 essentially handled deterministic CFGs of type LALR(1). Release 6.0 extends it by including RLR (an extension of LR parsing strategy in which an unbounded number of look-ahead terminal symbols may be used, if necessary), non-deterministic CF parsers based upon push-down automata of type LR, RLR or left-corner, and a parser generator for Range Concatenation Grammars (RCGs), hence the leap inn numbers from 3 to 6.

5.2. System DyALog

Participant: Éric Villemonte de la Clergerie [maintainer].

DYALOG*: http://atoll.inria.fr Rubrique « Logiciels »*

DYALOG provides an environment to compile and execute grammars and logic programs. It is essentially based on the notion of tabulation, i.e. of sharing computations by tabulating traces of them. DYALOG is mainly used to build parsers for Natural Language Processing (NLP). It may nevertheless be used as a replacement for traditional PROLOG systems in the context of highly ambiguous applications where sub-computations can be shared.

The current release **1.10.6** of DYALOG is freely available by FTP under an open source license and runs on Linux platforms for x86 architectures.

The current release handles logic programs, DCGs (*Definite Clause Grammars*), FTAGs (*Feature Tree Adjoining Grammars*), FTIGs (*Feature Tree Insertion Grammars*) and XRCGs (*Range Concatenation Grammars* with logic arguments). Several extensions have been added to most of these formalisms such as intersection, Kleene star, and interleave operators. Typed Feature Structures (TFS) as well as finite domains may be used for writing more compact and declarative grammars. Cyclic terms are correctly handled by DYALOG.

C libraries can be used from within DYALOG to import APIs (`mysql`, `libxml`, ...) and modules with namespaces are now available.

DYALOG is largely used within ATOLL to build parsers but also derivative softwares, such as a compiler of MetaGrammars (cf. 5.3). It has also been used for building a parser from a large coverage French TIG/TAG grammar derived from a MetaGrammar. This parser has been used for the Parsing Evaluation campaign EASY (cf. 6.7).

DYALOG is also an essential component in the development of a robust Portuguese parser at the New University of Lisbon. It is occasionally used by several people at LORIA (Nancy), University of Coruña (Spain) and University of Pennsylvania.

5.3. MetaGrammar related tools

Participants: Éric Villemonte de la Clergerie [correspondant], François Thomasset.

MGCOMP*,* MGTOOLS*, and* FRMG*: http://atoll.inria.fr Rubrique « Catalogue »*

DYALOG (cf. 5.2) has been used to implement MGCOMP, a compiler of MetaGrammar (cf. 6.3). Starting from an XML representation of a MG, MGCOMP produces an XML representation of its TAG expansion.

The current version **1.4.1** is freely available by FTP under an open source license. It is used within ATOLL and (occasionally) at LORIA (Nancy) and at University of Pennsylvania.

The current version adds the notion of namespace, to get more compact and less error-prone meta-grammars. It also provides other extensions of the standard notion of MetaGrammar in order to generate very compact TAG grammars. These extensions include the notion of *Guarded nodes*, i.e. nodes whose existence and non-existence depend on the truth value of a guard, and the use of the regular operators provided by DYALOG on nodes, namely disjunction, interleaving and Kleene star.

The current version of MGCOMP has been used to compile a wide coverage MetaGrammar FRMG to get a grammar of around 100 TAG trees. Without the use of guarded nodes and regular operators, this grammar would have more than several thousand trees and would be almost intractable. FRMG has been packaged and is freely available.

To ease the design of metagrammars, a set of tools have been implemented by É. de la Clergerie and F. Thomasset, and collected in MGTOOLS (version **1.0.1**). This package includes a converter from a compact format to a XML pivot format, an Emacs mode for the compact and XML formats, a graphical viewer interacting with Emacs and XSLT stylesheets to derive HTML views.

5.4. Morphosyntactic processing tools

Participants: Éric Villemonte de la Clergerie, Lionel Clément, Guillaume Rousse, Benoît Sagot.

List of tools: http://atoll.inria.fr Rubrique « Catalogue »

ATOLL develops several tools that may be used for the first levels of linguistic processing preceding parsing, in particular morpho-syntax (cf. 6.5). They are freely available under open source licenses, keeping in mind that most of these tools are still beta versions.

LEXED (4.5.1) a C software originally developed by L. Clément to build efficient and compact lexica from lists of words (completed with additional information).

TOKENIZER (5.2.1) a C tokenizer originally developed by L. Clément for French that may be easily adapted for other languages. It can output an XML stream of tokens.

LINGPIPE (0.1.0) a small set of Perl modules originally developed by É. de la Clergerie to setup and configure a linguistic pipeline. The current version of lingpipe comes with a basic set of wrappers for the various linguistic tools we use for the morpho-syntactic processing of French (tokenizer, tagger, lexicon lookup, ...)

Guillaume Rousse has updated and completed most of these tools for BIOTIM. He has also done an important work to package them. Because of the growing number of tools developed and maintained by ATOLL, we have designed a small XML database which allows us to derive an online catalog.

Other alternate tools for segmentation and morphosyntactic processing have been developed this year (cf. 6.4) and should soon be distributed.

5.5. Lexicon Lefff

Participants: Lionel Clément, Benoît Sagot.

French morphological lexicon LEFFF*: http://www.lefff.net*

LEFFF is a freely available French morphological lexicon for verbs that has been automatically extracted from a very large corpus [13].

A (not distributed) new version of Lefff, used internally in ATOLL, covers all grammatical categories (not just verbs) and includes syntactic information (such as verb categorization frames).

6. New Results

6.1. Contextual Parsing

Keywords: *Context-sensitive grammatical formalisms*, *finite transducers*, *grammatical modularity*, *lexical functional grammars*, *polynomial parse time*, *range concatenation grammars*, *shared parse forests*.

Participants: Pierre Boullier, Benoît Sagot.

MCS *Mildly Context-sensitive Grammars*

RCG *Range Concatenation Grammars*

TAG *Tree Adjoining Grammars*

This year, our work mainly concentrates along three axes:

- the design and implementation of a spelling error corrector based on finite transduction techniques;
- the design and implementation of (a first version of) a Lexical Functional Grammar parser which concentrates on the sharing of identical computations;
- and the improvement of our Range Concatenation Grammars parser.

6.1.1. SXSPELL: A spelling error corrector

In the framework of a linguistic processing chain, one mandatory link is the spelling error corrector. The basis upon which such a corrector may rely is the well known regular languages formalism. However, implementations with real size static lexicon and numerous dynamic (i.e., for each word) spelling correction possibilities have been seldom attempted.

On the one hand we have the lexicon which may be seen as a finite automaton (FA) $\mathcal{F}$. Transitions between two consecutive states are performed on letters. A word w is in the lexicon iff there is a path through $\mathcal{F}$ which spells w, from its initial state to one of its final states. If such a complete path does not exist, w is assumed to be a spelling error of some (other) word(s) of the lexicon. The spelling error correction problem is the following: Find all the words in $\mathcal{F}$ which are *close to* w. One classical way to express this closeness is to define a distance between two words that we call *cost*. A word may be transformed into another word by sequences of atomic simple operations (insertion, deletion, substitution, swap, ...) on letters. Each such operation can by specified by a *spelling error correction rule* $\mathcal{R}_i$ of the form $\text{lhs}_i \to \text{rhs}_i, c_i, k_i$ in which lhs_i specifies a pattern (say substring in the simpler case) of the input word which can be changed into rhs_i at cost c_i. Moreover, if this spelling error correction rule is performed several times, an additional composition penalty cost of k_i is applied. This means that applying two simple corrections of cost c may cost more than $2c$. The total cost of a correction is the sum of the individual and composition costs of its atomic simple operations. Of course, between two corrections, the *better* one is supposed to be the one with the lower total cost.

Initially, the (erroneous) word $w = a_1 ... a_n$ is transformed into a deterministic FA (DFA) whose transition function δ is such that $\delta(i, a_i) = i + 1$ (1 is the initial state and $n + 1$ is a final state). More precisely, we built a finite transducer (FT) $\mathcal{T}_0$ with three tapes. In the general case we have $\delta(i, a_i/b_i/c_i) = i + 1$ which means that the cost of transforming a_i into b_i is c_i. Of course, here we have $a_i = b_i$ and $c_i = 0$, the transformation of a_i into itself is performed at null cost. The output of $\mathcal{T}_0$ on w is the n-tuple $((b_1, c_1), ..., (b_n, c_n))$ which is interpreted as follows: the transformation of $a_1...a_n$ into $b_1...b_n$ costs $c_1 + ... + c_n$. Note that, in the general case, the a_i's and b_j's may both be ε transitions.

It is a well-known result of the FT theory that each spelling error correction rule $\mathcal{R}_i$ can be transformed into a FT say $\mathcal{T}_i$. This means that we can build a FT $\mathcal{T}^w$ by composing $\mathcal{T}_0$ with $\mathcal{T}_1$, ..., $\mathcal{T}_i$, ..., $\mathcal{T}_l$, if there are l correction rules. The final result $\mathcal{T}^w$ is a transducer which specifies all the corrections (and their costs), that we can apply, according to the rules $\mathcal{R}_i$, to a given word w.

The final step is to perform the intersection of the lexicon $\mathcal{F}$ with $\mathcal{T}^w$, the result of which is itself a FT. The language of this FT is a set of couples (w_i, t_i) meaning that correcting w into w_i costs T_i where $T_i = \sum_{j=1}^{l} c_j$ if $t_i = c_1...c_l$. Of course if t_m is the minimal cost of all the t_i's, the *best* correction of w is w_m.

The difficulty of this approach is not the underlying theory which is well known, but comes from the sizes of the automata that we have to handle in real size applications. For example, our French lexicon contains more than 400 000 inflected forms. Typically, the number l of spelling error correction rules $\mathcal{R}_i$ reaches several hundreds. This means that, for each misspelled word w, the final transducer $\mathcal{T}^w$, is the (dynamic) composition of hundreds of elementary FTs. The number of corrections specified by $\mathcal{T}^w$ may be over billions and billions. Finally, the best correction(s) is computed in intersecting (theoretically a Cartesian product) these two *monsters*.

One must admit that the feasibility of such an approach was not a priori clear. However, the results are so encouraging that we have decided to use SXSPELL for the EASY campaign and to put it as a module in the SYNTAX library.

6.1.2. SXLFG: A Lexical Functional Grammar Parser

Lexical Functional Grammar (LFG) is a grammatical theory assuming two parallel levels of syntactic representation: constituent structure (c-structure) and functional structure (f-structure).

- C-structures have the form of context-free phrase structure trees;
- F-structures are sets of pairs of attributes and values; attributes may be features, such as tense and gender, or functions, such as subject and object.

At least at a conceptual level, we may see an LFG parser as a two-phase process: the first phase is a CF parser which builds the C-structure while the second phase evaluates the F-structure on the tree built by the first phase. However, the CF-backbone of real linguistic grammars (including LFG) are usually massively ambiguous. For example, for a sentence, we have exceeded the capacity of a single floating point 32 bit word in counting its number of parse trees. In ATOLL, we know how to handle such a combinatorial explosion of resulting tree structures. In the LFG context, this means that, for any given sentence w, we can compute in polynomial time a polynomial size parse forest which represents all the possible C-structures of w (See for example [3]). However, the efficient evaluation of F-structures on parse forests is still a research problem. Of course, the unfolding of the parse forest into single trees upon which F-structures are evaluated is not a viable method. We have designed and implemented a method which evaluates F-structures directly on a parse forest and which shares common [sub-]computations.

The coupling of our guided Earley parser with the previous shared computation of F-structures results in a new LFG parser called SXLFG.

Though this parser still needs to be improved, it is sufficiently mature to support full natural language descriptions. SXLFG is one of the three parsers used by ATOLL in the EASY campaign (cf. 6.7).

6.1.3. Range Concatenation Grammars

Of course, our work on Range Concatenation Grammars (RCG) is still active. Their theoretical basis has been published this year as a book chapter ([11]). From a practical point of view, the usage of RCGs which is done by B. Sagot (Section 6.4) puts new challenges on the corresponding parser. In this case, RCGs are not used as a *specification* formalism but as an *implementation* formalism: this means that the grammar is not written by hand but is automatically generated from a higher specification level which is close to a linguistic view of natural languages. As an example, our RCG parser, is now able to handle predicates whose number of arguments (linear or non linear) can exceed several tens (there is a predicate with 70 arguments in the current French specification).

6.2. Automata and Tabulation for Parsing

Keywords: *Dynamic Programming*, *Logic Programming*, *Parsing*, *Push-Down Automata*, *TAG*, *Tabulation*, *coordination*.

Participants: Éric Villemonte de la Clergerie, Tatiana Samoussina, Alexandra Mounier.

TAG *Tree Adjoining Grammars*

TIG *Tree Insertion Grammars*

We have started investigating a more active use of tabulation to handle some complex linguistic phenomena. The basic idea is that because derivations are tabulated in a system like DyALog, it is possible at parsing time to take decisions based on the examination of some sub-derivation. Furthermore, DyALog provides some logic predicates that may be used to follow derivations, which means these derivations can be handled (almost) as first-class citizens. More concretely, during her internship, Tatiana Samoussina has done some preliminary experiments based on these ideas to handle some cases of coordinations. Coordinations are complex phenomena in NLP because they break the “normal” pattern of sentence constructions by introducing many kinds of ellipsis like in “Jean eats an apple and John [] an orange”. However, many cases of coordination may intuitively be understood as duplicating similar derivations before and after the coordination word, with the possibility of ellipsis on shared parts between these derivations. The experiments have shown the potential interest of this idea but have also shown that more support has to be added to DyALog to handle derivations, in particular to duplicate a derivation. Alexandra Mounier has started working on these issues, also bringing a stronger linguistic expertise to understand what kinds of coordination may be handled by the proposed techniques. It should be noted that we believe that this active use of tabulated derivations during parsing has a large range of applications, for instance to handle grammatical errors.

Regarding the development of DYALOG, this year has been mostly devoted to complete the implementation of regular operators (disjunction, interleaving, Kleene star) for all formalisms (logic predicates, DCGs, TAGs and TIGs, partially RCGs). From a theoretical point of view, these regular operators may be applied to most grammatical formalisms without modifying their complexity. From a practical point of view, they allow the design of much more compact grammars with better performances when parsing if these operators are correctly implemented (not by expansion). In particular, the interaction of interleaving (free ordering between sequences of non-terminals or tree nodes) and Kleene star proved to be difficult to implement. The use of these operators for TAG and TIG was essential in the design of parsers for our wide coverage French grammar FRMG.

In the context of the Parsing Evaluation campaign EASY (cf. 6.7), we had to convert the output of FRMG parser to the XML format expected by the organizers. It was also necessary to setup a desambiguation algorithm to select some best parse in the shared set of answers returned by our parser. These two tasks were the occasion to explore the design of algorithms working on shared forests, namely shared dependency forests. During his internship, Mehdi Ben Hmida developed a first prototype based on XML technology. For ease of development and for efficiency reasons, É. de la Clergerie has moved to DYALOG. The EASY campaign has nevertheless shown very serious algorithmic problems, related to the complexity of disambiguation on shared dependency forests. Last minute solutions have been implemented but a more complete investigation has to been done.

6.3. Designing grammars using MetaGrammars

Participants: Éric Villemonte de la Clergerie, François Thomasset.

MG *MetaGrammars*

The exact formalization of MetaGrammars (MG) is still a subject of research that we explore through cooperations with Project-Teams "Langues & Dialogue" and "Calligramme" (LORIA).

Roughly speaking, a metagrammar is a list of classes expressing constraints. A class may inherit constraints from one or more parents and is used to describe some elementary linguistic phenomena. Constraints express existence of nodes, relationships between these nodes (ancestor, parent, sibling, equality, ...) and content as feature structures attached to nodes or to the class. A class can also states that it provides or needs some functionality. The role of a MG compiler is to combine classes in order to get neutral classes (all needs filled by providers and conversely), to check that constraints are satisfied and to use these constraints to generate the (minimal) structures of the grammars (trees in the case of TAGs).

É. de la Clergerie has developed, with DYALOG, a prototype of MG compiler, called MGCOMP. This new prototype is quite efficient and allow the exploration of new features for MetaGrammars. In particular, this prototype has been extended in 2004 with features allowing the design of more compact MG and the generation of very compact TAG grammars. To reduce MG size, a class can now require a functionality (express a need) in some namespace. A final neutral class can therefore embed several instances of a same class in different namespaces, removing potential name clashes. To reduce the size of the resulting grammar, the regular operators available in DYALOG such as interleaving, disjunction and Kleene star may be used at the level of the MG. Actually the interleaving operator (denoting free order between sequences of nodes) is implicitly activated when ordering between sibling nodes is underspecified. Another powerful extension, related to the functionalities of DYALOG, is the notion of *guarded node*: the existence (resp. non-existence) of a node may be conditioned to the truth values of a positive (resp. negative) guard, expressed as a boolean expression over path equations in feature structures. Whenever possible, these guards are reduced during the compilation process. The remaining guards are checked during parsing.

All these extensions were quite essential in the development of the large coverage French MetaGrammar FRMG. In a few months, we developed this metagrammar to be used for the Parsing Evaluation campaign EASY. At the date of the campaign, the metagrammar was formed of 191 classes, with a resulting TAG grammar of 126 trees corresponding to several thousand trees (estimated to be more than 5000 trees) if fully expanding guards and regular operators. The TAG grammar was then compiled by DyALog as an hybrid TIG/TAG grammar (i.e. an automatic analysis was done to identify TIG parts of the grammar).

Although the current prototype of MGCOMP was powerful enough to compile our large coverage metagrammar, several problems of efficiency due to combinatorial explosions were detected. Because of a very limited time, only partial solutions have been implemented but better ones have to be investigated to handle larger metagrammars.

The development of FRMG was facilitated by the parallel development with F. Thomasset of an edition environment for MG. The original XML format for MG has been completed by a more convenient one for edition, with an associated Emacs style. A graphical viewer interacting with the editor has also been developed.

Almost since its beginning, the development of FRMG has been tested using testsuites of sentences to measure coverage, efficiency and rate of ambiguity. Parsing results could also be visualized through different graphical and non graphical formats using tools already developed within ATOLL (a server of parsers and forest converters). A WEB interface may also be used to browse the trees of the grammar.

An essential point for FRMG was the interaction with our lexicon Lefff (6.4). The lexicon entry of a word in the lexicon should state which trees of the grammar may be anchored by this word. Practically, this is achieved by unifying two features structures, called *Hypertags*, one for the word and one for the tree. For a word, its hypertag gives information about the linguistic properties of the word (for a instance, that it is a ditransitive verb) while the hypertag attached to a tree (and built during the compilation of the underlying metagrammar) states what linguistic phenomena are covered by this tree (for instance ditransitivity). The problem is the coherence between the information provided by the lexicon and those provided by the trees. Information are often missing, incomplete or even false in a very large lexicon (more than 400000 inflected forms in LEFFF) and feedback techniques were needed to detect some of these problematic entries. By parsing tens of thousands of sentences, we tried, for instance, to identify unknown words and measure “parsability rates” (i.e., for a given word, the ratio between the number of occurrences in successfully parsed sentences and the total number of occurrences). A word with a parsability rate much lower than the average coverage rate of the grammar generally indicates a lexicon entry to be modified or completed (or may hint for some linguistic phenomena not yet covered by the grammar).

The current version of FRMG achieves a coverage rate of 99.69% on a subset of 361 sentences from Eurotra testsuite and of 93.38% on the 1661 sentences of TSNLP testsuite. The average parsing time is of 1.81s on Eurotra sentences and of .72s on TSNLP sentences on a 2GHz laptop, with the fact that execution times were twice faster before last minute changes tried to (very slightly) increase coverage but whose application should be better controlled. Results about parsing correctness (precision) are much more difficult to provide but will be (partially) provided by the evaluation results from EASY.

6.4. Acquisition and use of semantic lexica

Participants: Benoît Sagot, Lionel Clément.

French morphological lexicon Lefff: http://www.lefff.net

This year was for Benoît Sagot’s second year of PhD research on "acquisition and use of semantic lexica" (co-directed by Laurence Danlos, Lattice/TALaNa, Paris 7, and Eric de La Clergerie, from Atoll). During this period, his work focused mainly on four different albeit related topics.

6.4.1. RCG as a linguistic formalism

First of all, he went further in the elaboration of a new linguistic formalism, based on Range Concatenation Grammars [11], but designed to take into account the specificities of natural language with respect to standard formal grammars. This formalism allows to merge at the grammar level both syntax and semantics, which is very interesting both for computational reasons (less ambiguity, polynomial parsing) and for linguistic reasons: in particular, it shows both the feasibility of an efficient syntactico-semantical grammar, and the fact that usual approaches to linguistic grammars, such as dependency, constituency, topology or predicate-arguments semantics, can be seen of approximations/projections of global analyzes, as presented at the “Journée Portes Ouvertes sur les Sciences du Langage of the CNRS”. A preliminary and yet outdated form of these reflexions has been published in [18], and a more advanced state has been presented during a “Journée TALaNa”.

The second topic of research was the effective use of this formalism in the development of a syntactico-semantical grammar for French, which has not yet a very large coverage, but which implements some of the most computationally complex phenomena, such as long distance extraction, coordination, or control verbs. In addition to a compiler that converts the formalism to pure RCG (making it possible to use Pierre Boullier’s RCG parser) this has led to several ideas and realizations to exploit the analyzes that are produced. This includes, among others, a collaboration with Adil El Ghali (Lattice, Paris 7) about the coupling of Sagot’s system with an ontology based on description logic [19], works on the projection of the analyzes to visualize them as dependency graphs, constituents tree, or topological boxes, and works with Laurence Delort on the integration of SDRT in the proposed formalism to model discourse structure (in progress).

6.4.2. Crafting a large coverage lexicon

The third topic of the activities of Benoît Sagot concerns the acquisition and production of a large-coverage lexicon. This has led to different kinds of approaches: automatic statistical acquisition of a large-coverage lexicon of French verbs (published in [13], and made freely available on www.lefff.net), and semiautomatic improvement of an already existing lexicon. The latter approach was based on SXSPELL, a very efficient spelling correction tool developed by Pierre Boullier in collaboration with Benoît Sagot.

This lexicon has been completed with initial syntactic information, in particular about verb subcategorization. Since the original version, it has been deeply modified by adopting inheritance mechanisms to achieve a better factorization of information. Feedback provided by parsers has also be used to improve the quality of the lexicon.

6.4.3. Exploiting lexicons

The last topic concerns a very strong participation during the last quarter of 2004 to the Parsing evaluation campaign EASY. Benoît Sagot developed a complete input pipe based on SXSPEL. for providing segmentation, tokenization and recognition of named entities. Of course, this pipe relies on the above-mentioned lexicon in order to provide information to the parsers.

For EASY, Benoît Sagot has also collaborated with Pierre Boullier and Lionel Clément to design and implement two flavors of LFG parsers (SXLFG and XLFG). It should be noted the strong interactions that exist between grammars and lexicons: the coverage of some linguistic phenomena is only possible if some pertinent informations are available in the lexicon but, of course, these informations may be too difficult to obtain.

6.5. NLP Infrastructure and standardization

Participants: Lionel Clément, Guillaume Rousse, Benoît Sagot, Éric Villemonte de la Clergerie.

ATOLL tries to design and setup an XML-based linguistic pipeline, making easier the integration of new components by wrapping them if necessary. The pipeline mainly covers the first layers of linguistic processing, namely morpho-syntactic processing (segmentation, tagging, lexicon lookup, named entities, ...). It integrates several tools which are developed within ATOLL by L. Clément (cf. 5.4) and which have been improved.

The main role of the pipeline is to feed entry to our parsers. In particular, this pipeline has been partially rewritten and completed by G. Rousse to be used for handling botanical corpus in the context of the action BIOTIM. A recurrent problem is the issue of the various formats produced or expected by the different tools. An important effort has therefore been done to be able to convert different morphosyntactic tagsets (for several variants of MULTEXT, for TreeTagger, for FASTER, for ACABIT) to and from a pivot XML representation using feature structures. Because several tools may provide similar information (for instance a tagger and a lexicon), (simple) mediation algorithms have been investigated to determine which information to keep. This mediation is of course made possible because information may be compared.

For the EASY campaign and because of choices we made, several alternate tools have been developed to handle word and sentence segmentation and named entity detection (dates, proper names, numbers, URLs, abbreviations, ...). For questions of time, these tools are not yet part of our pipeline but should be soon incorporated.

This work on the first layers of NL processing feeds our reflexion by testing and demoing propositions for standardizing morpho-syntactic annotations in the context of French action Normalangue (cf 7.1) and of ISO subcommittee TC37SC4 for the normalization of linguistic resources.

6.6. Implicit Information in Natural Language

Participant: Areski Nait Abdallah.

Areski Nait Abdallah has been investigating the formalization and algorithmic processing of implicit information in natural language. He is collaborating with Alain Lecomte in the development of a partial information-based model of implicitness including implicatures and presuppositions [16]. Background knowledge is assumed to consist of a central kernel (corresponding to hard knowledge) together a "protective" belt (corresponding to soft knowledge). A partial model is defined by a triple (i_0, J, I_1), where i_0 and i_1 are partial assignments that are coherent with one another, and J is a set of "expectations" corresponding to justifications. Hard truths are those statements that receive the truth value "true" under i_0, whereas "soft" truths are those statements that receive the truth value "true" under the combination of i_0 and i_1. The set of conventional implicits is defined as being the set of formulae which are true under all minimal models corresponding to the sentences being uttered. Our formalization allows accommodation as well as resolution by the context to be accounted for. It still needs to be enriched, by the introduction of dynamic features, as well as the possibility of handling conversational implicatures by means of an adequate formalization of Gice's conversational maxims.

6.7. Parsing Evaluation campaign EASY

Keywords: *Evaluation*, *Parsing*.

Participants: Éric Villemonte de la Clergerie, Benoît Sagot, Lionel Clément, Pierre Boullier, Guillaume Rousse, Philippe Deschamp, Mehdi Ben Hmida, François Thomasset.

EASY Action: http://atoll.inria.fr Rubrique « Projets »

This year's work has been strongly oriented by our participation to the Parsing Evaluation campaign EASY organized by French Technolangue action EVALDA. The protocol for this campaign was to parse around 35000 sentences distributed in 43 corpora covering very different styles (journalistic, literacy, oral, mail, medical, questions/answers). The sentences were available for one week and we had to return XML files providing information about *chunks* (non recursive constituents such as nominal phrases) and *dependencies* (such as subject-verb dependencies). This campaign evaluates much more than just parsing, because its completion requires the availability of a full set of NLP tools and resources (segmentation, lexicons, grammars, parsers, ...). It was also needed to adjust our results to the format expected by the organizers (by disambiguating, filtering and converting). In order to be ready, a very important effort of development has therefore been necessary.

We were able to present three parsers for EASY, namely XLFG (Lionel Clément), SXLFG (Pierre Boullier and Benoît Sagot, Section 6.1) and FRMG (É. de la Clergerie, Section 6.3). XLFG and SXLFG depends on a LFG grammar designed by Lionel Clement (and modified for SXLFG). FRMG depends on a metagrammar strongly inspired by a previous one designed by Lionel Clément. All parsers are using the same lexicon LEFFF (with different modalities), originally developed by Lionel Clément and Benoît Sagot [13], and lastly modified and re-designed by Benoît Sagot (Section 6.4).

Several other tools have to be completed or developed to handle segmentation, in particular for detecting named entities (dates, numbers, proper names, abbreviations, URLs, ...) (cf. 6.4).

The campaign was run during the week starting December 8th. We used two clusters of computers and had to develop last minute scripts to dispatch and balance loads. Because Lionel Clément has left ATOLL in July, he has no easy access to enough computer power and XLFG was unfortunately unable to complete EASY (among other reasons).

After all kinds of difficulties, results for SXLFG and FRMG have been returned to the organizers for almost all the sentences. The organizers have now to compare our results with reference hand annotations. The evaluation results should be known during the first quarter of 2005.

Many tools and resources developed and used by ATOLL for the campaign EASY have been packaged (tar.gz and rpm) and are already freely available (or will be shortly).

6.8. Processing Botanical Corpora

Participants: Guillaume Rousse, Éric Villemonte de la Clergerie.

BIOTIM Action: http://atoll.inria.fr Rubrique « Projets »

In the context of French action BIOTIM (cf. 7.2), ATOLL is involved in processing botanical corpora. Guillaume Rousse was recruited at the end of 2003 to work in the action BIOTIM. His work is centered on extracting knowledge from OCRized botanical corpora [22].

In order to setup a complete linguistic processing chain, he had to integrate many linguistic tools, developed internally in ATOLL or externally (Section 6.5). Whenever possible, these tools have also be submitted as contributions to MandrakeLinux, in order to ensure for these tools a larger diffusion and an easier installation.

Because of many spelling problems due to OCR, spelling correction techniques have been tried. However, further investigations of the corpora have shown much deeper problems that could not be automatically corrected and that were due to the way the OCR has been done. It was decided to ask for a new OCRization of the corpus with a much higher quality, hence almost removing the need for spelling corrections. The new corrected corpora have been received at the end of 2004.

On the original corpora, and using the above mentionned NLP pipeline, we tried experiements to extract terminology (using the external tools FASTER and ACABIT). A compound term such as "nervure latérale" also provides a relation between a governor “nervure” and a governee “latérale”. These relations may used to classify simple (or compound) terms, using the distributional hypothesis (Harris) stating that two terms sharing a similar set of governors or governees may often be considered as semantically close. Preliminary experiments have been tried on our extracted terminology. However, these experiments raise the issue of evaluation, to assess the quality of a terminology or of a classification. Many parameters and thresholds may be tried during the experiments, giving different results that difficult to evaluate for such a specialized domain.

6.9. Free Software

Keywords: *Copyright*, *Economy*, *Free Software*, *Linux*, *Open Source*, *Patent*.

Participant: Bernard Lang.

The problem raised by the open availability of linguistic resources, whether linguistic processing software (such as taggers, parsers, etc.) or linguistic data (such as lexicons, grammars, or corpora) has raised our interest in the development of free scientific resources. There is a wide consensus that the limited availability of the results produced by earlier research, due to excessive use of intellectual property, has been a major impediment to the progress of computational linguistics research, especially in Europe.

It is a policy of our group to make our results freely available.

B. Lang has taken a strong interest in these issues and has become very active in understanding better the legal and economic aspects of the production, dissemination and use of intangible goods. Much of the work is observing the evolution of the free economy of intangibles, how it develops, and how it relates to the evolution of the legal system. One important aspect is the impact on research practice, on communication between researchers, and on the valorization of research results.

7. Contracts and Grants with Industry

7.1. Action Normalangue/RNIL

Participants: Éric Villemonte de la Clergerie, Lionel Clément.

Normalangue Home Page: http://www.normalangue.org/
RNIL Home Page:http://atoll.inria.fr/RNIL/
TC37SC4 Home Page: http://www.tc37sc4.org/ MAF demonstrator: http://atoll.inria.fr/mafdemo

ATOLL is a leader participant in the RNIL subpart of action Normalangue, funded by French program Technolangue. This action promotes the emergence of standardized representations for linguistic resources, in parallel with the definition of API for the corresponding linguistic tools. The action supports the French mirror group of ISO sub-committee TC37 SC4 for the normalization of linguistic resources.

É. de la Clergerie chairs this mirror group, which has organized several meetings in 2004. L. Clément and É. de la Clergerie are the promoters of a French proposition of a morpho-syntactic annotation framework (MAF), which has been accepted as a new work item by ISO TC37SC4.

We have refined our proposal in a ISO Working draft [21] and are moving toward a Committee Draft in the next few months, after incorporating remarks made during the last ISO meeting on MAF (Pisa, November 2004). A small demonstrator for French, based on ATOLL’s tools, has been recently activated to illustrate our proposal.

É. de la Clergerie is also strongly involved in the standardization of feature structures using an XML representation.

7.2. Action BIOTIM

Participants: Éric Villemonte de la Clergerie, Benoît Sagot, Guillaume Rousse.

BIOTIM home page: http://www-rocq.inria.fr/imedia/biotim/

Funded by ACI program on “Masses de données” (Data Warehouses), action BIOTIM has started end of 2003 for 3 years. Its thematic is the processing of botanical textual corpora and image collections in order to extract knowledge and establish bridges between texts and images for more intelligent navigations at a semantic level. ATOLL is essentially concerned with the linguistic processing of textual corpora with generic methods to extract terminologies, ontologies and knowledge bases.

The other participants to BIOTIM are INRIA project-team IMEDIA (leader), CNAM team Vertigo, INRA team URGV, IRD, and LIFO (University of Orléans).

7.3. Action EVALDA/EASY

Participants: Éric Villemonte de la Clergerie, Pierre Boullier, Lionel Clément, Benoît Sagot, Mehdi Ben Hmida.

ATOLL participates to the parsing evaluation campain EASY of action EVALDA of French program Technolangue.

The campaign took place mi-december, after a strong effort of developpement of tools and resources by the member of ATOLL in order to be ready (Section 6.7).

Results should be known in the first quarter of 2005.

7.4. Action eCOTS

Participant: Bernard Lang.

Though INRIA is not a member of the eCOTS association resulting from a former collaboration with industry, B. Lang is still having occasional collaborations with this association (founded by Thales, Bull and EDF). Its purpose is the development of an open information site on software components. He participates in this context to the ICCBSS 2005 conference.

8. Other Grants and Activities

8.1. National Actions

Ph. Deschamp is a member of the French “Commission spécialisée de terminologie de l’informatique et des composants électroniques” (terminology committee for Computer Science and Electronic), and distributes on-line the glossary http://www-rocq.inria.fr/who/Philippe.Deschamp/CMTI/ resulting of his work (more than 130 000 downloads). Ph. Deschamp is also a member of the French “Commission spécialisée de terminologie et de néologie des télécommunications” (terminology committee for telecommunication).

B. Lang is vice-president of AFUL (http://www.aful.org), “Association Francophone des Utilisateurs de Linux et des Logiciels Libres”, and member of the administration board of ISoc-France, the Internet Society French branch. He is also a member of the scientific board of association SOISSON Informatique Libre.

8.1.1. Open Source Software

B. Lang has presented the notion of open source software in several workshops, talks and conferences, organized by local collectivities and administrations.

8.2. International networks and working groups

8.2.1. Open Source Software

B. Lang has been several times invited to talk on Open Source Software.

B. Lang is a member of an expert committee on Open Source Software for the European Commission General Direction for Information Society (ex DG 13) (http://eu.conecta.it/).

8.2.2. Action INRIA-ICTTI FASTLING

We had no official funding in 2004 for this long-lasting cooperation between ATOLL and team CENTRIA of Lisbon New University. However, we have continued to collaborate and a new PAI named KLING has been recently accepted for 2005.

8.2.3. PAI PICASSO CATALINA-2

Funding for visits has been granted by the French-Spanish PAI (Programme d'actions intégrées) PICASSO to renew a cooperation named CATALINA-2 between ATOLL and team COLE at University of La Coruña. We cooperate on parsing techniques (in particular for TAGs) and are interested by establishing a more ambitious project on information extraction.

8.2.4. ISO subcommittee TC37SC4

The participation of ATOLL to French Technolangue action Normalangue has resulted in a strong implication in ISO subcommittee TC37 SC4 on the normalization of linguistic resources (http://www.tc37sc4.org/). É. de la Clergerie and L. Clément have participated to ISO events and have played a role of experts (in particular on Morpho-Syntax and Feature Structures).

8.3. Visits and invitations

Visits of Gabriel Pereira Lopes in January and May 2004, and visit of Vitor Rocio in January (action INRIA-ICTII FASTLING).

Two one month visits of Francisco Jose Ribadas Pena in June and December 2004 and a two week visit of Manuel Vilares Ferro in September (action PICASSO CATALINA-2).

9. Dissemination

9.1. Animation at INRIA

B. Lang is an elected member of INRIA's “Conseil Scientifique”.

É. de la Clergerie has participated to the INRIA Rocquencourt “section d'audition” for the 2004 CR2 recruitment campaign.

G. Rousse has delivered two INRIA internal formations, the first one on packaging softwares under Linux and the second one on using software development tools (CVS, AUTOTOOLS, RPM, ...). B. Lang also contributed a session on Free Software.

G. Rousse has participated to the realization of the compilation "CD-ROM des logiciels libres", edition 2004, in collaboration with Pierre Weis. He also participates to the supervising of a student working on the future automatization of this task. He was a member of an INRIA working group on “Software Development at INRIA” and a contributor of the final document produced by this group.

B. Lang has made some contributions to the design of the Free Software license CeCILL (http://www.cecill.info/), created by INRIA, CEA, and CNRS.

9.2. Supervising

É. de la Clergerie has supervised the internships of Tatiana Samoussina [23], and Mehdi Ben Hmida [20]. He also co-supervises the PhD thesis of Benoît Sagot with Laurence Danlos (TALaNa/LATTICE, University Paris 7).

9.3. Jury

- B. Lang is a member of the CNAM expert committee in computer science.

- É. de la Clergerie is a member of the recruitment committee of University of Orléans.
- B. Lang was the opponent at the defence of Peter Ljunglöf's PhD dissertation entitled "Grammatical Framework and Generalized Context-Free Grammars", at Chalmers University, Göteborg, Sweden, December 6th.
- É. de la Clergerie has been a member of the PhD jury for Djamé Seddah at LORA, Nancy, November 5th.

9.4. Teaching

Starting December 2003, L. Clément has delivered courses (40h) on XML in DESS "Gestion de documents électroniques et de flux d'information" (management of electronic documents and information flows) at University Paris X.

G. Rousse has done a presentation on free softwares at the "Ecole Polytechnique de Nantes".

B. Lang has delivered a talk on legal and economic issues of free software in the "Information Technologies" workshop, organized by the "Institut d'Économie Industrielle" of "Université des Sciences Sociales", Toulouse I.

9.5. Committees

- Participation of É. de la Clergerie to the editorial board of French journal T.A.L. http://www.atala.org/tal/tal.html and Guest Editor of T.A.L. issue 44/3 on "Evolutions in Parsing"
- Participation of É. de la Clergerie to program committees for TALN'04, the French national conference on NLP and for LREC'04 workshop on "A Registry of Linguistic Data Categories within an Integrated Language Resource Repository Area". Reviews for COLING'04, TALN'04, ACL'04 Students, ESSLLI'04,
- Participation of P. Boullier to the Program Committee of FG-MOL 2005 (10th conference on Formal Grammars and 9th meeting of Mathematics of Language).
- É. de la Clergerie has reviewed a proposal for the French program ACI "Masses de données'.
- É. de la Clergerie has participated as expert to a working group of the SGDN (Secrétariat Général de la Défense Nationale) to evaluate a proposal.
- B. Lang is organizing a panel entitled "Free and Proprietary software in COTS-based software development" for the conference ICCBSS 2005 (4th International Conference on COTS-Based Software Systems), 7-11 february 2005, Bilbao, Spain.
- B. Lang is vice-president of the SIL-CETRIL association for the economic development of the Soisson area (http://www.sil-cetril.org/article.php3?id_article=35).
- B. Lang has participated to the working group PIETA (Prospective de la propriété intellectuelle) of the Commissariat Général du Plan.
- B. Lang was audited by the Senate Socialist Group in preparation of the law "Confiance dans l'Économie Numérique" (March 30th).
- B. Lang has advised local governments about the use of Free Software (Bezon, April 1st; Conseil régional d'Ile de France, May 28th).

9.6. Softwares

G. Rousse is a contributor for MandrakeLinux, helping the packaging and diffusion of many scientific softwares (including ATOLL's softwares).

9.7. Participation to workshops, conferences, and invitations

- Participation of É. de la Clergerie and L. Clément to ISO TC37SC4 meeting (Jeju, Korea, February 2004). Participation of É. de la Clergerie to other ISO TC37 meetings in Lisbon (May), Paris (August) and Pisa (December). Participation of É. de la Clergerie to LREC'04
- One week visit of É. de la Clergerie at Universities of La Coruña and Vigo (PAI PICASSO CATALINA-2).
- É. de la Clergerie has presented DYALOG at LORIA (Nancy), University of Vigo (Spain) and university of Lisbon (Portugal).
- Participation with presentations of B. Sagot at TALN'04 (Fes, Marocco), LREC'04 (Lisbon) and TSD (Brno, Czech Republic). B. Sagot also coordinates the seminar TALaNa.
- Participation with presentations of Areski Nait Abdallah to the workshop TCAN "La construction du savoir scientifique dans la langue" (Grenoble) and to "Logic Colloquium 2004", the European Conference of the Association for Symbolic Logic (Torino, Italy).
- Participation with presentation of Guillaume Rousse to the workshop TDWG'04 (New Zealand).
- B. Lang was an invited speaker at the "Colloque International de la Chaire Arcelor" of the Université Catholique de Louvain, on the theme "Brevet - Innovation - Intérêt général", 11-13 mars 2004.
- B. Lang participated to a panel "Protection électronique de l'Innovation", at the Forum Européen de l'Administration Électronique, Paris, 15-16 december.
- B. Lang was invited to speak at the 5èmes Rencontres Parlementaires sur la Société de l'Information et de l'Internet, « Pour une politique de l'Internet », December 2nd, on the topic « Vers une "politique numérique" européenne ».
- Participation and contribution of B. Lang to several meetings on the potential of Free Software, and on économic, legal and political issues:

  - Professional conference on Free Software, organized by Thalix. June 16th
  - Professional conference on Free Software, organized by the local economy association "Mêlée Numérique", Toulouse, June 22nd.
  - Panel "Qui peut encore concurrencer Microsoft ?" organized by FNAC, October 22nd.
  - Panel "Enjeux éducatifs et sociétaux du logiciel libre" at salon EDUCATEC 2004, Paris, Novembre 17-19.
  - Presentation on Free Software for the group "Documentation" of the Conférence des Grandes Ecoles, December 10th, Ecole des Mines de Paris.
  - Presentation on "Les Logiciels Libres d'un point de vue économique" at the event Libr'east of Paris at IUT de Marne la Vallée, April 23-25.
  - Presentation on Free Resources at the working group "Espace numérique fédérateur pour les CPGE et les formations équivalentes" (ENF-CPGE) at the Colloque international ePrep 2004, INT, Evry, May 6-7.

- B. Lang was interviewed on several occasions by national radios about the evolution of software industry, and specifically on Free Software, competition, and software patents.

10. Bibliography

Major publications by the team in recent years

[1] P. Boullier. *A Cubic Time Extension of Context-Free Grammars*, in "Grammars", vol. 3, nº 23, 2000.

[2] P. Boullier. *On TAG Parsing*, in "Traitement Automatique des Langues (T.A.L.)", issued June 2001, vol. 41, nº 3, 2000, p. 111-131.

[3] P. Boullier. *GUIDED EARLEY PARSING*, in "Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03), Nancy, France", April 2003, p. 43–54, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Pierre.Boullier/earley_final.

[4] L. Clément, A. Kinyon. *Generating parallel multilingual LFG-TAG grammars from a MetaGrammar*, in "Proc. of ACL'03", 2003.

[5] B. Lang. *Complete Evaluation of Horn Clauses: an Automata Theoretic Approach*, Technical report, nº 913, INRIA, Rocquencourt, France, November 1988, http://www.inria.fr/rrrt/rr-0913.html.

[6] B. Lang. *Towards a Uniform Formal Framework for Parsing*, in "Current issues in Parsing Technology", M. Tomita (editor)., also appear in the Proc. of Int. Workshop on Parsing Technologies - IWPT89, chap. 11, Kluwer Academic Publishers, 1991.

[7] É. Villemonte de la Clergerie. *Parsing Mildly Context-Sensitive Languages with Thread Automata*, in "Proc. of COLING'02", August 2002, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/COLING02.pdf.

[8] É. Villemonte de la Clergerie. *Automates à Piles et Programmation Dynamique. DyALog : Une application à la programmation en Logique*, Ph. D. Thesis, Université Paris 7, 1993.

[9] É. Villemonte de la Clergerie, M. A. Alonso Pardo. *A tabular interpretation of a class of 2-Stack Automata*, in "Proc. of ACL/COLING'98", August 1998, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/SD2SA.ps.gz.

Articles in referred journals and book chapters

[10] M. A. Alonso, É. Villemonte de la Clergerie, V. J. Diaz, M. Vilares. *Relating Tabular Parsing Algorithms for LIG and TAG*, G. S. John Carroll, H. Bunt (editors)., Text, Speech and Language Technology, revised notes of a paper for IWPT2000, vol. 23, chap. 8, Kluwer Academic Publishers, 2004, p. 157–184.

[11] P. Boullier. *New Developments in Parsing Technology*, Text, Speech and Language Technology, vol. 23, chap. Range Concatenation Grammars, Kluwer Academic Publishers, 2004, p. 269–289.

[12] B. Lang. *Quel modèle économique pour les créations immatérielles ?*, in "Les Nouveaux dossiers de l'Audiovisuel", Institut National de l'Audiovisuel, nº 1, septembre-octobre 2004, p. 66-67.

Publications in Conferences and Workshops

[13] L. CLÉMENT, B. SAGOT, B. LANG. *Morphology Based Automatic Acquisition of Large-coverage Lexica*, in "proc. of LREC'04", May 2004, p. 1841–1844.

[14] B. LANG. *Brevetabilité du Logiciel : le point de vue d'un chercheur en informatique*, in "Actes du Colloque "Brevet - Innovation - Intérêt général"", B. REMICHE (editor)., to appear, Chaire Arcelor, 2004.

[15] K. LEE, H. BUNT, S. BAUMAN, L. BURNARD, L. CLÉMENT, E. DE LA CLERGERIE, T. DECLERCK, L. ROMARY, A. ROUSSANALY, C. ROUX. *Towards an international standard on feature structure representation*, in "proc. of LREC'04", May 2004, p. 373–376, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/fs-lrec04.pdf.

[16] M. NAIT ABDALLAH, A. LECOMTE. *Implicatures scalaires, logique de l'information partielle et programmation logique*, in "Workshop TCAN La construction du savoir scientifique dans la langue, Grenoble", 2004.

[17] M. NAIT ABDALLAH. *An algebraic approach to commonsense reasoning*, in "Proc. Logic Colloquium 2004, European Conference of the Association for Symbolic Logic, Torino (Italy)", 2004.

[18] B. SAGOT, P. BOULLIER. *Les RCG comme formalisme grammatical pour la linguistique*, in "Actes de TALN'04, Fès, Maroc", 2004, p. 403-412.

[19] B. SAGOT, A. EL GHALI. *Coupling grammar and knowledge base: Range Concatenation Grammars and Description Logics*, in "Proceedings of TSD'04, Brno, Tchéquie", 2004.

Miscellaneous

[20] M. BEN HMIDA. *Traitements de sorties d'analyseurs syntaxi*, Technical report, DEA Informatique et Systèmes Intelligents Univ. Paris Dauphine, September 2004.

[21] L. CLÉMENT, É. VILLEMONTE DE LA CLERGERIE. *Terminology and other language resources – Morpho-Syntactic Annotation Framework (MAF)*, ISO TC37SC4 WG2 Working Draft, 2004, http://atoll.inria.fr/RNIL/TC37SC4-docs/draft-MAF-en.pdf.

[22] G. ROUSSE. *Automatized knowledge extraction from paper documents*, Poster presented at TDWG'04 (ChristChurch, NZ), October 2004.

[23] T. SAMOUSSINA. *Traitement linguistique et informatique de la coordination*, Stage X, DIX – École Polytechnique, July 2004.

[24] É. VILLEMONTE DE LA CLERGERIE. *Designing efficient parsers with DyALog*, Slides presented at GLINT, Universidade Nova de Lisboa, June 2004, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/DyALogLisbon04-small.ps.gz.

Bibliography in notes

[25] M.-H. CANDITO. *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées*, Ph. D. Thesis, Université Paris 7, January 1999.

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

[27] S. EARLEY. *An Efficient Context-Free Parsing Algorithm*, in "Communications ACM 13(2)", ACM, 1970, p. 94-102.

[28] R. M. KAPLAN, J. BRESNAN. *Lexical-Functional Grammar: A formal system for grammatical representation*, in "The Mental Representation of Grammatical Relations, Cambridge, MA", J. BRESNAN (editor)., Reprinted in Mary Dalrymple, Ronald M. Kaplan, John Maxwell, and Annie Zaenen, eds., Formal Issues in Lexical-Functional Grammar, 29-130. Stanford: Center for the Study of Language and Information. 1995., The MIT Press, 1982, p. 173-281.

[29] F. PEREIRA, D. WARREN. *Parsing as Deduction*, in "Proc. of the 21st Annual Meeting of the Association for Computationnal Linguistic, Cambridge (Massachussetts)", 1983, p. 137-144.

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