INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Atoll

Atelier d'Outils Logiciels pour le Langage naturel

Rocquencourt

THEME SYM

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]

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]
Alexis Nasr [Professor, TALANA, Univ. of Paris 7, starting October 2005]

Visiting scientists
Manuel Vilares Ferro [1 week, August 2005, University of Vigo]
Gabriel Pereira Lopes [2 weeks, September 2005, New University of Lisbon]
Francisco Riberra [November 2005 til March 2006, University of La Coruña]
Djamé Seddah [2 weeks, July 2005, LORIA]
Yannick Parmentier [1 week, May 2005, LORIA]

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

Technical staff
Guillaume Rousse

Student intern
Alexandra Mounier [CNAN Engineer Internship, til September 2005]

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 Meta-Grammars, 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) [48] and Head-Driven Phrasal Structure Grammars (HPSGs) [52] rely on more expressive Typed Feature Structures (TFS) [45] 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 [10]. 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 [11], [1]. More recently, *Thread Automata* (TA) [9] 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* [47] or *parsing as deduction* [51] 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 [8]. 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 in the context of the parsing evaluation campaign EASY. Disambiguation algorithms can also be applied on such shared structures.

3.2. Linguistic Infrastructure and Normalization

Participants: Éric Villemonte de la Clergerie, Guillaume Rousse, Benoît Sagot, Pierre Boullier, Philippe Deschamp, 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 and for processing raw texts before sending them to our parsers (cf. 6.7).

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.

MG *Meta-Grammars*

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, supplement and correct linguistic resources. Successful experiments have been conduced with different languages for the automatic acquisition of morphological knowledge from raw corpora. We would like to investigate also a higer 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. 6.10).

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 *Meta-Grammar* (MG) (cf. 6.4). Introduced by [44], a Meta-Grammar 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 [7].

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 and has continued working on these issues in 2005 (cf. 7.3). 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.

Translation Parsing is an important step in translations based on the transfer between language at a deep abstract syntactic level (or possibly at a semantic level).

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 in 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 **1.10.7** of DYALOG is freely available by FTP under an open source license and runs on Linux platforms for x86 architectures. A port for PowerPC, initiated by Djamé Seddah, should be available before the end of 2005.

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 [39].

C libraries can be used from within DYALOG to import APIs (`mysql`, `libxml`, `sqlite`, ...).

DYALOG is largely used within ATOLL to build parsers but also derivative softwares, such as a compiler of Meta-Grammars (cf. 5.3). It has also been used for building a parser from a large coverage French TIG/TAG grammar derived from a Meta-Grammar. This parser has been used for the Parsing Evaluation campaign EASY (cf. 7.3 and [37]).

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

5.3. Tools and resources for Meta-Grammars

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 Meta-Grammar (cf. 6.4). 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 Meta-Grammar 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 Meta-Grammar FRMG to get a grammar of around 100 TAG trees [37]. 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. The current version of FRMG has been completed to handle *support verbs* (such as *prendre garde [à]*).

To ease the design of meta-grammars, 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. A new version is under development to provide an even more compact syntax and some checking mechanisms to avoid frequent typo errors.

5.4. Morpho-syntactic processing tools

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

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. They are freely available under open source licenses, keeping in mind that most of these tools are still beta versions.

SxPipe (1.0.0) a container package, developed by B. Sagot, that includes many scripts for morpho-syntactic processing. It also includes a spelling corrector (SXSPELL) and a segmenter (for sentences and tokens) that rely on Syntax. The deployment of the various component is handled by LINGPIPE.

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, ...)

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).

Guillaume Rousse has updated and completed most of these tools for BIOTIM. He has also done an important work to package them.

5.5. Lexicon Lefff

Participant: Benoît Sagot.

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

LEFFF 1 is a freely available French morphological lexicon for verbs that has been automatically extracted from a very large corpus.

A new version of Lefff, Lefff 2 (currently partially distributed), used 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 with LFGs

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 on the improvement of our Lexical Functional Grammar parser SXLFG which focus on the sharing of identical computations: recent advances include (among others) techniques known as cyclic functional structures, lazy evaluation, internal disambiguation.

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 [5]). 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. It is now able to handle cyclic F-structures, implements a lazy unification to optimise the computation of these structures, and allows the grammar writer to specify disambiguation heuristics that can be applied on F-structures associated with any node of the forest. This improvements w.r.t. the first version of SXLFG have resulted in parsers that run approximately 5 to 10 times faster.

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. 7.3). We are starting to use the new version of SXLFG to parse large corpora, in order to validate our parsing techniques but also to learn information from the resulting analyses.

6.2. Spelling error correction

Keywords: *finite state transducer*, *spelling correction*.

Participants: Pierre Boullier, Benoît Sagot.

Following the development of our spelling correction techniques in 2004, based on finite transduction techniques, we have improved in 2005 the quality of our spelling rules in SXSPELL.

6.3. Automata and Tabulation for Parsing

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

Participants: Éric Villemonte de la Clergerie, Alexandra Mounier, Benoit Sagot.

TAG *Tree Adjoining Grammars*

TA *Thread Automata*

We have continued our exploration of 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, Alexandra Mounier 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 following in parallel two derivations, before and after the coordination word, with the possibility of ellipsis on shared parts between these derivations. Some preliminary support has been added to DYALOG, to be able to follow simultaneously two derivations (a completed one before the coordination word acting as the reference, and the other one after the coordination word). A few experiments have been tried and we plan to deploy and test these ideas on a larger scale within our wide coverage grammar FRMG. However, the parsing mechanisms that are used alter the shared derivation forests that are produced (with in particular the sharing of ellipsis) and, therefore, we need to update several tools that rely on shared forests. This work on handling coordinations has also been the main subject of discussions during the 2 week visit of Djamé Seddah.

At a more theoretical level, we are trying to extend Thread Automata. Originally introduced to ensure dynamic programming interpretations for Mildly Context-Sensitive [MCS] formalisms, TAs are not powerful enough to cover some extreme cases of scrambling (that are outside the scope of MCSs), as illustrated, for instance, with the MIX language of all sequences on alphabet a, b, c with an equal number of occurrences for each letter ($\{w \in \{a, b, c\}^* \| |w|_a = |w|_b = |w|_c\}$). While details remain to be checked, a possible extension for TAs would be to allow, under conditions, to displace threads. We also thinking about the implementation of TAs within DYALOG and at their use, in conjunction with Meta-Grammars, for handling Multi-Component TAGs.

6.4. Designing grammars using Meta-Grammars

Participants: Éric Villemonte de la Clergerie, François Thomasset.

MG *Meta-Grammars*

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

Roughly speaking, a meta-grammar 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 Meta-Grammars. In parallel, a French Meta-Grammar FRMG was quickly developed in 2004, relying on a development environment MGTOOLS.

In 2005, we have done some cleaning of FRMG and improved the handling of cleft constructions. We have also devised and implemented a way to handle support verbs (such as "*prendre garde à*") without adding new trees. The key idea is (a) to use an optional predicative noun as a co-anchor and (b) to be able to use the sub-categorization frame carried by the noun in place of the sub-categorization carried by the verbal anchor.

We have also implemented in MGTOOLS a new format for representing MGs in a even more compact way. The new format also allows for type and feature declarations, hence providing a way to detect quickly frequent typo errors when designing MGs.

6.5. Syntactico-semantic grammars and parsing

Participants: Benoît Sagot, Pierre Boullier.

TAG *Tree Adjoining Grammars*

LFG *Lexical Functional Grammars*

HPSG *Head-driven Phrasal Structure Grammars*

RCG *Range Concatenation Grammars*

Most current linguistic formalisms rely on a low-complexity syntactic backbone on top of which unification-based decorations are computed. This is for example the case in TAG and LFG. The case of HSPG is even more extreme, since it has (in theory) no backbone.

However, although it has some advantages (see 3.1.2), this architecture has several linguistic and computational drawbacks. Moreover, lexical semantics can only be applied as an extra unification-based layer, or be hidden in a intellectually non-satisfying probabilistic model. The idea is therefore to use a more powerful backbone to be able to get rid of the unification-based layer.

In ATOLL, P. Boullier has developed such a formalism, namely RCGs, as well as a very efficient parser for this formalism (see 3.1). However, RCGs as such are not suitable to encode linguistic knowledge. Therefore, we developed a linguistic formalism on top of RCGs, named Meta-RCGs [36], which strongly relies on the non-linearity of RCGs (closure by intersection), and is hence able to make all kinds of linguistic constraints interact simultaneously, including morphological, syntactic and lexical semantic constraints. The Meta-RCG formalism, thanks to a tool developed by B. Sagot, can be compiled into (very complex) standard RCGs, and therefore use P. Boullier's RCG parser. B. Sagot developped a middle-coverage Meta-RCG grammar for French as well as a toy syntactico-semantic Meta-RCG lexicon. Although the resulting parser can not yet be compared with FRMG or SXLFG in terms of coverage, it has already led to promising results [35]. In particular, it shows that traditional linguistic points of view on a sentence (constituency, dependancy, topology, predicate-argument semantics) can be extracted from the full Meta-RCG parse.

6.6. Acquisition of morphological and syntactic lexical information

Participant: Benoît Sagot.

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

Among the different resources that are needed for Natural Language Process- ing tasks, the lexicon plays a central role. However, the development or enrichment of a large and precise lexicon, even restricted to morphological information, is a difficult task, in particular because of the huge amount of data that has to be collected. Therefore, most large-coverage morphological lexicons for NLP concern only a few languages, such as English. Moreover, these lexicons are usually the result of the careful work of human lexicographers who develop them manually over years, and for this reason they are often not freely available.

Therefore, we currently investigate methods to automatically acquire lexical knowledge, in particular morphological and syntactic knowledge [25]. These methods, that may involve manual validation to guarantee the quality of the resources that are produced, had been succesfully applied to supplement the LEFFF lexicon for French in 2004, and have been used to acquire from scratch a lexicon for Slovak [34], language that lack large-coverage resources. Our method involves now also derivational morphology, which is a link to the acquisition of syntactic knowledge.

Direct acquisition of syntactic knowledge has been also performed for French and used to add syntactic information to the new version of the LEFFF, concerning in particular verbal lemmas. This new version, the LEFFF 2 [41], [33], is now a large-coverage formalism-independant syntactic lexicon for French, and is currently used in all our parsers.

6.7. NLP Infrastructure and standardization

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

French morpho-syntax demo: http://atoll.inria.fr/mafdemo

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.

Primarily developed for the EASY, several tools have been developed and grouped within package SXPIPE [32] to handle word and sentence segmentation and named entity detection (dates, proper names, numbers, URLs, abbreviations, ...). More components have been added in 2005, in particular in relation with the processing of botanical corpus.

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 [27], [40].

6.8. Implicit Information in Natural Language

Participant: Areski Nait Abdallah.

This work is done in collaboration with Alain Lecomte (Univ. Grenoble) and is based on [50]. Its aim is to formalize the use of implicit meanings in Natural Language. The Logic of Partial Information allows us to treat presupposed meanings and other implicit meanings like “soft” deductions. From a semantic point of

view, “soft” truths are members of minimal partial models. Such models are triples (i_0, J, i_1) where i_0 and i_1 are partial functions on P, the set of propositional variables, and J is a set of justifications (ie. of partial functions from P to $\{0, 1\}$). Function i_0 gives the kernel knowledge and i_1 gives the “belt” knowledge, in a conception of knowledge inherited from Lakatos (“The Methodology of Scientific Research Programmes”, Cambridge University Press, 1977).

In *“On expressing vague quantification and scalar implicatures in the logic of partial information”* [30], we use the logic of partial information to re-examine some early analyses of vague quantifiers in French such as *quelques, peu, beaucoup* that are found in particular in the work of O. Ducrot [46]. Our approach is based on the paradigm offered by the logical formalization of the sorites paradox. We argue that this paradox offers a general scheme along which the argumentation structure of all vague quantifiers in French may be expressed. We also offer a variational principle approximating Grice’s maxims in the case of vague quantification.

6.9. Evaluation

Keywords: *Evaluation, Parsing*.

Participants: Éric Villemonte de la Clergerie, Benoît Sagot.

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

ATOLL has recently developed several important linguistic resources and tools, such as LEFFF, SXPIPE, FRMG, and SXLFG. More and more, we need to evaluate and assess their quality before going further.

The participation of ATOLL to the French evaluation campaign EASy (in December 2004) has been a first step in this direction [37], [22], [26], [23]. The participants to EASY were expected to return information about 6 kinds of non recursive constituents (Nomimal chunks, Adjectival chunks, Adverbial chunks, verbal kernels, ...) and 14 kinds of dependencies (verb-subject, verb-object, ...). The evaluation was done on a set of around 35000 sentences covering various kinds of style (journalistic, literacy, mail, medical, speech, questions).

We are still waiting for the full results of this campaign, but have already received preliminary results about constituents for a subset of 4262 sentences. We have started using these results to analyze the weaknesses of ours tools and resources and have developed a few scripts to be able to synthesize our own statistics. We are now able to replay the EASY experiment and have actually started to do it. Our new results already show a clear improvement w.r.t. the original campaign (statistics for FRMG at http://atoll.inria.fr/results6/index.html).

We are also testing our parsers (FRMG and SXLFG) on other corpora, in particular a journalistic corpus “*le monde diplomatique*” (17Mwords) and some of the botanical corpus used for Biotim. More than 300 Ksentences have been parsed with FRMG, with a coverage rate (for full parsing) of 42% and more statistics may be found at http://atoll.inria.fr/results5/distrib.html. Almost 400 Ksentences have been parsed with the new version of SXLFG, with a coverage rate of 53%, although the precision of these parses are slightly lower than those of FRMG [25].

We also investigate error mining techniques to find, in the parsed corpora, the words that have significantly low parsability rates. Such words usually denotes incorrect or incomplete entries in our lexicon LEFFF, or errors in the grammar.

We have recently received the treebank developed at Univ. of Paris 7 by Anne Abeillé and are planning several experiments to further evaluate the quality of our parsers w.r.t. the information provided by this treebank.

6.10. Processing Botanical Corpora

Participants: Guillaume Rousse, Éric Villemonte de la Clergerie, François Role.

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.

The work effectued this year on BIOTIM is twofold.

First, the continuation of last year effort on NLP pipeline. We had to rework it fully for integrating tools developed during the EASY campaign by other team members, and for ensuring a better compliance to MAF.

Experiments on terminology extraction have been presented at TIA 2005 [31]. One of the problen we had was the poor typographic quality of the OCRized corpora and a new OCRization was done. However, despite the improved quality of the newly OCRized corpora, they are still issues with spelling errors and noise induced by the formatting (layout) of the original documents (pagination, illustrations, numerisation artifacts, etc...). Hence the need for some kind of input filtering.

Therefore, we started to work on retrieving the logical structuring of the corpora, to adress this very issue. With a generic regular expression based chunker, and corpus-specific configurations, we are able to segregate and label the various parts of interest in the document, mainly taxon descriptions. Domain specific integrity rules allow some automatic error correction for undetected patterns, such as missing taxons in taxonomic hierarchy. Coupled with morpho-syntactic processing with our our NLP pipeline, a preliminary study has been done by François Role to assess the possibility to extract an ontology and to represent it in OWL [19].

We have also started parsing some corpora, exploiting the logical structuring to remove the non pertinent parts (such as the bibliographical notices). We now need to assess the quality of the parsing, in order to eventually tune the meta-grammar to the very specific style of these botanical corpora. We also need to complete and enrich a domain specific lexicon. The next step, in collaboration with LIFO (Univ. of Orléans) will be to exploit the dependencies, produced during parsing, to extract a small lexical ontology.

6.11. Morphology and finite state transducers

Participant: François Barthélemy.

François Barthélemy worked on Finite-State Morphology, the approach of Natural Language Processing which consist in describing the morphology of human languages into various formalisms which are compiled into finite-state machines (automata and transducers). There are two main approaches:

- describing arbitrary relations between word components and actual forms using cascading contextual rewrite rules;
- describing same-length relations, where each symbol of the abstract component is mapped to exactly one symbol of actual forms and conversely, using two-level rules applied in parallel.

Following previous propositions (e.g. the work by Kiraz [49]), we investigated an intermediate approach where there is a mapping between substrings instead of single symbols, within the tuples of the relations. We defined a new class of transducers which is closed under intersection, whereas arbitrary transducers are not [21]. Intersection allows for modular descriptions and the morphology of a language may therefore be described as the intersection of local constraints. The transducers that are used are n-tape transducers which describe n-ary relations. They have the same theoretical power as Finite-State Automata, but are more convenient to describe morphology. We are currently working on some applications which demonstrate the interest of the formalism.

6.12. 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 (2003-2005)

Participant: Éric Villemonte de la Clergerie.

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 2005. É. de la Clergerie is project leader for a proposition of a morpho-syntactic annotation framework (MAF), which has been accepted as a new work item by ISO TC37SC4.

We have recently submitted a revised version of MAF for Committee Draft (CD) ballot [40], after incorporating remarks made during the last ISO meeting on MAF (Pisa, November 2004; Berlin, April 2005; Warsaw, August 2005). A small demonstrator for French, based on ATOLL's tools, has been activated and updated to illustrate our proposal.

É. de la Clergerie is also strongly involved in the standardization of feature structures using an XML representation, now reaching the DIS stage.

7.2. Action BIOTIM (2003 – 2006)

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

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, Benoît Sagot.

EASy Home page http://www.limsi.fr/Recherche/CORVAL/easy/

ATOLL has participated to the parsing evaluation campain EASY of action EVALDA of French program Technolangue. The campaign took place on mi-december 2004 with the participation of 14 parsers. The participants to EASy were expected to return information about 6 kinds of non recursive constituents (Nomimal chunks, Adjectival chunks, Adverbial chunks, verbal kernels, ...) and 14 kinds of dependencies (verb-subject, verb-object, ...). The evaluation was done on a set of around 35000 sentences covering various kinds of style (journalistic, literacy, mail, medical, speech, questions).

ATOLL has provided results for two parsers, namely FRMG and SXLFG [23]. We have received a preliminary evaluation for the constituents on a subset of 4262 sentences. We are still waiting for a complete set of results from the organizers, in particular regarding the dependencies.

7.4. Action LexSynt (2005 – ??)

Participants: Benoît Sagot, Éric Villemonte de la Clergerie.

LexSynt Home page http://lexsynt.inria.fr/

The action, funded by ILF (*Institut de Linguistique Française*), groups 13 teams, including INRIA teams ATOLL, Calligramme, Langue & Dialogue, and Signes. The main objective of this action is to design a reference syntactic-semantic lexicon for French. The work should take place in coordination with existing producers of lexicons (for merging resources), with grammar designers (to ensure usability in parsers), and with the current proposal **LMF** for the standardization of lexicons (ISO TC37 SC4).

7.5. Former action eCOTS

Participant: Bernard Lang.

Though INRIA is not a member of the eCOTS association resulting from a former collaboration with industry (closed in 2004), 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 participated 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. PAI Pessoa KLING (2005 - 2006)

Funding for visits has been granted by the French-Portuguese PAI (Programme d’actions intégrées) PESSOA to continue a long-lasting collaboration between ATOLL and team CENTRIA of Lisbon New University, led by Gabriel Pereira Lopes.

8.2.3. Former PAI PICASSO CATALINA-2

For administrative reasons, it was no longer possible to submit a new French-Spanish PAI (Programme d’actions intégrées) PICASSO between ATOLL and team COLE at Universities of La Coruña and of Vigo, led by Manuel Vilares Ferro. However, we have continued a program of visits, with a short visit by Manuel Vilares and a longer one by Francisco Jose Riberra Pena, planned from November 2005 to March 2006.

8.2.4. XTAG Collaboration

We have renewed some contacts with the group XTAG at University of Pennsylvania, in relation with Meta-Grammars.

8.2.5. ISO subcommittee TC37 SC4 on "Language Resources Management"

The participation of ATOLL to French Technolangue action Normalangue has resulted in a strong implication in ISO subcommittee TC37 SC4 on "Language Resources Management" (http://www.tc37sc4.org/). É. de la Clergerie has participated to ISO events (Berlin, April 2005; Warsaw, August 2005) and has played a role of expert (in particular on morpho-syntactic annotations [MAF], feature structures [FSR], and on the new work item on syntactic annotations [SynAF]).

8.3. Visits and invitations

A two weeks visit of Gabriel Pereira Lopes (New Univ. of Lisbon, Portugal) in September 2005 (PAI).

A one week visit of Manuel Vilares Ferro (from Univ. of Vigo, Spain) in August 2005.

A 3 months visit of Francisco Jose Ribadas Pena (Univ. of La Coruna) is planned from November 2005 to March 2006.

9. Dissemination

9.1. Animation at INRIA

B. Lang is an elected member of INRIA's "Conseil Scientifique" and a member of CUR (Research Unit Committee).

É. de la Clergerie is member of the GTAI subcommittee of COST committee.

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 co-supervises the PhD thesis of Benoît Sagot with Laurence Danlos (TALaNa/LATTICE, University Paris 7). He has also supervised the internship of Alexandra Mounier (CNAM Caen) on the handling of coordination (cf. 6.3).

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.
- É. de la Clergerie was a jury member for the PhD of Jesus Vilares Ferro, May 20, (La Coruña, Spain; supervisors: Miguel Alonso Pardo & José Luis Freire Nistal); thesis titled "Applicaciones del procesamiento del lenguaje natural en la recuperación en español".
- É. de la Clergerie was a thesis referee and jury member for the PhD of Sylvain Salvati, June 13rd, (CALLIGRAMME/LORIA, Nancy; supervisor: Ph. de Groote) thesis titled "Problèmes de filtrage et problèmes d'analyse pour les grammaires catégorielles abstraites".
- É. de la Clergerie is a thesis referee for the PhD of Tristan VanRullen (Univ. of Aix en Provence; supervisor: Ph. Blache); thesis titled "Vers une analyse syntaxique à granularité variable".

9.4. Committees

- Participation of É. de la Clergerie to the editorial board of French journal T.A.L. http://www.atala.org/rubrique.php3?id_rubrique=1.
- Participation of É. de la Clergerie to the program committees of IWPT'05 (*International Workshop on Parsing Technologies*, Vancouver, October, http://bulba.sdsu.edu/iwpt05/), for EPIA workshop TEMA'05 (*Text Mining and Applications*, Covihã, December, http://tema.epia05.di.ubi.pt/), and for TALN'06 (*Traitement Automatique des Langues Naturelles*, Leuven, April 2006, http://www.taln.be). He has also reviewed papers for Language Resources and Evaluation (journal), IJCNLP'05, ESSLLI'05 students conference, ICALP'05, and EACL'06.
- Participation of P. Boullier to the program committee of FG-MOL 2005 (10th conference on Formal Grammar and The 9th Meeting on Mathematics of Language, Edinburgh, Scotland, 5-7 August 2005). He has also reviewed papers for TCS (journal), and EACL'06.
- Participation of B. Sagot to the student program committee of LACL'05 (*Logical Aspects of Computational Linguistics*, Bordeaux, http://www.labri.fr/projet/signes/LACL/aac.htm). He has also reviewed papers for IJCNLP'05.
- 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.

9.5. Softwares

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

9.6. Participation to workshops, conferences, and invitations

- Participation of É. de la Clergerie to ISO TC37SC4 meetings (Berlin, Germany, April, and Warsow, Polland, August).
- É. de la Clergerie has presented DYALOG at TALANA seminar (Januray) and Meta-Grammars at the "Institut Universitari de Linguistica Aplicada" (IULA, Univ. Pompeu Fabra, Barcelona, Spain, October), at LORIA (Nancy, October), and at SIGNES (INRIA Futurs, Bordeaux, December). He has presented BIOTIM action at Microsoft Research Center (Cambridge, September).
- Participation with presentations of É. de la Clergerie at TALN'05 (*Traitement Automatique des Langues Naturelle*, Dourdan), LT&C'05 (2nd *Language and Technology Conference*, Pozdan, Polland), TIA'05 (6th meeting on *Terminologie et Intelligence Artificielle*, Rouen), CSLP'05 (2nd International Workshop on *Constraint Solving and Language Processing*, Barcelona, Spain), IWPT'05 (9th *International Workshop on Parsing Technologies*, Vancouver, CA).
- Participation with presentations of B. Sagot at LACL'05 (*Logical Aspects of Computational Linguistics*, Bordeaux), TALN'05 (Dourdan), LT&C'05 (Poznań, Poland), TSD (8th int. conference on *Text, Speech, and Dialogue*, Karlovy Vary, Czech Republic), and IWPT'05 (Vancouver, CA). He has also presented LEFFF at the ATALA meeting on *Interface lexique-grammaire et lexiques syntaxiques et sémantiques* (Paris, March 12). B. Sagot also coordinates the seminar TALaNa.
- Participation with presentation of Areski Nait Abdallah to LACL'05 (Bordeaux).

- Presentation of “Traitement du langage naturel en Perl” (NLP Processing in Perl) by G. Rousse to the “Journées Francophones de Perl 2005”, Luminy (http://lis.snv.jussieu.fr/~rousse/recherche/JFP2005/html/tables/index.html Participation to TIA’05.
- Participation of P. Boullier, Ph. Deschamp, and F. Thomasset to TALN’05.
- B. Lang was invited to participate to a working group on Free Software organized in Brussels by the Information Society Directorate, on march 22nd.
- Participation of B. Lang to a meeting organized by CGTI (Conseil général des technologies de l’information) on the theme : "Les logiciels libres dans les administrations publiques européennes : état des lieux" (march 24th).
- B. Lang has organized and moderated a panel on "Free and Proprietary software in COTS-based software development" at the ICCBSS conference in Bilbao (Spain) on February 9th. http://www.iccbss.org/2005/iccbss_program.html
- B. Lang gave two présentations at the annual conference of ATUL, the Tunisian Free Software association in Tunis (February 24-25).
- B. Lang has been invited by INTIF to speak at the "Rencontres Africaines des Logiciels Libres (RALL 2005) in Libreville (Gabon), october 19th to 21st. This includes a session at the Gabonese parliament.
- B. Lang gave a presentation on issues related to software patenting at the LacFREE 2005 conference (II Conference on Development and Use of Free Software in Latin America and the Caribbean) in recife (Brazil) on December 5-8. http://www.lacfree2005.org/?q=en
- B. Lang has delivered a presentation on intellectual property, and on the use of "free" licensing and trademarking applicable to linguistic resources. LexSynt meeting, Paris, September 12th.
- Participation and contribution of B. Lang to several meetings on the potential of Free Software, and on économic, legal and political issues:

  – Conference on free software in education and enterprises at the university of Valencienne (40th anniversary), on january 25th.
  – Presentation of free software at the Workshop on Statistical Cartography organized by the Ministère de l’équipement, Paris, on April 12th.
  – Presentation at the debate "Nouvelles technologies et démocratie : le cas des logiciels libres", organized by SNES-FSU (Syndicat National des Chercheurs Scientifiques) in Nanterre on May 27th. http://www.sncs.cnrs-bellevue.fr/IMG/pdf/logiciels-libres_27-05-051.pdf
  – Presentation on the theme "Le savoir et les formes de son appropriation" at the FSU seminar. http://institut.fsu.fr/chantiers/eco_connaissance/pres.htm
  – Participation to a panel on economic issues of free softwre at Educatec, in Paris, on November 25th

10. Bibliography

Major publications by the team in recent years

[1] M. A. ALONSO PARDO, É. VILLEMONTE DE LA CLERGERIE, V. J. DIAZ, M. VILARES. *New Developments in Parsing Technology*, J. CARROLL, G. SATTA, H. BUNT (editors). , Text, Speech and Language Technology, revised notes of a paper for IWPT2000, vol. 23, chap. Relating Tabular Parsing Algorithms for LIG and TAG, Kluwer Academic Publishers, 2004, p. 157–184, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/KAP_chapter8.Giorgio_John.pdf.

[2] P. BOULLIER. *A Cubic Time Extension of Context-Free Grammars*, in "Grammars", vol. 3, n° 23, 2000.

[3] P. BOULLIER. *On TAG Parsing*, in "Traitement Automatique des Langues (T.A.L.)", issued June 2001, vol. 41, n° 3, 2000, p. 111-131.

[4] P. BOULLIER. *Counting with Range Concatenation Grammars*, in "Theoretical Computer Science", vol. 293, 2003, p. 391–416.

[5] 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.

[6] P. BOULLIER. *New Developments in Parsing Technology*, G. S. JOHN CARROLL, H. BUNT (editors). , Text, Speech and Language Technology, vol. 23, chap. Range Concatenation Grammars, Kluwer Academic Publishers, 2004, p. 269–289, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/KAP_chapter12.Giorgio_John.pdf.

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

[8] B. LANG. *Towards a Uniform Formal Framework for Parsing*, M. TOMITA (editor). , Kluwer Academic Publishers, 1991, p. 153-171, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Bernard.Lang/framework.ps.Z.

[9] 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.

[10] 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.

[11] 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 refereed journals and book chapters

[12] P. BOULLIER. *a non-statistical parsing-based approach*, A. JOSHI, S. BANGALORE (editors). , to appear, MIT Press, 2005.

[13] P. BOULLIER, B. SAGOT. *Analyse syntaxique profonde à grande échelle: SxLFG*, in "Traitement Automatique des Langues (T.A.L.)", Submitted, October 2005.

[14] B. LANG. *Brevet, Logiciel libre et Innovation*, in "La vie de la recherche scientifique", n° 360, February 2005, p. 18-21.

[15] B. LANG. *Libres, les logiciels !*, in "CAES magazine", to appear, n° 77, hiver 2005.

[16] B. LANG. *Logiciels libres et développement*, in "la revue nouvelle", n° 6-7, juin-juillet 2005, p. 56-62.

[17] B. LANG. *Matching with multiplication and exponentiation*, in "Mathematical Structures in Computer Science", vol. 15, n° 05, October 2005, p. 959-968.

[18] B. LANG, M. LINGLET. *Réflexion sur les mécanismes de contrôle de la communication et de l'économie*, in "Expertises", n° 296, October 2005, p. 330-335.

[19] F. ROLE, G. ROUSSE. *Construction incrémentale d'une ontologie par analyse du texte et de la structure des documents*, in "Document numérique", To appear, 2005.

[20] B. SAGOT. *From raw corpus to word lattices: robust pre-parsing processing*, in "Archives of Control Sciences", Submited to the special issue of selected papers from LTC'05, October 2005.

Publications in Conferences and Workshops

[21] F. BARTHÉLEMY. *Partitioning Multitape Transducers*, in "Finite State Methods in Natural Language Processing (FSMNLP), Helsinki (Finland)", September 2005.

[22] P. BOULLIER, L. CLÉMENT, B. SAGOT, É. VILLEMONTE DE LA CLERGERIE. *Chaînes de traitement syntaxique*, in "Proceedings of TALN'05, Dourdan, France", ATALA, June 2005, p. 103–112, http://atoll.inria.fr/~sagot/pub/TALN05easy+sxpipe.pdf.

[23] P. BOULLIER, L. CLÉMENT, B. SAGOT, É. VILLEMONTE DE LA CLERGERIE. *« Simple comme EASy :-) »*, in "Proceedings of TALN'05 EASy Workshop (poster), Dourdan, France", ATALA, June 2005, p. 57–60, http://atoll.inria.fr/~sagot/pub/TALN05easyworkshop.pdf.

[24] P. BOULLIER, B. SAGOT. *Efficient and robust LFG parsing: SxLfg*, in "Proceedings of IWPT'05, Vancouver, Canada", October 2005, p. 1–10, http://atoll.inria.fr/~sagot/pub/IWPT05.pdf.

[25] P. BOULLIER, B. SAGOT. *Efficient parsing of large corpora with a deep LFG parser*, in "Proc. of LREC'06", submitted, October 2005.

[26] P. BOULLIER, B. SAGOT, L. CLÉMENT. *Un analyseur LFG efficace pour le Français: SXLFG*, in "Proceedings of TALN'05 (poster), Dourdan, France", ATALA, June 2005, p. 403–408, http://atoll.inria.fr/~sagot/pub/TALN05sxlfg.pdf.

[27] L. CLÉMENT, É. VILLEMONTE DE LA CLERGERIE. *MAF: a Morphosyntactic Annotation Framework*, in

"proc. of the 2nd Language & Technology Conference (LT'05), Poznan, Poland", April 2005, p. 90–94, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/LTC05.pdf.

[28] 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, 2005.

[29] B. LANG. *Software Patentability: a computer research scientist's view*, in "Comptes-rendus de la Conférence LacFree 2005, Recife-Olinda (Brésil)", Extended version of a French version, December 2005.

[30] A. NAIT ABDALLAH, A. LECOMTE. *On expressing vague quantification and scalar implicatures in the logic of partial information*, in "Lecture Notes in Computer Science 3492 (© Springer-Verlag), Proceedings of LACL'05, Bordeaux, France", April 2005, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Areski.NaitAbdallah/LACL05.ps.

[31] G. ROUSSE, É. VILLEMONTE DE LA CLERGERIE. *Analyse automatique de documents botaniques: le projet Biotim*, in "proc. of TIA'05, Rouen, France", April 2005, p. 95–104, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/tia2005.pdf.

[32] B. SAGOT, P. BOULLIER. *From raw corpus to word lattices: robust pre-parsing processing*, in "proc. of the 2nd Language & Technology Conference (LT'05), Poznan, Poland", Selected for potential journal publication, April 2005, p. 348–351, http://atoll.inria.fr/~sagot/pub/LTC05.pdf.

[33] B. SAGOT, L. CLÉMENT, É. VILLEMONTE DE LA CLERGERIE, P. BOULLIER. *The Lefff 2 syntactic lexicon for French: architecture, acquisition, use*, in "Proc. of LREC'06", submitted, October 2005.

[34] B. SAGOT. *Automatic acquisition of a Slovak lexicon from a raw corpus*, in "Lecture Notes in Artificial Intelligence 3658 (© Springer-Verlag), Proceedings of TSD'05, Karlovy Vary, Czech Republic", September 2005, p. 156–163, http://atoll.inria.fr/~sagot/pub/TSD05.pdf.

[35] B. SAGOT. *Les Méta-RCG: description et mise en oeuvre*, in "Proceedings of TALN'05 (poster), Dourdan, France", ATALA, June 2005, p. 493–498, http://atoll.inria.fr/~sagot/pub/TALN05metarcg.pdf.

[36] B. SAGOT. *Linguistic facts as predicates over ranges of the sentence*, in "Lecture Notes in Computer Science 3492 (© Springer-Verlag), Proceedings of LACL'05, Bordeaux, France", April 2005, p. 271–286, http://atoll.inria.fr/~sagot/pub/LACL05.pdf.

[37] F. THOMASSET, É. VILLEMONTE DE LA CLERGERIE. *Comment obtenir plus des Méta-Grammaires*, in "Proceedings of TALN'05, Dourdan, France", ATALA, June 2005, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/mg05.pdf.

[38] É. VILLEMONTE DE LA CLERGERIE. *From Metagrammars to Factorized TAG/TIG Parsers*, in "Proceedings of IWPT'05 (poster), Vancouver, Canada", October 2005, p. 190–191, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/IWPT05sub.pdf.

[39] É. VILLEMONTE DE LA CLERGERIE. *DyALog: a Tabular Logic Programming based environment for NLP*, in "Proceedings of 2nd International Workshop on Constraint Solving and Language Processing (CSLP'05), Barcelona, Spain", October 2005, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/CSLP05.pdf.

Internal Reports

[40] É. VILLEMONTE DE LA CLERGERIE. *Terminology and other language resources – Morpho-Syntactic Annotation Framework (MAF)*, Submitted for CD ballot, Committe Draft, n° ISO TC37 SC4 WG2 N119 Rev3, ISO, October 2005.

Miscellaneous

[41] B. SAGOT, L. CLÉMENT, É. VILLEMONTE DE LA CLERGERIE, P. BOULLIER. *Vers un méta-lexique pour le français : architecture, acquisition, utilisation*, Journée d'étude de l'ATALA sur l'Interface lexique-grammaire et lexiques syntaxiques et sémantiques, March 2005, http://atoll.inria.fr/~sagot/pub/JourneeATALA.pdf.

[42] É. VILLEMONTE DE LA CLERGERIE. *Concevoir des analyseurs syntaxiques avec DyALog à partir de méta-grammaires*, Slides presented at TALaNa, University Paris 7, January 2005, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/TALANA05-small.ps.gz.

[43] É. VILLEMONTE DE LA CLERGERIE. *From Meta-Grammars to Factorized grammars*, Slides presented at IULA, Universitat Pompeu Fabra, Barcelona, October 2005.

Bibliography in notes

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

[45] 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.

[46] O. DUCROT. *Les échelles argumentatives*, Éditions de Minuit, 1980.

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

[48] 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.

[49] G. A. KIRAZ. *Computational Nonlinear Morphology*, Cambridge University Press, 2001.

[50] A. NAIT ABDALLAH. *The Logic of Partial Information*, Springer, 1995.

[51] 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.

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