INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Alpage

Analyse Linguistique Profonde À Grande Échelle

Paris - Rocquencourt

THEME SYM

Table of contents

Alpage is a common project with University Paris 7. The team was created on July the 1st, 2007 and became an INRIA project on February the 1st, 2008. Starting January 1st, 2009, Alpage will be an UMR-I (University Paris 7 & Inria) registered in the Paris 7 quadriennal plan.

1. Team

Research Scientist

Pierre Boullier [Research Director (DR) Inria, HdR]
Pascal Denis [Research Associate (CR) Inria]
Eric de La Clergerie [Research Associate (CR) Inria]
Benoît Sagot [Research Associate (CR) Inria]

Faculty Member

François Barthelemy [Associate Professor (MC) CNAM]
Marie Candito [Associate Professor (MC) Univ. Paris 7]
Benoît Crabbé [Associate Professor (MC) Univ. Paris 7]
Laurence Danlos [Full Professor (PR) Univ. Paris 7, Member of IUF, Scientific leader of Alpage, HdR]
Sylvain Kahane [Full Professor (PR) Univ. Paris X, HdR]
Djamé Seddah [Associate Professor (MC) Univ. Paris 4]

Technical Staff

Isabelle Cabrera [Associate Engineer (IA) Inria (until September)]

PhD Student

André Bittar [PhD student (allocataire) Univ. Paris 7 (since 2007)]
François-Régis Chaumartin [PhD student Univ. Paris 7]
Elżbieta Gryglicka [PhD student (CIFRE) Thales & Univ. Paris 7]
Pierre Hankach [PhD student (CIFRE) France Telecom & Univ. Paris 7 (since 2006)]
Juliette Thuillier [PhD student (allocataire) Univ. Paris 7 (since September)]

Visiting Scientist

Milagros Fernandez Gavilanes [PhD student Univ. of Vigo (Spain), 1-month visit in November and December]
Darja Fišer [PhD student Univ. of Ljubljana (Slovenia), 1-month visit from January to February]
Miguel Ángel Molinero Álvarez [PhD student Univ. of Ourense (Spain), 2-month visit in November and December]
Giorgio Satta [Full professor Univ. de Padua (Itay), visitor from April to June, HdR]
Sahil Thapa [3-month internship]

Administrative Assistant

Christelle Guiziou [Secretary (SAR) Inria]

2. Overall Objectives

2.1. Overall Objectives

Alpage is a joint team with University Paris 7 (Department of Linguistics) that was created in July 2007, with members coming in majority from the former Paris 7 Talana team (member of the Lattice UMR) and INRIA former project-team Atoll. Both teams were specialized in Natural Language Processing (NLP, in French: TAL, for *Traitement Automatique des Langues*), the former with a strong linguistic background, the latter with a strong computational background. Since February 2008, Alpage is a full Inria project-team. Starting January 1st, 2009, Alpage will be an UMR-I (University Paris 7 & Inria) registered in the Paris 7 quadriennal plan.

The Alpage team is specialized in **Language modeling**, **Computational linguistics** and **Natural Language Processing (NLP)**. These fields are considered central in the new Inria strategic plan, and are indeed of crucial importance for the new information society. Applications of this domain of research include the numerous technologies grouped under the term of "language engineering" (information retrieval, information extraction, spelling, grammatical and semantic correction, automatic summarizing, machine translation, man machine communication, etc).

NLP, the domain of Alpage, is a subfield of both artificial intelligence, linguistics, and cognition. It studies the problems of automated understanding and generation of natural human languages. Natural language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate. Natural language generation systems convert information from computer databases into human language. Alpage focuses on *text* understanding and, to a lesser extent, generation (by opposition to speech processing and generation).

NLP applications are numerous, and include machine translation, question answering, information retrieval, information extraction, text simplification, automatic or computer-aided translation, automatic symmetrization, foreign language reading and writing aid, and others.

NLP is a multidisciplinary domain. Indeed, it requires an expertise in formal and descriptive linguistics (to develop linguistic models of human languages), in computer science and algorithmics (to design and develop efficient programs that can deal with such models), in applied mathematics (to acquire automatically linguistic or general knowledge) and in other related fields. It is one of the specificities of Alpage to put together NLP specialists with a strong background in all these fields (in particular, linguistics for Paris 7 Alpage members, previously in the Lattice UMR, computer science and algorithmics for Inria members).

One specificity of NLP is the diversity of human languages it has to deal with. Alpage focuses on French and English, but does not ignore other languages, through collaborations, in particular with those that are already studied by its members or by long-standing collaborators (e.g., Arabic, English, Polish, Slovak, Spanish). This is of course of high relevance, among others, for language-independant modeling and multi-lingual tools and applications.

Alpage's overall objective is to develop linguistically relevant *and* computationally efficient tools and resources for natural language processing and its applications. More specifically, Alpage focuses on the following topics:

- Research topics:
  - deep syntactic modeling and parsing. This topic includes, but is not limited to, development of advanced parsing technologies, development of large-coverage and high-quality adaptive linguistic resources, and use of hybrid architectures coupling shallow parsing, deep parsing, and (probabilistic and symbolic) disambiguation techniques;
  - modeling and processing of language at a supra-sentential level (discourse modeling and parsing, anaphora resolution, etc);
  - NLP-based knowledge acquisition techniques
- Application domains:
  - automatic information acquisition (both linguistic information, inside a bootstrapping scheme for linguistic resources, and document content, with a more industry-oriented perspective);
  - text mining;
  - automatic generation;
  - with a more long-term perspective, automatic or computer-aided translation, which is an historical domain of expertise for Talana.

2.2. Highlights

In 2008, Alpage has carried out numerous achievements in each of its domains of expertise (formal languages, linguistic modeling, lexical resources, surface processing, symbolic and statistic parsing, discourse modeling). Among them, the following can be highlighted:

- Alpage has managed to parse very large corpora (hundreds of millions of words) with its comprehensive syntactic processing chain (pre-processing, deep parsing, disambiguation and syntactic-semantic relations extraction);
- Alpage has developed a state-of-the-art probabilistic parser for French that performs almost as well as symbolic parsers developed for a long time (including within Alpage) on journalistic corpora;
- Alpage's lexical resources have reached a new level of maturity, thanks to the Alexina framework, hence allowing:
  - to initiate the fast development of resources for other languages;
  - a qualitative and quantitative improvement of its syntactic lexicon for French (the Le*fff*);
  - the development of a new semantic lexicon (wordnet) for French, the WOLF;
- Alpage's formalism for modeling discourse structures, D-STAG, has been fully specified and motivated, hence allowing for a preliminary implementation, whose development is already ongoing.
- finally, some of Alpage's tool have been applied, or are about to be applied, in operational industrial contexts, within the *vera* software and the TEXT-ELABORATOR tool.

3. Scientific Foundations

3.1. From programming languages to linguistic grammars

Keywords: *CFG*, *MCS formalisms*, *unification-based formalisms*.

Participants: Pierre Boullier, Éric de La Clergerie, Benoît Sagot, Giorgio Satta.

CFG context-free grammars

MCS formalisms Mildly Context-Sensitive formalisms are a class of formalisms that is stricly more powerful than CFGs, but stricly less powerful than formalisms that cover the class of all languages recognizable in polynomial time

Historically, several members of Alpage were originally specialists in the domain of modeling and parsing for programming languages, and are working for more than 10 years on the generalization and extension of the techniques involved to the domain of natural language. 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 consensual grammatical formalism has yet been accepted by the whole 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 three following large families:

Mildly Context-Sensitive (MCS) 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) and Multi-component TAGs with trees as elementary structures, Linear Indexed Grammars (LIGs). Although they are strictly more powerful than MCS formalisms, Range Concatenation Grammars (RCGs, introduced and used by Alpage members, such as Pierre Boullier and Benoît Sagot [47], [72], [75]) are also parsable in polynomial time.

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) and Head-Driven Phrasal Structure Grammars (HPSGs) rely on more expressive Typed Feature Structures (TFS) or constraints.

Unification-based formalisms with an MCS backbone The two above-mentioned characteristics may be combined, for instance by adding logic arguments or constraints to non-terminals in TAGs.

However, despite this diversity, convergences may be found between these formalisms and most of them 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.

3.2. Metagrammars

Keywords: *Metagrammar*, *TAG*.

Participants: Éric de La Clergerie, Benoît Crabbé, Marie Candito.

Metagrammar a metagrammar is a grammatical description that is an abstraction of the grammar level; a metagrammar is composed of classes that include elements of grammatical description and combination constraints; classes are combined, and their elements of grammatical description are merged, according to these combination contraints into final classes; the combination of grammatical descriptions contained in final classes constitute a grammar in the usual sense of the term

TAG Tree-Adjoining Grammar

LFG Lexical-Functional Grammar

For hand-crafted grammars, some Alpage members try to design adequate tools and adequate levels of representation for linguists, and in particular Meta-Grammars [83], [80]. Meta-Grammars 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. 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.

Inside Alpage, both Éric de La Clergerie (`mgcomp` system, FRMG metagrammar) and Benoît Crabbé (XMG system, Benoît Crabbé's metagrammar) are foreground actors of the development and implementation of these notions. It is also worth noting that this emergence of the MG notion is a good illustration of this cross-fertilization between ex-Talana members (the birth place of MGs) and ex-Atoll members.

3.3. Symbolic parsing techniques

Keywords: *Parsing*, *shared parse forest*.

Participants: Pierre Boullier, Éric de La Clergerie, Benoît Sagot.

The existence of a continuum of grammatical formalisms, from CFGs and TAGs to LFGs, RCGs, and even Meta-Grammars, 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 various kinds of automata to describe parsing strategies for complex formalisms. Dynamic Programming interpretation of automata derivations are then used to handle large scale level of ambiguities.

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.

3.3.1. Multi-pass approach

As is usually done for programming language parsing, NLP parsing can be broken into several successive phases of increasing complexity : lexical analysis, shallow parsing (e.g., chunk parsing), parsing (e.g., building LFG constituency trees/forests), "semantics" (in the sense of compilation theory, i.e., attributes computation, such as so-called LFG functional structures, or n-best computation based on probabilistic models),...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 constituency structures, which requires, at least, Context-Free Grammars. Similarly, CFGs are not powerful enough to describe some contextual phenomena needed in dependencies computation. Beside a better efficiency (each phase being handled with the best level of complexity), decomposing increases modularity.

Indeed, most formalisms found in the above-mentioned Horn continuum are structured by a non-contextual backbone (this includes not only CFG-equivalent formalisms as well as LFG, but also many variants of HPSG, and many grammars developed in the TAG framework). This backbone may be first parsed with SYNTAX , a very efficient and generic non-contextual parser generator developed mostly by Pierre Boullier and distributed as an open-source software[1] [45], [46]. More formalism-specific treatment can then be applied to check additional constraints, as done by Pierre Boullier and Benoît Sagot for chunk-level parsing and LFG functional structures computation [48], [50], [49].

3.3.2. 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. 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 allowed us to handle formalisms such as TAGs and LIGs. More recently, Thread Automata (TA) 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 or parsing as deduction and generalizes several approaches found in Parsing but also in Logic Programming. The DYALOG system, developed by Éric de La Clergerie [82] implements this approach for Logic Programming and several grammatical formalisms. It is used by Alpage members to develop efficient TAG parsers (e.g., Éric de La Clergerie's FRMG and Benoît Crabbé's French TAG parser), but also by several French and foreign teams [80], [83].

3.3.3. Shared parse and derivation forests

Both previously presented approaches share several characteristics, for instance the use of dynamic programming ideas and 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. 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).

[1] SYNTAX is also used in project-team VASY in the domain it has been first developed for, namely programming languages.

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, ...), especially after conversion to dependency forests providing dependency information directly between words. Disambiguation algorithms, both symbolic and probabilistic (if quantitative data is available) can also be applied on such shared structures.

3.4. Dynamic wide coverage lexical resources

Keywords: *Lexical resources*, *dynamic resources*.

Participants: Benoît Sagot, Laurence Danlos, Éric de La Clergerie.

Grammatical formalisms and associated parsing generators are useful only when used together with linguistic resources (lexicons, grammars) so as to build operational parsers, especially when considering modern lexically oriented grammatical formalisms. Hence, linguistic resources are the topic of the following section.

However, wide coverage linguistic resources are scarce and expensive, because they are difficult to build, especially when hand-crafted. This observation motivates us to investigate methods, along to manual development techniques, to automatically or semi-automatically acquire, supplement and correct linguistic resources.

Linguistic expertise remains a very important asset to benefit efficiently from such techniques, including those described below. Moreover, linguistically oriented environments with adequate collaborative interfaces are needed to facilitate the edition, comparison, validation and maintenance of large scale linguistic resources. Just to give some idea of the complexity, a syntactic lexicon, as described below, should provide rich information for several tens of thousands of lemma and several hundreds of thousands of forms.

Successful experiments have been conduced by Alpage members with different languages for the automatic acquisition of morphological knowledge from raw corpora. At the syntactic level, work has been achieved on automatic acquisition of atomic syntactic information. All such techniques need of course to be followed by manual validation, so as to ensure high-quality results.

For French, these techniques, and others, have lead some Alpage members (both Inria and Paris 7) to develop one of the main syntactic resources for French, the Le*fff* .

In the last 2 years, Alpage members have shown how to benefit from other more linguistically-oriented resources, such as the Lexique-Grammaire and DICOVALENCE , in order to improve the coverage and quality of the Le*fff* . This work is a good example of how Inria and Paris 7 members of Alpage fruitfullt collaborate: this collaboration between NLP computer scientists and NLP linguists have resulted in significant advances which would have not been possible otherwise.

3.5. Treebanks development and exploitation

Keywords: *Treebank*.

Participants: Benoît Crabbé, Marie Candito, Éric de La Clergerie.

Treebank a treebank is a set of sentences whose syntactic analysis has been performed manually (it is called a “treebank” in reference to the fact that in most cases, these analyses are represented as trees, be them constituency or dependency trees)

At the international level, the last decade has seen the emergence of a very strong trend of researches on statistical methods in NLP. This trend results from several reasons but one of them, in particular for English, is the availability of large annotated corpora, such as the Penn Treebank (1M words extracted from the Wall Street journal, with syntactic annotations) or the the British National Corpus (100M words covering various styles annotated with parts of speech). Such annotated corpora are very valuable to extract stochastic grammars or to parametrize disambiguation algorithms.

These successes have lead to many similar proposals of corpus annotations. A long (but non exhaustive) list may be found on the internet[2] and includes mostly resources for languages other than French, apart from the French Treebank, developed in Anne Abeillé's team at University Paris 7.

However, the development of such treebanks is very costly from a human point of view and represents a long standing effort. The volume of data that can be manually annotated remains limited and is generally not sufficient to learn very rich information (sparse data phenomena). Furthermore, designing an annotated corpus involves choices that may block future experiments to acquire new kinds of linguistic knowledge. Last but not least, it is worth mentioning that even manually annotated corpora are not error prone.

Hence, two directions are investigated by Alpage members, and will be of increasing importance. First, Alpage members are working actively on the exploitation of the French Treebank for developing probabilistic parsers, as described in section 6.4.

Second, a bootstrapping approach is also investigated, where corpora can be parsed by many different parsing systems, so as to build automatically a consensual treebank which can reach a very large size (typically 100-million words); such a treebank (or parsing results from individual parsers) can be used to acquire linguistic information so as to enrich lexica, leading to better parsers. This has been achieved for example at Alpage thanks to error mining techniques in parsing results, and the Passage ANR project, lead by Éric de La Clergerie, applies this bootstrapping approach at a national level. Such an approach leads to resources and parsers that co-evolve, in a virtuous circle: resources are used by tools on corpus to improve resources and prepare the next generation of resources (by adding richer information). This constitutes the first steps towards the definition of generic learning algorithms, not relying on costly manually annotated corpora.

3.6. Building and evaluating full-featured parsing systems

Keywords: *Parsing systems*, *Pre-processing*.

Participants: Éric de La Clergerie, Benoît Sagot, Pierre Boullier.

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, Alpage is the leader of the Passage ANR project, which is the follow-up of the EASy parsing evaluation campaign held in 2004 and conducted by team LIR at LIMSI.

However, although developing parsing techniques, grammars, and lexica constitute obviously the key efforts towards deep large-scale linguistic processing, these components need to be included inside a full and robust processing chain, able to handle any text from any source. The development of such linguistic chains, such as SXPipe , is not a trivial task. Moreover, when used as a preliminary step before parsers, the quality of parsers' results strongly depends on the quality of such chains.

3.7. Standardization

Keywords: *Standardization*.

Participants: Éric de La Clergerie, Benoît Sagot.

Standardization the process of developing and agreeing upon technical standards, including formats, e.g., for storing corpora or lexicons.

Both evaluation and integration of parsing systems raise the general problem of standardization. Interoperability between software components and linguistic resources is vital so as to be able to improve and enrich them by collaborating with other teams, be them French or not. This pushed the community to get involved in standardization efforts, both at a national and international level. Some Alpage members are committed in several AFN OR and ISO standardization committees (Technolangue action Normalangue; ISO TC37SC4: work on MAF "Morphosyntactic Annotation Framework", FSR/FSD "feature Structures" and SynAF "Syntactic Annotation Framework").

[2] http://www.ims.uni-stuttgart.de/projekte/TIGER/related/links.shtml

3.8. Discourse structures

Keywords: *Discourse structures*, *RST*, *SDRT*, *TAG*.

Participants: Laurence Danlos, Pascal Denis, Benoît Sagot.

Collaboration with Nicholas Asher (IRIT, Toulouse).

SDRT Segmented Discourse Representation Theory

RST Rhetorical Structure Theory

TAG Tree-Adjoining Grammar

Until now, the linguistic modeling and automatic processing of sentences has been the main focus of the community. However, many applications would benefit from more large-scale approaches which go beyond the level of sentences. This is not only the case for automatic translation: information extraction/retrieval, summarizing, and other applications do need to resolve anaphoras, which in turn can benefit from the availability of hierarchical discourse structures induced by discourse relations (in particular through the notion of right frontier of discourse structures). Moreover, discourse structures are required to extract sequential (chronological, logical,...) or hierarchical representations of events. It is also useful for topic extraction, which in turns can help syntactic and semantic disambiguation.

Although supra-sentential problematics received increasing attention in the last years, there is no satisfying solution to these problems. Among them, anaphora resolution and discourse structures have a far-reaching impact and are domains of expertise of Alpage members. But their formal modeling has now reached a maturity which allows to integrate them, in a near future, inside future Alpage tools, including parsing systems inherited from Atoll.

It is well known that a text is not a random sequence of sentences: sentences are linked the ones to the others by "discourse relations", which give to the text a hierarchical structure. Traditionally, it is considered that discourse relations are lexicalized by connectors (adverbial connectors like *ensuite*, conjunctions like *parce que*), or are not lexicalized. This vision is however too simple:

- first, some connectors (in particular conjunctions of subordination) introduce pure modifiers and must not be considered as bearing discourse relations,
- second, other elements than connectors can lexicalize discourse relations, in particular verbs like *précéder / to precede* or *causer / to cause*, which have facts or fact eventualities as arguments [58].

There are three main frameworks used to model discourse structures: RST, SDRT , and, more recently, D-LTAG. Inside Alpage, Laurence Danlos has introduced D-STAG (Discourse Synchronous TAGs, [59]), which subsumes in an elegant way both SDRT and RST, to the extent that SDRT and RST structures can be obtained by two different partial projections of D-STAG structures. As done in D-LTAG, D-STAG extends a lexicalized TAG analysis so as to deal with the level of discourse. D-STAG has been fully formalized, and is hence possible to implement (thanks to Synchronous TAG, or even TAG parsers), provided one develops linguistic descriptions in this formalism.

3.9. Coreference resolution

Keywords: *Coreference*.

Participants: Pascal Denis, Elżbieta Gryglicka, Laurence Danlos, Benoît Sagot.

Collaboration with Nicholas Asher (IRIT, Toulouse).

Coreference coreference occurs when multiple expressions in a sentence or document have the same referent.

An important challenge for the understanding of natural language texts is the correct computation of the *discourse entities* that are mentioned therein —persons, locations, abstract objects, and so on. In addition to identifying individual referential expressions (e.g., *Nicolas Sarkozy*, *Neuilly*, *l'UMP*) and properly typing them (e.g. *Nicolas Sarkozy* is a PERSON, *Neuilly* is a LIEU), the task is also to determine the other mentions with which these expressions are coreferential. Part of the difficulty of this task is that natural languages provide many ways to refer to the same entity (including the use of pronouns such as *il*, *ses* and definite descriptions such as *le prÈsident*, making them highly ambiguous. The identification of coreferential links and other anaphoric links (such as "associative anaphora") plays a key role for various applications, such as extraction and retrieval of information, but also the summary or automatic systems question-answer. This central role of coreference resolution has been recognized by the inclusion of this task in different international evaluation campaigns, beginning with the campaigns *Message Understanding Conference* (in particular, MUC-6 and MUC-7)[3], and more recently *Automatic Content Extraction* (ACE)[4] and *Anaphora Resolution Evaluation* (ARE)[5]. The creation and distribution of corpora developped as part of these campaigns have significantly boosted research in automatic coreference resolution. In particular, they have made possible the application of machine learning techniques (mostly supervised ones) to the problem of coreference resolution. This in turn has led to the development of systems that were both more robust and more precise, thus making more realistic their integration within these larger systems. Some of the best systems based on supervised learning methods are described in [79], [68], [67], [69], [65], [61]. Note that a few attemtps were also made at using unsupervised techniques (mostly clustering methods) for the task [52], [70], but these systems are still far from rivaling their supervised counterparts.

4. Application Domains

4.1. Panorama

Keywords: *Automatic summarization*, *Automatic translation*, *Computer-aided translation Information extraction*, *Information acquistion*, *Information retrieval*, *Language writing aid*, *Man-machine communication*, *Question-answering systems*, *Spelling correction*, *Text mining*.

NLP tools and methods have many possible application domains, some of which are already mature enough to be commercialized. They can be roughly classified in three groups:

Man-Machine Communication : mostly speech processing and text-to-speech, often in a dialogue context; today, commercial offers are limited to restricted domains (train tickets reservation...);

Language writing aid : spelling, grammatical and stylistic correctors for text editors, controlled-language writing aids (e.g., for technical documents), memory-based translation aid, foreign language learning tools, as well as vocal dictation;

Access to information : tools to enable a better access to information present in huge collections of texts (e.g., the Internet): automatic document classification, automatic document structuring, automatic summarizing, information acquisition and extraction, text mining, question-answering systems, as well as surface machine translation. Information access to speech archives through transcriptions is also an emerging field.

Alpage focuses on some applications included in the two last points.

4.2. Information extraction and knowledge acquisition

Keywords: *Information extraction*, *knowledge acquisition*.

[3] See, respectively: http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html and http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.

[4] http://www.nist.gov/speech/tests/ace/

[5] http://clg.wlv.ac.uk/events/ARE/

Participants: Éric de La Clergerie, François-Régis Chaumartin.

The first domain of application for Alpage parsing systems will be information extraction, and in particular knowledge acquisition, be it linguistic or not, and text mining.

Knowledge acquisition for a given restricted domain is something that has already been studied by some Alpage members for several years (ACI Biotim, biographic information extraction from the Maitron corpus). François-Régis Chaumartin, PhD student at Talana and businessman, is working on information extraction from the English Wikipedia. Indeed, chunking or, better, syntactic (and semantic) parsing gives an access, through learning techniques, to useful information present in documents. Obviously, the progressive extension of Alpage parsing systems to a full syntactic *and* semantic parsing will increase the quality of the extracted information, as well as the scope of information that can be extracted. Such knowledge acquisition efforts bring solutions to current problems related to information access and take place into the emerging notion of *semantic web*. The transition from a web based on data (textual documents,...) to a web based on knowledge requires linguistic processing tools which are able to provide fine grained pieces of information, in particular by relying on high-quality deep parsing. For a given domain of knowledge (say, tourism), the extraction of a domain ontology that represents its key concepts and the relations between them is a crucial task, which has a lot in common with the extraction of linguistic information.

The automatic acquisition of linguistic information lies half-way between applications and resources development, already described above. Hence, we shall not repeat here our objectives concerning this domain.

All these applications in the domain of information extraction raise exciting challenges that will require altogether ideas and tools coming from the domains of computational linguistics, machine learning and knowledge representation.

4.3. vera

Keywords: *Opinion mining*.

Participant: Benoît Sagot.

vera is a joint project with a world-wide leader in the domain of employee research (opinion mining among the employees of a company or organization). The aim of *vera* is to provide an all-in-one environment for editing (i.e., normalizing the spelling and typography), understanding and classifying answers to open-ended questions, and relating them with closed-ended questions, so as to extract as much valuable information as possible from both types of questions. The editing part relies in part on SxPipe 5.5 and Alexina morphological lexicons. Other parts of *vera* are not directly related to NLP, and therefore fall outside the scope of Alpage's work.

4.4. TNS

Keywords: *NLP (Natural Language Generation)*.

Participant: Laurence Danlos.

NLG in a given technical domain has reached a sufficient level of maturity for real applications. The development of such applications is based on G-TAG, a formalism based on Tree Adjoining Grammar, [57]. This formalism, dedicated to the "tactical component" is enriched with a document structuring module taking ideas from SDRT (Segmented Discourse Representation Theory, [42]), [60].

5. Software

5.1. Syntax

Keywords: *CFG*, *Parser generator*, *Parsing*, *RCG*.

Participants: Pierre Boullier [correspondant], Benoît Sagot.

See also the web page http://syntax.gforge.inria.fr/.

The (currently beta) version 6.0 of the SYNTAX system (freely available on INRIA GForge) incules various deterministic and non-deterministic CFG parser generators, including an efficient implementation of the Earley algorithm, with many original optimizations, that is used in several of Alpage's NLP tools, including the pre-processing chain SXPipe and the LFG deep parser SXLFG. SYNTAX 6.0 also includes parsers for various contextual formalisms, including a parser for Range Concatenation Grammars (RCG) that can be used among others for TAG and MC-TAG parsing.

During year 2008, this version of SYNTAX has been successfuly ported to many 32-bit and 64-bit architectures, in collaboration with project-team VASY (INRIA Rhône-Alpes), one of SYNTAX' user for non-NLP applications. Their expertise in software porting has helped SYNTAX developers to enhance the quality, portability, organization and distribution of the system. This should lead in the very near future to a full distribution of a non-beta version of SYNTAX 6.0. Other current or former direct users of SYNTAX , outside Alpage, include Alexis Nasr (Marseille) as well as (indirectly) all SXPipe and/or SXLFG users.

5.2. SxLfg

Keywords: *LFG*, *Parsing*.

Participants: Benoît Sagot [correspondant], Pierre Boullier.

SXLFG is a parser generator based on SYNTAX for Lexical-Functional Grammars (LFG) [49], [49], [48]. Functional structures are efficiently computed on top of the CFG shared forest generated by SYNTAX . The efficiency is achieved thanks to computation sharing, lazy evaluation, compact data representation, rule-based and/or n-best disambiguation. It can be helped by a chunk-based module which, when used without f-structures computation, constitutes a state-of-the-art chunker. SXLFG uses various error recovery techniques in order to build a robust parser.

With our grammar for French (written in a meta-formalism of LFG and compiled automatically into pure LFG), it leads to the SXLFG-fr parsing system for French, which relies on the Lefff and takes SXPipe outputs as input. It constitutes a very efficient deep parser, which can parse several million-word corpus in only several hours [48], [51]

5.3. DyALog

Keywords: *Dynamic programming*, *Logic programming*, *Metagrammar*, *Parsing*, *TAG*.

Participants: Éric de La Clergerie [correspondant], Sahil Thappa, Djamé Seddah.

See also the web page http://dyalog.gforge.inria.fr/.

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 **1.12.0** of DYALOG is freely available by FTP under an open source license and runs on Linux platforms for x86 architectures and on Mac OS intel. A port for PowerPC, initiated by Djamé Seddah, should be soon available.

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 [82].

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

DyALog is largely used within Alpage to build parsers but also derivative softwares, such as a compiler of Meta-Grammars (cf. 5.4). 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 and the last Passage campaign (Dec. 2007), cf. 8.1.1 and [80]

DyALog is used at LORIA (Nancy), University of Coruña (Spain), Instut Gaspard Monge (Univ. Marne La Vallée), and University of Nice.

DyALog and other companion modules are available on INRIA GForge.

5.4. MgKit: tools and resources for Meta-Grammars

Keywords: *Metagrammar*.

Participants: Éric de La Clergerie [correspondant], Isabelle Cabréra.

See also the web page http://mgkit.gforge.inria.fr/

DyALog (cf. 5.3) has been used to implement `mgcomp`, a compiler of Meta-Grammar (cf. 6.9). Starting from an XML representation of a MG, `mgcomp` produces an XML representation of its TAG expansion.

The current version **1.4.3** is freely available by FTP under an open source license. It is used within Alpage 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 release provide a dump/restore mechanism for faster compilations on incremental changes of a meta-grammars.

The current version of `mgcomp` has been used to compile a wide coverage Meta-Grammar FRMG (version 1.2.0) to get a grammar of around 160 TAG trees [80]. 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 meta-grammars, a set of tools have been implemented by Éric de La Clergerie and collected in MGTools (version **2.2.1**). 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.

The various tools on Metagrammars have available on INRIA GForge.

5.5. Alpage's linguistic workbench, including SxPipe

Keywords: *Named entites*, *Segmentation*, *Spelling correction*, *Surface processing*, *Tokenization*.

Participants: Benoît Sagot [correspondant], Pierre Boullier, Éric de La Clergerie.

Alpage's linguistic workbench is a set of packages for corpus processing and parsing. Among these packages, the SxPipe package is of a particular importance

SxPipe, now in version 2 [12] is a modular and customizable chain aimed to apply to raw corpora a cascade of surface processing steps. It is used

- as a preliminary step before Alpage's parsers (FRMG, SxLfg);
- for surface processing (named entities recognition, text normalization...).

Developed for French and for other languages, SXPipe 2 includes, among others, various named entities recognition modules in raw text, a sentence segmenter and tokenizer, a spelling corrector and compound words recognizer, and an original context-free patterns recognizer, used by several specialized grammars (numbers, impersonal constructions...).

5.6. The syntactic lexicon Lefff and the Alexina framework

Keywords: *Alexina*, *Lefff*, *Morphological lexicon*, *Syntactic lexicon*.

Participants: Benoît Sagot [correspondant], Laurence Danlos.

See also the web page http://gforge.inria.fr/projects/alexina/.

Alpage's freely available syntaxic lexicon for French, the Le*fff*, is now in version 3. It is developed within Alpage's Alexina framework for the acquisition and modeling of morphological and syntactic lexical information. Other Alexina lexicons do exist, in particular for Polish, Slovak, English and now Spanish (see 6.12).

Historically, the Le*fff* 1 was a freely available French morphological lexicon for verbs that has been automatically extracted from a very large corpus. Since version 2, the Le*fff* covers all grammatical categories (not just verbs) and includes syntactic information (such as subcategorization frames); Alpage's tools, including Alpage's parsers, rely on the Le*fff*. The version 3 of the Le*fff*, which has been released in 2008, improves the linguistic relevance and the interoperability with other lexical models (see 6.11).

5.7. EasyRef syntactic annotation tool

Participant: Éric de la Clergerie [correspondant].

PASSAGE action
[31]

A collaborative WEB service EASYREF has been developed, in the context of ANR action Passage, to handle syntactically annotated corpora. EASYREF may be used to view annotated corpus, in both EASY or PASSAGE formats. The annotations may be created and modified. Bug reports may be emitted. The annotations may be imported and exported. The system provides standard user right management. The interface has been designed with the objectives to be intuitive and to speed edition.

EASYREF relies on an Model View Controller design, implemented with the Perl Catalyst framework. It exploits WEB 2.0 technologies (i.e. AJAX and JavaScript).

The current version has been used by ELDA to annotate a new corpus of several thousands words for PASSAGE. EASYREF is maintained under INRIA GForge.

6. New Results

6.1. Theoretical advances on polynomial contextual formalisms

Keywords: *MCS formalisms*, *RCG*, *formal grammars*.

Participants: Pierre Boullier, Giorgio Satta, Benoît Sagot.

Collaboration with Giorgio Satta, in particular during his 3-month visit at Alpage. Giorgio Satta is full Professor at the University of Padua (Italy) and chair of the European Chapter of the Association for Computational Linguistics (EACL).

RCG an RCG (Range Concatenation Grammar) is a rewriting grammar in which rewriting rules are Horn clauses whose variables denote ranges over the input string; RCGs define exactly the set of languages that can be parsed in polynomial time.

In 2008, Alpage members have pursued their researches on contextual languages. The work on Mildly-Context Sensitive grammars, in collaboration with Giorgio Satta, has not yet led to definitive results. We are still working on difficult problems whose solutions should be published in 2009. On the other hand, the work on Range Concatenation grammars (RCGs) has been pursued in two main directions:

- The creation of counters
- The parsing of DAGs

6.1.1. The creation of counters

We have already shown in [44] that RCGs have the necessary formal power *to count*. Let us recall that Context-Free Grammars only knows to count up to 2, while such number reaches 4 for Tree-Adjoining Grammars. Nevertheless, the handling of numbers in RCGs is rather artificial since any number was denoted by the size of a range, while the operations of incrementation and decrementation are simulated by the scan of terminal symbols. In the new version, counters have been promoted as first class objects. Some predicate arguments may be specialized as *counters*. A counter is a string of variable symbols or non negative integers whose value is a non negative integer (and not a range). The decrement (resp. increment) operation is denoted by a string concatenation operation occurring in LHS (resp. RHS). As for ranges, the equality of values is denoted by the string equality of their counters. Of course, this new possibility does not add any formal power which is still exactly PTIME, but it may allow to define some features in a much more easy, pleasant and understandable way.

6.1.2. The parsing of DAGs

In many natural language processing (NLP) applications the source text cannot be considered as a (linear) string of terminal symbols, but rather as a finite set of finite strings, conveniently represented as a DAG. They make it possible to represent an exponential number of strings w.r.t. their lengths n in $\mathcal{O}(n)$ space.

On the other hand, RCGs were only defined to handle linear sentences, we thus have studied whether they could be extended in order to process DAGs as inputs.

For the subpart of the RCGs which is linear,[6] this extension is very simple, we only have to say that the bounds of the ranges are the states of the DAG, and the standard definitions and parser then apply without (almost) any changes.

However, the previous result does not hold when we consider the full class of RCGs. The problem comes from the non-linearity of the RCG formalism, more specifically from the meaning of a non-linear variable, say X, in an instantiated clause. If the input is a string, all the occurrences of X denote the same range, i.e., the same substring occurrence. If the input is a DAG, each occurrence of X which denotes the same pair (p, q) of DAG states way well have been validated (by the derivation mechanism) by non-identical substrings starting from p and leading to q. If this is the case, this instantiated clause must clearly be rejected.

We have defined an algorithm which handles input DAGs on the full class of RCGs and which leaves the standard RCG parsing algorithm (almost) unchanged because it works at the shared parse forest level. Moreover, we have shown that this pruning algorithm works in polynomial time and space. In other words, even non-linear RCGs can parse in polynomial time some exponential number of sentences.

Moreover, the basic structures involved during its implementation will be reused by an other extension of RCGs which are the Synchronous RCGs and which are an important part of Boullier's emeritus program.

6.2. Discourse Synchronous TAGs: a formalism for discourse structure representation

Keywords: *Discourse structure*, *d-stag*.

Participants: Laurence Danlos, Pascal Denis, Benoît Sagot.

[6] This is the part which defines the mildly context-sensitive formalisms.

D-STAG is a new formalism for the automatic analysis of the discourse structure of texts. The analyses computed by D-STAG are hierarchical discourse structures annotated with discourse relations, that are compatible with discourse structures computed in SDRT, [42]. The discourse analysis extends the sententiall analysis, without modifying it, which simplifies the realization of the system.

This formalism has reached a sufficient level of maturity for starting implementation. The architecture of D-STAG consists of three modules :

1. the sentential analysis, which gives for each sentence of the input discourse a syntactic and semantic analysis;
2. the sentence–discourse interface, which is a module that is necessary if one wants (and it is what we want) not to modify the sentential analysis;
3. the discourse analysis, which computes discourse structure.

For the first step, the French parser we use is FRMG developed within Alpage by Éric de la Clergerie [83]. The second step consists in getting a "normalized form for discourse" from the syntactic analysis of a suite of sentences. This normalized form is a sequence of "discourse words" where each discourse word is either a discourse connective, or a label S_i for a clause, or a punctuation sign marking the end of a sentence or surrounding an adverbial subordinate clause. This phase is currently implemented by two master students with the help of Laurence Danlos and BenoÓt Sagot. The last phase will be carried out in 2009.

6.3. Finite-state multi-tape transducers

Keywords: *Finite-state automaton*, *Multi-tape transducer*.

Participant: François Barthelemy.

Alpage has been working in the definition of finite-state multi-tape transducers using typed Cartesian Product. Tapes are identified using a unique name and the Cartesian Product is an operator which allows the combination of several components which are either a language on a given tape or an embedded Cartesian Product on several tapes. The components of a Cartesian Product must be independent, namely they do not share any tape. The types are implemented in tapes using auxiliary symbols which are used to obtain a closure under intersection (and also difference and complementation) of the transducers.

François Barthélemy developped a system called Karamel devoted to the development and execution of finite-state multi-tape transducers. The system comprises a language and a Integrated Development Environment. The language uses three ways for defining finite state machines:

- regular expressions extended with typed Cartesian product
- operators applied to previously defined machines. These operators are the usual rational operators and extensions, but also intersection, complementation and difference which are in general not internal operations on rational transducers. They are however for the subclass of transducers used in Karamel. There are also two special operations which respectively recognize and extract an untyped language on a given tape of a typed description.
- contextual rules called Generalized Restriction rules by Yli-Jyrä and Koskenniemi [85]. They are a powerful and abstract mean to express constraints.

The IDE is written in HTML/CSS/Javascript. It provides some basic edition functions, some test facilities and an interface to execute the descriptions. Karamel uses a C++ library from AT&T called FSM which implements efficiently finite-state algorithms. At the moment, Karamel is still a prototype. We plan to complete its development and begin to distribute it in the near future.

6.4. Applying semi-supervised learning algorithm to learn a French probabilistic parser

Keywords: *PCFG*, *Probabilistic grammar induction*.

Participants: Benoît Crabbé, Marie Candito, Djamé Seddah.

We have been working on how to use syntactically-annotated data (the French TreeBank [40]) for parsing French, both with lexicalized and unlexicalized models. Indeed, due to the lack of previous work (only two papers on the subject before 2007, [41], [77]) on probabilistic parsing of French and in order to be able to confirm the accuracy of this type of approach, we carryed out a important phase of adapting and porting various probabilistic models to French.

Unlexicalized models We investigated the use of a semi-supervised learning algorithm [71] that acquires a probabilistic CFG augmented with latent annotations. The work has focused on how to best instantiate a treebank in order to improve performance. They specifically studied the impact of the following French treebank features: (i) compound words representation, (ii) preterminal symbol set, and (iii) word inflection. They have shown that some treebank transformations have positive impact on the results [15]. We are currently working on finding tree structure transformations to optimize the learned parser.

Lexicalized models We adapted and ported various lexicalized probabilistic model to French, namely the Collin's Model 1 [56] in its [43]'s implementation and the Stochastic Tree Adjunct Grammars model introduced by [76] through [55]'s parser. As the French Treebank has evolved since the first reports on French parsing, we had to instanciate all our results on every version of the treebank, including the deeply modified version of [77]. Given that lexicalized parsing rely on two sets of linguistic heuristics (Head percolation table and Argument adjoint distinction table), those had to be created and evaluated for each treebank-parser pair.

We obtained for all techniques state-of-the-art results on French. The best combination so far is the [71] adapted for French trained on the [15]'s modified treebank version.

6.5. From syntagmatic trees to dependency trees

Keywords: *Dependency*, *Parser evaluation*, *Probabilistic parsing*.

Participants: Benoît Crabbé, Marie Candito, Djamé Seddah.

Syntagmatic trees are generally not the right level of syntactic representation for many tasks. Dependency trees are generally preferred for tasks such as information extraction or lexical acquisition, because the links between words are made explicit and typed with functional roles such as e.g., subject, object, modifier etc...Dependency trees are also more neutral with respect to particular syntagmatic annotation schemes (in a dependency tree, each word has exactly one governor except the head of the whole sentence). This being given we worked on a further step of translating syntagmatic trees into dependency trees, starting from our [71]-based parser [15] (see 6.4). This is a two step procedure, first the syntactic trees obtained with the probabilistic parser can be further enriched with functional annotation, with a functional role labeler (see below). Second functionnaly-annotated trees can be translated into a labeled dependency tree (following a procedure described for instance by [66]).

In order to measure our dependency extracting procedure, we have manually validated a reference corpus of dependency structures for 120 sentences from the French Treebank. We have defined a "pivot" surfacic dependency format, designed to facilitate transformations into international standards such as GR ([53]) or Parc700 ([64]), and also into the EASY format. This last format is essential to compare to parsers in the French community. This is a rare occasion given to compare on the same data and evaluation framework a probabilistic parser and the symbolic parsers that compete in the EASy campaign. A further known advantage of dependency structures is to allow the expression of non-projective dependencies, such as the one existing between *en* and *efficacité* in the sentence *réformer le système, pour en améliorer l'efficacité*. These non-projective dependencies cannot directly be obtained via a constituent-to-dependency transformation. In the 120-sentence reference dependency corpus, we found 2% of non-projective dependencies. This allows for an architecture that first derive projective dependencies (as the parser does today), and then renders non-projective certain dependencies. This last step is currently under study.

As we are aiming toward probabilistic deep parsing, it is worth noting that a certain amount of linguistic phenomena are not taken into account by the treebank we use for training. For instance, coordination with ellipsis has been proven difficult for any linguistic theory and thus for any treebank which does not handle discontinuous phrases. That is why, following preliminary efforts by [78], a formal modeling of elliptic coordination has been achieved by [28] that is meant to be applied either into our training data or in a post-parsing stage analysis.

Finally we plan to engage the statistical parser in the EASy parsing campaign. EASy enforces setting up parsers running on multiple domains (written journalistic, oral, medical, email...) it requires to carry research activities on domain adaptation : we need to adapt a parser trained on journalistic data to other domains.

6.6. Probabilistic TIG-based dependency parsing

Keywords: *Dependency parsing*, *PCFG*, *TAG*, *TIG*, *supertagging*.

Participants: Pierre Boullier, Benoît Sagot.

Collaboration with Alexis Nasr (LIF, Université de Marseille-Provence), Owen Rambow (Cornell University, New York, USA) and Srinivas Bangalore (AT&T labs, USA).

PCFG (Probabilistic Context-Free Grammar) a Context-Free Grammar (CFG) with probabilities associated with each production.

Two members of Alpage, in collaboration with other teams in France and USA, developed a state-of-the-art dependency parser for English, named MICA (this acronym recalls the four different affiliations of the developers: (University of) Marseille, Inria, Cornell University and AT&T). It relies on a grammar (TIG) extraction algorithm initially developed by [54] and applied on the Penn TreeBank. The grammar extraction step allows to learn a supertagger, which is the first step of the full parsing process. The output of the supertagger, partially pruned, is given as an input to a parser generated by SYNTAX from the extracted grammar.

Results are approximatively state-of-the-art as far as precision and recall is concerned, and significantly better in terms of parsing speed. The work on MICA will directly benefit to the SEQUOIA project (see 8.1.3), as soon as all underlying techniques are transfered to French.

6.7. Optimized reduction of probabilized shared parse forests

Keywords: *PCFG*, *forest pruning*, *forest unfolding*, *n-best trees*, *shared parse forest*.

Participants: Pierre Boullier, Benoît Sagot.

Collaboration with Alexis Nasr (LIF, Université de Marseille-Provence), within the ANR funded-project SEQUOIA (see 8.1.3).

PCFG (Probabilistic Context-Free Grammar) a Context-Free Grammar (CFG) with probabilities associated with each production.

The output of a CFG parser such as parsers created with SYNTAX is a shared parse forest, which is an acyclic graph that represents all the syntactic parses of the parsed sentence. Such a graph can represent an exponential number (with respect to the length of the sentence) of parses as a cubic object. Therefore, when probabilistic information is associated with the rules of the CFG (Probabilistic CFG, PCFG), it is necessary to extract from the forest the n most likely parses with respect to the PCFG. Standard state-of-the-art algorithms that extract the n best parses (Huang 2005) produce a collection of trees, losing the factorization that have been realized by the parser, and reproduce some identical sub-trees in several parses. This situation is not satisfactory since the post-parsing processes (such as reranking) will not take advantage of the factorization and will reproduce some identical work on common sub-trees. One way to solve the problem is to prune the forest by eliminating sub-forests that do not contribute to any of the n most likely trees. Such techniques usually over-generate: the pruned forest contains more than the n most likely trees.

The new direction that we explored in 2008 is the production of shared forests that contain *exactly* the n most likely trees, avoiding the explicit construction of n different trees and the over-generation of pruning techniques. This process can be seen as a forest transduction which is applied on a forest and produces another forest. The transduction applies some local transformations on the structure of the forest, developing some parts of the forest when necessary. If n is not very small, the forest produced is generally larger than the input forest even if it contains less trees. We developed two types of algorithms for building such a forest containing exactly n trees, which try to minimize its size. Quantitative results should be published in early 2009.

6.8. Tabulation and probabilities

Keywords: *Dynamic Programming*, *Parsing*, *Probabilities*, *Tabulation*.

Participants: Éric de La Clergerie, Sahil Thappa.

During his 2-month internship, Sahil Thappa has started to extend DyALog in order to handle weight or probabilities during parsing. An analysis of DyALog has shown that very few modifications seems to be necessary. A few of them deal with the compiler part of DyALog to allow the representation and compilation of weighted grammars. The other modifications are in DyALog virtual machine, essentially on (a) the representation and handling of the backpointers attached to items and (b) the agenda to allow for more flexible weight-based dynamic scheduling policies.

6.9. Designing efficient parsers using Meta-Grammars and DyALog

Participant: Éric de La Clergerie.

MG *Meta-Grammars*

An effort has been done to improve the efficiency of the FRMG based parser for French, leading to parsing times divided by 10 on average over the year (as shown by logs on the EASy corpus). Some modifications have been done at the level of the meta-grammar, essentially to add more constraints. However, the main gains come from generic optimization within DyALog to handle Tree Insertion Grammars. These optimizations include, among others, a better identification of the variables to be propagated when traversing an elementary tree or susceptible to be modified through adjoining, a better identification of items that need not to tabulated, a better use of the left-corner relation, a better indexing mechanism for finite-domain terms, ...

6.10. Large scale corpus processing

Participants: Éric de La Clergerie, Isabelle Cabrera.

In the context of the PASSAGE action, we have worked on large scale corpus processing, specially by moving to GRID 5000, hence being able to use several tens of computers. Improving the efficiency of parsing was a first but non sufficient step in this direction (cf. 6.9). The ALPI installer script was completed to ease the installation of the Alpage processing chain on new computers, and specially on GRID 5000. Another step was to design a more efficient dispatcher Perl script, able to dispatch sentences to parse to several hundred grid nodes with minimal communication costs. Finally, several side scripts (used for instance to convert parse forests) have also been improved in terms of efficiency. Over the various experiments tried on GRID 5000, more than 100 million words have been parsed. The latest results show that is possible to parse a 20 million words corpus in 3 hours on 80 dual-core computers.

6.11. Merging syntactic lexical resources for improving the Lefff

Keywords: *Alexina*, *Merging lexical resources*, *Morphological lexicon*, *Syntactic lexicon*.

Participants: Benoît Sagot, Laurence Danlos.

Alpage's morphological and syntactic lexicon for French, the Le*fff* (*Lexique des formes fléchies du français*), has been released under a new version, the Le*fff* 3, based on the new version of the Alexina model.[7] This lexical developement and modeling architecture is based on two representation levels:

- The intensional lexicon factorizes the lexical information by associating each lemma with a morphological class and deep syntactic information (a deep subcategorization frame, a list of possible restructurations, and other syntactic features such as information on control, attributes, mood of sentencial complements, etc.);
- The extensional lexicon, which is generated automatically by *compiling* the intensional lexicon, associates each inflected form with a detailed structure that represents all its morphological and syntactic information: morphological tag, surface subcategorization frame corresponding to one particular redistribution, and other syntactic features.

The intensional representation is used for an efficent description, while the extensional is directly used by NLP tools such as parsers.

The Le*fff* has been converted into this new Alexina model, hence leading to the release of the Le*fff* 3 (under the free LGPL-LR license, like previous versions of the Le*fff*). Moreover, this new model enabled Alpage to convert other freely available lexical resources for French, such as DICOVALENCE, into the same model. This allowed to compare different resources, and to merge lexical information coming from different sources within the Le*fff*. In particular, a careful interpretation and merging task has lead to a much better treatment of pronominal verb structures in the Le*fff*, both from the point of view of both the coverage and the linguistic relevance.

6.12. Developing a lexicon and a parser for Spanish

Keywords: *Leffe*, *Spanish*, *Syntactic lexicon*.

Participants: Benoît Sagot, Éric de La Clergerie.

Collaboration with Miguel Ángel Molinero Álvarez (University of Ourense, Galicia, Spain) and Lionel Nicolas (University of Nice).

As a preliminary work for the Victoria Spanish-French project (see 8.2.1), some of Alpage's members have began to develop a syntactic lexicon and a metagrammar for Spanish, in collaboration with other members of the Victoria project. In particular, the 2-month visit of Miguel Ángel Molinero Álvarez at Alpage, in November and December 2008, has lead to the publication (under the LGPL license) of a first version of the Le*ffe* (*Léxico de formas flexionadas del español*), a syntactic lexicon for Spanish which relies on the same framework than the Le*fff*, namely Alexina.

Several other lexical resources for Spanish exist, but none of them was satisfying in terms of coverage (all words, including rare ones, in all categories should be included), quality (manually and automatically developed resources contain various errors) and richness (applications such as parsing require at least morphological and syntactic information, including subcategorization frames). Nevertheless, each of these existing resource is a provider of valuable lexical information. Merging these resources and expanding them thanks to semi-automatic techniques is therefore a promising idea. However, it requires to be able to interpret all input resources despite partly incompatible lexical models, to convert them into a common model and format, and then to merge these converted lexicons. None of these three steps is trivial. Actualy, this approach is being successfuly applied by Alpage for developing the Le*fff*, and we extended it for developing the Le*ffe*.

In parallel with the development of the syntactic lexicon Le*ffe*, the development of a meta-grammar for Spanish has been initiated, using FRMG as a starting point, thus taking advantage of the close proximity of French and Spanish. Thanks to this metagrammar and to the Le*ffe*, a deep DYALOG-based parser for Spanish should be released by Alpage in the near future.

[7] http://gforge.inria.fr/projects/alexina

Within the Victoria project, these efforts will be pursued, and extended to Galician[8] and French.

6.13. Towards multilingual tools and resources

Keywords: *Language-independant lexical model*, *Multilingual surface processing*.

Participant: Benoît Sagot.

In 2008, a particular effort has been achieved within Alpage, and notably by Benoît Sagot, for improving the support of various languages in two different series of tools. First, the lexical development framework on which the Le*fff* is based, Alexina, has been clearly modularized, which lead to the development of morphological and even syntactic lexical resources for languages other than French. Apart from French, Alpage has now its own morphological and syntactic lexicon for Spanish (see 6.12), its own morphological lexicon for Polish (large-coverage) and Slovak (medium-coverage), and is able to integrate into the Alexina framework morphological resources such as thoses developed within the MULTEXT and MULTEXT-East projects, or DELA lexicons developed at University of Marne-la-Vallée.

Thanks to these lexicons for other languages, Alpage has been able to turn its pre-processing chain SxPipe into a multilingual tool [12]. Indeed, SxPipe is now able to handle French and English with a high quality level, as well as other languages such as Polish, Slovak, Spanish or Italian. The French and English versions of SxPipe are already used in the operational system *vera* (see 4.3).

6.14. The WOLF, a new French Wordnet

Keywords: *Dependency parsing*, *TAG*, *TIG*, *supertagging*.

Participant: Benoît Sagot.

Collaboration with Darja Fišer (University of Ljubljana, Slovenia), Karën Fort (University Paris 13) and Fabienne Venant (LORIA, Nancy).

Website of the WOLF: http://wolf.gforge.inria.fr

wordnet a wordnet is a semantic resource in which each entry represents a meaning, and is filled by words (“litterals”) that can express this meaning: these words constitute a set of synonyms, or *synset*.

The first wordnet was developed for English at Princeton University (PWN). Over time it has become one of the most valuable resources in applications for natural language understanding and interpretation, such as word-sense disambiguation, information extraction, machine translation, document classification and text summarisation, which initiated the development of wordnets for many other languages apart from English [84], [81]. Currently, wordnets for more than 50 languages are registered with the Global WordNet Association (http://www.globalwordnet.org/). While it is true that manual construction of each wordnet is the most reliable and produces the best results as far as linguistic soundness and accuracy is concerned, such an endeavor is highly time-consuming and expensive. This is why alternative, semi- or fully automatic approaches have been proposed. By taking advantage of the existing resources they facilitate faster and easier development of a wordnet.

Apart from the knowledge acquisition bottleneck, another major problem in the wordnet community is the availability of the developed wordnets. Currently, only a handful of them are freely available (Arabic, Hebrew, Irish and Princeton). Although a wordnet for French has been created within the EuroWordNet (EWN) project [84], the resource has not been widely used mainly due to licensing issues. In addition, there has been no follow-up project to further extend and improve the core French WordNet since the EWN project has ended [63].

[8] A co-official language in north-west Spain.

This is why Alpage initiated the development of a new French Wordnet, the WOLF (Wordnet Libre du Français), freely avaible under the LGPL-compatible Cecill-C license [74], [73]. A baseline has been built thanks to automatic techniques that leverage freely available multilingual resources. Further work involving other French lexical resources and manual validation has enabled to speed-up the improvement of the WOLF in terms of quality and coverage (for now, this step has been applied only on adverbial synsets [27]).

6.15. Error mining in parsing results and beyond

Keywords: *Error mining*, *automatic acquisition of lexical information*, *lexicon development*.

Participants: Benoît Sagot, Éric de La Clergerie.

Collaboration with Lionel Nicolas (University of Nice) and Miguel Ángel Molinero Álvarez (University of Ourense, Galicia, Spain).

The coverage of a parser depends mostly on the quality of the underlying grammar and lexicon. The development of a lexicon both complete and accurate is an intricate and demanding task. In 2008, the technology developed at Alpage for detecting automatically missing, incomplete and erroneous entries in a morphological and syntactic lexicon has been used extensively, and proven efficient and useful in practice[8][13].

Moreover, it is the basis of a more complete framework that is able to detect such dubious entries with different techniques, and suggest corrections hypotheses for these entries. The detection of dubious lexical entries is now tackled by two different techniques; the first one is based on a specific statistical model, the other one benefits from information given by a part-of-speech tagger. The generation of correction hypotheses for dubious lexical entries is achieved by studying which modifications could improve the successful parse rate of sentences in which they occur. This process brings together various techniques based on different tools such as taggers, parsers and statistical models.

We applied this technique for improving the Le*fff*, and more generally Alpage's tools. It will also be used for helping and speeding up the developement of the Spanish morphological and syntactic lexicon Le*ffe* (see 6.12).

6.16. Processing of temporal information in French texts

Keywords: *Named entities*, *Temporal information*.

Participants: André Bittar, Benoît Sagot, Laurence Danlos.

Content-based indexing the process of extracting from a document (here a picture) compact and structured significant visual features that will be used and compared during the interactive search.

Over the past year or so, André Bittar has written an annotation guide for the marking up of French texts according to the ISO-TimeML annotation specification, developed modules for the automatic annotation of French texts in accordance with these guidelines and produced a Gold Standard annotated corpus of journalistic and biographical texts for evaluation purposes. The annotation guide for French was written based on linguistic inquiry, in tandem with the manual annotation of the Gold Standard corpus, as well as research in the domain of theoretical and formal linguistics - notably in syntax and semantics.

Modules have been developed for the automatic annotation of temporal information in French texts according to the ISO-TimeML standard. These modules annotate several types of linguistically-realised entities in French texts, namely events and states, temporal expressions and relational markers. They rely on a pre-processing of the text which is carried out by the modules SXPipe as well as Macaon, developed by Alexis Nasr and Alejandro Acosta (former members of the Paris 7 Talana team). Input to the annotation modules is a text having undergone shallow syntactic analysis (chunking), as well as part-of-speech tagging and morphological analysis. The modules output an annotated text enriched with temporal annotations according to the ISO-TimeML specification language.

6.17. Identification and semantic analysis of expressions referring to people

Keywords: *Named entities*, *coreference*.

Participants: Elżbieta Gryglicka, Benoît Sagot.

In the context of semantic text analysis, Elżbieta Gryglicka is working on coreference et anaphora resolution. The motivation of her PhD thesis (Cifre PhD in collaboration with Thales) is fully automatic identification of expressions referring to people and making explicit their referentials links. The aim is to develop an automatic independent module, which will be able to identify and to analyse various sorts of expressions such as pronouns and definite noun phrases (the goal of most systems), but also plural or collective nouns and indefinite noun phrases. The method is inspired by recent work in the information extraction domain and particularly the named entities recognition and classification task. The first step of our approach and its evaluation is described in [33]. This version of the module uses a set of local grammars to annotate and to collect information about the people. For example : "Laurent Gbagbo, President of Côte d'Ivoire" provides the information that the entity typed as PERSON (identified by "Laurent Gbgabo") has_fonction of "president" in set of COUNTRY (entity identified as "Côte d'Ivoire"). This information is stored in a XML knowledge base which is used further for the process of reference resolution. This approach deals mainly with the class of definite nous phrases, especially those which cannot be resolved with syntactic and linguistic methods.

6.18. Formal modeling of the syntax-semantic interface: minimal syntactic units and dependency structures

Keywords: *Linguistic formalisms*.

Participant: Sylvain Kahane.

Within Alpage, Sylvain Kahane focuses on formal syntax modeling, which is both relevant from a linguistic an NLP point of view. Indeed, the nature of the syntactic representation is a crucial question for improving parsing: what is the syntactic structure that a parser must build, notably for using it as the entry of the syntax-semantics interface and to get a semantic representation more easily.

As far as written French is concerned, this problem has been tackled in Kahane's recent works in several ways. [22] tries to characterize the minimal units of syntax, that is the minimal linguistic signs which can freely combine with others signs. That includes lexemes, inflectional morphemes and various particles between lexicon and grammar like clitics, articles and grammatical prepositions. One of the difficulty for defining the minimal syntactic units comes from the fact that they do not match with the semantic units due to various case of phraseologisation.

Today main formal systems are based on phrase structures. [38] shows that a formalism like HPSG do not really need phrases from a theoretical point of view and can be view as a dependency grammar. This is proved by modeling extraction, one of the cornerstones of all the contemporary formalisms. It results that phrases in HPSG rather play a computational role in the combination of lexical descriptions (and more generally of the descriptions associated to the minimal syntactic units) and that the same dependency grammar can be implemented with various phrase structures in HPSG.

6.19. Modeling and parsing the syntax of spoken French

Keywords: *Spoken language transcripts*, *spoken language syntax*.

Participants: Sylvain Kahane, Benoît Sagot, Éric de La Clergerie, Marie Candito, Benoît Crabbé.

Alpage plays an important role within the syntactic part of the ANR project Rhapsodie (see 8.1.4) lead by Anne Lacheret (University Paris X). The aim of the project is to study the matching of prosody and syntax on a 30 hours corpus of spoken French by providing prosodic and syntactic annotations. Sylvain Kahane is the coordinator of the syntax workpackage, but other alpage members do participate actively as well.

One of the major challenge of spoken language is to analyse utterances which are syntacticly cohesive without functional relation, like in the so-called two-points effect : *vous avez donné quelque chose de plus à la femme des armes de persuasion* [39]. In [62], based on the Aix School grid analysis of spoken French, the notion of "pile" is introduced, allowing for an elegant description of various paradigmatic phenomena like disfluency, reformulation, apposition, two-points effect, question-answer relationships, and different types of coordination. Piles naturally complete dependency annotations by modeling non-functional relations between phrases.

7. Contracts and Grants with Industry

7.1. TEXT-ELABORATOR (2008–2009)

Keywords: *NLG (Natural Language Generation).*

Participant: Laurence Danlos.

TEXT-ELABOATOR is an NLG (Natural Language Generation) project funded by TNS-Sofres. It is leaded by the startup Watch System Assistance for whom Laurence Danlos works as a scientific consultant. The NLG system should be operational within TNS in the spring of 2009. There is some confidentiality around this project since TNS wants to control the schedule of their announcing the customers that the comments on the statistical data are automatically generated.

8. Other Grants and Activities

8.1. National Initiatives

8.1.1. ANR project PASSAGE (2006 – 2008)

Participants: Éric de La Clergerie [project leader], Benoît Sagot, Pierre Boullier.

PASSAGE Homepage: http://atoll.inria.fr/passage
EASy homepage: http://www.limsi.fr/Recherche/CORVAL/easy/

PASSAGE is an action in ANR MDCA program (*Masse de Données Connaissance Ambiantes*) started in 2007. The participants are Alpage (coordinator), LIR (LIMSI, Orsay), "Langue & Dialogue" (LORIA, Nancy), LI2CM (CEA-LIST), plus several contractors (ELDA, TAGMATICA and several providers of parsing systems).

PASSAGE stands for "*Large Scale Production of Syntactic Annotations to move forward*" . Its main objectives are to parse a large corpus (100 to 200 million words) with several parsers (around 10 systems), combine the results provided by these parsers and use the resulting annotations to acquire new linguistic knowledge (semantic classes, subcategorization frames, disambiguation probabilities, ...). A small part of the corpus (around 400000 words) will be manually validated to be used as a reference treebank. Two evaluation campaigns based on the work done during the Technolangue action EASy will be conducted during PASSAGE to assess the performances of the parsing systems. The annotations and derived linguistic resources will be made available.

8.1.2. Action Scribo (2007 – 2009)

Participants: Éric de La Clergerie, Benoît Sagot, Pierre Boullier, Pascal Denis, André Bittar.

SCRIBO Homepage: http://www.scribo.ws/xwiki/bin/view/Main/WebHome

Scribo aims at algorithms and collaborative free software for the automatic extraction of knowledge from texts and images, and for the semi-automatic annotation of digital documents. SCRIBO has a total budget of 4.3M Euros and is funded by the French "Pôle de compétivité" Systematic from Mid 2008 til Mid 2010. It brings 9 participants together: AFP, CEA LIST, INRIA, LRDE (Epita), Mandriva, Nuxeo, Proxem, Tagmatica and XWiki.

8.1.3. ANR project SEQUOIA (2009 – 2011)

Participants: Benoît Sagot, Pierre Boullier, Marie Candito, Benoît Crabbé, Pascal Denis, Éric de La Clergerie, Djamé Seddah.

Alpage play a major role in the ANR-funded project SEQUOIA, lead by Alexis Nasr (LIF, University of Marseille-Provence, former member of the Talana team at University Paris 7). This project, which started informally before its official launching date (January 2009) aims at developing or adapting probabilistic parsing techniques in order to release a high-performance parser for French based on SYNTAX. It brings together specialists of NLP and specialists of Machine Learning, in a very fruitful way.

8.1.4. ANR project Rhapsodie (2008 – 2010)

Participants: Sylvain Kahane, Benoît Sagot, Éric de La Clergerie, Marie Candito, Benoît Crabbé.

Rhapsodie is an ANR project headed by Anne Lacheret (University Paris X). The aim of the project is to study the matching of prosody and syntax on a 30 hours corpus of spoken French by providing prosodic and syntactic annotations. Alpage participates to the project at two different levels: the specification of the transciption and syntactic annotation framework and the use of parsers for preparing the manually validated syntactic corpus annotation.

8.2. European Initiatives

8.2.1. Galician government research project Victoria (2008 – 2010)

Participants: Éric de La Clergerie, Benoît Sagot.

As a followup of a long lasting collaboration with Galician universities, Alpage is strongly involved as associate researchers in the Galician government research project Victoria on the development of Spanish and Galician linguistic resources by adapting tools, methods and resources developed by Alpage. Section 6.12 describes the preliminary results obtained in this direction in 2008.

8.2.2. French-German ANR project Pergram (2009 – 2011)

Participant: Benoît Sagot.

The Pergram project (French-German ANR/DFG project) is lead by Pollet Samvelian (University Paris 3). Its goal is the description of central phenomena in Persian and the development of a non-trivial grammar fragment in the framework of HPSG. The development of this grammar will benefit from the expertise of the German side on phenomena that are not found in French or English, such as scrambling, but will also deal with Persian-specific phenomena such as complex noun-verb predicates. In parallel, the project includes the development of various lexical resources, thanks in part to techniques and tools developed by Alpage members within the Alexina framework: (i) a full form lexicon of verbs and common nouns, (i) valency frames for verbs (iii) the most common Light Verb Constructions (LVCs) and including idiomatic preverb light verb combinations.

8.3. International Initiatives

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

Participant: Éric de La Clergerie.

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

8.3.2. NSF project "CAREER: Automaton Theories of Human Sentence Comprehension" (2009 – 2010)

Participant: Éric de La Clergerie.

Éric de La Clergerie is involved in a new collaboration in the recently funded NSF project "CAREER: Automaton Theories of Human Sentence Comprehension" led by John Hale from Cornell University. This project aims to explore plausible psycholinguistic models, in particular based on automata such as Thread Automata.

8.4. Exterior research visitors

A 3-month visit of Prof. Giorgio Satta from Univ. of Padua (Italy) from April to June 2008.

A 4-month visit of Miguel Molinero-Alvarez from Univ. of La Coruña (Spain) from September to December 2008.

A one-month visit of Darja Fišer from Univ. of Ljubljana (Slovenia) from January to February 2008.

A one-month visit of Milagros Fernandez Gavilanes from Univ. of Vigo (Spain) in November 2008.

9. Dissemination

9.1. Animation at INRIA and University Paris 7

- Alpage, and more specifically Benoît Crabbé, is organizing the NLP seminar of the Linguistics *École Doctorale* of University Paris 7. In 2008, the following speakers gave a talk in this seminar:
  - Pollet Samvelian and Kim Gerdes (Paris 3)
  - Philippe Muller (IRIT)
  - Pascal Denis (INRIA)
  - Maud Ehrmann (XRCE)
  - Helge Dyvik (University of Bergen, Norway)
  - Giorgio Satta (University of Padova, Italy)
  - Josef van Genabith (National Centre for Language Technology NCLT, Dublin City University, Ireland)
  - Erhard W. Hinrichs (Eberhard-Karls University Tübingen, Germany)
  - Piet Mertens (KU Leuven, Belgium)
  - Didier Bourigault (ERSS-CNRS Toulouse)
  - Philippe Langlais (RALI IRO Montreal, Canada)
  - Natalie Schluter (NCLT, Dublin City University, Dublin, Ireland)
  - David Reitter (ICCS/HCRC Edinburgh, United Kingdom)
  - Laurence Danlos (Paris 7 / INRIA)
- Laurence Danlos was the director of the CNRS UMR 8094 (LATTICE) until the 31st of December;

- Laurence Danlos is member of the scientific council of the Linguistic department of University Paris 7;
- Éric de La Clergerie is an elected substitute member of INRIA's "Conseil scientifique";
- The whole Alpage team met in Marseilles for a 2-day team workshop (*journées au vert*) in October, in collaboration with Alexis Nasr (University of Marseille).

9.2. Supervising

- Laurence Danlos was the PhD advisor for Céline Raynal and Laurence Delort who defended respectively in June and December 2008 within LATTICE;
- Laurence Danlos is the PhD advisor for four Alpage students: Pierre Hankach (France Telecom) who should finish in February 2009, André Bittar (allocataire Paris 7) who should finish in December 2009, Elżbieta Gryglicka (Cifre Thales) who should finish in March 2010 and Juliette Thullier (allocataire Paris 7) who started in October 2008;
- Éric de La Clergerie has supervised the internship of Sahil Thapa on the handling of probabilities within DyALog;
- Benoît Crabbé has supervized the Master 2 internship of François Guérin on probabilistic parsing for French and the conversion of the resulting parses into the EASy format for evaluation purposes;
- Benoît Sagot has supervized the Master 1 reasearch internship of Guillaume Lechien on the development of an web-based edition interface for the WOLF.

9.3. Jury

- Laurence Danlos was a reviewer for the HDR dissertation of Myriam Bras (Université de Toulouse);
- Laurence Danlos was a reviewer for the PhD dissertation of Alexandros Tantos (University of Konstanz, Germany) and a member of the committee for the PhD dissertation of Maud Ehrmann (Xerox-Grenoble and University Paris 7) and François Lareau (Université du Québec à Montréal, Canada, and Université Paris 7);
- Éric de La Clergerie was a reviewer for the Phd dissertation of Jean-Philippe Prost (Macquarie University, Sydney) and examiner for the French defense (Univ. de Provence, Dec.);
- Benoît Sagot was an examiner for the Phd dissertation of Laurence Delort (Univ. Paris 7).
- Éric de La Clergerie is a member of the recruitment committee in Section 27 of University Paris 13, University Paris 11, and University of Orléans;

9.4. Committees

- Alpage is involved in the French journal T.A.L. (AERES linguistic rank: A). Éric de La Clergerie, who is a member of the editorial board, has been nominated as "Redacteur en chef". Laurence Danlos has been nominated as member of the editorial board. Benoît Sagot is "Secrétaire de rédaction" of the journal; Pierre Boullier and Benoît Sagot were also external reviewers for the volume 49-1;
- Participation of Laurence Danlos to the program committee of TALN 2008;
- Participation of Éric de La Clergerie to the program committees of TALN'08, TAG+9, LGC'08, IGCL'08, CSLP'08 and scientific committee of LREC 2008. He has also reviewed for ACL'08 (areas: "Syntax and Parsing" and "Phonology/Morphology, FS, POS tagging, and word segmentation") and EACL'2009;
- Participation of Pierre Boullier to the program committees of ACL'08 (area "Syntax and Parsing"), TAG+9 (International Workshop on Tree Adjoining Grammars), FG 2008 (Formal Grammars); he was reviewer for the Journal of Information and Computation (special issue) and for the Journal on Research On Language and Computation (ROLC, vol. 24);

- Participation of Pascal Denis to the program committees of CILCING 2008 (9th International Conference on Intelligent Text Processing and Computational Linguistics, Haifa, Israel) and SuB 2008 (Sinn und Bedeutung, Stuttgart, Germany);
- Participation of Benoît Sagot to the program committees of TALN 2008, IIS 2008 (Intelligent Information Systems, Warsaw, Poland) and ALTW 2008 (Australasian Language Technology Workshop, Tasmania, Australia);
- Participation of Benoît Crabbé to the program committees of TALN 2008;
- Éric de La Clergerie is program chair for the next edition of the International Workshop on Parsing Technologies (Paris, 2009);
- Evaluation by Laurence Danlos of two projects for ANR Program CONTINT (STIC),
  Evaluation by Laurence Danlos of three CIFRE (ANRT) applications,
  Evaluation by Laurence Danlos of an EPSRC (Engineering and Physical Sciences Research Council, UK) project,
  Evaluation by Laurence Danlos of a project from thr research council of the University of Leuven (Belgium).

9.5. Participation to workshops, conferences, and invitations

Note: Partitipation of associate members to workshops and conferences are not mentioned.

- Participation of Éric de La Clergerie to ISO TC37SC4 meetings (Marrakech, May; Pisa, Sept.)
- Laurence Danlos was invited speaker to Constraints in Discourse'08 (Postdam, Germany) and to the annual Stuttgart University workshop (Bleubeuren, Germany);
- Participation with presentation of Laurence Danlos and Pierre Hankach at CID'08;
- Participation with presentation of Laurence Danlos to a 1-day ATALA workshop on teaching NLP in France;
- Laurence Danlos was invited for a seminar at the University of Leuven (Belgique);
- Éric de La Clergerie was invited to deliver a tutorial on "TAG parsing" at TAG+9 (Tübingen, Germany, June)
- Éric de La Clergerie was invited to deliver a talk on "Mining the concept of error mining" at the NATAL workshop (Nancy, LORIA, June)
- Laurence Danlos and Éric de La Clergerie were invited to a 1-day working meeting at Univ. of Santiago (June) with presentations on the Le*fff* and on Meta-grammars.
- Éric de La Clergerie has presented his work on DyALog and Meta-Grammars at Institut Gaspard Monge (University Marne la Vallée).
- Participation with presentations of Éric de La Clergerie at the First Workshop on Automated Syntactic Annotations for Interoperable Language Resources [29], [31], LREC'2008 [30] and TALN 2008.
- Participation with joint presentations of Laurence Danlos and Benoît Sagot at the Lexicon-Grammar conference [18] and the Workshop "Lexicographie et informatique : bilan et perspectives" (Nancy) [35].
- Participation with presentations of Benoît Sagot at TALN 2008 [26] and at the Lexicon-Grammar conference [27].
- Participation with presentations of Marie Candito and Benoît Crabbé at TALN 2008 [15].
- Participation of all members of Alpage to TALN 2008.
- Participation of Laurence Danlos to IJCNLP'08 (Hyderabad, India) and JSM'08 (Toulouse);

9.6. Teaching

Alpage, following Talana, is in charge of the prestigious cursus of Computational Linguistics of Paris 7, historically the first cursus in France in this domain. This cursus, which starts in License 3 and includes a Master 2 (research) and a professional Master 2, is lead by Laurence Danlos. Benoît Crabbé is in charge of the License 3, and Laurence Danlos is in charge of the both Master 2. All faculty members of Alpage are strongly involved in this cursus, but some Inria members also participated in teaching and supervizing internships. Unless otherwise specified, all teaching done by Alpage members belong to this cursus. Teaching by associate members in other universities are not indicated.

Laurence Danlos[9]:

- Introduction to NLP (3rd year of License, 24h);
- Discourse, NLU and NLG (2nd year of Master, 39h).

Marie Candito:

- French syntax (2nd year of Licence, 21h, License of Linguistics of University Paris 7)
- Formal languages theory and parsing (1st year of Master, 24h)
- Information retrieval (2nd year of professional Master, 12h)
- Machine translation (1st year of Master, 48h)
- Lexical Functional Grammar (3rd year of Licence, 48h)

Benoît Crabbé:

- Finite-state techniques for information extraction (2nd year of Master, 30h)
- Probabilistic techniques for NLP (1st year of Master, 60h)
- Introduction to programming I (3rd year of Licence, 60h)
- Introduction to programming II (3rd year of Licence, 30h)
- Corpus linguistics (3rd year of Licence, 30h)

Éric de La Clergerie:

- Éric de La Clergerie Prolog and NLP (3rd year of Licence, 12 hours)

10. Bibliography

Major publications by the team in recent years

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

[2] P. BOULLIER, B. SAGOT. *Are very large grammars computationnaly tractable?*, in "Proceedings of IWPT'07, Prague, Czech Republic", (selected for publication as a book chapter), 2007.

[3] B. CRABBÉ, M. CANDITO. *Expériences d'analyse syntaxique statistique du français*, in "Actes de la 15ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'08), Avignon", 2008, p. 45–54.

[9] Since her nomination to IUF (September 2004), Laurence Danlos teaches only 65h "équivalent TD".

[4] L. DANLOS. *Capacité générative forte de RST, SDRT et des DAG de dépendances pour le discours*, in "Traitement Automatique des Langues (TAL)", (an English version has been published as bookchapter in 2008 under the title Strong generative capacity of RST, SDRT and discourse dependency DAG), vol. 47, n^{o} 1, 2006.

[5] L. DANLOS. *D-STAG : A discourse formalism using synchronous TAG*, in "Language, Representation and Reasoning", University of the Basque country Press, 2008.

[6] L. DANLOS, B. SAGOT. *Constructions pronominales dans Dicovalence et le lexique-grammaire – Intégration dans le Lefff*, in "Proceedings of the 27th Lexicon-Grammar Conference, L'Aquila, Italy", 2008.

[7] B. SAGOT, P. BOULLIER. *SxPipe 2: architecture pour le traitement prÃ©syntaxique de corpus bruts*, in "Traitement Automatique des Langues (T.A.L.)", to appear, vol. 49, n^{o} 2, 2008.

[8] B. SAGOT, É. VILLEMONTE DE LA CLERGERIE. *Error Mining in Parsing Results*, in "Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia", Association for Computational Linguistics, July 2006, p. 329–336, http://www.aclweb.org/anthology/P/P06/P06-1042.

[9] É. VILLEMONTE DE LA CLERGERIE, O. HAMON, D. MOSTEFA, C. AYACHE, P. PAROUBEK, A. VILNAT. *PASSAGE: from French Parser Evaluation to Large Sized Treebank*, in "Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco", may 2008.

[10] É. VILLEMONTE DE LA CLERGERIE. *From Metagrammars to Factorized TAG/TIG Parsers*, in "Proceedings of IWPT'05, Vancouver, Canada", October 2005, p. 190–191, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/IWPT05mg.pdf.

Year Publications

Articles in International Peer-Reviewed Journal

[11] P. DENIS, J. BALDRIDGE. *Global joint models for coreference resolution and named-entity classification*, in "Revista de Procesamiento de Lenguaje Natural", to appear, vol. 42, 2008.

[12] B. SAGOT, P. BOULLIER. *SxPipe 2: architecture pour le traitement présyntaxique de corpus bruts*, in "Traitement Automatique des Langues (T.A.L.)", to appear, vol. 50, n^{o} 1, 2009.

[13] B. SAGOT, É. VILLEMONTE DE LA CLERGERIE. *Fouille d'erreurs sur des sorties d'analyseurs syntaxiques*, in "Traitement Automatique des Langues (T.A.L.)", vol. 49, n^{o} 1, 2008.

International Peer-Reviewed Conference/Proceedings

[14] F. BARTHÉLEMY. *Typage, produit cartésien et unités d'analyse pour les modèles à états finis*, in "In Traitement Automatique de la Langue Naturelle (TALN), poster session, Avignon (France)", June 2008.

[15] B. CRABBÉ, M. CANDITO. *Expériences d'analyse syntaxique statistique du français*, in "Actes de la 15ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'08), Avignon", 2008, p. 45–54.

[16] L. DANLOS. *D-STAG: Parsing Discourse with synchronous TAG and SDRT background*, in "Proceedings of the Third International Workshop on Constraints in Discourse (CID'2008), Postdam, Germany", 2008.

[17] L. DANLOS, P. HANKACH. *Right Frontier Constraint for discourses in non canonical order*, in "Proceedings of the Third International Workshop on Constraints in Discourse (CID'2008), Postdam, Germany", 2008.

[18] L. DANLOS, B. SAGOT. *Constructions pronominales dans Dicovalence et le lexique-grammaire – Intégration dans le Lefff*, in "Proceedings of the 27th Lexicon-Grammar Conference, L'Aquila, Italy", 2008.

[19] P. DENIS, J. BALDRIDGE. *Specialized models and ranking for coreference resolution*, in "Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2008), Hawaii, USA", 2008.

[20] M. FERNÁNDEZ, É. VILLEMONTE DE LA CLERGERIE, M. VILARES. *Mining Conceptual Graphs for Knowledge Acquisition*, in "Proc. of acm cikm Workshop on Improving Non-English Web Searching (inews'08), Napa Valley, USA", J. V. FOTIS LAZARINIS, J. TAIT (editors), 2008, p. 25–32.

[21] G. FRANCOPOULO, T. DECLERCK, V. SORNLERTLAMVANICH, É. VILLEMONTE DE LA CLERGERIE, M. MONACHINI. *Data Category Registry: Morpho-syntactic and Syntactic Profiles*, in "LREC-2008 Workshop on Uses and usage of language resource-related standards", May 2008.

[22] S. KAHANE. *Les unités de la syntaxe et de la sémantique : le cas du français*, in "Proceedings of the "1er Congrès Mondial de Linguistique Française", Paris, France".

[23] L. NICOLAS, B. SAGOT, É. VILLEMONTE DE LA CLERGERIE, J. FARRÉ. *Computer aided correction and extension of a syntactic wide-coverage lexicon*, in "Proc. of CoLing 2008, Manchester, UK", August 2008.

[24] P. PAROUBEK, É. VILLEMONTE DE LA CLERGERIE, S. LOISEAU, A. VILNAT, G. FRANCOPOULO. *PASSAGE Syntactic Representation*, in "The 7th International Workshop on Treebanks and Linguistic Theories (TLT 2009)", jan 2009.

[25] B. SAGOT, D. FISER. *Building a free French wordnet from multilingual resources*, in "Actes de Ontolex 2008, Marrakech, Maroc", 2008.

[26] B. SAGOT, D. FISER. *Construction d'un wordnet libre du français à partir de ressources multilingues*, in "Actes de TALN 2008, Avignon, France", 2008.

[27] B. SAGOT, K. FORT, F. VENANT. *Extension et couplage de ressources syntaxiques et sémantiques sur les adverbes*, in "Proceedings of the 27th Lexicon-Grammar Conference, L'Aquila, Italy", 2008.

[28] D. SEDDAH. *The Use of MCTAG to Process Elliptic Coordination*, in "In Proceeding of the Ninth International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+9), Tüebingen, Germany", June 2008.

[29] É. VILLEMONTE DE LA CLERGERIE, C. AYACHE, G. DE CHALENDAR, G. FRANCOPOULO, C. GARDENT, P. PAROUBEK. *Large scale production of syntactic annotations for French*, in "proc. of The First Workshop on Automated Syntactic Annotations for Interoperable Language Resources, Hong-Kong", January 2008.

[30] É. VILLEMONTE DE LA CLERGERIE, O. HAMON, D. MOSTEFA, C. AYACHE, P. PAROUBEK, A. VILNAT. *PASSAGE: from French Parser Evaluation to Large Sized Treebank*, in "Proceedings of the Sixth International

Language Resources and Evaluation (LREC'08), Marrakech, Morocco", European Language Resources Association (Ed.), may 2008.

[31] É. VILLEMONTE DE LA CLERGERIE. *A collaborative infrastructure for handling syntactic annotations*, in "proc. of The First Workshop on Automated Syntactic Annotations for Interoperable Language Resources, Hong-Kong", January 2008.

[32] A. VILNAT, G. FRANCOPOULO, O. HAMON, S. LOISEAU, P. PAROUBEK, É. VILLEMONTE DE LA CLERGERIE. *Large Scale Production of Syntactic Annotations to Move Forward*, in "COLING'08 Workshop on Cross-Framework and Cross-Domain Parser Evaluation", August 2008.

National Peer-Reviewed Conference/Proceedings

[33] E. GRYGLICKA. *Un système d'annotation des entités nommées du type personne pour la résolution de la référence*, in "Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)", Juin 2008.

[34] L. NICOLAS, B. SAGOT, M. Á. MOLINERO, J. FARRÉ, É. VILLEMONTE DE LA CLERGERIE. *Extensión y corrección semi-automática de léxicos morfo-sintácticos*, in "24th edition of the conference of the Spanish Society for Natural Language Processing (SEPLN 2008)", sep 2008.

[35] B. SAGOT, L. DANLOS. *Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français*, in "Proceedings of the workshop "Lexicographie et informatique : bilan et perspectives", Nancy, France", 2008.

Workshops without Proceedings

[36] D. SEDDAH. *Coordination and Control processing using Multi-Component TAG grammar: is this realistic ?*, in "In Proceeding of the first workshop on Natural Language Processing in Nancy (NATAL), Nancy, Loria, France", June 2008.

Scientific Books (or Scientific Book chapters)

[37] L. DANLOS. *Strong generative capacity of RST, SDRT and discourse dependency DAGs*, in "Constraints in Discourse", A. BENZ, P. KÜHNLEIN (editors), John Benjamins, Amsterdam/Philadelphia, 2008.

[38] S. KAHANE. *On the Status of Phrases in Head-Driven Phrase Structure Grammar: Illustration by a Fully Lexical Treatment of Extraction*, in "Dependency in Linguistic Description", A. POLGUÈRE, I. MEL'CUK (editors), to appear, John Benjamins (Language Companion Series), Amsterdam/Philadelphia, p. 111–150.

[39] S. KAHANE. *Le rôle des structures et représentations dans l'évolution des théories syntaxiques*, G. LECOINTRE, J. PAIN (editors), Université Paris X - Nanterre, Nanterre, France.

References in notes

[40] A. ABEILLÉ, L. CLÉMENT, F. TOUSSENEL. *Building a treebank for French*, Kluwer, Dordrecht, 2003.

[41] A. ARUN. *Statistical Parsing of the French Treebank*, Masters thesis, School of Informatics, University of Edinburg, 2004.

[42] N. ASHER, A. LASCARIDES. *Logics of Conversation*, Cambridge University Press, Cambridge, 2003.

[43] D. BIKEL. *Design of a multi-lingual, parallel-processing statistical parsing engine*, in "Proceedings of the second international conference on Human Language Technology Research", Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2002, p. 178–182.

[44] P. BOULLIER. *Counting with Range Concatenation Grammars*, in "Theoretical Computer Science", vol. 293, n^{o} 2, feb 2003, p. 391–416.

[45] 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.pdf.

[46] P. BOULLIER. *Supertagging: A Non-Statistical Parsing-Based Approach*, in "Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03), Nancy, France", April 2003, p. 55–65, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Pierre.Boullier/supertaggeur_final.pdf.

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

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

[49] 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.

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

[51] P. BOULLIER, B. SAGOT. *Deep non-probabilistic parsing of large corpora*, in "Proc. of LREC'06", 2006, http://atoll.inria.fr/~sagot/pub/LREC06a.pdf.

[52] C. CARDIE, K. WAGSTAFF. *Noun phrase coreference as clustering*, in "Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, University of Maryland, MD", Association for Computational Linguistics, 1999, p. 82–89.

[53] J. CARROLL, T. BRISCOE, A. SANFILIPPO. *Parser evaluation: a survey and a new proposal*, in "First International Conference Language Resources and Evaluation (LREC'98), Granada, 1998", 1998.

[54] J. CHEN, V. K. SHANKER. *Automated extraction of tags from the penn treebank*, 2004, p. 73–89.

[55] D. CHIANG. *Statistical Parsing with an Automatically Extracted Tree Adjoining Grammar*, in "Data-Oriented Parsing", 2003.

[56] M. COLLINS. *Head driven statistical models for natural language parsing*, Ph. D. Thesis, University of Pennsylvania, Philadelphia, 1999.

[57] L. DANLOS. *A Lexicalized formalism for Text Generation inspired from TAG*, in "TAG Grammar", A. ABEILLÉ, O. RAMBOW (editors), CSLI, 2001.

[58] L. DANLOS. *Discourse Verbs and Discourse Periphrastic Links*, in "Second International Workshop on Constraints in Discourse, Maynooth, Ireland", 2006.

[59] L. DANLOS. *D-STAG : un formalisme pour le discours basé sur les TAG synchrones*, in "Proceedings of TALN 2007, Toulouse, France", to appear, 2007.

[60] L. DANLOS, B. GAIFFE, L. ROUSSARIE. *Document structuring à la SDRT*, in "International workshop on text generation - ACL, Toulouse", 2001, p. 94–102.

[61] P. DENIS, J. BALDRIDGE. *Joint determination of anaphoricity and coreference resolution using integer programming*, in "Proceedings of HLT-NAACL 2007, Rochester, NY", 2007.

[62] K. GERDES, S. KAHANE. *Speaking in piles*, submitted.

[63] C. JACQUIN, E. DESMONTILS, L. MONCEAUX. *French EuroWordNet Lexical Database Improvements*, in "Proc. of CICLing'07 (LNCS 4394)", 2007.

[64] T. KING, R. CROUCH, S. RIEZLER, M. DALRYMPLE, R. KAPLAN. *The PARC 700 dependency bank*, in "4th International Workshop on Linguistically Interpreted Corpora (LINC-03)", 2003.

[65] X. LUO. *Coreference or not: a twin model for coreference resolution*, in "Proceedings of HLT-NAACL 2007, Rochester, NY", 2007, p. 73-80.

[66] D. M. MAGERMAN. *Natural Language Parsing as statistical pattern recognition*, 1994.

[67] A. MCCALLUM, B. WELLNER. *Conditional Models of Identity Uncertainty with Application to Noun Coreference*, in "Proceedings of NIPS 2004", 2004.

[68] V. NG, C. CARDIE. *Improving Machine Learning Approaches to Coreference Resolution*, in "Proceedings of ACL 2002", 2002, p. 104–111.

[69] V. NG. *Machine Learning for Coreference Resolution: From Local Classification to Global Ranking*, in "Proceedings of ACL 2005, Ann Arbor, MI", 2005, p. 157–164.

[70] V. NG. *Unsupervised Models for Coreference Resolution*, in "Proceedings of EMNLP 2008", 2008.

[71] S. PETROV, L. BARRETT, R. THIBAUX, D. KLEIN. *Learning Accurate, Compact, and Interpretable Tree Annotation*, in "Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia", Association for Computational Linguistics, July 2006.

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

[73] B. SAGOT, D. FISER. *Building a free French wordnet from multilingual resources*, in "Actes de Ontolex 2008, Marrakech, Maroc", 2008.

[74] B. SAGOT, D. FISER. *Construction d'un wordnet libre du français à partir de ressources multilingues*, in "Actes de TALN 2008, Avignon, France", 2008.

[75] 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.

[76] Y. SCHABES. *Stochastic lexicalized tree-adjoining grammars*, in "Proceedings of the 14th conference on Computational linguistics, Morristown, NJ, USA", Association for Computational Linguistics, 1992, p. 425–432.

[77] N. SCHLUTER, J. VAN GENABITH. *Preparing, Restructuring, and Augmenting a French Treebank: Lexicalised Parsers or Coherent Treebanks?*, in "Proceedings of PACLING 07", 2007.

[78] D. SEDDAH, B. SAGOT. *Modeling and Analysis of Elliptic Coordination by Dynamic Exploitation of Derivation Forests in LTAG parsing*, in "Proceedings of TAG+8, Sydney, Australia", July 2006, p. 147-152.

[79] W. M. SOON, H. T. NG, D. LIM. *A machine learning approach to coreference resolution of noun phrases*, in "Computational Linguistics", vol. 27, n^{O} 4, 2001, p. 521–544.

[80] 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.

[81] D. TUFIS. *BalkaNet Design and Development of a Multilingual Balkan WordNet*, in "Romanian Journal of Information Science and Technology", vol. 7, n^{O} 1–2, 2000.

[82] 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.

[83] É. VILLEMONTE DE LA CLERGERIE. *From Metagrammars to Factorized TAG/TIG Parsers*, in "Proceedings of IWPT'05, Vancouver, Canada", October 2005, p. 190–191, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/IWPT05mg.pdf.

[84] VOSSEN, P.. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*, Kluwer, Dordrecht, 1999.

[85] A. M. YLI-JYRÄ, K. KOSKENNIEMI. *Compiling contextual restrictions on strings into finite-state automata*, in "Proceedings of the Eindhoven FASTAR Days 2004 (September 3–4), Eindhoven, The Netherlands", December 2004.