INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Alpage

Analyse Linguistique Profonde À Grande Échelle

Paris - Rocquencourt

Theme : Audio, Speech, and Language Processing

Table of contents

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. Since January 1st, 2009, Alpage an UMR-I (University Paris 7 & Inria) registered in the Paris 7 quadriennal plan as the UMR-I 001.

1. Team

Research Scientist

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

Faculty Member

François Barthélemy [Associate Professor (MC) CNAM]
Marie Candito [Associate Professor (MC) Univ. Paris 7]
Benoit Crabbé [Associate Professor (MC) Univ. Paris 7]
Laurence Danlos [Full Professor (PR) Univ. Paris 7, Member of IUF, Team leader, HdR]
Sylvain Kahane [Full Professor (PR) Univ. Paris X, Associate member, HdR]
Philippe Muller [delegation from Université Paul Sabatier, Toulouse, since September 2009]
Djamé Seddah [Associate Professor (MC) Univ. Paris 4]

Technical Staff

Caroline Benoît [funded by the PASSAGE ANR Project, localized at LIMSI for scientific reasons]
François Guérin [funded by the ANR Project PASSAGE]

PhD Student

André Bittar [PhD student Univ. Paris 7 (since 2007), now ATER at Université Paris-Est Marne-la-Vallée]
Luc Boruta [PhD student (allocataire) (since October 2009)]
François-Régis Chaumartin [PhD student Univ. Paris 7]
Elżbieta Gryglicka [PhD student (CIFRE) Thales & Univ. Paris 7]
Enrique Henestroza Anguiano [PhD funded by the SEQUOIA ANR project (since November 2009)]
Charlotte Roze [PhD student (allocataire) Univ. Paris 7 (since October 2009)]
Rosa Stern [7-month internship, then PhD student (CIFRE) AFP & Univ. Paris 7 (since November 2009)]
Juliette Thuilier [PhD student (allocataire) Univ. Paris 7 (since 2008)]

Post-Doctoral Fellow

Marie-Laure Guénot [6 months, funded by the ANR Project PASSAGE]
Sattisvar Tandabany [funded by the ANR Project SEQUOIA]

Visiting Scientist

Kuppusamy Lakshmanan [9-month ERCIM visitor]

Administrative Assistant

Christelle Guiziou [Secretary (SAR) Inria]

Other

Vanessa Combet [ARITT contract with Kwaga]
Frédéric Meunier [ARITT contract with Kwaga]
Victor Mignot [4-month internship]
Gaëlle Recourcé [funded by the Scribo project]

2. Overall Objectives

2.1. Overall Objectives

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 text 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., Spanish Polish, Slovak, Persian, Galician, and others). This is of course of high relevance, among others, for language-independant modeling and multilingual 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, (probabilistic and symbolic) 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.

3. Scientific Foundations

3.1. From programming languages to linguistic grammars

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

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 [65], [106], [111]) 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

Participants: Éric Villemonte de La Clergerie, Benoit 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 [118], [114]. 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

Participants: Pierre Boullier, Éric Villemonte 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] [63], [64]. 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 [66], [68], [67].

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 [117] 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 [114], [118].

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).

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. Probabilistic parsing approaches

Participants: Marie Candito, Benoit Crabbé, Pascal Denis, Djamé Seddah, Benoît Sagot, Pierre Boullier.

The development of large scale symbolic grammars has long been a lively topic in the French NLP community. Surprisingly, the acquisition of probabilistic grammars aiming at stochastic parsing, using either supervised or unsupervised methods, has not attracted much attention despite the availability of large manually syntactic annotated data for French. Nevertheless, the availability of the Paris 7 French Treebank [57], allowed [86] to carry out the extraction of a Tree Adjoining Grammar [89] and led [58] to induce the first effective lexicalized parser for French. Yet, as noted by [112], the use of the treebank was "challenging". Indeed, before carrying out successfully any experiment, the authors had to perform a deep restructuring of the data to remove errors and inconsistencies.

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

On the other hand, [79] showed that with a new released and corrected version of the treebank. it was possible to train statistical parsers from the original set of trees. This path has the advantage of an easier reproducibility and eases verification of reported results.

Before that, it is important to describe the characteristics of the parsing task. In the case of statistical parsing, two different aspects of syntactic structures are to be considered : their capacity to capture regularities and their interpretability for further processing.

3.4.1. Generalizability

Learning for statistical parsing requires structures that capture best the underlying regularities of the language, in order to apply these patterns to unseen data.

Since capturing underlying linguistic rules is also an objective for linguists, it makes sense to use supervised learning from linguistically-defined generalizations. One generalization is typically the use of phrases, and phrase-structure rules that govern the way words are grouped together. It has to be stressed that these syntactic rules exist at least in part independently of semantic interpretation.

3.4.2. Interpretability

But the main reason to use supervised learning for parsing, is that we want structures that are as *interpretable* as possible, in order to extract some knowledge from the analysis (such as deriving a semantic analysis from a parse). Typically, we need a syntactic analysis to reflect how words *relate* to each other. This is our main motivation to use supervised learning : the learnt parser will output structures as defined by linguists-annotators, and thus interpretable within the linguistic theory underlying the annotation scheme of the treebank. It is important to stress that this is more than capturing syntactic regularities : it has to do with the *meaning* of the words.

It is not certain though that both requirements (generalizability / interpretability) are best met in the same structures. In the case of supervised learning, this leads to investigate different instantiations of the training trees, to help the learning, while keeping the maximum interpretability of the trees. As we will see with some of our experiments, it may be necessary to find a trade-off between generalizability and interpretability.

Further, it is not guaranteed that syntactic rules infered from a manually annotated treebank produce the best language model. This leads to methods that use semi-supervised techniques on a treebank-infered grammar backbone, such as [94], [104].

3.4.3. The Alpage's statistical parsing architecture

In order to carry out the task of building a statistical parser for French, we started by exploring the state-of-the-art in statistical parsing technology for others languages. However, as much of the work being done in statistical parsing has been carried out by English speaking teams, most of the parser publicly available is specifically tuned for the English language underlying current practice in the field by mostly training and parsing the Wall Street Journal sections of the Penn Treebank.

That is why we decided to adapt to French, the state-of-the-art parsers available in the two phrase structures parsing paradigms : lexicalized and unlexicalized parsers. We found out that in order to get the best performance from our annotated data, the annotation scheme has to be modified to include some important morpho-syntactic information [79],[3]. This led the unlexicalized parser we adapted ([104], [79]) to offer the best performance for French so far. Meanwhile, as we are working on a very small data set, we explored various means (lemmatization, clustering) of reducing the data sparseness issues originating from a somewhat small lexicon. In the context of working with various phrase structure based parsers, we have been naturally inclined to design a data driven phrase structure to dependency parsing process that remains generic whatever the parser being used. Overall, this architecture exhibits state-of-the-art results, even though recent work on adapting pure statistical dependency parser to French, which was carried out in our team, for the sake of thoroughness, shows that a model *à la* McDonald [96] exhibits a slight improvement over our main architecture [28].

3.5. Dynamic wide coverage lexical resources

Participants: Benoît Sagot, Laurence Danlos, Éric Villemonte 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 [110]. At the syntactic level, work has been achieved on automatic acquisition of atomic syntactic information and automatic detection of errors in the lexicon [121],[8]. At the semantic level, automatic wordnet development tools have been described [105], [119], [88], [87]. 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 Lefff [108],[42], as well as a wordnet for French, the WOLF [109], the first freely available resource of the kind.

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 Lefff and the WOLF. This work is a good example of how Inria and Paris 7 members of Alpage fruitful collaborate: this collaboration between NLP computer scientists and NLP linguists have resulted in significant advances which would have not been possible otherwise.

3.6. Treebanks development and exploitation

Participants: Benoit Crabbé, Marie Candito, Éric Villemonte 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 [57].

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

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.

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 [116]. 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.

Nevertheless, members of Alpage are involved in the Rhapsodie ANR project (see 8.1.4). One of the tasks of this project, coordinated by Sylvain Kahane, is to develop a dependency Treebank for a little corpus of Spoken French (3 hours = 36,000 words). The corpus, orthographically transcripted, is manually segmented by linguists in rectional units, where words are linked by dependency relations. These units will be parsed by the Alpage team. A difficulty comes form the fact, that due to disfluencies, reformulation, and so on, rectional unit are not disjoint, and the syntactic trees we obtain must be patched up. This is a first step in the direction of the parsing of spoken languages. The next step would be to see how to obtain automatically a segmentation in rectional units.

3.7. Building and evaluating full-featured parsing systems

Participants: Éric Villemonte 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 [7]. Moreover, when used as a preliminary step before parsers, the quality of parsers' results strongly depends on the quality of such chains.

3.8. Standardization

Participants: Éric Villemonte 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").

3.9. Discourse structures

Participants: Laurence Danlos, Pascal Denis, Benoît Sagot.

SDRT Segmented Discourse Representation Theory

RST Rhetorical Structure Theory

TAG Tree-Adjoining Grammar

Collaboration with Nicholas Asher (IRIT, Toulouse).

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 [81].

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, [82],[4]), 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.

3.10. Coreference resolution

Participants: Pascal Denis, Philippe Muller, Elżbieta Gryglicka, Laurence Danlos.

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

Collaboration with Nicholas Asher (IRIT, Toulouse).

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 président*, 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 question-answering systems. 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 [113], [100], [95], [101], [93];[6]. Note that a few attemtps were also made at using unsupervised techniques (mostly clustering methods) for the task [72], [102], but these systems are still far from reaching the performance of their supervised counterparts.

4. Application Domains

4.1. Panorama

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

Human-computer interaction : 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, such as information extraction and knowledge acquisition (4.2 and 4.5), text mining (4.3), text generation (4.4).

4.2. Information extraction and knowledge acquisition

Participants: Éric Villemonte de La Clergerie, François-Régis Chaumartin, Elżbieta Gryglicka, Rosa Stern, Benoît Sagot.

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

[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/

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, Scribo project). François-Régis Chaumartin, PhD student at Alpage and CEO of Proxem, 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.

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

4.3. Processing answers to open-ended questions in surveys: vera

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 (see section 5.6) 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. Generation of textual reports about statistical data: EASYTEXT

Participant: Laurence Danlos.

In 2009, the generation system EASYTEXT has been achieved. Is is now an operational system used at TNS-SOFRES. Its is based on G-TAG, a formalism based on Tree Adjoining Grammar, [80], enriched with a document structuring module taking ideas from SDRT (Segmented Discourse Representation Theory, [59]), [83]. This formalism has been implemented in .net by WatchSystem Assistance. EASYTEXT is fully integrated in TNS modules.

An example of text generated by EASYTEXT along with the source data is shown at http://www.linguist.univ-paris-diderot.fr/~danlos/EvolVar.htm.

As TNS-SOFRES was pleasantly surprised by the quality of the automatically generated texts, they ask for further extensions of EASYTEXT which are currently worked on.

Another application of NLG we foresee is the automatic production of captions for photos. There is ongoing discussions with AFP (*Agence france Presse*) on the topic.

4.5. Processing e-mails: Kwaga

Participants: Laurence Danlos, Benoît Sagot, Frédéric Meunier, Vanessa Combet.

Kwaga is an online service available for different webmails and email clients *via* an asynchronous plugin. Its aim is to enrich (and not to replace) existing email applications, helping users to decide which emails they should read and in which order, and providing additional features related to emails such as incoming email categorization, priority level assignment and automatic calendar events creation based on the content of the email, and others.

Such a service necessarily relies on NLP tools and techniques. In its first version, Kwaga uses the Unitex tool, developed at the University of Marne-la-Vallée. The SXPipe procesing chain, developed at Alpage, is another option that has several comparative advantages: it has been decided to grant SXPipe a stability and technical characteristics that are suitable for industrial use, and to have Kwaga benefit from this technology.

To achieve this goal, an ARITT contract was signed between Alpage and the Kwaga company, that develops and distributes this service, with objectives and results that benefit to both sides:

- For Alpage, this allowed to (start and) integrate SXPipe within the UIMA standard, which is a common requirement from industrial partners. Moreover, linguistic information embedded in SXPipe has been extended for improving the quality of its results on Kwaga's domain, namely email corpora.
- For Kwaga, this allowed for a technology transfer of recent research outcomes. Replacing Unitex by SXPipe, which is more recent and based on the last advances in both computer science and linguistics, appears to be a reasonable and promising choice, that should be evaluated shortly.

5. Software

5.1. Syntax

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) includes various deterministic and non-deterministic CFG parser generators. It includes in particular 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. This implementation of the Earley algorithm has been recently extended to handle probabilistic CFG (PCFG), by taking into account probabilities both during parsing (beam) and after parsing (n-best computation). 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 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 (Marseilles) and other members of the SEQUOIA ANR project (see section 8.1.2), as well as (indirectly) all SXPipe and/or SXLFG users.

5.2. SxLfg

Participants: Benoît Sagot [correspondant], Pierre Boullier.

SXLFG is a parser generator based on SYNTAX for Lexical-Functional Grammars (LFG) [67], [67], [66]. 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 Le*fff* and takes SXPipe outputs as input. It constitutes a very efficient deep parser, which can parse several million-word corpus in only several hours [66], [69]

5.3. System DyALog

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

DYALOG on INRIA GForge: 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 **1.13.0** of DYALOG is freely available by FTP under an open source license and runs on Linux platforms for x86 architectures and on Mac OS X (ppc and intel). Two ports for 64bit architectures on Mac OS Intel and Linux have been added this year, with a speed-up of 2 wrt to their 32bits counterparts.

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 [117].

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 two PASSAGE campaigns (Dec. 2007 and Nov. 2009), cf. 8.1.1 and [114],[54].

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. Tools and resources for Meta-Grammars

Participant: Éric Villemonte de La Clergerie [correspondant].

On INRIA GForge: http://mgkit.gforge.inria.fr/

DYALOG (cf. 5.3) has been used to implement `mgcomp`, a compiler of Meta-Grammar (cf. 6.10). 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 [114]. 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. The Bonzai PCFG-LA parser

Participants: Benoit Crabbé [correspondant], Marie Candito, François Guérin, Pascal Denis, Djamé Seddah.

Alpage has developped as support of the research papers [79],[26],[3], [9] a statistical parser for French, named Bonzai, trained on the French Treebank. This parser provides both a phrase structure and a projective dependency structure specified in [28] as output. This parser operates sequentially : (1) it first outputs a phrase structure analysis of sentences reusing the Berkeley implementation of a PCFG-LA trained on French by Alpage (2) it applies on the resulting phrase structure trees a process of conversion to dependency parses using a combination of heuristics and classifiers trained on the French treebank. The parser currently outputs several well known formats such as Penn treebank phrase structure trees, Xerox like triples and CONLL-like format for dependencies. The parsers also comes with basic preprocessing facilities allowing to perform elementary sentence segmentation and word tokenisation, allowing in theory to process unrestricted text. However it is believed to perform better on newspaper-like text. This parser is to be released in 2010 under a GPL license.

5.6. Alpage's linguistic workbench, including SxPipe

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

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

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 [107] 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...).

5.7. MElt

Participants: Pascal Denis [correspondant], Benoît Sagot.

As described in 6.18, MElt (called MElt_{fr} in [30]) is a newly developed part-of-speech tagger, trained for French on the French TreeBank and coupled with the Le*fff*. It is distributed freely as a part of the Alpage linguistic workbench.

5.8. The syntactic lexicon Lefff and the Alexina framework

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 6.7).

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 6.6).

5.9. System EasyRef

Participants: Éric Villemonte de La Clergerie [correspondant], François Guérin.

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).

Version 2 has been used by ELDA and LIMSI to annotate a new corpus of several thousands words for PASSAGE. A preliminary version 3 has been developed by François Guérin, relying on Berkeley DB XML to handle very large annotated corpora and to provide a complete query language expanded as XQuery expressions. EASYREF is maintained under INRIA GForge.

6. New Results

6.1. Mildly Context-Sensitive formalisms

Participants: Pierre Boullier, Benoît Sagot.

Pierre Boullier and Benoît Sagot have worked on formal aspects of Mildly Context-Sensitive formalisms in three directions. First, Pierre Boullier and Benoît Sagot have published an algorithm for parsing an input DAG (or lattice) with a general Range Concatenation Grammar (RCG) [24]. The feasability of this task, as well as its possible exponential time complexity, were known, but no algorithm had been published before (to our best knowledge).

Second, Pierre Boullier and Benoît Sagot have put together the TIG (Tree Insertion Grammars) formalism and the multiple components paradigm underlying the MC-TAG (Multiple-Component Tree Adjoining Grammars), and therefore defined and studied the properties of MC-TIGs (Multiple-Component Tree Insertion Grammars) [23]. This hierarchy of formalisms define a hierarchy of languages that exhibit interesting properties. For example, 2-MC-TIGs are strictly more powerful than TAGs, they are parsable in time $O(n^6)$ just as TAGs, and allow for grammatical descriptions using bi-component structures — the drawback being that so-called wrapping auxiliary trees are prohibited.

Third, Benoît Sagot has been working in collaboration with Giorgio Satta (University of Padova, Italy), on the optimal reduction of the rank of 2-LCFRS (Linear Context-Free Rewriting Systems), that are equivalent to simple (linear) RCGs with at most 2 arguments per predicate. This work has led Benoît Sagot to go to Padova for a 1-month stay. Publications should appear on that topic in 2010.

6.2. Membrane Grammars and Bracketed Contextual Grammars

Participants: Éric Villemonte de La Clergerie, Kuppusamy Lakshmanan.

In the context of Dr Lakshmanan's ERCIM fellowship at INRIA, we have explored several syntactic formalisms such as Membrane Grammars and Bracketed Contextual Grammars, exploring their relationships with existing formalism classes such as the class of Mildly Context-Sensitive (MCS) formalisms. We also explored the design of parsing strategies for them, in particular with Thread Automata.

6.3. Finite-State Multi-Tape Transducers

Participant: François Barthélemy.

François Barthélemy 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 [120]. 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. Karamel implements an original unit test framework inspired from the JUnit framework for Java [17]. Tests of finite-state transducers are performed using assertions, namely evaluable boolean predicates. Tests may involve auxiliary finite-state machines called *fixtures* (e.g.: a given input to a transducer and the corresponding expected output are fixtures). At the moment, Karamel is still a prototype. We plan to complete its development and begin to distribute it in the near future.

The relevance of multi-tape transducers for Natural Language Processing has been exemplified in a case study in Semitic Morphology: a comprehensive verbal grammar of the Akkadian language has been written using Karamel [18].

6.4. Automatic tools for developping, improving and correcting lexical resources

Participants: Benoît Sagot, Éric Villemonte de La Clergerie.

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

In collaboration with Lionel Nicolas (University of Nice) and Miguel Ángel Molinero Álvarez (University of Ourense, Galicia, Spain), within the Victoria project funded by the government of Galicia [53], [33], we worked towards the development of a more complete lexical development framework that is able to detect missing and dubious entries in existing lexicons with different techniques, and suggest respectively addition and corrections hypotheses for these entries. This is achieved thanks to 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 [34], [35].

We applied this technique for improving the Le*ff*e, an Alexina lexicon for Spanish (see 6.7).

6.5. Merging syntactic lexical resources for improving the Lefff: continued

Participants: Benoît Sagot, Laurence Danlos.

The development of the Le*fff* has been pursued. First of all, some work on a particular class of verbs [39] has allowed us to improve the quality of the Le*fff* but also to start and exploit an additional very valuable resource, namely the Lexique des Verbes Français from Dubois and Dubois-Charlier. This will allow us to benefit from one more lexical resource for improving the Le*fff*, apart from DicoValence and Lexicon-Grammar Tables.

A new version of the Le*fff* is under preparation, that is the result of a full conversion and merging of Dicovalence with the current version of the Le*fff*, following the work described in [5]. The semi-automatic validation of the result of this merging is in progress, and should lead to the release of a greatly improved version of the Le*fff*, version 3.5, that will be quantitatively compared to version 3.0 by comparing the results of the FRMG parser when it uses the one version or the other.

More generally, we have realized during this year, and in particular during the ATALA workshop on French parsing (a satellite event of IWPT'09), that the Le*fff* is now a widely used resource within the French NLP community.

6.6. Comparing lexical resources for parsing: Lexicon-Grammar vs. Lefff

Participants: Benoît Sagot, Éric Villemonte de La Clergerie.

Lexicon-Grammar tables are currently one of the major sources of syntactic lexical information for the French language. Moreover, several Lexicon-Grammar tables exist for other languages. However, current tables suffer from various types of inconsistency and incompleteness. In particular, defining features are not represented in the tables. To remedy this situation, tables of classes are being developed at IGM (Université Paris-Est) for each category, and notably for verbs, which associate the set of their defining features with each class. Preliminary results of this long-term effort allowed Benoît Sagot, in collaboration with Elsa Tolone from IGM, to convert verb tables into the Alexina format, the format of the Le*fff*, hence turning it into a lexicon usable by the FRMG parser, named LgLex [44], [43], [49].

This allowed Éric de La Clergerie and Benoît Sagot to build a variant of FRMG that doesn't uses the Le*fff*, but that uses a lexicon derived from the Le*fff* in which lexical entries for standard verbs are replaced by those from LgLex. We evaluated both FRMG variants on the EASy corpus with the EASy metrics. For now, results are still slightly better with the original Le*fff*. However, in this process, we identified many possible improvements, in particular in the tables themselves and in the conversion process. Moreover, a very promising complementarity between both resources has been found (the set of parsable sentences strongly vary w.r.t. the lexicon used), which confirms Alpage's approach to lexical development that emphasizes, among other techniques, the conversion, comparison and merging of existing lexical resources.

6.7. The Leffe and SpMg: a lexicon and a parser for Spanish

Participants: Benoît Sagot, Éric Villemonte de La Clergerie.

Collaboration with Miguel Ángel Molinero Álvarez (University of Ourense, Galicia, Spain) and Lionel Nicolas (University of Nice). In relation with the Victoria Spanish-French project (see 8.2.1), some of Alpage's members have worked on the development of a syntactic lexicon and a metagrammar for Spanish, in collaboration with other members of the Victoria project. In particular, improvements have been made in 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 [32], [52], [31]. Many techniques for lexical information acquisition have been improved as well, such as converting and merging of syntactic lexicons, corpus-based extraction of morphological lexicons, and others.

In parallel with the development of the syntactic lexicon Le*ffe*, the development of a meta-grammar for Spanish is ongoing. This metagrammar, SpMg, uses 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*, preliminary versions of a deep DyALog-based parser for Spanish has been built.

Within the Victoria project, these efforts will be pursued, and extended to Galician[6], and possibly adapted to other languages.

6.8. PerLex, a morphological lexicon for the Persian language

Participant: Benoît Sagot.

In the context of the PerGram project 8.2.2, Benoît Sagot collaborated with Géraldine Walther (LLF, Université Paris 7) and Pollet Samvélian (Université Paris 3) to begin the development of a morphological and syntactic lexicon for the Persian language, as well as a processing chain (i.e., a Persian version of SxPipe). In 2009, the first step towards this goal has been achieved, and the first version of the PerLex lexicon has been released [38]. It only contains morphological information (valency frames and complex predicates are planned for 2010), has not been manually validated yet apart from verbs and some specific entries (a full validation by native speakers is planned in the first half of 2010), and is still to be completed and augmented thanks to techniques described in 6.4. But this is the first large-coverage freely available lexicon for Persian.

6.9. Lexconn: French Lexicon of Discourse Connectives

Participants: Laurence Danlos, Charlotte Roze, Philippe Muller.

LEXCONN is a French lexicon of 330 discourse connectives, collected with their syntactic category and the discourse relation(s) they express [56], [37]. Such a resource already exists for English, Spanish and German, but LEXCONN is the first one for French. The lexicon aims at being exhaustive. It has been constructed manually, applying systematic connective identification criteria, associating a SDRT relation, and the type (coordinating or subordinating) of this relation with each connective. This work leads to a reflexion on the set of relations defined in SDRT and the distinction between implicit relations (i.e. not marked by a connective) and explicit relations (i.e. marked by a connective).

Building a French lexicon of discourse connectives brought several results. It implied a systematic methodology to identify discourse connectives and associate them discourse relations, resting on various studies about connectives. In addition, it shows which connectives remain to be studied in detail (especially connectives marked as "unknown", to which we couldn't associate any discourse relation). A statistical analysis of the resulting lexicon permitted to quantify several things, like importance of the various discourse relations in terms of number of connectives, and count of ambiguous connectives (i.e. connectives that can establish more than one relation). LEXCONN contains 330 connectives. About 70% are non-ambiguous, which is an encouraging result, and only 3% establish more than one relation. Concerning ambiguous connectives, we think that there is two cases: the case where a connective establish relations of the same type (coordinating or subordinating), and the case where a connective establish relations of the two types. The first case seems less problematic than the second in an NLP perspective, because it doesn't implies structural ambiguity. Only 22 connectives are in the second case.

Despite these results, LEXCONN has to be improved: some information has to be added. For example, some information about ambiguity between discourse usage and non discourse usage has to be introduced. This improvement will be possible with other linguistic analysis, but also with automatic analysis on ANNODIS corpus: we could examine the link between position in the host clause and discursive/non-discursive role for adverbials. However, LEXCONN already constitute a precious resource for NLP. It might help for discourse markers annotation in ANNODIS project, in which connectives are not yet marked. A statistical analysis of the connectives on corpus can also be useful, for example concerning connective's frequency. Such analysis could help answering the following question: are ambiguous connectives the most frequent ones?

6.10. Designing efficient parsers using Meta-Grammars and DyALog

Participants: Éric Villemonte de la Clergerie, Marie-Laure Guénot.

MG *Meta-Grammars*

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

In the context of the PASSAGE action and of last parsing evaluation campaign, we have tried to improve the coverage and quality of FRMG by exploring various approaches. Beyond the use of error mining techniques on large raw corpora, we have tried more supervised based approaches relying on the repeated processing of the 4000 reference EASy treebank. Each run provides detailed information about the errors of the analysers, in particular through the use of confusion matrices for chunks and dependencies. It is also possible to build confusion matrices tracing the changes between two runs, useful to quickly detect unexpected and unwanted consequences of modifications in FRMG or companion modules. A more linguistic evaluation of the phenomena pointed by the matrices, through the examination of corresponding sentences (with the help of logs and EasyRef) was useful to detect all kinds of problems in the processing chains, some of them being also errors in the treebank (and then correction of these errors using EasyRef). The process iterated over a few months allowed a several points increase of the quality of FRMG (+2% to reach 87.7% for chunks and +4.5% to reach 64.1% for dependencies), proving the importance of good methodologies and good tools (feedback, visualization, query, ...) to improve a linguistic processing chain.

We increased the coverage of FRMG by adding new classes in the underlying meta-grammar, in particular to handle causative constructions, more cases of superlative constructions, adjective subcategorization, to cite a few of them. Such extensions tend to slow parsing, in particular because ambiguity increases. Various optimizations have been tried at different levels of the processing chain in order to contain this effect. The disambiguation algorithm on the shared dependency forests has been revised to be more efficient and better weights for the disambiguation rules have been searched through trial and errors, using the above mentioned feedback techniques. More automatic machine-learning based techniques have been tried, not leading yet to better results.

6.11. Probabilistic TIG-based dependency parsing

Participants: Pierre Boullier, Benoît Sagot.

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

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

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) [16]. It relies on a grammar (TIG) extraction algorithm initially developed by [75] 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 8.1.2), as soon as all underlying techniques are transfered to French.

The MICA parser is distributed freely (http://mica.lif.univ-mrs.fr/).

6.12. Optimized reduction of probabilized shared parse forests

Participants: Pierre Boullier, Benoît Sagot.

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

Collaboration with Alexis Nasr (LIF, Université de Marseille-Provence), within the ANR funded-project SEQUOIA (see 8.1.2).

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 since 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.

The integration of these algorithms within the system SYNTAX has been achieved, thus allowing to get very interesting quantitative results [22]: in general, the size of the resulting forest, for reasonable values of n (say, 100), has the same order of magnitude as that of the pruned forest, but it contains only the best n trees.

6.13. Introducing beam search techniques in the Earley algorithm

Participants: Pierre Boullier, Benoît Sagot.

In the context of the SEQUOIA project, Pierre Boullier and Benoît Sagot have been working on various techniques for reducing the search space of the Earley CFG parsing algorithm when using Probabilistic CFGs (PCFGs). These techniques have been implemented in the SYNTAX system, but have not been fully evaluated yet, nor published (this should be done in 2010).

In short, beam search techniques or variants thereof can be introduced at different stages of the Earley algorithm. In particular, given an Earley item, estimations and/or exact figures for the best probability of the prefix and the suffix of the item as well as for the parts at the right and at the left of the dot within the item, can be computed. This allows for various types of online item pruning, some of them being exact (i.e., the overall best parse will always be retained), some of them not. This work is crucial when dealing with huge grammars with a huge ambiguity, such as grammars generated by the Berkeley split-merge algorithm [103].

6.14. Dependency parsing

Participants: Marie Candito, Benoit Crabbé, Pascal Denis, François Guérin.

Dependency trees are often preferred to syntagmatic trees for many NLP tasks, such as information extraction, question answering, lexical acquisition. We started in 2008, and continued in 2009, to work on the conversion of the syntagmatic trees of the French treebank into surface dependency trees. We have now a stabilized version of a dependency treebank : the French treebank converted to dependencies [26].

The constituent-to-dependency conversion procedure can also be applied to syntagmatic trees as output by a parser trained on the syntagmatic treebank. Hence, we have various ways to obtain a parser outputting dependency trees : (i) training a parser on syntagmatic trees, and converting the output of this parser into dependencies [26]. And (ii) directly using existing algorithms to train a dependency parser on the treebank converted to dependencies. We have begun a comparison of the two approaches. First bare results [28] show for now that this second approach leads to better results : directly training a dependency parser with the MST algorithm [97] outperforms the architecture where a parser is trained on the French treebank (using Petrov's algorithm), and output trees from this parser are converted to dependencies. We plan to work on a more qualitative comparison of the strength and weaknesses of both approaches.

6.15. Improving the lexical coverage of statistical parsers

Participants: Marie Candito, Benoit Crabbé, Djamé Seddah, Enrique Henestroza Anguiano.

Probabilistic parsers are trained on treebanks, namely syntactically annotated sentences, and this training allows to capture syntactic regularities. Yet, though lexical information is known to play a crucial role in determining the syntactic structure of a sentence, many lexical phenomena cannot be learned simply by training on a treebank of a few thousands of sentences (the French treebank we use contains about 12000 sentences). First because treebanks cover only a small part of the French vocabulary. Second, because lexical data is very sparse : a corpus contains a few very frequent words, and a lot of rare words. Compared to English, this is even truer for French, or more generally inflected languages : morphological marks for gender, number, tense etc... drastically augment the vocabulary size.

6.15.1. Word clustering

To cope with this inherent limitation of statistical parsing techniques, we have investigated the use of word clusters instead of words as input to the parser. Our work was inspired by [91], who have shown that word clusters obtained with unsupervised techniques could improve statistical dependency parsing, when used as features for classifiers determining the weights of dependency arcs. We tried to use word clusters within the framework of generative statistical parsing. We have first defined an algorithm to get rid of morphological marks for gender, number, tense and mood, without resorting to part-of-speech tagging. It makes use of the Lefff lexicon [108], and allows to cluster forms on a morphological basis, still keeping the morpho-syntactic ambiguities of input words. We applied this process to the *L'Est Républicain* corpus, a 125 million word journalistic corpus, freely available at CNRTL (http://www.cnrtl.fr/corpus/estrepublicain). Then, we applied Brown's algorithm for unsupervised word clustering ([71]) on the resulting corpus. Using the resulting word to cluster mapping, we were able to train a parser (using Petrov's algorithm) on a modified treebak, where word forms are replaced by their cluster. This has led to a significant improvement of parsing performance [25], when tested on part of the treebank used as a test set. The method has two advantages. First, because the reduction of the vocabulary size (to clusters) leads to better probability estimations, that explains the improvement on a test set taken from the treebank. Second, this reduced vocabulary (the set of clusters) corresponds in fact to an augmented set of word forms known at training time. There are then less totally unknown word forms at parsing time. This suggests that parsing performance should also be better for parsing text of a domain different from that of the treebank.

6.15.2. Data driven lemmatization

In conjunction with the work being done in the team on word clusterization, where the goal is to obtain better probability estimates (cf. previous section), we are also working on the integration of a lemmatization process into our parsing chain. Let us recall that the lemmatization is the process of getting the canonical form of a given word form (ie. *mangerions* is lemmatized as *manger* for instance), therefore it is a mean to reduce data sparseness issues (common when one is working with very small amount of annotated data). A collaboration between Gzregorz Chrupała, a postdoctoral researcher from the Saarland University, and our team was initiated via an invitation offered to Djamé Seddah to work for a week on the adaptation of Chrupala's state-of-the-art data driven morphology learner tool (Morfette, [77]) to the French language. This was done by the integration of the Alpage's wide coverage lexicon (Lefff, [108]) into the Morfette's training set. This fruitful collaboration led to the development of Morfette's module aimed toward French that exhibits the best results so far for French both in POS tagging and in lemmatization. The POS tagging state-of-the-art, for example, is 97.88% with Morfette (Overall accuracy in MElt [30] 97.70% — but let us recall here that the MElt models does not use lemmatization information during training); on unseen words, Morfette reaches 92.50% (90.01% for MElt). Therefore the inclusion of a tuple lemma+POS instead of a simple word form in one of our parsers will help to improve parsing results. Papers on this topic are in preparation to be submitted to ACL and to COLING 2010.

6.16. Comparing various models for statistical parsing

Participants: Djamé Seddah, Marie Candito, Benoit Crabbé.

In parallel with the effort to reduce data sparseness issues coming from small treebank with rich morphology, we are also experimenting various parsing models for the French Treebank. This work started in fall of 2008 and is still on going. It involved adapting the famous Charniak's parser [74] to a romance language, which had never been done before, to various instances of the Collins' model [78] and to two types of Stochastic Tree Insertion Grammar [76], one of those being a very promising formalism (spinal-stig, see [48]). For the latter, the idea is to consider sequence of unary branching rules as fragment of trees (called spines) instead of seeing those trees as set of CFG rules. For one instance of the French Treebank, the grammar is thus very compact, being made of 83 unlexicalized spines instead of 14 000 CFG trees for the same treebank. Some attention is raised by the parsing community on this topic: using a similar formalism, [73] achieved state of the art results on parsing the WSJ. Seeing this bubbling on this topic, one can consider that a paradigm shift is actually on its way in the parsing community: working in a horizontal (CFG) way means data sparseness whereas switching to vertical grammars (spines) implies working with very compact grammars. A preliminary paper has been submitted last November [48].

6.17. Large scale corpus processing

Participant: Éric Villemonte de La Clergerie.

In the context of the PASSAGE action, we have continued the explore the use of distributed computing for processing of large corpora, largely using GRID 5000 and a local cluster at INRIA Rocquencourt. We use more and more such resources also for the post-parsing phases and the ambition is to use them for machine-learning phases.

GRID5000 and the local cluster were specially useful for the parsing evaluation campaign (October and November 2009), even such real life experiments tend to show that scripts in such complex environments are never robust enough.

6.18. State-of-the-art French tagging with MElt

Participants: Pascal Denis, Benoît Sagot, Djamé Seddah.

Pascal Denis and Benoît Sagot worked on a new MaxEnt-based tagger, MElt, trained on the French TreeBank for building a tagger for French. This baseline, which makes no use of an external lexical resource, can be significantly improved by coupling it with the French morphosyntactic lexicon Le*fff*. The resulting tagger, MElt_{fr} , reaches a 97.7% accuracy that is, to our best knowledge state-of-the-art for that task (i.e., tagging with no lemmatization information). More precisely, the addition of lexicon-based features yield error reductions of 23.3% overall and of 27.5% for unknown words (corresponding to accuracy improvements of .7% and 3.9%, respectively) compared to the baseline tagger [30].

Pascal Denis and Benoît Sagot also showed that the use of a lexicon improves the quality of the tagger at any stage of lexicon and training corpus development. Moreover, they approximately estimated development times for both resources, and show that the best way to optimize human work for tagger development is to work on the development of both an annotated corpus and a morphosyntactic lexicon.

Moreover, Djamé Seddah has initiated a collaboration with Grzegorz Chrupała (University of Saarbrücken, Germany), who independently proposed a system called Morfette [77] based on the same machine learning techniques than MElt but that benefits from lemmatization information in the training data for improving tagging accuracy and providing lemmas in addition to tags in the output. This collaboration should lead to joint efforts between MElt and Morfette, in order to improve tagging and lemmatization accuracy, applying these techniques to other languages (including resource-scarse languages), and studying the influence of tagging and lemmatization on parsers' performances when used as pre-processing steps.

6.19. Towards a better understanding of frequency effects in syntax

Participants: Benoit Crabbé, Juliette Thuilier.

Some members of Alpage are involved in the statistical parsing of French, the idea of using probabilistic devices for parsing is rather new in France. Alpage has shown earlier [79] that such parsers are performing quite well on French.

Since earlier non statistical parsers were inspired by a trend in linguistics that rejects the idea of granting any importance to frequency effects, it remains largely unknown which are these probabilistic factors that help parsing. This question has almost never been adressed for French. We decided to launch a theoretical investigation aimed at identifying which factors come into play when we take frequency effects into account.

In collaboration with Gwen Fox (Université Pars 3), the first investigation in this direction has been led towards identifying the importance of constraints that drive the placement of adjectives wrt the noun in the noun phrase in French. This study brings an additional element to Bresnan's thesis [70], according to which the syntactic competence of human beings is indeed probabilistic. Further ongoing studies on adjectives will try to bring evidences for the facts (1) that the grammar of a natural language is intrisically redundant and (2) that we indeed store in our mind not only words of the language but also highly frequent grammatically compositional sequences.

As can be seen from the outline above, this line of research brings us closer to cognitive sciences and more specifically to frameworks inspired by construction grammar. We hope in the very long run that these investigations will bring further insights on the design of probabilistic parsers. In NLP the framework that is closest to implementing construction grammar is Data Oriented Parsing [62].

6.20. Discourse Synchronous TAGs: a formalism for discourse analysis

Participant: Laurence Danlos.

D-STAG is a new formalism for the automatic analysis of the discourse structure of texts [4]. The analyses computed by D-STAG are hierarchical discourse structures annotated with discourse relations, that are compatible with discourse structures computed in SDRT, [59]. The discourse analysis extends the sentential analysis, without modifying it, which simplifies the realization of the system. More precisely, it is based on the following architecture with 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.

The second step consists in getting a "normalized form for discourse" (DNF) from the syntactic analysis of a suite of sentences. It turns out that the results of the (French) syntactic analyzers are not good enough to obtain satisfactory DNFs. This negative findings can be explained by the following data: in the evaluation campaigns of French syntactic analysers, namely EASy next PASSAGE, the metrics that are used give the same importance to short-rang and long-rage relations (dependencies). The former are much more numerous that the latter and so are quite relevant to be highly ranked. Moreover the former are much more easy to compute. As a result, the long-range relations are somehow neglected. Unfortunately, the DNF for a discourse can only be obtained with a high quality tool for segmenting sentences into clauses, which requires to detect long-range dependencies.

For this reason, we postpone the implementation of D-STAG waiting for best results from French syntactic analyzers. However, we are enhancing the coverage of D-STAG by studying how to handle quotations and the quotation incidents that introduce them. This work was initiated in the project Scribo (see 6.23 and 8.1.5). This lead to an inventory of "quotation verbs" extracted from an AFP corpus, half of them being not reported speech verbs. We start exploring the structure of discourses with quotations, which may question some basic principles in SDRT.

6.21. Processing of temporal information in French texts

Participants: André Bittar, Laurence Danlos, Pascal Denis, Philippe Muller.

TempEval-2: André Bittar, Pascal Denis and associated member Philippe Muller (delegation at Alpage from the Université Paul Sabatier, Toulouse), in collaboration with Michel Gagnon (École Polytechnique de Montréal), are currently participating in the TempEval-2 campaign for the evaluation of systems designed for the annotation of temporal information in natural language texts (http://www.timeml.org/tempeval2). This group of three researchers and one PhD student was responsible for the creation of an evaluation corpus for French. The corpus has recently been finished and submitted to the campaign organisers. It will be used as the gold standard against which to gauge the performance of automated systems designed for the annotation of temporal information in French language texts.

It relies on the previous work of André Bittar on the adaptation of the TimeML annotation framework for temporal expressions to French language [20], [19]. André Bittar and Laurence Danlos also worked on the integration of light verb constructions in TimeML [21].

6.22. Coreference Resolution

Participant:

Early machine learning approaches to coreference resolution rely on local, discriminative pairwise classifiers [113], [100], [98] made considerable progress in creating robust coreference systems, but their performance still left much room for improvement. This stems from two main deficiencies:

- **Decision locality**. Decisions are made independently of others; a separate clustering step forms chains from pairwise classifications. But, coreference clearly should be conditioned on properties of an entity as a whole.
- **Knowledge bottlenecks**. Coreference involves many different factors, e.g., morphosyntax, discourse structure and reasoning. Yet most systems rely on small sets of shallow features. Accurately predicting such information and using it to constrain coreference is difficult, so its potential benefits often go unrealized due to error propagation.

More recent work has sought to address these limitations. For example, to address decision locality, McCallum and Wellner [95] use conditional random fields with model structures in which pairwise decisions influence others. Denis [85] and Klenner [90] use integer linear programming (ILP) to perform global inference via transitivity constraints between different coreference decisions. Denis and Baldridge [84] use a ranker to compare antecedents for an anaphor simultaneously rather than in the standard pairwise manner. To address the knowledge bottleneck problem, Denis and Baldridge [6] use ILP for joint inference using a pairwise coreference model and a model for determining the anaphoricity of mentions. Also, Denis and Baldridge [84] and Bengston and Roth [61] use models and features, respectively, that attend to particular types of mentions (e.g., full noun phrases versus pronouns). Furthermore, Bengston and Roth [61] use a wider range of features than are normally considered, and in particular use predicted features for later classifiers, to considerably boost performance.

In [13], we use ILP to extend the joint formulation of Denis and Baldridge [6] using named entity classification and combine it with the transitivity constraints [85], [90]. Intuitively, we only should identify antecedents for the mentions which are likely to have one [99], and we should only make a set of mentions coreferent if they are all instances of the *same* entity type (eg, PERSON or LOCATION). ILP enables such constraints to be declared between the outputs of independent classifiers to ensure coherent assignments are made. It also leads to global inference via both constraints on named entity types and transitivity constraints since both relate multiple pairwise decisions.

We show that this strategy leads to improvements across the three main metrics proposed for coreference: the MUC metric [115], the b^3 metric [60], and CEAF metric [92]. In addition, we contextualize the performance of our system with respect to cascades of multiple models and oracle systems that assume perfect information (e.g. about entity types). We furthermore demonstrate the inadequacy of using only the MUC metric and argue that results should *always* be given for all three.

6.23. Sapiens: visualizing quotations in news wires

Participants: Benoît Sagot, Éric Villemonte de La Clergerie, Rosa Stern, Pascal Denis, Victor Mignot, Laurence Danlos, Gaëlle Recourcé.

In relation to the Scribo project (see 8.1.5), several Alpage members were involved in the development of a demonstration environment for linguistic processing. This environment, named Sapiens, is a platform of quotations visualization in news wires associated with its author and context [50]. It has been applied to a corpus provided by the Agence France-Presse (AFP). Sapiens demonstrates how named entities can be related to events, here to quotations in news wires from AFP (Agence France Presse). demonstrated during the annual System@tic meeting, in front of a large audience including the State Secretary for Research.

The orginality of this environment is that it relies on a deep linguistic processing chain that includes SXPipe processing chain (that includes named entity recognition), the FRMG parser and a coreference resolution module, which allows for extracting quotations with a wide coverage and an extended definition, including quotations which are only partially quotes-delimited verbatim transcripts. It is an example of an application based on information extraction, which can be useful to final users as journalists looking for relevant information in news archives. The resulting information is stored in a database and can thus be reused, for instance in the development of an ontology.

From a more linguistic perspective, this work led us to try and study the syntactic and discursive properties of so-called quotation verbs (such as "say", "laugh" or "conclude") that can head constructions such as *"It is a wonderful idea", laughed Peter*. This raises very important NLP issues, since such constructions are in contradiction with many assumtions made by most parsers, although they are very common and very intersting from an applicative point of view. This linguistic study has been described in two submitted publications, one that focuses on the syntactic level [38] and another that includes the discursive level.

7. Contracts and Grants with Industry

7.1. TEXT-ELABORATOR (2008–2009)

Participant: Laurence Danlos.

TEXT-ELABOATOR is a NLG (Natural Language Generation) project funded by TNS Sofres. It is leaded by the startup Watch System Assistance for whom L. Danlos works as a scientific consultant. The NLG system is operational within TNS since the fall of 2009.

7.2. Kwaga (ARITT contract, 2009)

Participants: Benoît Sagot, Laurence Danlos.

Kwaga is a start up which develops a product to help anyone getting through her abundance of e-mails. This product is based on a (superficial) semantic analysis of e-mails.

The objective of the contract between Kwaga and Alpage is twofold: first, to give SXPipe an industrial aspect by integrating it into a processing chain that relies on the UIMA standard; second, to evaluate the results of SXPipe on e-mails. If these results are good enough, SXPipe— in its UIMA version — could be used instead of Unitex in the product developed by Kwaga.

8. Other Grants and Activities

8.1. National Initiatives

8.1.1. ANR project PASSAGE (2006 – 2008)

Participants: Éric Villemonte de La Clergerie, Benoît Sagot, Pierre Boullier, François Guérin, Caroline Benoît, Marie-Laure Guénot.

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 and extended till mid 2010. 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.

This year is essentially the participation of two ALPAGE parsers to the 2nd parsing evaluation campaign organized by PASSAGE (Fall 2009).

8.1.2. ANR project Sequoia (2009 – 2011)

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

Alpage plays 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 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.

8.1.3. ANR project EDyLex (Nov. 2009 – Oct. 2011)

Participants: Benoît Sagot [principal investigator], Gaëlle Recourcé, Rosa Stern, Laurence Danlos, Pascal Denis.

EDyLex is an ANR project (STIC/CONTINT) headed by Benoît Sagot. The focus of the project is the dynamic acquisition of new entries in existing lexical resources that are used in syntactic and semantic parsing systems: how to detect and qualify an unknown word or a new named entity in a text? How to associate it with phonetic, morphosyntactic, syntactic, semantic properties and information? Various complementary techniques will be explored and crossed (probabilistic and symbolic, corpus-based and rule-based...). Their application to the contents produced by the AFP news agency (Agence France-Presse) constitutes a context that is representative for the problems of incompleteness and lexical creativity: indexing, creation and maintainance of ontologies (location and person names, topics), both necessary for handling and organizing a massive information flow (over 4,000 news wires per day).

The participants of the project, besides Alpage, are the LIF (Université de Méditerrannée), the LIMSI (CNRS team), two small companies, Syllabs and Vecsys Research, and the AFP.

8.1.4. ANR project Rhapsodie (2008 – 2010)

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

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.

8.1.5. Action Scribo (2007 – 2009, extended until 2010)

Participants: Éric Villemonte de La Clergerie, Benoît Sagot, Rosa Stern, Pascal Denis, Gaëlle Recourcé, Victor Mignot.

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 end 2010. It brings 9 participants together: AFP, CEA LIST, INRIA, LRDE (Epita), Mandriva, Nuxeo, Proxem, Tagmatica and XWiki.

8.2. European Initiatives

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

Participants: Éric Villemonte de La Clergerie, Benoît Sagot.

As a follow-up of a long lasting collaboration with Galician universities, ALPAGE, Éric de La Clergerie and Benoît Sagot are 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 6.7 describes the results obtained in this direction in 2009.

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, for which a first version is now available, (ii) 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 Villemonte 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 Villemonte 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 9-month visit of Dr Kuppusamy Lakshmanan (Vellore Institute of Technology, India) in the context of the ERCIM fellowship programme.

A 2 month visit of Daniel Fernandez (Univ. of Vigo) from November to December 2009.

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 2009, the following speakers gave a talk in this seminar: Laura Kallmeyer (Univ. Tuebingen), Asaf Bachrach (MIT/Inserm/CEA), Brian Roark (OSU, Oregon State University), Joakim Nivre (Univ Uppsala), Alexis Nasr (Université de la Méditerrannée), Jennifer Foster (Dublin City University), Reut Tsarfaty (University van Amsterdam), Kenji Sagae (University of Southern California), Roberto Basili (Univ Roma II), Antonio Balvet, L. Barque, R Marin (Univ. Lille, STL), Paola Merlo (Univ. Geneva).
- 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";

9.2. Supervising

- Laurence Danlos is the official PhD advisor for all the ALPAGE students (except Luc Boruta, whose advisor is Emmanuel Dupoux) since she is the only member of the team with an HDR:
  - André Bittar (allocataire Paris 7) who should finish in 2010. Pascal Denis is co-advisor as well as Pascal Amsili
  - Elżbieta Gryglicka (Cifre Thales) who should finish in 2010. Frédéric Landragin (CNRS, LATTICE) is co-advisor
  - Juliette Thullier (allocataire Paris 7) who started in October 2008. Benoît Crabbé is co-advisor
  - Charlotte Roze (allocataire Paris 7) who started in October 2009. Philippe Muller (MC à Toulouse 2, Délégation INRIA for 2009-2010) is co-advisor
  - Rosa Stern (cifre AFP 7) who started in October 2009. Benoît Sagot is co-advisor
  - Enrique Henestroza Anguiano (ANR Sequoia funding) who started in October 2009. A. Nasr (Prof. Université de la Méditerranée) is co-advisor as well as M. Candito.
- Laurence Danlos was the PhD advisor for Pierre Hankach (France-Télécom) who defended his thesis in March 2009.
- Benoît Crabbé was an examiner for the PhD defense of Nicolas Barrier who defended his thesis in 2009.
- Laurence Danlos has supervized the Master 2 internship of:
  - Charlotte Roze on a lexical data base of French connectives which records their syntactic type and the discourse relation(s) they convey.
  - Grégoire Detrez on a chunker which aims at segmenting complex sentences into clauses.
- Marie Candito and Sylvain Kahane supervized the Master 2 internship of Ugo Jardonnet, on the qualitative and quantitative description of the 'flow' of syntactic dependencies;

- François-Régis Chaumartin supervized the Master 2 internship of Joanne Boisson;
- Djamé Seddah was the supervized the Master 1 internship of Louise Bouchseche, on semi-automatic annotation of elliptic coordinations on Treebank (Master 1 ILGII, Université Paris 4 Sorbonne), and the Master 2 internship of Yuanyuan XU, on Lexicalized Tree adjoining grammars extraction from the Chinese Penn Treebank (Master 1 ILGII, Université Paris 4 Sorbonne).
- Éric de La Clergerie has supervised the internship of Victor Mignot on the development of WEB service SAPIENS.
- Pascal Denis supervised the internship of Alexis Vanacker, Ecole des Mines de Nancy, June-August 2009, on the development of a collaborative annotation tool for anaphoric relations.

9.3. Committees

- Alpage is involved in the French journal T.A.L. (AERES linguistic rank: A). Éric de La Clergerie is "Rédacteur en chef". Laurence Danlos has been nominated as member of the editorial board. Benoît Sagot is "Secrétaire de rédaction" of the journal; Benoît Crabbé, Pascal Denis and Benoît Sagot did external reviews for T.A.L. in 2009.
- Éric de La Clergerie was program chair for the 2009 edition of the International Workshop on Parsing Technologies organized in Paris by Alpage, for which Laurence Danlos was local chair;
- Éric de La Clergerie was co-organizer of the ATALA worshop on "French Parsing Systems" (October 10th, 2009), as a satellite event of IWPT'09.
- Participation of Laurence Danlos to the program committee of TALN'09
- Participation of Éric de La Clergerie to the program committees of TALN'09, TEMA'09 and CLA'09. He has also reviewed for ACL-IJCNL'09 (areas: "Syntax and Parsing") and EACL'09.
- Participation of Pierre Boullier to the program committees of ACL 2009
- Participation of Pascal Denis to the program committees of the journals *ACM Transactions on Speech and Language Processing*, *ACM Computing Surveys* and the conferences ACL 2009, EMNLP 2009, IEEE-ICSC 2009, the NAACL/HLT 2009 Workshop on Integer Linear Programming for NLP and TLS XII.
- Participation of Benoît Sagot to the program committees of ACL 2009 and TALN 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 INALCO team for AERES.
- Evaluation by Éric de La Clergerie of one project for ANR CONTINT Program
- Evaluation by Benoît Sagot of one project for ANR White Program
- Éric de La Clergerie participated to an AERES evaluation committee.
- Djamé Seddah and Benoît Sagot are elected board member of the French NLP society (ATALA); Djamé Seddah is Program Chair of the "Journées ATALA" (one day long workshops in NLP, 4 or 5 per year).
- Laurence Danlos was the head of the organizing committee for the 50th birthday of ATALA (Association du Traitment Automatique du Langage Naturel), of which Djamé Seddah was a member. This one-day event, which took place in the grand Amphithéâtre de la Sorbonne, gathered 300 participants.

9.4. Participation to workshops, conferences, and invitations

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

- Participation with presentations of Marie Candito, Benoît Crabbé and Djamé Seddah at the EACL'09 workshop on Computational Aspects of Grammatical Inference [3],[47].
- Participation with presentation of François Barthélemy at the EACL'09 workshop on Computational Approaches to Semitic Languages [18].
- Participation with presentation of François Barthélemy at the Conference on Implementation and Application of Automata [17].
- Participation with presentations of Pierre Boullier, André Bittar, Marie Candito, Benoît Crabbé, Laurence Danlos, Pascal Denis, François Guérin, Benoît Sagot and Éric de La Clergerie at TALN'09 [26], [34], [44], [21].
- Participation with presentation of Pascal Denis at PACLIC'09 [30].
- Participation with presentations of Benoît Sagot at the 28th Lexis and Grammar Conference [43], [39].
- Participation with presentations of Benoît Sagot at LTC'09 [49], [50].
- Participation with presentations of Benoît Sagot and Laurence Danlos at Formal Grammars [23], [29].
- Participation with presentation of André Bittar at the 19th Meeting of Computational Linguistics in The Netherlands [20].
- Participation with presentation of André Bittar at LAW III [19].
- Participation with presentations of Pierre Boullier, Marie Candito, Benoît Crabbé, Benoît Sagot, Djamé Seddah and Éric de La Clergerie at IWPT'09 [25], [55], [22], [24],[9]; Djamé Seddah organized and moderated a panel session during IWPT'09; note that the whole IWPT'09 conference was organized by Alpage and hosted by Université Paris 7, as well as the ATALA workshop on French Parsing organized the following day; Éric de La Clergerie gave a presentation during that workshop [54].
- As a follow-up to the IWPT'09 panel session, Djamé Seddah has put together researchers from various universities and countries to write a workshop proposal on Parsing for Morphologically-Rich Languages. This proposal has been accepted by a joint commitee of the most prestigious conference in worlwide NLP, and the workshop will be held in 2010 as a satellite event to the NAACL conference. Djamé Seddah is heading the scientific committee of the workshop, of which Marie Candito is a member. Benoît Crabbé and Benoît Sagot are members of the reviewing committee.
- Participation of several Alpage members to the ESSLLI'09 summer school in Bordeaux, France (this concerns Laurence Danlos, Benoît Sagot, Rosa Stern, Juliette Thuilier, François Guérin).
- Benoît Sagot did a one week stay in Padova University (December 2009), work with Giorgio Satta on formal aspects of Linear Context-Free Rewriting Systems (LCFRS).
- Djamé Seddah did a one week stay in Saarland University (December 2009), work with Grzegorz Chrupala on Data driven morphology acquisition.
- Participation of Éric de La Clergerie to ISO TC37SC4 meetings (June 2009, Boulder, USA).
- Benoît Crabbé gave 3 invited talks: one at DCU (Ireland), one at UCL (Belgium), one at Orléans (Lifo).
- Benoît Crabbé and Marie Candito gave 1 invited talk at IGM (Paris Est).
- Pascal Denis gave 3 invited talks at Universität Heidelberg, Loria (Nancy) and Alpage.
- Djamé Seddah gave 1 invited talk at the Saarland University.
- Éric de La Clergerie was invited to deliver an invited talk on "Feature Structures" at LexiPraxi 2009 (Paris, October 2009).

9.5. Teaching

Alpage 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 directed by Laurence Danlos. Benoît Crabbé and then Marie Candito were 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 participate 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[7]: Introduction to NLP (3rd year of License, 24h); Discourse, NLU and NLG (2nd year of Master, 36h).

Marie Candito: Information retrieval (2nd year of professional Master, 12h); Probabilistic methods for Natural language processing (1st year of Master, 48h); Machine translation (1st year of Master, 48h); Probabilities and statistics for Natural language processing (3rd year of Licence, 24h); French syntax (2nd year of Licence, 21h, License of Linguistics of University Paris 7).

Benoît Sagot: Parsing systems (2nd year of Master, 24h).

Éric de La Clergerie: Prolog and NLP (2rd year of Master, 12h).

Benoît Crabbé (INRIA delegation): Probabilistic methods for NLP (1st year of Master, 48h); Introduction to programming II (3rd year of Licence, 24h).

Pascal Denis: Computational Semantics (2nd year of Master, 12h).

Charlotte Roze: Introduction to Programming (3rd year of License, 24h); Algorithmics (3rd year of License, 24h).

François-Régis Chaumartin: Modélisation (UML) et bases de données (SQL) (2rd year of professional Master, 24h).

Djamé Seddah, as an Assistant Professor in CS in the University Paris 4 Sorbonne, member of the UFR ISHA, mainly teaches “Generic Programming and groupware”, “Distributed Application and Object Programming”, “Syntaxic tools and text Processing for NLP”, “Machine Translation Seminars” in both years of the Master “Ingénierie de la Langue pour la Gestion Intelligente de l’Information”. Djamé Seddah is also the “Directeur des études” of a CS transversal module for the Sorbonne’s undergraduate students (ie “Certificat Informatique et Internet”).

André Bittar is an ATER at Université Paris-Est Marne-la-Vallée, where he taught “Introduction to Operating Systems” (1st year of DUT, 52h), “Unix/HTML” (1st year of License, 48h) and “Programming with Python” (1st year of Master, 12h) during the first semester of the university year 2009-2010.

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.

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

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

[3] M. CANDITO, B. CRABBÉ, D. SEDDAH. *On statistical parsing of French with supervised and semi-supervised strategies*, in "EACL 2009 Workshop Grammatical inference for Computational Linguistics, Athens, Greece", 2009.

[4] L. DANLOS. *D-STAG : un formalisme d'analyse automatique de discours fondé sur les TAG synchrones*, in "Traitement Automatique des Langues", vol. 50, n^{o} 1, 2009.

[5] L. DANLOS, B. SAGOT. *Constructions pronominales dans Dicovalence et le lexique-grammaire – Intégration dans le Lefff*, in "Linguisticæ Investigationes", vol. 2, n^{o} 32, 2009.

[6] P. DENIS, J. BALDRIDGE. *Joint Determination of Anaphoricity and Coreference Resolution using Integer Programming*, in "HLT-NAACL", 2007, p. 236-243.

[7] B. SAGOT, P. BOULLIER. *SxPipe 2: architecture pour le traitement présyntaxique de corpus bruts*, in "Traitement Automatique des Langues (T.A.L.)", 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.

[9] D. SEDDAH, M. CANDITO, B. CRABBÉ. *Cross Parser Evaluation and Tagset Variation: a French Treebank Study*, in "Proceedings of the 11th Internation Conference on Parsing Technologies (IWPT'09), Paris, France", 2009, p. 150-161.

[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.

Year Publications

Doctoral Dissertations and Habilitation Theses

[11] P. HANKACH. *Génération automatique de textes par satisfaction de contraintes*, Université Paris VII Denis-Diderot, Paris, France, 2009, Supervizor: Laurence Danlos, Ph. D. Thesis.

Articles in International Peer-Reviewed Journal

[12] L. DANLOS. *D-STAG : un formalisme d'analyse automatique de discours fondé sur les TAG synchrones*, in "Traitement Automatique des Langues", vol. 50, n^{o} 1, 2009.

[13] P. DENIS, J. BALDRIDGE. *Global joint models for coreference resolution and named entity classification*, in "Procesamiento del Lenguaje Natural", vol. 43, 2009.

[14] B. SAGOT, L. DANLOS. *Constructions pronominales dans Dicovalence et le lexique-grammaire — Intégration dans le Lefff*, in "Linguisticæ Investigationes", vol. 32, n^{o} 2, 2009, http://atoll.inria.fr/~sagot/pub/LI09pron.pdf.

[15] B. SAGOT, K. FORT, F. VENANT. *Extension et couplage de ressources syntaxiques et sémantiques sur les adverbes*, in "Linguisticæ Investigationes", vol. 32, n^{o} 2, 2009, http://atoll.inria.fr/~sagot/pub/LI09adv.pdf.

International Peer-Reviewed Conference/Proceedings

[16] S. BANGALORE, P. BOULLIER, A. NASR, O. RAMBOW, B. SAGOT. *MICA: A Dependency Parser Based on Tree Grammars*, in "Proceedings of NAACL 2009 (application note), Boulder, Colorado, USA", 2009, http://atoll.inria.fr/~sagot/pub/NAACL09appnote.pdf US .

[17] F. BARTHÉLEMY. *A Testing Framework for Finite-State Morphology*, in "Proceedings of the 14th International Conference on Implementation and Application of Automata (CIAA'09), Sydney, Australia", 2009, p. 75-83.

[18] F. BARTHÉLEMY. *The Karamel System and Semitic Languages: Structured Multi-Tiered Morphology*, in "Proceedings of the EACL'09 Workshop on Computational Approaches to Semitic Languages (CASL'09), Athens, Greece", 2009.

[19] A. BITTAR. *Annotation of Events and Temporal Expressions in French Texts*, in "Proceedings of the 3rd Linguistic Annotation Workshop (LAW III), Singapore", 2009.

[20] A. BITTAR. *Annotation of Temporal Information in French Texts*, in "Proceedings of CLIN: The 19th Meeting of Computational Linguistics in The Netherlands, Groningen, Netherlands", 2009.

[21] A. BITTAR, L. DANLOS. *Intégration des constructions à verbe support dans TimeML*, in "Proceedings of TALN 2009, Senlis, France", 2009.

[22] P. BOULLIER, A. NASR, B. SAGOT. *Constructing parse forests that include exactly the n-best PCFG trees*, in "Proceedings of IWPT 2009, Paris, France", 2009, http://atoll.inria.fr/~sagot/pub/iwpt09exactnbest.pdf.

[23] P. BOULLIER, B. SAGOT. *Multi-Component Tree Insertion Grammars*, in "Proceedings of Formal Grammars 2009, Bordeaux, France", 2009, http://atoll.inria.fr/~sagot/pub/FG09.pdf.

[24] P. BOULLIER, B. SAGOT. *Parsing Directed Acyclic Graphs with Range Concatenation Grammars*, in "Proceedings of IWPT 2009, Paris, France", 2009, http://atoll.inria.fr/~sagot/pub/iwpt09rcg.pdf.

[25] M. CANDITO, B. CRABBÉ. *Improving generative statistical parsing with semi-supervised word clustering*, in "Proceedings of IWPT'09), Paris, France", 2009.

[26] M. CANDITO, B. CRABBÉ, P. DENIS, F. GUÉRIN. *Analyse syntaxique du français : des constituants aux dépendances*, in "Proceedings of TALN'09, Senlis, France", 2009.

[27] M. CANDITO, B. CRABBÉ, D. SEDDAH. *On statistical parsing of French with supervised and semi-supervised strategies*, in "EACL 2009 Workshop Grammatical inference for Computational Linguistics, Athens, Greece", 2009.

[28] M. CANDITO, B. CRABBÉ, P. DENIS. *Statistical French dependency parsing: treebank conversion and first results*, in "Submitted to LREC'10, La Valetta, Malta", 2010.

[29] L. DANLOS. *D-STAG: a Formalism for Discourse Analysis based on SDRT and using Synchronous TAG*, in "Proceedings of the 14th Conference on Formal Grammar (FG'09), Bordeaux, France", 2009, p. 1-20.

[30] P. DENIS, B. SAGOT. *Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort*, in "Proceedings of PACLIC 2009, Hong Kong, China", 2009, http://atoll.inria.fr/~sagot/pub/paclic09tagging.pdf.

[31] M. A. MOLINERO, B. SAGOT, L. NICOLAS. *A morphological and syntactic wide-coverage lexicon for Spanish: the Leffe*, in "Proceedings of Recent Advances in Natural Language Processing (RANLP)", 2009 ES .

[32] M. A. MOLINERO, B. SAGOT, L. NICOLAS. *Building a morphological and syntactic lexicon by merging various linguistic ressources.*, in "Proceedings of NODALIDA 2009, Odense, Danemark", 2009, http://atoll.inria.fr/~sagot/pub/Nodalida09.pdf ES .

[33] L. NICOLAS, M. A. MOLINERO, B. SAGOT, E. SÁNCHEZ TRIGO, É. VILLEMONTE DE LA CLERGERIE, M. A. PARDO, J. FARRÉ, J. MIQUEL VERGÉS. *Towards efficient production of linguistic resources: the Victoria Project*, in "EuroConference Recent Advances in Natural Language Processing (RANLP)", 2009 ES .

[34] L. NICOLAS, B. SAGOT, M. A. MOLINERO, J. FARRÉ, É. VILLEMONTE DE LA CLERGERIE. *Trouver et confondre les coupables : un processus sophistiqué de correction de lexique*, in "Proceedings of TALN'09, Senlis, France", 2009, http://atoll.inria.fr/~sagot/pub/TALN09lexfix.pdf ES .

[35] L. NICOLAS, B. SAGOT, M. A. MOLINERO, J. FARRÉ, É. VILLEMONTE DE LA CLERGERIE. *Mining parsing results for lexical correction: toward a complete correction process of wide-coverage lexicons*, in "LNAI 5603, selected papers presented at the LTC 2007 conference", Springer, 2009 ES .

[36] 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)", 2009.

[37] C. ROZE, L. DANLOS, P. MULLER. LEXCONN *: a French Lexicon of Discourse Connectives*, in "Signalling Text Organisation - 8th International Workshop on Multidisciplinary Approaches to Discourse 2010 (MAD 10), Moissac, France", 2010.

[38] B. SAGOT, L. DANLOS, R. STERN. *A Lexicon of French Quotation Verbs for Automatic Quotation Extraction*, in "Submitted to LREC'10, La Valetta, Malta", 2010.

[39] B. SAGOT, K. FORT. *Description et analyse des verbes désadjectivaux et dénominaux en -ifier et -iser*, in "Proceedings of the 28th Lexis and Grammar Conference, Bergen, Norway", 2009, http://hal.archives-ouvertes.fr/docs/00/40/24/51/PDF/clg09ifier.pdf.

[40] B. SAGOT, K. FORT, F. VENANT. *Extending the adverbial coverage of a French wordnet*, in "Proceedings of the NODALIDA 2009 workshop on WordNets and other Lexical Semantic Resources, Odense, Danemark", 2009, http://hal.archives-ouvertes.fr/docs/00/40/23/05/PDF/BSKFFV09.pdf.

[41] B. SAGOT. *Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish*, in "LNAI 5603, selected papers presented at the LTC 2007 conference", Springer, 2009.

[42] B. SAGOT. *The Lefff, a freely available, accurate and large-coverage lexicon for French*, in "Submitted to LREC'10, La Valetta, Malta", 2010.

[43] B. SAGOT, E. TOLONE. *Exploitation des tables du Lexique-Grammaire pour l'analyse syntaxique automatique*, in "Proceedings of the 28th Lexis and Grammar Conference, Bergen, Norway", 2009, http://atoll.inria.fr/~sagot/pub/clg09lglex.pdf.

[44] B. SAGOT, E. TOLONE. *Intégrer les tables du lexique-grammaire à un analyseur syntaxique robuste à grande échelle*, in "Proceedings of TALN 09 (poster session), Senlis, France", 2009, http://atoll.inria.fr/~sagot/pub/TALN09lglex.pdf.

[45] B. SAGOT, G. WALTHER. *A morphological lexicon for the Persian language*, in "Submitted to LREC'10, La Valetta, Malta", 2010.

[46] D. SEDDAH, M. CANDITO, B. CRABBÉ. *Cross Parser Evaluation and Tagset Variation: a French Treebank Study*, in "Proceedings of the 11th Internation Conference on Parsing Technologies (IWPT'09), Paris, France", 2009, p. 150-161.

[47] D. SEDDAH, M. CANDITO, B. CRABBÉ. *Adaptation de parsers statistiques lexicalisés pour le français : Une évaluation complète sur corpus arborés*, in "Proceedings of Traitement automatique des Langues Naturelles (TALN 2009), Senlis, France", 2009.

[48] D. SEDDAH. *Exploring the Spinal-Tig Model for parsing French*, in "Submitted to LREC'10, Valetta, Malta", 2010.

[49] E. TOLONE, B. SAGOT. *Using Lexicon-Grammar tables for French verbs in a large-coverage parser*, in "Proceedings of LTC 2009, Poznań, Poland", 2009, http://atoll.inria.fr/~sagot/pub/ltc09lglex.pdf.

[50] É. VILLEMONTE DE LA CLERGERIE, B. SAGOT, R. STERN, P. DENIS, G. RECOURCÉ, V. MIGNOT. *Extracting and Visualizing Quotations from News Wires*, in "Proceedings of L&TC 2009, Poznań, Poland", 2009, http://atoll.inria.fr/~sagot/pub/ltc09sapiens.pdf.

National Peer-Reviewed Conference/Proceedings

[51] L. DANLOS. *Extension de la notion de verbe support*, in "Actes du Colloque International Supports et prédicats non verbaux dans les langues du monde, Paris, France", 2009.

[52] M. A. MOLINERO, B. SAGOT, L. NICOLAS. *Construcción y extensión de un léxico morfológico y sintáctico para el Español: el Leffe*, in "Proceedings of SEPLN 09, San Sebastian, Spain", 2009, http://atoll.inria.fr/~sagot/pub/SEPLN09leffe.pdf ES .

[53] L. NICOLAS, M. A. MOLINERO, B. SAGOT, E. SÁNCHEZ TRIGO, É. VILLEMONTE DE LA CLERGERIE, M. A. PARDO, J. FARRÉ, J. MIQUEL VERGÉS. *Producción eficiente de recursos lingüísticos: el proyecto Victoria*, in "Proceedings of SEPLN 09, San Sebastian, Spain", 2009, http://atoll.inria.fr/~sagot/pub/SEPLN09victoria.pdf ES .

[54] É. VILLEMONTE DE LA CLERGERIE, B. SAGOT, L. NICOLAS, M.-L. GUÉNOT. *FRMG: évolutions d'un analyseur syntaxique TAG du français*, in "Actes électroniques de la Journée ATALA sur "Quels analyseurs syntaxiques pour le français ?"", ATALA, October 2009.

Books or Proceedings Editing

[55] É. VILLEMONTE DE LA CLERGERIE (editor). *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, Association for Computational Linguistics, October 2009.

Other Publications

[56] C. ROZE. LEXCONN *: Base lexicale des connecteurs discursifs du français*, Université Paris 7 Denis Diderot, Paris, France, 2009, Masters thesis.

References in notes

[57] A. ABEILLÉ, L. CLÉMENT, F. TOUSSENEL. *Building a treebank for French*, in "Treebanks: building and using parsed corpora", A. ABEILLÉ (editor), Kluwer academic publishers, 2003, p. 165-188.

[58] A. ARUN, F. KELLER. *Lexicalization in Crosslinguistic Probabilistic Parsing: The Case of French*, in "Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI", 2005, p. 306–313.

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

[60] A. BAGGA, B. BALDWIN. *Algorithms for scoring coreference chains*, in "Proceedings of LREC 1998", 1998, p. 563-566.

[61] E. BENGSTON, D. ROTH. *Understanding the Value of Features for Coreference Resolution*, in "Proceedings of EMNLP 2008, Honolulu, Hawaii", 2008, p. 294-303.

[62] R. BOD. *The Data-Oriented Parsing Approach: Theory and Application*, Springer, 2008, p. 307–342.

[63] P. BOULLIER. *Guided Earley Parsing*, in "Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03), Nancy, France", April 2003, p. 43–54.

[64] 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.

[65] 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.

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

[67] P. BOULLIER, B. SAGOT. *Efficient and robust LFG parsing: SxLfg*, in "Proceedings of IWPT'05, Vancouver, Canada", October 2005, p. 1–10.

[68] 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://hal.archives-ouvertes.fr/docs/00/41/30/77/PDF/TALN05sxlfg.pdf.

[69] P. BOULLIER, B. SAGOT. *Deep non-probabilistic parsing of large corpora*, in "Proc. of LREC'06", 2006.

[70] J. BRESNAN. *Is syntactic knowledge probabilistic? Experiments with the English dative alternation*, in "In Roots: Linguistics in Search of Its Evidential Base, Berlin", S. FEATHERSTON, W. STERNEFELD (editors), Series: Studies in Generative Grammar, Mouton de Gruyter, 2007, p. 77–96.

[71] P. F. BROWN, V. J. DELLA, P. V. DESOUZA, J. C. LAI, R. L. MERCER. *Class-based n-gram models of natural language*, in "Computational linguistics", vol. 18, n^{o} 4, 1992, p. 467–479.

[72] 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.

[73] X. CARRERAS, M. COLLINS, T. KOO. *TAG, dynamic programming, and the perceptron for efficient, feature-rich parsing*, in "Proc. of CoNLL-2010", 2008, p. 9–16.

[74] E. CHARNIAK. *A maximum-entropy-inspired parser*, in "Proceedings of the first conference on North American chapter of the Association for Computational Linguistics", 2000, p. 132–139.

[75] J. CHEN, V. K. SHANKER. *Automated extraction of tags from the penn treebank*, in "New developments in parsing technology, Norwell, MA, USA", Kluwer Academic Publishers, 2004, p. 73–89.

[76] D. CHIANG. *Statistical parsing with an automatically-extracted tree adjoining grammar*, in "Proceedings of the 38th Annual Meeting on Association for Computational Linguistics", 2000, p. 456–463.

[77] G. CHRUPAŁA, G. DINU, J. VAN GENABITH. *Learning Morphology with Morfette*, in "Proceedings of LREC2008", 2008.

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

[79] 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.

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

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

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

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

[84] P. DENIS, J. BALDRIDGE. *Specialized models and ranking for coreference resolution*, in "Proceedings of EMNLP 2008, Honolulu, Hawaï", 2008.

[85] P. DENIS. *New Learning Models for Robust Reference Resolution*, University of Texas at Austin, 2007, Ph. D. Thesis.

[86] A. DYBRO-JOHANSEN. *Extraction automatique de grammaires á partir d'un corpus français*, Université Paris 7, 2004, Masters thesis.

[87] D. FISER. *Leveraging Parallel Corpora and Existing Wordnets for Automatic Construction of the Slovene Wordnet*, in "Proceedings of L&TC'07, Poznań, Poland", 2007.

[88] N. IDE, T. ERJAVEC, D. TUFIS. *Sense Discrimination with Parallel Corpora*, in "Proc. of ACL'02 Workshop on Word Sense Disambiguation", 2002.

[89] A. K. JOSHI. *Introduction to Tree Adjoining Grammar*, in "The Mathematics of Language", A. MANASTER-RAMER (editor), J. Benjamins, 1987.

[90] M. KLENNER. *Enforcing coherence on coreference sets*, in "Proceedings of RANLP 2007", 2007.

[91] T. KOO, X. CARRERAS, M. COLLINS. *Simple Semi-supervised Dependency Parsing*, in "Proc. of ACL-08, Columbus, USA", 2008.

[92] X. LUO. *On coreference resolution performance metrics*, in "Proceedings of HLT-NAACL 2005", 2005, p. 25-32.

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

[94] T. MATSUZAKI, Y. MIYAO, J. TSUJII. *Probabilistic Cfg with Latent Annotations*, in "Proc. of ACL-05, Ann Arbor, USA", 2005, p. 75-82.

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

[96] R. T. MCDONALD, K. CRAMMER, F. C. N. PEREIRA. *Online Large-Margin Training of Dependency Parsers*, in "Proc. of ACL'05, Ann Arbor, USA", 2005.

[97] R. T. MCDONALD, F. C. N. PEREIRA. *Online Learning of Approximate Dependency Parsing Algorithms*, in "Proc. of EACL'06", 2006.

[98] T. MORTON. *Coreference for NLP applications*, in "Proceedings of ACL 2000, Hong Kong", 2000.

[99] V. NG, C. CARDIE. *Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution*, in "Proceedings of COLING 2002", 2002.

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

[101] 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.

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

[103] 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.

[104] S. PETROV, L. BARRETT, R. THIBAUX, D. KLEIN. *Learning Accurate, Compact, and Interpretable Tree Annotation*, in "Proc. of ACL-06, Sydney, Australia", 2006.

[105] P. RESNIK, D. YAROWSKY. *A perspective on word sense disambiguation methods and their evaluation*, in "ACL SIGLEX Workshop Tagging Text with Lexical Semantics: Why, What, and How?, Washington, D.C., USA", 1997.

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

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

[108] B. SAGOT, L. CLÉMENT, É. VILLEMONTE DE LA CLERGERIE, P. BOULLIER. *The Lefff 2 syntactic lexicon for French: architecture, acquisition, use*, in "Proc. of LREC'06", 2006, http://hal.archives-ouvertes.fr/docs/00/41/30/71/PDF/LREC06b.pdf.

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

[110] B. SAGOT. *Automatic acquisition of a Slovak lexicon from a raw corpus*, in "Lecture Notes in Artificial Intelligence 3658 (© Springer-Verlag), Proceedings of TSD'05, Karlovy Vary, Czech Republic", September 2005, p. 156–163.

[111] 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.

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

[113] 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.

[114] F. THOMASSET, É. VILLEMONTE DE LA CLERGERIE. *Comment obtenir plus des Méta-Grammaires*, in "Proceedings of TALN'05, Dourdan, France", ATALA, June 2005.

[115] M. VILAIN, J. BURGER, J. ABERDEEN, D. CONNOLLY, L. HIRSCHMAN. *A model-theoretic coreference scoring scheme*, in "Proceedings fo the 6th Message Understanding Conference (MUC-6), San Mateo, CA", Morgan Kaufmann, 1995, p. 45–52.

[116] 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 (ELRA) (editor), may 2008.

[117] 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.

[118] É. VILLEMONTE DE LA CLERGERIE. *From Metagrammars to Factorized TAG/TIG Parsers*, in "Proceedings of IWPT'05, Vancouver, Canada", October 2005, p. 190–191.

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

[120] 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.

[121] G. VAN NOORD. *Error Mining for Wide-Coverage Grammar Engineering*, in "Proc. of ACL 2004, Barcelona, Spain", 2004.