IN PARTNERSHIP WITH:
Université Denis Diderot (Paris 7)

Activity Report 2012

Project-Team ALPAGE

Large-scale deep linguistic processing

IN COLLABORATION WITH: Analyse Linguistique Profonde A Grande Echelle (ALPAGE)

RESEARCH CENTER
Paris - Rocquencourt

THEME
Audio, Speech, and Language Processing

Table of contents

Project-Team ALPAGE

Keywords: Natural Language, Linguistics, Semantics, Knowledge Acquisition, Knowledge

This project is a common project with University Paris Diderot–Paris 7. The team has been created on July 1, 2007 and became an UMR-I on January 1, 2009 (UMR-I 001).

Creation of the Project-Team: January 01, 2008 .

1. Members

Research Scientists

Pierre Boullier [Emeritus Senior Researcher (DR-E) Inria, HdR]
Éric Villemonte de La Clergerie [Junior Researcher (CR) Inria]
Benoît Sagot [Junior Researcher (CR) Inria]

Faculty Members

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

Engineers

Mickael Morardo [Inria DTI-funded Engineer in collaboration with Lingua et Machina]
Géraldine Walther [Research Engineer funded by the ANR project EDyLex]
Virginie Mouilleron [Research Engineer funded by the ANR project EDyLex (since July 2012)]
Damien Nouvel [Research Engineer funded by the ANR project EDyLex (since October 2012)]
Marion Richard [Master 1 student, annotator funded by the ANR project EDyLex (July 2012)]
Sarah Beniamine [Master 1 student, annotator funded by the ANR project EDyLex (June 2012)]
Paul Bui Quang [Inria ADT-junior Engineer (since October 2012)]

PhD Students

Luc Boruta [PhD student (allocataire) (since October 2009)]
Chloé Braud [PhD student (allocataire) (since September 2011)]
François-Régis Chaumartin [PhD student Univ. Paris 7 (until Sept. 2012)]
Valérie Hanoka [PhD student (CIFRE) at Verbatim Analysis & Univ. Paris 7]
Enrique Henestroza Anguiano [PhD funded by the ANR project SEQUOIA]
Emmanuel Lassalle [PhD student (ENS stipendium) Univ. Paris 7]
Pierre Magistry [PhD student (allocataire) Univ. Paris 7]
Charlotte Roze [PhD student (allocataire) Univ. Paris 7]
Rosa Stern [PhD student (CIFRE) AFP & Univ. Paris 7 (November 2009–October 2012), then Research Engineer funded by the ANR project EDyLex (since October 2012)]
Juliette Thuilier [PhD student (until Sept. 2012)]
Marion Baranes [PhD student, hired at *viavoo* (since February 2012)]
Corentin Ribeyre [PhD student (allocataire) Univ. Paris 7 (since November 2012)]
Marianne Djemaa [PhD student (ANR ASFALDA) (since October 2012)]
Isabelle Dautriche [PhD Student (since September 2012)]

Post-Doctoral Fellows

Margaret Grant [postdoc Labex EFL (since October 2012)]
Yves Scherrer [Post-doc funded by the LabEx EFL, Strand 6, operation LR 2.2 (since October 2012)]

Administrative Assistant

Assia Saadi [Secretary (SAR) Inria]

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". This includes domains such as machine translation, question answering, information retrieval, information extraction, text simplification, automatic or computer-aided translation, automatic summarization, foreign language reading and writing aid. From a more research-oriented point of view, experimental linguistics can be also viewed as an "application" of NLP.

NLP, the domain of Alpage, is a multidisciplinary domain which studies the problems of automated understanding and generation of natural human languages. 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, computer science and algorithmics for Inria members).

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 generation (by opposition to *speech* processing and generation).

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, Persian and others). This is of course of high relevance, among others, for language-independant modeling and multi-lingual tools and applications.

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

- Research topics:
  - deep syntactic modeling and parsing. This topic includes, but is not limited to, development of advanced parsing technologies, development of large-coverage and high-quality adaptive linguistic resources, and use of hybrid architectures coupling shallow parsing, (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:
  - experimental linguistics;
  - automatic information extraction (both linguistic information, inside a bootstrapping scheme for linguistic resources, and document content, with a more industry-oriented perspective);
  - text normalization, automatic and semi-automatic spelling correction;
  - text mining;
  - automatic generation;
  - with a more long-term perspective, automatic or computer-aided translation.

2.2. Highlights of the Year

- A statistical parsing architecture for Italian using MELT in a pre-processing step has obtained the best results in the EVALITA shared task on Italian parsing [35] (cf. 5.7).
- Two different instances of Alpage parsing architectures were ranked 2nd and 3rd at the SANCL shared task on parsing user-generated content, organized by Google [38] (cf. 5.7 and 6.4).
- Release of two freely available out-of-domain treebanks for French: the SequoiaBank focusing on well-edited texts such as Wikipedia, Europarl, ...; the French Social Media Bank, focusing on noisy user-generated content (Facebook, Twitter, ...), the latter being the first available dataset for Facebook in any language – cf 6.4.

3. Scientific Foundations

3.1. From programming languages to linguistic grammars

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

Historically, several members of Alpage were originally specialists in the domain of modeling and parsing for programming languages, and are working for more than 15 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 [58], [96], [103]) 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.

An efficient way to develop large-coverage hand-crafted symbolic grammars is to use adequate tools and adequate levels of representation, and in particular Meta-Grammars, one of Alpage's areas of expertise [121], [117]. 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.

3.2. Statistical Parsing

Contrary to symbolic approaches to parsing, in statistical parsing, the grammar is extracted from a corpus of syntactic trees : a treebank. The main advantage of the statistical approach is to encode within the same framework the parsing and disambiguating tasks. The extracted grammar rules are associated with probabilities that allow to score and rank the output parse trees of an input sentence. This obvious advantage of probabilistic context-free grammars has long been counterbalanced by two main shortcomings that resulted in poor performance for plain PCFG parsers: (i) the generalization encoded in non terminal symbols that stand for syntagmatic phrases is too coarse (so probabilistic independence between rules is too strong an assertion) and (ii) lexical items are underused. In the last decade though, effective solutions to these shortcomings have been proposed. Symbol annotation, either manual [84] or automatic [91], [92] captures inter-dependence between CFG rules. Lexical information is integrated in frameworks such as head-driven models that allow lexical heads to percolate up the syntagmatic tree [72], or probabilistic models derived from lexicalized Tree Adjoining grammars, such as Stochastic Tree Insertion Grammars [70].

In the same period, totally different parsing architectures have been proposed, to obtain dependency-based syntactic representations. The properties of dependency structures, in which each word is related to exactly one other word, make it possible to define dependency parsing as a sequence of simple actions (such as read buffer and store word on top of a stack, attach read word as dependent of stack top word, attach read word as governor of stack top word ...) [129], [89]. Classifiers can be trained to choose the best action to perform given a partial parsing configuration. In another approach, dependency parsing is cast into the problem of finding the maximum spanning tree within the graph of all possible word-to-word dependencies, and online classification is used to weight the edges [85]. These two kinds of statistical dependency parsing allow to benefit from discriminative learning, and its ability to easily integrate various kinds of features, which is typically needed in a complex task such as parsing.

Statistical parsing is now effective, both for syntagmatic representations and dependency-based syntactic representations. Alpage has obtained state-of-the-art parsing results for French, by adapting various parser learners for French, and works on the current challenges in statistical parsing, namely (1) robustness and portability across domains and (2) the ability to incorporate exogenous data to improve parsing attachment decisions. We review below the approaches that Alpage has tested and adapted, and the techniques that we plan to investigate to answer these challenges.

In order to investigate statistical parsers for French, we have first worked how to use the French Treebank [55] and derive the best input for syntagmatic statistical parsing [74]. Benchmarking several PCFG-based learning frameworks [11] has led to state-of-the-art results for French, the best performance being obtained with the split-merge Berkeley parser (PCFG with latent annotations) [92].
In parallel to the work on dependency based representation, presented in the next paragraph, we also conducted a preliminary set of experiments on richer parsing models based on Stochastic Tree Insertion Grammars as used in [70] and which, besides their inferior performance compared to PCFG-LA based parser, raise promising results with respect to dependencies that can be extracted from derivation trees. One variation we explored, that uses a specific TIG grammar instance, a *vertical* grammar called *spinal* grammars, exhibits interesting properties wrt the grammar size typically extracted from treebanks (a few hundred unlexicalized trees, compared to 14 000 CFG rules). These models are currently being investigated in our team [113].
Pursuing our work on PCFG-LA based parsing, we investigated the automatic conversion of the treebank into dependency syntax representations [65], that are easier to use for various NLP applications such as question-answering or information extraction, and that are a better ground for further semantic analysis. This conversion can be applied on the treebank, before training a dependency-based parser, or on PCFG-LA parsed trees. This gives the possibility to evaluate and compare on the same gold data, both syntagmatic- and dependency-based statistical parsing. This also paved the way for studies on the influence of various types of lexical information. Results are sketched in section 6.4.

3.3. Dynamic wide coverage lexical resources

Participants: Benoît Sagot, Laurence Danlos, Rosa Stern, 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 [102]. At the syntactic level, work has been achieved on automatic acquisition of atomic syntactic information and automatic detection of errors in the lexicon [130],[10]. At the semantic level, automatic wordnet development tools have been described [95], [123], [82], [80]. 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 to develop one of the main syntactic resources for French, the Le*fff* [98],[8], developed within the Alexina framework, as well as a wordnet for French, the WOLF [7], the first freely available resource of the kind.

In the last few years, Alpage members have shown how to benefit from other more linguistically-oriented resources, such as the Lexique-Grammaire and DICOVALENCE , in order to improve the coverage and quality of the Le*fff* 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.

Moreover, an increasing effort has been made towards multilingual aspects. In particular, Alexina lexicons developed in 2010 or before exist for Slovak [102], Polish [104], English, Spanish [87], [86] and Persian [108], not including freely-available lexicons adapted to the Alexina framework.

3.4. Shallow processing

Participants: Éric Villemonte de La Clergerie, Benoît Sagot, Rosa Stern.

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 was the leader of the PASSAGE ANR project (ended in June 2010), 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 (symbolic or probabilistic), 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 [6]. Moreover, when used as a preliminary step before parsers, the quality of parsers' results strongly depends on the quality of such chains. In that regard, less-standard pre-processings such as word clustering have led to promising results [112].

In fact, such processing chains are mostly used as such, and not only as pre-processing tools before parsing. They aim at performing the basic tasks that produce immediately usable results for many applications, such as tokenization, sentence segmentation, spelling correction, and, most importantly, named entity detection, disambiguation and resolution.

3.5. Discourse structures

Participants: Laurence Danlos, Charlotte Roze.

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 anaphora, 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 [76].

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, [77],[5]), 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.

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.

Experimental linguistics : tools to explore language in an objective way (this is related, but not limited to corpus linguistics).

Alpage focuses on some applications included in the three last points, such as information extraction and (linguistic and extra-linguistic) knowledge acquisition (4.2), text mining (4.3), spelling correction (4.5) and experimental linguistics (4.6).

4.2. Information extraction and knowledge acquisition

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

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). Obviously, the progressive extension of Alpage parsing systems or even shallow processing chains to the semantic level 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, news or 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.

In the last years, such efforts have been targeted towards information extraction from news wires in collaboration with the Agence France-Presse (Rosa Stern is a CIFRE PhD student at Alpage and at AFP, and works in relation with the ANR project EDyLex) as well as in the context of the collaboration between Alpage and Proxem, a startup created by François-Régis Chaumartin, PhD student at Alpage (who has defended his PhD in 2012).

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

Participants: Benoît Sagot, Valérie Hanoka.

Verbatim Analysis is a startup co-created by Benoît Sagot from Alpage and Dimitri Tcherniak from Towers Watson, a world-wide leader in the domain of employee research (opinion mining among the employees of a company or organization). The aim of its first product, *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. Several other parts of *vera* are co-owned by Verbatim Analysis and by Inria.

4.4. Multilingual terminologies and lexical resources for companies

Participants: Éric Villemonte de La Clergerie, Mickael Morardo, Benoît Sagot.

Lingua et Machina is a small company now headed by François Brown de Colstoun, a former Inria researcher, that provides services for developing specialized multilingual terminologies for its clients. It develops the WEB framework Libellex for validating such terminologies. A formal collaboration with ALPAGE has been set up, with the recruitment of Mikael Morardo as engineer, funded by Inria's DTI. He works on the extension of the web platform *Libellex* for the visualization and validation of new types of lexical resources. In particular, he has integrated a new interface for handling monolingual terminologies, lexical networks, and bilingual wordnet-like structures.

4.5. Automatic and semi-automatic spelling correction in an industrial setting

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

NLP tools and resources used for spelling correction, such as large n-gram collections, POS taggers and finite-state machinery are now mature and precise. In industrial setting such as post-processing after large-scale OCR, these tools and resources should enable spelling correction tools to work on a much larger scale and with a much better precision than what can be found in different contexts with different constraints (e.g., in text editors). Moreover, such industrial contexts allow for a non-costly manual intervention, in case one is able to identify the most uncertain corrections. An FUI project on this topic has been proposed in collaboration with Diadeis, a company specialized in text digitaliNzation, and two other partners. It has been rerouted to the "Investissements d'avenir" framework, and has been accepted. It started in 2012.

4.6. Experimental linguistics

Participants: Benoît Crabbé, Juliette Thuilier, Luc Boruta.

Alpage is a team that dedicates efforts in producing ressources and algorithms for processing large amounts of textual materials. These ressources can be applied not only for purely NLP purposes but also for linguistic purposes. Indeed, the specific needs of NLP applications led to the development of electronic linguistic resources (in particular lexica, annotated corpora, and treebanks) that are sufficiently large for carrying statistical analysis on linguistic issues. In the last 10 years, pioneering work has started to use these new data sources to the study of English grammar, leading to important new results in such areas as the study of syntactic preferences [60], [128], the existence of graded grammaticality judgments [83].

The reasons for getting interested for statistical modelling of language can be traced back by looking at the recent history of grammatical works in linguistics. In the 1980s and 1990s, theoretical grammarians have been mostly concerned with improving the conceptual underpinnings of their respective subfields, in particular through the construction and refinement of formal models. In syntax, the relative consensus on a generative-transformational approach [71] gave way on the one hand to more abstract characterizations of the language faculty [71], and on the other hand to the construction of detailed, formally explicit, and often implemented, alternative formulation of the generative approach [59], [94]. For French several grammars have been implemented in this trend, among which the tree adjoining grammars of [63], [73] among others. This general movement led to much improved descriptions and understanding of the conceptual underpinnings of both linguistic competence and language use. It was in large part catalyzed by a convergence of interests of logical, linguistic and computational approaches to grammatical phenomena.

However, starting in the 1990s, a growing portion of the community started being frustrated by the paucity and unreliability of the empirical evidence underlying their research. In syntax, data was generally collected impressionistically, either as ad-hoc small samples of language use, or as ill-understood and little-controlled grammaticality judgements (Schütze 1995). This shift towards quantitative methods is also a shift towards new scientific questions and new scientific fields. Using richly annotated data and statistical modelling, we address questions that could not be addressed by previous methodology in linguistics. In this line, at Alpage we have started investigating the question of choice in French syntax with a statistical modelling methodology. Currently two studies are being led on the position of attributive adjectives w.r.t. the noun and the relative position of postverbal complement. This research has contributed to establish new links with the Laboratoire de Linguistique Formelle (LLF, Paris 7) and the Laboratoire de Psychologie et Neuropsychologie Cognitives (LPNCog, Paris 5).

On the other hand we have also started a collaboration with the Laboratoire de Sciences Cognitives de Paris (LSCP/ENS) where we explore the design of algorithms towards the statistical modelling of language acquisition (phonological acquisition). This is currently supported by one PhD project.

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.

Direct NLP users of SYNTAX for NLP, outside Alpage, include Alexis Nasr (Marseilles) and other members of the (now closed) SEQUOIA ANR project, Owen Rambow and co-workers at Columbia University (New York), as well as (indirectly) all SXPipe and/or SXLFG users. The project-team VASY (Inria Rhône-Alpes) is one of SYNTAX' user for non-NLP applications.

5.2. System DyALog

Participant: Éric Villemonte de La Clergerie [maintainer].

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 and architectures and on Mac OS intel (both 32 and 64bits architectures).

The current release handles logic programs, DCGs (*Definite Clause Grammars*), FTAGs (*Feature Tree Adjoining Grammars*), FTIGs (*Feature Tree Insertion Grammars*) and XRCGs (*Range Concatenation Grammars* with logic arguments). Several extensions have been added to most of these formalisms such as intersection, Kleene star, and interleave operators. Typed Feature Structures (TFS) as well as finite domains may be used for writing more compact and declarative grammars [120].

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. 5.3). It has also been used for building a parser from a large coverage French TIG/TAG grammar derived from a Meta-Grammar. This parser has been used for the Parsing Evaluation campaign EASy, the two Passage campaigns (Dec. 2007 and Nov. 2009), cf. [117], [118], and very large amount of data (700 millions of words) in the SCRIBO project.

DYALOG and other companion modules are available on Inria GForge.

5.3. Tools and resources for Meta-Grammars

Participant: Éric Villemonte de La Clergerie [maintainer].

`mgcomp`, MGTOOLS, and FRMG on Inria GForge: http://mgkit.gforge.inria.fr/

DYALOG (cf. 5.2) has been used to implement `mgcomp`, Meta-Grammar compiler. Starting from an XML representation of a MG, `mgcomp` produces an XML representation of its TAG expansion.

The current version **1.5.0** 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 provides 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 2.0.1) to get a grammar of around 200 TAG trees [122]. 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, mostly by Éric de La Clergerie, and collected in MgTools (version **2.2.2**). 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 are available on Inria GForge. FRMG is used directly or indirectly (through a Web service or by requiring parsed corpora) by several people and actions (ANR Rhapsodie, ANR Chronoline, ...)

5.4. The Bonsai PCFG-LA parser

Participants: Marie Candito [correspondant], Djamé Seddah.

Web page:
http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

Alpage has developed as support of the research papers [74], [65], [66], [11] a statistical parser for French, named Bonsai, trained on the French Treebank. This parser provides both a phrase structure and a projective dependency structure specified in [4] 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. The parser is available under a GPL license.

5.5. The MICA parser

Participants: Benoît Sagot [correspondant], Marie Candito, Pierre Boullier, Djamé Seddah.

Web page:
http://mica.lif.univ-mrs.fr/

MICA (Marseille-Inria-Columbia- AT&T) is a freely available dependency parser [57] currently trained on English and Arabic data, developed in collaboration with Owen Rambow and Daniel Bauer (Columbia University) and Srinivas Bangalore (AT&T). MICA has several key characteristics that make it appealing to researchers in NLP who need an off-the-shelf parser, based on Probabilistic Tree Insertion Grammars and on the Syntax system. MICA is fast (450 words per second plus 6 seconds initialization on a standard high-end machine) and has close to state-of-the-art performance (87.6% unlabeled dependency accuracy on the Penn Treebank).

MICA consists of two processes: the supertagger, which associates tags representing rich syntactic information with the input word sequence, and the actual parser, based on the Inria SYNTAX system, which derives the syntactic structure from the n-best chosen supertags. Only the supertagger uses lexical information, the parser only sees the supertag hypotheses.

MICA returns n-best parses for arbitrary n; parse trees are associated with probabilities. A packed forest can also be returned.

5.6. Alpage's linguistic workbench, including SxPipe

Participants: Benoît Sagot [correspondant], Rosa Stern, Marion Baranes, Damien Nouvel, Virginie Mouilleron, 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 [97] 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 (e.g., FRMG);
- for surface processing (named entities recognition, text normalization...).

Developed for French and for other languages, SxPipe 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, quotations...). In 2012, SxPipe has received a renewed attention in four directions:

- Support of new languages, and most notably German (although this is still at a very preliminary stage of development;
- Analysis of unknown words, in particular in the context of the ANR project EDyLex and of the collaboration with *viavoo*; this involves in particular (ii) new tools for the automatic pre-classification of unknown words (acronyms, loan words...) (ii) new morphological analysis tools, most notably automatic tools for constructional morphology (both derivational and compositional), following the results of dedicated corpus-based studies;
- Development of new local grammars for detecting new types of entities, such as chemical formulae or dimensions, in the context of the PACTE project.

5.7. MElt

Participants: Benoît Sagot [correspondant], Pascal Denis.

MElt is a part-of-speech tagger, initially trained for French (on the French TreeBank and coupled with the Le*fff*), English [78], Spanish, Kurmanji Kurdish [125] and Persian [106], [107]. It is state-of-the-art for French. It is distributed freely as a part of the Alpage linguistic workbench.

In 2012, MElt has underwent two major upgrades:

- It has been successfully trained and used on Italian [35], Spanish [26] and German data. In particular, a statistical parsing architecture for Italian that used MElt in a pre-processing step has obtained the best results in the EVALITA shared task on Italian parsing [35].
- MElt can now be called within a wrapper developed for handling noisy textual data such as user-generated content produced on Web 2.0 platforms (forums, blogs, social media); more precisely, this wrapper is able to "clean" such data, then tag it using MElt, and finally transfer MElt annotations from the "cleaned" data, which could be annotated more easily, to the original noisy data. This architecture has proved useful on French for creating the French Social Media Bank [37], [36]. On English, it has played an important role within both variants of the Alpage parsing architecture that were ranked 2nd and 3rd at the SANCL shared task on parsing user-generated content, organized by Google [38].

5.8. The Alexina framework: the Lefff syntactic lexicon, the Aleda entity database and other Alexina resources

Participants: Benoît Sagot [correspondant], Laurence Danlos.

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

Alexina is Alpage's Alexina framework for the acquisition and modeling of morphological and syntactic lexical information. The first and most advanced lexical resource developed in this framework is the Le*fff*, a morphological and syntactic lexicon for French.

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.

Other Alexina lexicons exist, at various stages of development, in particular for Spanish (the Le*ff*e), Polish, Slovak, English, Galician, Persian, Kurdish, Italian and since this year for German, as well as for Latin and Maltese verbs. These lexicons are used in various tools, including instances of the MElt POS-tagger.

Alexina also hosts *Aleda* [114], [33] a large-scale entity database currently developed for French but under development for English, Spanish and German, extracted automatically from Wikipedia and Geonames. It is used among others in the SXPipe processing chain and its NP named entity recognition, as well as in the NOMOS named entity linking system.

5.9. The free French wordnet WOLF

Participants: Benoît Sagot [correspondant], Marion Richard, Sarah Beniamine.

The WOLF (Wordnet Libre du Français) is a wordnet for French, i.e., a lexical semantic database. The development of WOLF started in 2008 [99], [100]. At this time, we focused on benefiting from available resources of three different types: general and domain-specific bilingual dictionaries, multilingual parallel corpora and Wiki resources (Wikipedia and Wiktionaries). This work was achieved in a large part in collaboration with Darja Fišer (University of Ljubljana, Slovenia), in parallel with the development of a free Slovene wordnet, sloWNet. However, it was also impacted by specific collaborations, e.g., on adverbial synsets [101].

2012 results concerning the WOLF are described in the corresponding section.

The WOLF is freely available under the Cecill-C license. It has already been used in various experiments, within and outside Alpage.

5.10. Automatic construction of distributional thesauri

Participant: Enrique Henestroza Anguiano [correspondant].

FREDISTis a freely-available (LGPL license) Python package that implements methods for the automatic construction of distributional thesauri.

We have implemented the context relation approach to distributional similarity, with various context relation types and different options for weight and measure functions to calculate distributional similarity between words. Additionally, FREDISTis highly flexible, with parameters including: context relation type(s), weight function, measure function, term frequency thresholds, part-of-speech restrictions, filtering of numerical terms, etc.

Distributional thesauri for French are also available, one each for adjectives, adverbs, common nouns, and verbs. They have been constructed with FreDist and use the best settings obtained in an evaluation. We use the *L'Est Republicain* corpus (125 million words), *Agence France-Presse* newswire dispatches (125 million words) and a full dump of the French Wikipedia (200 million words), for a total of 450 million words of text.

5.11. Tools and resources for time processing

Participant: Laurence Danlos [correspondant].

Alpage developed the *French TimeBank*, a freely-available corpus annotated with ISO-TimeML-compliant temporal information (dates, events and relations between events).

5.12. System EasyRef

Participants: Éric Villemonte de La Clergerie [maintainer], Corentin Ribeyre.

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 and revised by Éric de La Clergerie, 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. Advances in symbolic and hybrid parsing with DyALog and FRMG

Participants: Éric Villemonte de La Clergerie, François Barthélemy, Julien Martin.

Within the team is developed a wide-coverage French meta-grammar (FRMG) and a efficient hybrid TAG/TIG parser based on the DYALOG logic programming environment [120] and on the Le*fff* morphological and syntactic lexicon [105]. It relies on the notion of factorized grammar, themselves generated from a representation that lies at a higher level of abstraction, named Meta-Grammars [122]. At that level, linguistic generalizations can be expressed, which in turn makes it possible to transfer meta-grammars from one language to a closely related one. The hybrid TAG/TIG parser generator itself implements all kinds of parsing optimizations: lexicalization (in particular via hypertags), left-corner guiding, top/bottom feature analysis, TIG analysis (with multiple adjoining), and others. The recent evolutions go towards an hybridization with statistical approaches.

6.1.1. Tuning FRMG's disambiguation mechanism

Continuing works initiated in 2011 on the exploitation of the dependency version of the French TreeBank (FTB), Éric de La Clergerie has explored the tuning of FRMG's rule base disambiguation mechanism using a larger set of features and weight learned from the FTB. In 2011, this approach led to on improvement from 82.31% to 84.54% in terms of accuracy (LAS - Labelled Attachment Score) on the test part of the FTB. By increasing the set of features, in particularly using higher-order dependency features (on parent edge and sibling edges), and a better understanding of the iterative tuning mechanism, it was possible to reach 85.95% LAS. This tuning mechanism is based on the idea of adding or subtracting some weight to a disambiguation rule given some specific contexts (provided by the features), where the delta is progressively learned from the accuracy of the disambiguation rule in terms of edge selection or rejection. The learning algorithm presents some relationships with the perceptron approach, but the use of a more standard implementation of the perceptron led to less interesting gains.

During the same time, the coverage of FRMG was improved (to reach for instance 94% of full parses on the FTB).

6.1.2. Synchronous Tree-Adjoining Grammars

A preliminary work has been done to implement *Synchronous Tree-Adjoining Grammars* (STAGs) in DyALog, relying on the notion of *Thread Automata* [119]. Synchronous Tree Adjoining Grammars is an instance of formalism where the order of the components of a tree structure is not fully determined. This leads to combinatorial alternatives when parsing, while a tree-structure corresponding to the input string has to be build. A specific front-end has been written to implement STAGs. The work on the back-end is still in progress, with the goal to have a common intermediate representation for several mildly context-sensitive formalisms where some node operations non-deterministically pick a node out of a finite set of nodes. STAGs are an instance of such formalisms, Multi-Component Tree Adjoining Grammars (MCTAGs) are another instance. The intermediate representation consists in Thread Automata (TA), an extension of Push-Down Automata where several threads of computations are considered and only one is active at any time.

6.1.3. Adding weights and probabilities to DyALog

Weights can already be used during the disambiguation phase of the FRMG parser, implemented in DyALog. However, a deeper implementation of weights and probabilities in DyALog was initiated in 2012 by Julien Martin during his Master internship. By enriching the structure of the backpointers (relating the items to their parent items), it is now possible to maintain an ordered weighted list of derivations, to update the scheduling of items wrt their weight, to update the weights of all the descendants of an item I when updating I's weight. The motivation is of course to be able to favor the best analysis first during parsing. A second objective (which has been implemented) is the possibility to extract the n-best parses after parsing (but keeping a shared derivation forest). A third objective, remaining to be done, is related to the use of beam search techniques to prune the search space during parsing. A longer-term objective is the abstraction of this work to be able to work on semi-rings.

6.2. Tree transformation

Participants: Éric Villemonte de La Clergerie, Corentin Ribeyre, Djamé Seddah.

In 2011, the conversion of native FRMG dependencies into the CONLL dependency scheme was the occasion to explore new ideas about tree transformation (for dependencies), based on the notion of two-level transformation with a first level relying on local transformation rules and a second level being controlled by constraints carried by the first level edges. During his Master internship, Corentin Ribeyre has formalized and re-implemented this approach in a more systematic and generic way. This work was also completed by the use of example-based learning techniques to quickly learn the local transformation rules of the first level. The line of research is motivated by possibility to quickly develop a reduced set of transformation rules (thanks to the examples and the constraint level) for a large variety of applications, such as information extration but also conversion toward a deep syntax level or a shallow semantic level. A poster paper was presented at TAG+11 [29].

6.3. lexical knowledge acquisition and visualization

Participants: Éric Villemonte de La Clergerie, Mickael Morardo, Benoît Sagot.

In relation with our collaboration with Lingua & Machina (cf section 4.4), Mikael Morardo has enriched the interfaces of the WEB platform Libellex for the visualization and validation of more complex lexical resource. In particular, the focus has been on the development of a graph-based view with the javascript Library `d3.js` to represent large lexical networks. The current implementation is powerful enough to deal with large networks of several teens of thousands of connections, allowing the visualization of fragments of the network and an easy navigation. Because the graph-view proved to be both intuitive and efficient, the previous list-based view for terminology was partially re-implemented in the new graph-view. It was also extended for visualizing and validating more complex lexical networks, like the French Wordnet WOLF coupled with the original English WordNet (cf 5.9).

The graph-based view was used to explore several networks built using Harris' distributional hypothesis (through a clustering algorithm) on the output of FRMG for several corpora. Because terminology was now be visualized at the same time, the clustering algorithm was modified to be able to take into account a list of terms (also automatically extracted from the parsed corpora) .

6.4. Advances in statistical parsing

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

6.4.1. Statistical Parsing

We have achieved **state-of-the art results for French statistical parsing**, adapting existing techniques for French, a language with a morphology richer than English, either for constituency parsing [110], [113] or dependency parsing [68]. We made available The Bonsai parsing chain [1] (cf. 5.4), that gathers preprocessing tools and models for French dependency parsing into an easy-to-use parsing tool for French. We designed our parsing pipeline with modularity in mind: our parsing models are interchangeable. For instance, dependencies output can either be generated from a PCFG-LA based parser associated with a functional role labeler or from any dependency parsers trained on our dependency treebank [68]. Tokens can either be raw words, POS tagged lemmas or word clusters [69].

We have innovated in the tuning of tagsets to optimize both grammar induction and unknown word handling [75], thus providing the best parsing models for French [111]. Then we have contributed on three main points:

1. conversion of the French Treebank [55] used as constituency training data into a dependency treebank [4], which is now used by several teams for dependency parsing;
2. an original method to reduce lexical data sparseness by replacing tokens by unsupervised word clusters, or morphological clusters [64], [112];
3. a postprocessing step that uses specialized statistical models for parse correction [81].

For the last 18 to 12 months, we have been increasingly focused in increasing the robustness of our parsing models by (a) validating our approach on other morphologically-rich languages; (b) other domains and (c) on user generated content. All of those challenging the current state-of-the-art in statistical parsing.

6.4.2. Multilingual parsing

Applying the techniques we developed for reducing lexical data, which is commonly found in morphologically-rich languages (MRLs) and optimizing the POS tagset, we integrated lexical information through data driven lemmatisation [112] and POS tagging [79]. This provided state-of-the-art results in parsing Romance languages such as Italian [35] and Spanish [26]. In the latter case, we mixed the outputs of two morphological analyzers and generated a version of the treebank where each morphological gold information was replaced by a predicted one. Relying on a rich lexicon developed within the Alexina framework (cf. 5.8) and accurate morphological treatment (cf. 6.5), this method brings more robustness to treebank-based parsing models.

6.4.3. Out-of-domain parsing : resources and parsing techniques

Statistical parsing is known to lead to parsers that exhibit quite degraded performance on input text that varies from the sentences used for training. Alpage has devoted a major effort on providing both evaluation resources and parser adaptation techniques, to increase robustness of statistical parsing for French. We have investigated several degrees of distance between the training corpus, the French Treebank, which is made of sentences from the *Le Monde* newspaper: we first focused on parsing well-edited texts, but from domains with varying difference with respect to the national newspaper *Le Monde* type of text. We then turned our attention to parsing user-generated content, hence potentially not only from a different domain than news, but also with great "noise" with respect to well-edited texts, and extremely divergent linguistic phenomena (see next subsection). As far as out-of-domain well-edited text, we have supervised the annotation and release

[1] http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

of the **Sequoia Treebank** [47] (https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=CorpusSequoia), a corpus of 3200 sentences annotated for part-of-speech and syntactic structure, from four subdomains : sentences from the regional newspaper *L'Est Républicain*, from the French Wikipedia, from the Europarl Corpus (European parliamentary debates), and from reports of the European Medicine Agency. We have proposed a word clustering technique, with clusters computed over a "*bridge*" corpus that couples indomain and target domain raw texts, to improve parsing performance on target domain, without degrading performance on indomain texts (contrary to usual adaptation techniques such as self-training). Preliminary experiments were performed on the biomedical domain only [67] and confirmed on the whole Sequoia Treebank [47].

6.4.4. Robust parsing of user-generated content

Until very recently out-of-domain text genres that have been prioritized have not been Web 2.0 sources, but rather biomedical texts, child language and general fiction (Brown corpus). Adaptation to user-generated content is a particularly difficult instance of the domain adaptation problem since Web 2.0 is not really a domain: it consists of utterances that are often ungrammatical. It even shares some similarities with spoken language [116]. The poor overall quality of texts found on such media lead to weak parsing and even POS-tagging results. This is because user-generated content exhibits both the same issues as other out-of-domain data, but also tremendous issues related to tokenization, typographic and spelling issues that go far beyond what statistical tools can learn from standard corpora. Even lexical specificities are often more challenging than on edited out-of-domain text, as neologisms built using productive morphological derivation, for example, are less frequent, contrarily to slang, abbreviations or technical jargon that are harder to analyze and interpret automatically.

In order to fully prepare a shift toward more robustness, we started to develop a richly annotated corpus of user-generated French text, the French Social Media Bank, which includes not only POS, constituency and functional information, but also a layer of "normalized" text[37]. This corpus is fully available and constitutes the first data set on Facebook data and the first instance of user generated content for an MRL.

Besides delivering a new data set, our main purpose here is to be able to compare two different approaches to user-generated content processing: either training statistical models on the original annotated text, and use them on raw new text; or developing normalization tools that help improving the consistency of the annotations, train statistical models on the normalized annotated text, and use them on normalized texts (before un-normalizing them).

However, this raises issues concerning the normalization step. A good sandbox for working on this challenging task is that of POS-tagging. For this purpose, we did leverage Alpage's work on MElt, a state-of-the art POS tagging system [15]. A first round of experiments on English have already led to promising results during the shared task on parsing user-generated content organized by Google in May 2012 [93], as Alpage was ranked second and third [38]. For achieving this result, we brought together a preliminary implementation of a normalization wrapper around the MElt POS tagger followed by a state-of-the art statistical parser improved by several domain adaptation techniques originally developed for parsing edited out-of-domain texts (cf. previous section).

One of our objectives is to generalize the use of the normalization wrapper approach to both POS tagging and parsing, for English and French, in order to improve the quality of the output parses. However, this raises several challenges: non-standard contractions and compounds lead to unexpected syntactic structures. A first round of experiments on the French Social Media Bank showed that parsing performance on such data are much lower than expected. This is why, we are actively working to improve on the baselines we established on that matter.

6.4.5. Precise recovery of unbounded dependencies

We focused on a linguistic phenomena known as long-distance dependencies. These are dependencies involved a fronted element that depends on a head that is potentially embedded in the clause the element is in front of. This embedding make such dependencies very hard to recover for a parser. Though this phenomena is rare, the corresponding dependencies are generally part of predicate-argument structures, and are thus very

important to recover for downstream semantic applications. We have assessed the low parsing performance of long-distance dependencies (LDDs) for French, proposed an explicit annotation of such dependencies in the French Treebank and the Sequoia Treebank, and evaluated several parsing architectures with the aim of maintaining high general performance and good performance on LDDs [22]. We found that using a non-projective parser helps for LDDs but degrades overall performance, while using pseudo-projective parsing [88] which transforms in a reversible way a non-projective treebank into a projective one) is the best strategy, in order to take advantage of the better performance of projective parsers.

6.5. Computational morphology and automatic morphological analysis

Participants: Benoît Sagot [correspondant], Marion Baranes, Virginie Mouilleron, Damien Nouvel.

Since 2011 and, Alpage members have started interacting with formal morphologists for taking part in the development and implementation of new morphological models and resources. Concerning inflectional morphology, this work has led to new versions of the morphological layer of the ALEXINA formalism, to new ALEXINA lexicons for several languages of choice (Kurdish languages and German, as mentioned above, but also Maltese and Latin, see the section on ALEXINA), and to studies about the quantitative assessment of morphological complexity, currently an active area of research in morphology, have been pursued following previous work published in 2011 [109], [126]. Concerning constructional morphology (derivation, composition) and borrowings, studies and experiments have been carried out in the context of the ANR EDyLex project and that of the collaboration with *viavoo* [45], following here as well experiments carried out in 2011 [124], [115], [127].

6.6. Advances in lexical morphology and syntax

Participants: Benoît Sagot [correspondant], Laurence Danlos, Éric Villemonte de La Clergerie.

The Alexina framework (cf. 5.8) [105] has been developed and used for developing various lexicons, in particular the Le*fff*, that are used in many tools such as POS-taggers [15] and parsers.

In 2012, the new developments within Alexina have been fourfold:

- A large amount of work has been made for developing a new morphological layer to Alexina, in collaboration with a specialist of formal morphology.
- In the context of this collaboration, new Alexina lexicons have been developed with a special focus on linguistic relevance and exhaustivity within a well-defined subset of lexical entries (e.g., Latin verbs, 1st-binyan Maltese verbs).
- The development of a new large-scale NLP-oriented Alexina lexicon has been initiated, namely that of DeLex, an Alexina lexicon for German. It is currently restricted to the morphological layer (no valency information yet) but already generates 2 million inflected lexical entries. The underlying morphological grammar makes use of the new morphological layer mentioned above.
- Following previous work, merging experiments between syntactic resources and the Le*fff* [30] and comparison experiments between such resources and the Le*fff* as reference lexicon for the FRMG parser have been carried out [43]. In the latter series of experiments, the Le*fff* has proven better, or rather more suitable, that other (converted) resources.

6.7. Named Entity Recognition and Entity Linking

Participants: Rosa Stern, Benoît Sagot.

Identifying named entities is a widely studied issue in Natural Language Processing, because named entities are crucial targets in information extraction or retrieval tasks, but also for preparing further NLP tasks (e.g., parsing). Therefore a vast amount of work has been published that is dedicated to named entity *recognition*, i.e., the task of identification of named entity *mentions* (spans of text denoting a named entity), and sometimes *types*. However, real-life applications need not only identify named entity mentions, but also know which real entity they refer to; this issue is addressed in tasks such as knowledge base population with entity resolution and linking, which require an inventory of entities is required prior to those tasks in order to constitute a reference.

6.7.1. Cooperation of symbolic and statistical methods for named entity recognition and typing

Named entity recognition and typing is achieved both by symbolic and probabilistic systems. We have performed an experiment [62] for making the rule-based system NP, SXPipe's high-precision named entity recognition system developed at Alpage on AFP news corpora and which relies on the *Aleda* named entity database, interact with LIANE, a high-recall probabilistic system developed by Frédéric Béchet (LIF) and trained on oral transcriptions from the ESTER corpus. We have shown that a probabilistic system such as LIANE can be adapted to a new type of corpus in a non-supervised way thanks to large-scale corpora automatically annotated by NP. This adaptation does not require any additional manual annotation and illustrates the complementarity between numeric and symbolic techniques for tackling linguistic tasks.

6.7.2. Nomos, a statistical entity linking system

For information extraction from news wires, entities such as persons, locations or organizations are especially relevant in a knowledge acquisition context. Through a process of named entity recognition and entity linking applied jointly, we aim at the extraction and complete identification of these relevant entities, which are meant to enrich textual content in the form of *metadata*. In order to store and access extracted knowledge in a structured and coherent way, we aim at populating an ontological reference base with these metadata. We have pursued our efforts in this direction, using an approach where NLP tools have early access to Linked Data resources and thus have the ability to produce metadata integrated in the Linked Data framework. In particular, we have studied how the entity linking process in this task must deal with noisy data, as opposed to the general case where only correct entity identification is provided.

We use the symbolic named entity recognition system NP, a component of SXPipe, and use it as a mention detection module. Its output is then processed through our entity linking system, which is based on a supervised model learned from examples of linked entities. Since our named entity recognition is not deterministic, as opposed to other entity linking tasks where the gold named entity recognition results are provided, it is configured to remain ambiguous and non-deterministic, i.e., its output preserves a number of ambiguities which are usually resolved at this level. In particular, no disambiguation is made in the cases of multiple possible mentions boundaries (e.g., *{Paris}+{Hilton}* vs. *{Paris Hilton}*). In order to cope with possible false mention matches, which should be discarded as linking queries, the named entity recognition output is made more ambiguous by adding a *not-an-entity* alternative to each mention's candidate set for linking. The entity linking module's input therefore consists in multiple possible readings of sentences. For each reading, this module must perform entity linking on every possible entity mention by selecting their most probable matching entity. Competing readings are then ranked according to the score of entities (or sequence of entities) ranked first in each of them. The reading with no entity should also receive a score in order to be included in the ranking. The motivation for this joint task lies in the frequent necessity of accessing contextual and referential information in order to complete an accurate named entity recognition; thus the part where named entity recognition usually resolves a number of ambiguities is left for the entity linking module, which uses contextual and referential information about entities.

We have realized a first implementation of our system, as well as experiments and evaluation results. In particular, when using knowledge about entities to perform entity linking, we discuss the usefulness of domain specific knowledge and the problem of domain adaptation.

In 2012, improvements have been made to Nomos by combining the NP named entity detection module with LIANE, a probabilistic system developed by Frédéric Béchet (LIF) in order to better predict possible false matches. The linking step has also been enriched with the use of a more complete and autonomous knowledge base derived from Wikipedia, as well as new parameters and ranking functions for the prediction of the mention/entity alignment.

In the context of this linking task for the processing of AFP corpora and content enrichment with metadata, we conducted a deep study of Semantic Web recent developments and especially of the Linked Data initiatives in order to consider the integration of AFP metadata in these knowledge representation frameworks. On this topic as well as the enlarged view of entity linking for semantic annotation of textual content, discussions have taken place with Eric Charton (CRIM, Montréal, Canada) during 2012 Fall.

The Nomos system as well as the general process of content enrichment with metadata and reference base population has been presented at a dedicated workshop at NAACL in June 2012 (AKBC-WEKEX 2012).

6.8. Advances in lexical semantics

Participants: Benoît Sagot [correspondant], Marion Richard, Sarah Beniamine.

In 2012, several contributions to the WOLF have been finalized and/or published. In particular, various successful attempts to enhance the coverage of the WOLF have been integrated within the master resource [23], [19], [31], [24]. A more original work has also been achieved, targeted at improving the precision of the resource by automatically detecting probable outliers [32]. This latter work has been integrated within the dedicated sloWTool platform, and these outliers partly validated by Slovene students of Romance studies. In parallel, a medium-scale manual validation effort has been achieved at Alpage thanks to the work of two Master students funded by the ANR EDyLex project, which has led to the validation of a vast majority of so-called "basic" synsets, i.e., what can be expected to be the most useful part of the resource.

The result of all this work has been integrated in a preliminary first non-alpha version of the WOLF, version WOLF 1.0b.

6.9. Techniques for transferring lexical resources from one language to a closely-related one

Participants: Yves Scherrer, Benoît Sagot.

Developing lexical resources is a costly activity, which means that large resources only exist for a small number of languages. In our work, we address this issue by transferring linguistic annotations from a language with large resources to a closely related language which lacks such resources. This research activity, funded by the Labex EFL, has started in October 2012.

First results include the development of a method to create bilingual dictionaries without any parallel data, depending solely on surface form similarities and their regularities. The resulting bilingual dictionaries are used to transfer part-of-speech annotations from one language to the other. At the moment, our methods are being tested with Wikipedia texts from various languages and dialects closely related to German, such as Dutch and Pfälzisch. We plan to extend this work to data from other language groups and to other types of linguistic annotations, for instance syntactic or semantic resources.

6.10. Modelling the acquisition of linguistic categories by children

Participants: Benoît Crabbé, Luc Boruta, Isabelle Dautriche.

This task breaks in two sub-tasks: acquisition of phonemic categories, and acquisition of syntactic categories.

Although we are only able to distinguish between a finite, small number of sound categories – i.e., a given language's phonemes – no two sounds are actually identical in the messages we receive. Given the pervasiveness of sound-altering processes across languages – and the fact that every language relies on its own set of phonemes – the question of the acquisition of allophonic rules by infants has received a considerable amount of attention in recent decades. How, for example, do English-learning infants discover that the word forms [kæt] and [kat] refer to the same animal species (i.e. *cat*), whereas [kæt] and [bæt] (i.e. *cat* $\sim$ *bat*) do not? What kind of cues may they rely on to learn that [sɪŋkɪŋ] and [θɪŋkɪŋ] (*sinking* $\sim$ *thinking*) can not refer to the same action? The work presented in this dissertation builds upon the line of computational studies initiated by [90], wherein research efforts have been concentrated on the definition of sound-to-sound dissimilarity measures indicating which sounds are realizations of the same phoneme. We show that solving Peperkamp et al.'s task does not yield a full answer to the problem of the discovery of phonemes, as formal and empirical limitations arise from its pairwise formulation. We proceed to circumvent these limitations, reducing the task of the acquisition of phonemes to a partitioning-clustering problem and using multidimensional scaling to allow for the use of individual phones as the elementary objects. The results of various classification and

clustering experiments consistently indicate that effective indicators of allophony are not necessarily effective indicators of phonemehood. Altogether, the computational results we discuss suggest that allophony and phonemehood can only be discovered from acoustic, temporal, distributional, or lexical indicators when—on average—phonemes do not have many allophones in a quantified representation of the input. This subtask has seen the Phd defense of Luc Boruta whose Phd thesis : *"Indicators of allophony and phonemehood"* was successfully defended in September 2012.

As for syntactic categorization, the task is concerned with modelling and implementing psychologically motivated models of language treatment and acquisition. Contrary to classical Natural Language Processing applications, the main aim was not to create engineering solutions to language related tasks, but rather to test and develop psycholinguistic theories. In this context, the study was concerned with the question of learning word categories, such as the categories of Noun and Verb. It is established experimentally that 2-year-old children can identify novel nouns and verbs. It has been suggested that this can be done using distributional cues as well as prosodic cues. While the plain distributional hypothesis had been tested quite extensively, the importance of prosodic cues has not been addressed in a computational simulation. We provided a formulation for modelling this hypothesis using unsupervised and semi-supervised forms of Bayesian learning (EM) both offline and online. This activity started with the master thesis of A. Gutman and has seen this year the start of a new Phd student : I. Dautriche.

6.11. Modelling and extracting discourse structures

Participants: Laurence Danlos, Charlotte Roze.

6.11.1. Lexical semantics of discourse connectives

Discourse connectives are words or phrases that indicate senses holding between two spans of text. The theoretical approaches accounting for these senses, such as text coherence, cohesion, or rhetorical structure theory, share at least one common feature: they acknowledge that many connectives can indicate different senses depending on their context. LEXCONN is a lexical database for French connectives [16].

The French connectives "*en réalité*" and "*en effet*" have been the topic of numerous studies but none of them was formalized. [53] gives a formalization of the conditions the two arguments of these connectives should meet. This formalization is based on factivity information as modeled in the FactBank corpus developed by Roser Sauri.

Sometimes, the sense of connectives is unique but its arguments are hard to determine. In particular, the second argument of an adverbial connective is not always equivalent to its syntactic arguments. This raises problems at the syntax-semantics interface which are described in [52]. The method to handle theses problems in a discursive parser will be studied in the ANR project POLYMNIE, which is headed by Sylvain Podogolla (Inria Lorraine) and which started in October 2012.

6.11.2. Discursive annotation

We plan to annotate the French corpus FTB (French Tree Bank) at the discursive level, in order to obtain the FDTB (French Discourse Tree Bank). The methodology that will be used is close to the one used in the PDTB (Penn Discourse Tree Bank). The first steps of this long term project are presented in [48], [49], [51].

This work is based on a new hierarchy of discourse relations and this new hierarchy was presented at an European workshop organized by the project MULDICO.

6.12. Modelling word order preferences in French

Participants: Juliette Thuilier, Benoît Crabbé, Margaret Grant.

We study the problem of choice in the ordering of French words using statistical models along the lines of [60] and [61]. This work aims at describing and model preferences in syntax, bringing additional elements to Bresnan's thesis, according to which the syntactic competence of human beings can be largely simulated by probabilistic models. We previously investigated the relative position of attributive adjectives with respect to the noun.

This year has seen the Phd thesis defense of Juliette Thuilier in September 2012.

In collaboration with Anne Abeillé (Laboratoire de Linguistique Formelle, Université Paris 7), we extended our corpora study with psycholinguistic questionnaires, in order to show that statistical models are reflecting some linguistic knowledge of French speakers. The preliminary results confirm that animacy is not a relevant factor in ordering French complements.

As regards to corpus work, we are extending the database with spontaneous speech corpora (CORAL-ROM and CORPAIX) and a wider variety of verbal lemmas, in order to enhance sample representativeness and statistical modelling. This activity has lead to the development of an extension of the French Treebank for oral corpora (approx 2000 sentences).

In a cross-linguistic perspective, we plan to strengthen the comparison with the constraints observed in other languages such as English or German with the recruitment of a new postdoc arriving at the beginning of 2013.

As can be seen from the outline above, this line of research brings us closer to cognitive sciences. We hope, in the very long run, that these investigations will bring new insights on the design of probabilistic parsers or generators. In NLP, the closest framework implementing construction grammars is Data Oriented Parsing (DOP).

7. Bilateral Contracts and Grants with Industry

7.1. Contracts with Industry

Alpage has developed several collaborations with industrial partners. Apart from grants described in the next section, specific collaboration agreements have been set up with Verbatim Analysis (license agreement and "CIFRE" PhD, see section 4.3), Lingua et Machina (DTI-funded engineer, see section 4.4), Viavoo, and Diadeis (the "Investissements d'Avenir" project PACTE has started in 2012, see section 4.5).

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. LabEx EFL (Empirical Foundations of Linguistics) (2011 – 2021)

Participants: Laurence Danlos, Benoît Sagot, Chloé Braud, Marie Candito, Benoît Crabbé, Pascal Denis, Charlotte Roze, Pierre Magistry, Djamé Seddah, Juliette Thuilier, Éric Villemonte de La Clergerie.

Linguistics and related disciplines addressing language have achieved much progress in the last two decades but improved interdisciplinary communication and interaction can significantly boost this positive trend. The LabEx (excellency cluster) EFL (Empirical Foundations of Linguistics), launched in 2011 and headed by Jacqueline Vaissière, opens new perspectives by adopting an integrative approach. It groups together some of the French leading research teams in theoretical and applied linguistics, in computational linguistics, and in psycholinguistics. Through collaborations with prestigious multidisciplinary institutions (CSLI, MIT, Max Planck Institute, SOAS...) the project aims at contributing to the creation of a Paris School of Linguistics, a novel and innovative interdisciplinary site where dialog among the language sciences can be fostered, with a special focus on empirical foundations and experimental methods and a valuable expertise on technology transfer and applications.

Alpage is a very active member of the LabEx EFL together with other linguistic teams we have been increasingly collaborating with: LLF (University Paris 7 & CNRS) for formal linguistics, LIPN (University Paris 13 & CNRS) for NLP, LPNCog (University Paris 5 & CNRS) LSCP (ENS, EHESS & CNRS) for psycholinguistics, MII (University Paris 4 & CNRS) for Iranian and Indian studies. Alpage resources and tools have already proven relevant for research at the junction of all these areas of linguistics, thus drawing a preview of what the LabEx is about: experimental linguistics (see Section 4.6). Moreover, the LabEx should provide Alpage with opportunities for collaborating with new teams, e.g., on language resource development with descriptive linguists (INALCO, for example).

Benoît Sagot is in charge of one of the 7 scientific "strands" of the LabEx EFL, namely the strand on Language Resources. Several other project members are in charge of research operations within 3 of these 7 strands ("Experimental grammar from a cross-linguistic perspective", "Computational semantic analysis", "Language Resources").

8.2. National Initiatives

8.2.1. ANR

8.2.1.1. ANR project ASFALDA (2012 – 2015)

Participants: Marie Candito, Benoît Sagot, Éric Villemonte de La Clergerie, Laurence Danlos, Marianne Djemaa.

Alpage is principal investigator team for the ANR project ASFALDA, lead by Marie Candito. The other partners are the Laboratoire d'Informatique Fondamentale de Marseille (LIF), the CEA-List, the MELODI team (IRIT, Toulouse), the Laboratoire de Linguistique Formelle (LLF, Paris Diderot) and the Ant'inno society.

The project aims to provide both a French corpus with semantic annotations and automatic tools for shallow semantic analysis, using machine learning techniques to train analyzers on this corpus. The target semantic annotations are structured following the FrameNet framework [56] and can be characterized roughly as an explicitation of "who does what when and where", that abstracts away from word order / syntactic variation, and to some of the lexical variation found in natural language.

The project relies on an existing standard for semantic annotation of predicates and roles (FrameNet), and on existing previous effort of linguistic annotation for French (the French Treebank). The original FrameNet project provides a structured set of prototypical situations, called frames, along with a semantic characterization of the participants of these situations (called *roles*). We propose to take advantage of this semantic database, which has proved largely portable across languages, to build a French FrameNet, meaning both a lexicon listing which French lexemes can express which frames, and an annotated corpus in which occurrences of frames and roles played by participants are made explicit. The addition of semantic annotations to the French Treebank, which already contains morphological and syntactic annotations, will boost its usefulness both for linguistic studies and for machine-learning-based Natural Language Processing applications for French, such as content semantic annotation, text mining or information extraction.

To cope with the intrinsic coverage difficulty of such a project, we adopt a hybrid strategy to obtain both exhaustive annotation for some specific selected concepts (commercial transaction, communication, causality, sentiment and emotion, time), and exhaustive annotation for some highly frequent verbs. Pre-annotation of roles will be tested, using linking information between deep grammatical functions and semantic roles.

The project is structured as follows:

- Task 1 concerns the delimitation of the focused FrameNet substructure, and its coherence verification, in order to make the resulting structure more easily usable for inference and for automatic enrichment (with compatibility with the original model);
- Task 2 concerns all the lexical aspects: which lexemes can express the selected frames, how they map to external resources, and how their semantic argument can be syntactically expressed, an information usable for automatic pre-annotation on the corpus;
- Task 3 is devoted to the manual annotation of corpus occurrences (we target 20000 annotated occurrences);
- In Task 4 we will design a semantic analyzer, able to automatically make explicit the semantic annotation (frames and roles) on new sentences, using machine learning on the annotated corpus;
- Task 5 consists in testing the integration of the semantic analysis in an industrial search engine, and to measure its usefulness in terms of user satisfaction.

The scientific key aspects of the project are:

- an emphasis on the diversity of ways to express the same frame, including expression (such as discourse connectors) that cross sentence boundaries;
- an emphasis on semi-supervised techniques for semantic analysis, to generalize over the available annotated data.

8.2.1.2. ANR project EDyLex (2010 – 2013)

Participants: Benoît Sagot [principal investigator], Rosa Stern, Damien Nouvel, Virginie Mouilleron, Marion Baranes, Marion Richard, Sarah Beniamine, Laurence Danlos.

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 maintenance 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 Méditerranée), the LIMSI (CNRS team), two small companies, Syllabs and Vecsys Research, and the AFP.

In 2012, several important developments have been achieved:

- Large-scale improvements within the WOLF (Free French WordNet)
- Corpus-based studies targeted at qualitatively understanding and quantitatively modeling French morphological construction mechanisms (derivation, composition, borrowing and others)
- Development of modules for automatic detection, classification and morphological analysis of unknown words in French corpora [45];
- Adaptation and extension of the NewsProcess architecture, previously developed at Alpage, for meeting the expectations of the EDyLex project in terms of lexicon extension from dynamic corpora, here AFP news wires.

8.2.1.3. ANR project Polymnie (2012-2015)

Participants: Laurence Danlos, Éric Villemonte de la Clergerie.

Polymnie is an ANR research project headed by Sylvain Podogolla (Sémagramme Inria Lorraine) with Melodi (INRIT, CNRS), Signes (LABRI, CNRS) and Alpage as partners. This project relies on the grammatical framework of Abstract Categorial Grammars (ACG). A feature of this formalism is to provide the same mathematical perspective both on the surface forms and on the more abstract forms the latter correspond to. As a consequence:

- ACG allows for the encoding of a large variety of grammatical formalisms such as context-free grammars, Tree Adjoining grammars (TAG), etc.
- ACG define two languages: an abstract language for the abstract forms, and an object language for the surface forms.

The role of Alpage in this project is to develop sentential or discursive grammars written in TAG so as to study their conversion in ACG.

8.2.1.4. "Investissements d'Avenir" project PACTE (2012 – 2014)

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

PACTE (*Projet d'Amélioration de la Capture TExtuelle*) is an "Investissements d'Avenir" project sumbitted within the call "Technologies de numérisation et de valorisation des contenus culturels, scientifiques et éducatifs". It started in early 2012.

PACTE aims at improving the performance of textual capture processes (OCR, manual script recognition, manual capture, direct typing), using NLP tools relying on both statistical (n-gram-based, with scalability issues) and hybrid techniques (involving lexical knowledge and POS-tagging models). It addresses specifically the application domain of written heritage. The project takes place in a multilingual context, and therefore aims at developing as language-independant techniques as possible.

PACTE involves 3 companies (DIADEIS, main partner, as well as A2IA and Isako) as well as Alpage and the LIUM (University of Le Mans). It brings together business specialists, large-scale corpora, lexical resources, as well as the scientific and technical expertise required.

In 2012, the results obtained within PACTE are mostly related to SxPipe and to DeLex, the new Alexina lexicon for German (as well as the German instance of MElt trained among other on DeLex). These results are described in more details in the corresponding "software" sections.

8.3. International Initiatives

8.3.1. Participation In International Programs

8.3.1.1. 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" Éric de La Clergerie has participated to an ISO meeting in Madrid (June 2012) and has played a role of expert (in particular on morpho-syntactic annotations [MAF], feature structures [FSR & new FSD], and syntactic annotations [SynAF]). MAF has finally reached the level of an ISO standard (ISO/FDIS 24611, oct. 2012). A paper [21] promoting both SynAF and MAF was presented at TLT'11 (Lisbon, Dec. 2012).

8.4. International Research Visitors

8.4.1. Visits of International Scientists

Roser Sauri, research scientist at Media-Lab in Barcelonea (Spain), has been "professeur invitée" at Alpage between the 1st of april and the 15th of May 2012. Roser Sauri is well known for her work on event factuality for which she developed a formal model and an annotated corpus. During her stay in Paris, she has been working with Alpage members to extend her model to discourse. Moreover, she helped Alpage in launching the FDTB (French Discourse Tree Bank), a project to annotate the French Tree Bank for discourse. Her experience in annotating similar copora for Catalan and Spanish was very fruitful and collaboration with her is going on.

8.4.1.1. Internships

Thomas Roberts (from Jun 2012 until Aug 2012)

Subject: Lefff-like English syntactic lexicon

Institution: Massachusetts Institute of Technology (United States)

9. Dissemination

9.1. Scientific Animation

- Alpage is involved in the French journal Traitement Automatique des Langues (T.A.L., AERES linguistic rank: A). Éric de La Clergerie is "Rédacteur en chef" and was the editor of the special issue 53/2 (2012) on "Time and Space". Laurence Danlos is a member of the editorial board. Benoît Sagot is a guest editor, jointly with Núria Bel, for a special issue on Language Resources for which Benoît Crabbé is a member of the specific reviewing committee.

- Alpage members were involved in many Program, Scientific or Reviewing Committees for other journals and conferences. For example, Éric de La Clergerie participated to the program committees of CSLP'12, LGC'12, STAIRS'12, NAACL-HLT'12 (Parsing Area), TALN'12 (Scientific Committee), LREC'12 (Scientific Committee), WoLE'2012 (1st International Workshop on Web of Linked Entities) and reviewed for the journal *Computational Linguistics*; Djamé Seddah was a PC member of EMNL2012, TALN 2012, CICLING 2012; and Benoît Crabbé was a reviewer for the journals LRE and *Language Modeling*.
- Alpage, (and more specifically Éric de La Clergerie, Djamé Seddah, and Laurence Danlos) was in charge of the organization of TAG+11 (on Tree Adjoining Grammars and Related Formalisms, http://alpage.inria.fr/tagplus11), that was held in Paris (September 2012).
- Laurence Danlos co-organized (with Pierre Zweigembaum , LIMSI, CNRS) a one-day ATALA workshop on Artificial Intelligence and Natural Language Processing that was held in Paris in April 2012.
- Benoît Sagot is elected board member of the French NLP society (ATALA) and Secretary since September 2010.
- Laurence Danlos is a member of the Permanent Committee of the TALN conference organized by ATALA.
- Laurence Danlos is a member of the Scientific Committee of the Linguistics UFR of University Paris Diderot.
- Benoît Sagot is a member of the Governing Board and of the Scientific Board of the LabEx (excellency cluster EFL), as head of the research strand on language resources; Laurence Danlos is a member of the Scientific Board of the LabEx EFL, representing Alpage.
- Djamé Seddah is one of the founders of the statistical parsing of morphologically rich language initiative that started during IWPT'09. He was the program co-chair of the successful SPMRL 2010 NAACL-HLT Workshop and of its 2011 edition that took place during IWPT'11. He co-chaired an ACL 2012 workshop centered around the Syntaxic and Semantic Processing of Morphologicall Rich Languages. He and Marie Candito are also involved (both as core members, Djamé Seddah as co-chair) into the MRL statistical parsing shared task that will be organized in this context (data are to be released in February 2013 and the results will be presented at EMLP 2013; Marie Candito is in charge of the French data). Moreover, Benoît Sagot is a member of the reviewing committees. Finally, Alpage is a regular sponsor of this series of workshops.
- Djamé Seddah was one of the guest-editors of a Special issue of the Computational Linguistics Journal dedicated to the Parsing of Morphologically-Rich Language.
- Djamé Seddah was the program chairs of the French NLP society one-day workshops (until September 2012) and a member of its board.
- Marie Candito organized the Alpage research seminar.
- Benoît Crabbé co-organized the research seminar : le.ctures in experimental linguistics (Univ P7)
- Benoît Crabbé and Benoît Sagot were project reviewers for the National Research Agency (ANR).

9.2. Participation to workshops, conferences, and invitations

Invited talks and seminars:

- Benoît Sagot gave an invited talk at the workshop on "Computational Approaches to Morphological Complexity" in Surrey, UK and he presentated the French Social Media Bank within the Working Group n°9 "Shallow Annotation" of the Consortium Corpus Écrits within the Très Grande Infrastructure de Recherche (TGIR) CORPUS.
- Benoît Sagot was invited to a seminar at Düsseldorf to talk on "Semi-automatic development of lexical resources".

- Laurence Danlos gave invited seminars at Laboratoire d'Informatique de l'Université Pierre et Marie Curie, France, at Computer Science department of UDM (Université de Montréal) Canada, at ENS Lyon and University Paris Sorbonne in France.
- Djamé Seddah gave an invited seminar at the University of Dusseldorf (Abteilung für Computerlinguistik), was a panelist at the NAACL Workshop on Statistical Analysis of Non Canonical Languages, and gave an invited talk at the Paris 7 Linguistics department.
- Benoît Crabbé gave an invited tutorial at the "Experiences, empiricité, expérimentations en linguistique" at Saint Raphael (organized by the CNRS lab HTL) at gave a talk at the "demi heure de sciences" of Inria's Paris Rocquencourt center.
- Marie Candito gave an invited talk at the seminar series of the LIMSI's Groupe Information, Langue Ecrite et Signée (ILES)

Participation to conferences and workshops (in almost all cases, this is associated with at least a talk or a poster presentation):

- Almost all members of Alpage participated to TALN 2012, Genoble, France.
- Several members of Alpage participated to TAG+11, organized by Alpage in Paris, France.
- Laurence Danlos and Juliette Thuillier participated to CMLF'12 (Congrès Mondial de Linguistique française) in Lyon, France.
- Laurence Danlos and Charlotte Roze participated to the Muldico Exploratory Workshop "Towards a multilingual database of connectives" in Iena, Deutchland.
- Laurence Danlos participated in a Linguistic workshop in Bordeaux.
- Rosa Stern participated to the EACL workshop on Innovative hybrid approaches to the processing of textual data – Poster
- Rosa Stern participated to the AKBC-WEKEX workshop (http://akbcwekex2012.wordpress.com) – Poster
- Éric de La Clergerie participated to the 11th International Workshop on Treebanks and Linguistic Theories (TLT11) in Lisbon, Portugal – Paper
- Laurence Danlos, Charlotte Roze, Éric de La Clergerie, Marie Candito participated to LREC'12 in Istanbul, Turkey – Papers
- Laurence Danlos and Éric de La Clergerie participated to the Lexis-Grammar Conference in Nové Hrady, Czech Republic – Paper
- Benoît Sagot participated to LREC'12 in Istanbul, Turkey – 5 oral presentations, 1 poster, and a participation at a workshop.
- Djamé Seddah presented one poster at EVALITA 2012 (Rome), one at LREC 2012 (Istanbul), one talk at COLING 2012 (Mumbai) and one talk at NAACL SANCL 2012 (Montreal).

9.3. Teaching - Supervision - Juries

9.3.1. 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. Marie Candito is in charge of the License 3, and Laurence Danlos is in charge of 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 (Inria partial delegation): Introduction to NLP (3rd year of License, 28h); Discourse, NLU and NLG (2nd year of Master, 28h).

Marie Candito (Inria part time delegation, renewed in September 2012): Information retrieval (2nd year of professional Master, 12h); Clustering and Classification (2nd year of professional Master, 12h); Automatic semantic analysis (2nd year of Master, 12h); Machine translation (1st year of Master, 48h);

Benoît Crabbé: Language data analysis (24h Master 2 P7) ; Logical and Computational structures for language modelling (12h Master 2 MPRI) with S. Schmitz. Introduction to computer science (24h L3 P7); Introduction to Corpus Linguistics (24h L3 P7); Stochastic methods for NLP (24h Master 1 P7); Introduction to Computational Linguistics (24h UFR P7);

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

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

Juliette Thuilier: Syntactic theories : Lexical-Functional Grammar (3rd year of Licence, 48h); Introduction to syntax (2nd year of License, 24h) at University Paris Sorbonne; Implementation of LFG Grammar (2nd year of License, 12h) at University Paris Sorbonne;

Pierre Magistry: Object Oriented Programming, Java-II (3rd year of licence, 12/24h) ; Syntax : HPSG with LKB (1st year Master, 24h)

Luc Boruta: Introduction to Programming II (3rd year of License, 24h); Algorithmics (3rd year of License, 24h); Language & Computer Science (1st year of License, 12h);

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

Chloé Braud: Programmation 2 Java 24h (L3 LI CM/TP); Sémantique computationnelle 24h (M1 LI TD).

Djamé Seddah (half-delegation since September 2011, renewed for 2012): 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”).

Ongoing PhDs and PhDs defended in 2012:

PhDs in progress:

- Corentin Ribeyre, *vers la syntaxe profonde pour l’interface syntaxe-sémantique*, started in November 2012, supervized by Laurence Danlos, co-supervized by Djamé Seddah and Éric de La Clergerie.
- Isabelle Dautriche, *Exploring early syntactic acquisition: a experimental and computational approach*, started in September 2012, co-supervized by Benoît Crabbé.
- Marion Baranes, *Correction orthographique contextuelle de corpus multilingues et multicanaux*, started in January 2012, supervized by Laurence Danlos, co-supervized by Benoît Sagot, in collaboration with the *viavoo* company.
- Marianne Djemaa, *Création semi-automatique d’un FrameNet du français, via interface syntaxe-sémantique*, started in October 2012, supervized by Laurence Danlos, co-supervised by Marie Candito
- Chloé Braud, *Développement d’un système complet d’analyse automatique du discours à partir de corpus annotés, bruts et bruités*, started in September 2011, supervized by Laurence Danlos
- Valérie Hanoka, *Construction semi-automatique de réseaux lexicaux spécialisés multilingues*, started in January 2011, supervized by Laurence Danlos, co-supervized by Benoît Sagot

- Enrique Henestroza Anguiano, *Enhancing statistical parsing with lexical resources*, started in November 2009, supervized by Laurence Danlos and co-supervized by Marie Candito and Alexis Nasr.
- Emmanuel Lassalle, *Résolution automatique des anaphores associatives*, started in September 2010, supervised by Laurence Danlos
- Pierre Magistry, *Construction (semi)-automatique de lexiques dynamiques du mandarin*, started in September 2010, supervized by Sylvain Kahane and co-supervized by Benoît Sagot and Marie-Claude Paris
- Charlotte Roze, *Vers une algèbre des relations de discours*, started in October 2009, supervized by Laurence Danlos
- Rosa Stern, *Construction d'une base de référence pour les métadonnées de dépêches d'agence: reconnaisance et résolution jointes d'entités avec adaptation au domaine*, started in November 2009, supervized by Laurence Danlos and co-supervized by Benoît Sagot

PhD defended

- François-Régis Chaumartin, *Extraction automatisée de connaissances d'une encyclopédie*, started in October 2005, supervized by Sylvain Kahane, and defended in September 2012
- Luc Boruta, *Indicators of Allophony and Phonemehood*, started in September 2009, co-supervized by Emmanuel Dupoux (LSCP/ENS) and Benoît Crabbé, defended in September 2012
- Juliette Thuilier, *Contraintes préférentielles et ordre des mots en français*, started in September 2008, supervized by Laurence Danlos and co-supervized by Benoît Crabbé, defended in September 2012

9.3.2. Juries

- Laurence Danlos was a member of the HdR defense committee of Matthieu Constant (University Paris-Est Marne-la-Vallée, Nov. 2012); the title of his dissertation was "Mettre les expressions multi-mots au coeur du Traitement Automatique des Langues - sur l'exploitation de ressources lexicales".
- Éric de La Clergerie was a member of the PhD defense committee of Karen Fort (Université Paris 13, Dec. 2012); the title of her dissertation was "Les ressources annote´es, un enjeu pour l'analyse de contenu : vers une me´thodologie de l'annotation manuelle de corpus".
- Éric de La Clergerie was a member of the PhD defense committee of Nguyen Van Tien (Université de Pau, Nov. 2012). The title of his dissertation was "Méthode d'extraction d'informations géographiques à des fins d'enrichissement d'une ontologie de domaine".
- Éric de La Clergerie was a member of the PhD defense committee of Milagros Fernandez Gavilanes (Univ. of La Coruña, Spain, Oct. 2012) in quality of co-supervisor. The title of her dissertation was "Adquisicio´n y representacio´n del conocimiento mediante procesamiento del lenguaje natural".
- Djamé Seddah was a member of the PhD defense committee of Anthony Sigogne (University Paris-Est Marne-la-Vallée, Nov. 2012). The title was "Inte´gration de ressources lexicales riches dans un analyseur syntaxique probabiliste".
- Djamé Seddah was a member of the a research fellow hiring committee at the University of Copenhagen (Denmark). Position funded by the Andeers Sogaard's ERC young researcher grant.
- Djamé Seddah was a member of the "Comité de Sélection" for an Assistant Professor position in Computer Science (CNU 27) at Université Paris Sorbonne.
- Benoît Sagot was a member of the "Comité de Sélection" for an Assistant Professor position in Computer Science (CNU 27) at University Paris-Est Marne-la-Vallée (IUT).

  Marie Candito was a member of the "Comité de Sélection" for an Assistant Professor position in Computer Science (CNU 27) at Université Aix-Marseille II.

9.4. Popularization

- Éric de La Clergerie and Mikael Morardo have presented some results about NLP and Knowledge Acquisition during “Les journées de la science” (Oct. 2012).
- Éric de La Clergerie was invited to talk to the Inria Sophia seminar “C@fe-in” and to the CIV (*Valbonne International Center*) about “Les ordinateurs comprennent-ils le langage ?” (*Do computers understand language ?*).

9.5. AERES Evaluation

The UMR-I formed by Alpage was evaluated by AERES on December 4th, 2012.

10. Bibliography

Major publications by the team in recent years

[1] A. BITTAR, P. AMSILI, P. DENIS, L. DANLOS. *French TimeBank: an ISO-TimeML Annotated Reference Corpus*, in "ACL 2011 - 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies", Portland, OR, United States, Association for Computational Linguistics, June 2011, http://hal.inria.fr/inria-00606631/en.

[2] P. BOULLIER. *Range Concatenation Grammars*, in "New Developments in Parsing Technology", H. BUNT, J. CARROLL, G. SATTA (editors), Text, Speech and Language Technology, Kluwer Academic Publishers, 2004, vol. 23, p. 269–289.

[3] 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).

[4] M. CANDITO, B. CRABBÉ, P. DENIS. *Statistical French dependency parsing: treebank conversion and first results*, in "Seventh International Conference on Language Resources and Evaluation - LREC 2010", Malte La Valletta, European Language Resources Association (ELRA), May 2010, p. 1840-1847, http://hal.inria.fr/hal-00495196/en.

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

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

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

[8] B. SAGOT. *The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French*, in "7th international conference on Language Resources and Evaluation (LREC 2010)", Malte Valletta, 2010, http://hal.inria.fr/inria-00521242/en.

[9] B. SAGOT, G. WALTHER. *Non-Canonical Inflection: Data, Formalisation and Complexity Measures*, in "SFCM 2011 - The Second Workshop on Systems and Frameworks for Computational Morphology", Zürich, Switzerland, C. MAHLOW, M. PIOTROWSKI (editors), Communications in Computer and Information Science, Springer, August 2011, vol. 100, p. 23-45 [*DOI :* 10.1007/978-3-642-23138-4], http://hal.inria.fr/inria-00615306/en.

[10] 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.

[11] 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.

[12] É. VILLEMONTE DE LA CLERGERIE. *Building factorized TAGs with meta-grammars*, in "The 10th International Conference on Tree Adjoining Grammars and Related Formalisms - TAG+10", New Haven, CO États-Unis, 2010, p. 111-118, http://hal.inria.fr/inria-00551974/en/.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[13] L. BORUTA. *Indicateurs d'allophonie et de phonémicité*, Université Paris-Diderot - Paris VII, September 2012, http://hal.inria.fr/tel-00746163.

Articles in International Peer-Reviewed Journals

[14] B. CRABBÉ, D. DUCHIER, C. GARDENT, J. LE ROUX, Y. PARMENTIER. *XMG : eXtendible MetaGrammar*, in "Computational Linguistics", November 2012, vol. 39, n^{o} 3, p. 1-66, http://hal.inria.fr/hal-00768224.

[15] P. DENIS, B. SAGOT. *Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging*, in "Language Resources and Evaluation", 2012, vol. 46, n^{o} 4, p. 721-736 [*DOI :* 10.1007/s10579-012-9193-0], http://hal.inria.fr/inria-00614819.

[16] C. ROZE, L. DANLOS, P. MULLER. *LEXCONN: a French lexicon of discourse connectives*, in "Discours", 2012, http://hal.inria.fr/hal-00702542.

[17] J. THUILIER, G. FOX, B. CRABBÉ. *Prédire la position de l'adjectif épithète en français : approche quantitative*, in "Lingvisticae Investigationes", June 2012, vol. 35, n^{o} 1, http://hal.inria.fr/hal-00698896.

[18] R. TSARFATY, D. SEDDAH, S. KÜBLER, J. NIVRE. *Parsing Morphologically Rich Languages: Introduction to the Special Issue*, in "Computational Linguistics", November 2012 [*DOI :* 10.1162/COLI_A_00133], http://hal.inria.fr/hal-00780897.

International Conferences with Proceedings

[19] M. APIDIANAKI, B. SAGOT. *Applying cross-lingual WSD to wordnet development*, in "LREC 2012 - Eighth International Conference on Language Resources and Evaluation", Istanbul, Turkey, May 2012, http://hal.inria.fr/hal-00703126.

[20] L. BORUTA, J. JASTRZEBSKA. *A Phonemic Corpus of Polish Child-Directed Speech*, in "LREC 2012 - Eighth International Conference on Language Resources and Evaluation", Istanbul, Turkey, May 2012, http://hal.inria.fr/hal-00702437.

[21] S. BOSCH, S. CHOI, É. VILLEMONTE DE LA CLERGERIE, A. CHENGYU FANG, G. FAASS, K. LEE, A. PAREJA-LORA, L. ROMARY, A. ZELDES, F. ZIPSER. *As a standardized serialisation for ISO 24615 - SynAF*, in "TLT11 - 11th international workshop on Treebanks and Linguistic Theories - 2012", Lisbon, Portugal, I. HENDRICKX, S. KÜBLER, K. SIMOV (editors), Ediçoes Colibri, November 2012, p. 37-60, http://hal.inria.fr/hal-00765413.

[22] M. CANDITO, D. SEDDAH. *Effectively long-distance dependencies in French : annotation and parsing evaluation*, in "TLT 11 - The 11th International Workshop on Treebanks and Linguistic Theories", Lisbon, Portugal, November 2012, http://hal.inria.fr/hal-00769625.

[23] K. GÁBOR, M. APIDIANAKI, B. SAGOT, É. VILLEMONTE DE LA CLERGERIE. *Boosting the Coverage of a Semantic Lexicon by Automatically Extracted Event Nominalizations*, in "LREC 2012 - Eighth International Conference on Language Resources and Evaluation", Istanbul, Turkey, May 2012, http://hal.inria.fr/hal-00703127.

[24] V. HANOKA, B. SAGOT. *Wordnet creation and extension made simple: A multilingual lexicon-based approach using wiki resources*, in "LREC 2012 : 8th international conference on Language Resources and Evaluation", Istanbul, Turkey, 2012, 6, http://hal.inria.fr/hal-00701606.

[25] E. HENESTROZA ANGUIANO, M. CANDITO. *Probabilistic lexical generalization for French dependency parsing*, in "SP-Sem-MRL 2012 - Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages", Jeju Island, Korea, Republic Of, 2012, http://hal.inria.fr/hal-00699675.

[26] J. LE ROUX, B. SAGOT, D. SEDDAH. *Statistical Parsing of Spanish and Data Driven Lemmatization*, in "Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012)", Corée, République De, 2012, 6, http://hal.archives-ouvertes.fr/hal-00702496.

[27] P. MAGISTRY. *Segmentation non supervisée : le cas du mandarin*, in "RECITAL - Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues - 2012", Grenoble, France, ATALA, June 2012, http://hal.inria.fr/hal-00701197.

[28] P. MAGISTRY, B. SAGOT. *Unsupervized Word Segmentation: the case for Mandarin Chinese*, in "ACL - Annual Meeting of the Association for Computational Linguistics - 2012", Jeju, Korea, Republic Of, ACL, July 2012, http://hal.inria.fr/hal-00701200.

[29] C. RIBEYRE, D. SEDDAH, É. VILLEMONTE DE LA CLERGERIE. *A Linguistically-motivated 2-stage Tree to Graph Transformation*, in "TAG+11 - The 11th International Workshop on Tree Adjoining Grammars and Related Formalisms - 2012", Paris, France, C.-H. HAN, G. SATTA (editors), Inria, September 2012, http://hal.inria.fr/hal-00765422.

[30] B. SAGOT, L. DANLOS. *Merging syntactic lexica: the case for French verbs*, in "LREC'12 Workshop on Merging Language Resources", Istanbul, Turkey, May 2012, http://hal.inria.fr/hal-00703128.

[31] B. SAGOT, D. FIŠER. *Automatic Extension of WOLF*, in "GWC2012 - 6th International Global Wordnet Conference", Matsue, Japan, Global Wordnet Association + Toyohashi University of Technology + National Institute of Japanese Language and Linguistics, January 2012, http://hal.inria.fr/hal-00655774.

[32] B. SAGOT, D. FIŠER. *Cleaning noisy wordnets*, in "LREC 2012 - Eighth International Conference on Language Resources and Evaluation", Istanbul, Turkey, May 2012, http://hal.inria.fr/hal-00703125.

[33] B. SAGOT, R. STERN. *Aleda, a free large-scale entity database for French*, in "LREC 2012 : eighth international conference on Language Resources and Evaluation", Istanbul, Turkey, 2012, 4, http://hal.inria.fr/hal-00699300.

[34] D. SEDDAH, M. CANDITO, B. CRABBÉ, H. ANGUIANO ENRIQUE. *Ubiquitous Usage of a French Large Corpus: Processing the Est Republicain Corpus*, in "LREC 2012 - The eighth international conference on Language Resources and Evaluation", Istambul, Turquie, May 2012, http://hal.inria.fr/hal-00780899.

[35] D. SEDDAH, J. LE ROUX, B. SAGOT. *Data Driven Lemmatization for Statistical Constituent Parsing of Italian*, in "Proceedings of EVALITA 2011", Roma, Italy, Italy, Springer, 2012, http://hal.inria.fr/hal-00702618.

[36] D. SEDDAH, B. SAGOT, M. CANDITO, V. MOUILLERON, V. COMBET. *Building a treebank of noisy user-generated content: The French Social Media Bank*, in "TLT 11 - The 11th International Workshop on Treebanks and Linguistic Theories", Lisbonne, Portugal, December 2012, Cet article constitue une version réduite de l'article "The French Social Media Bank : a Treebank of Noisy User Generated Content" (mêmes auteurs), http://hal.inria.fr/hal-00780898.

[37] D. SEDDAH, B. SAGOT, M. CANDITO, V. MOUILLERON, V. COMBET. *The French Social Media Bank: a Treebank of Noisy User Generated Content*, in "COLING 2012 - 24th International Conference on Computational Linguistics", Mumbai, Inde, Kay, Martin and Boitet, Christian, December 2012, http://hal.inria.fr/hal-00780895.

[38] D. SEDDAH, B. SAGOT, M. CANDITO. *The Alpage Architecture at the SANCL 2012 Shared Task: Robust Pre-Processing and Lexical Bridging for User-Generated Content Parsing*, in "SANCL 2012 - First Workshop on Syntactic Analysis of Non-Canonical Language , an NAACL-HLT'12 workshop", Montréal, Canada, May 2012, http://hal.inria.fr/hal-00703124.

[39] R. STERN, B. SAGOT, F. BÉCHET. *A Joint Named Entity Recognition and Entity Linking System*, in "EACL 2012 Workshop on Innovative hybrid approaches to the processing of textual data", Avignon, France, April 2012, http://hal.inria.fr/hal-00699295.

[40] R. STERN, B. SAGOT. *Population of a Knowledge Base for News Metadata from Unstructured Text and Web Data*, in "AKBC-WEKEX 2012 - The Knowledge Extraction Workshop at NAACL-HLT 2012", Montréal, Canada, 2012, http://hal.inria.fr/hal-00699297.

[41] J. THUILIER, L. DANLOS. *Semantic annotation of French corpora: animacy and verb semantic classes*, in "LREC 2012 - The eighth international conference on Language Resources and Evaluation", Istanbul, Turkey, European Language Resources Association (ELRA), May 2012, http://hal.inria.fr/hal-00698907.

[42] J. THUILIER. *Lemme verbal et classe sémantique dans l'ordonnancement des compléments postverbaux*, in "CMLF 2012 - Congrès Mondial de Linguistique Française", Lyon, France, Institut de Linguistique Française (CNRS, FR 2393), July 2012, http://hal.inria.fr/hal-00698909.

[43] E. TOLONE, B. SAGOT, É. VILLEMONTE DE LA CLERGERIE. *Evaluating and improving syntactic lexica by plugging them within a parser*, in "LREC 2012 - Eighth International Conference on Language Resources and Evaluation", Istanbul, Turkey, May 2012, http://hal.inria.fr/hal-00703195.

[44] É. VILLEMONTE DE LA CLERGERIE. *Etude du traitement de certains compléments de phrase dans le cadre d'une méta-grammaire*, in "LGC'12 - 31ème 30ème Colloque international sur le Lexique et la Grammaire", Nové Hrady, Czech Republic, J. RADIMSKY (editor), Intitut de langues Romanes, University of South Bohemia, September 2012, http://hal.inria.fr/hal-00765384.

National Conferences with Proceeding

[45] M. BARANES. *Vers la correction automatique de textes bruités: Architecture générale et détermination de la langue d'un mot inconnu*, in "RECITAL'2012 - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues", Grenoble, France, 2012, p. 95-108, http://hal.inria.fr/hal-00701400.

[46] C. BENZITOUN, K. FORT, B. SAGOT. *TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe*, in "JEP-TALN 2012 - Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles", Grenoble, France, June 2012, p. 99-112, http://hal.inria.fr/hal-00709187.

[47] M. CANDITO, D. SEDDAH. *Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical*, in "TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles", Grenoble, France, June 2012, http://hal.inria.fr/hal-00698938.

[48] L. DANLOS, D. ANTOLINOS-BASSO, C. BRAUD, C. ROZE. *Vers le FDTB : French Discourse Tree Bank*, in "TALN 2012 : 19ème conférence sur le Traitement Automatique des Langues Naturelles", Grenoble, France, G. ANTONIADIS, H. BLANCHON, G. SÉRASSET (editors), ATALA/AFCP, 2012, vol. 2, p. 471-478, http://hal.inria.fr/hal-00703407.

[49] L. DANLOS. *Méthodologie pour le FDTB (French Discourse Tree Bank)*, in "La linguistique de corpus à l'heure de la confrontation entre concepts, techniques et applications", Bordeaux, France, 2012, 2, http://hal.inria.fr/hal-00755329.

[50] B. SAGOT, M. RICHARD, R. STERN. *Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées*, in "Traitement Automatique des Langues Naturelles (TALN)", Grenoble, France, G. ANTONIADIS, H. BLANCHON, G. SÉRASSET (editors), June 2012, vol. 2 - TALN, http://hal.inria.fr/hal-00703108.

Conferences without Proceedings

[51] L. DANLOS, D. ANTOLINOS-BASSO, C. BRAUD, C. ROZE. *Vers le FDTB : French Discourse Tree Bank*, in "JAD'12 - Journée Atala Discours", Paris, France, ATALA et revue Discours, May 2012, http://hal.inria.fr/hal-00704705.

[52] L. DANLOS. *Connecteurs de discours adverbiaux : Problémes à l'interface syntaxe-sémantique*, in "31th International Conference on Lexis and Grammar", Nové Hrady, Czech Republic, Institut de Langues Romanes, University of South Bohemia, 2012, 7, http://hal.inria.fr/hal-00755367.

[53] L. DANLOS. *Formalisation des conditions d'emploi des connecteurs en réalité et (et) en effet*, in "Congrès Mondial de Linguistique Française", Lyon, France, Institut de Linguistique Française, 2012, http://hal.inria.fr/hal-00755413.

Books or Proceedings Editing

[54] M. APIDIANAKI, I. DAGAN, J. FOSTER, Y. MARTON, D. SEDDAH, R. TSARFATY (editors). *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, Association for Computational Linguistics, 2012, 113, http://hal.inria.fr/hal-00702616.

References in notes

[55] A. ABEILLÉ, N. BARRIER. *Enriching a French Treebank*, in "Proceedings of LREC'04", Lisbon, Portugal, 2004.

[56] C. BAKER, C. FILLMORE, J. LOWE. *The berkeley framenet project*, in "Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1", Association for Computational Linguistics, 1998, p. 86–90.

[57] S. BANGALORE, P. BOULLIER, A. NASR, O. RAMBOW, B. SAGOT. *MICA: A Probabilistic Dependency Parser Based on Tree Insertion Grammars*, in "NAACL 2009 - North American Chapter of the Association for Computational Linguistics (Short Papers)", Boulder, Colorado, États-Unis, 2009, http://hal.inria.fr/inria-00616695/en/.

[58] P. BOULLIER. *Range Concatenation Grammars*, in "New Developments in Parsing Technology", H. BUNT, J. CARROLL, G. SATTA (editors), Text, Speech and Language Technology, Kluwer Academic Publishers, 2004, vol. 23, p. 269–289.

[59] J. BRESNAN. *The mental representation of grammatical relations*, MIT press, 1982.

[60] J. BRESNAN, A. CUENI, T. NIKITINA, H. BAAYEN. *Predicting the Dative Alternation*, in "Cognitive Foundations of Interpretation", Amsterdam, Royal Netherlands Academy of Science, Amsterdam, 2007, p. 69-94.

[61] J. BRESNAN, M. FORD. *Predicting syntax: Processing dative constructions in American and Australian varieties of English*, in "Language", 2010, vol. 86, n^{o} 1, p. 168–213, http://muse.jhu.edu/content/crossref/journals/language/v086/86.1.bresnan.html.

[62] F. BÉCHET, B. SAGOT, R. STERN. *Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées*, in "TALN'2011 - Traitement Automatique des Langues Naturelles", Montpellier, France, 2011, http://hal.inria.fr/inria-00617068.

[63] M. CANDITO. *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées*, Université Paris 7, 1999.

[64] M. CANDITO, B. CRABBÉ. *Improving generative statistical parsing with semi-supervised word clustering*, in "Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)", Paris, France, October 2009, p. 169-172, short paper (4 pages), http://hal.archives-ouvertes.fr/hal-00495267/en/.

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

[66] 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.

[67] M. CANDITO, E. HENESTROZA ANGUIANO, D. SEDDAH. *A Word Clustering Approach to Domain Adaptation: Effective Parsing of Biomedical Texts*, in "IWPT'11 - 12th International Conference on Parsing Technologies", Dublin, Irlande, October 2011, http://hal.inria.fr/hal-00659577.

[68] M. CANDITO, J. NIVRE, P. DENIS, E. HENESTROZA ANGUIANO. *Benchmarking of Statistical Dependency Parsers for French*, in "23rd International Conference on Computational Linguistics - COLING 2010", Chine Beijing, Coling 2010 Organizing Committee, Aug 2010, p. 108-116, 9 pages, http://hal.inria.fr/hal-00514815/en.

[69] M. CANDITO, D. SEDDAH. *Parsing word clusters*, in "NAACL/HLT-2010 Workshop on Statistical Parsing of Morphologically Rich Languages - SPMRL 2010", États-Unis Los Angeles, Association for Computational Linguistics, Jun 2010, p. 76-84, http://hal.inria.fr/hal-00495177/en.

[70] 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.

[71] N. CHOMSKY. *Aspects of the theory of Syntax*, MIT press, 1965.

[72] M. COLLINS. *Head Driven Statistical Models for Natural Language Parsing*, University of Pennsylvania, Philadelphia, 1999.

[73] B. CRABBÉ. *Grammatical Development with XMG*, in "Logical Aspects of Computational Linguistics (LACL)", Bordeaux, 2005, p. 84-100, Published in the Lecture Notes in Computer Science series (LNCS/LNAI), vol. 3492, Springer Verlag.

[74] 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, France, 2008, p. 45–54.

[75] 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, France, June 2008, p. 44-54, http://hal.archives-ouvertes.fr/hal-00341093/en/.

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

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

[78] 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.

[79] P. DENIS, B. SAGOT. *Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français*, in "Traitement Automatique des Langues Naturelles : TALN 2010", Canada Montréal, 2010, http://hal.inria.fr/inria-00521231/en.

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

[81] E. HENESTROZA ANGUIANO, M. CANDITO. *Parse correction with specialized models for difficult attachment types*, in "EMNLP 2011 - The 2011 Conference on Empirical Methods in Natural Language Processing", Edinburgh, United Kingdom, 2011, http://hal.inria.fr/hal-00602083/en.

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

[83] F. KELLER. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*, University of Edinburgh, 2000.

[84] D. KLEIN, C. D. MANNING. *Accurate Unlexicalized Parsing*, in "Proceedings of the 41st Meeting of the Association for Computational Linguistics", 2003.

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

[86] 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.

[87] 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.

[88] J. NIVRE, J. NILSSON. *Pseudo-projective dependency parsing*, in "Proc. of ACL 2005", 2005, p. 99–106.

[89] J. NIVRE, M. SCHOLZ. *Deterministic Dependency Parsing of English Text*, in "Proceedings of Coling 2004", Geneva, Switzerland, COLING, Aug 23–Aug 27 2004, p. 64–70.

[90] S. PEPERKAMP, R. LE CALVEZ, J.-P. NADAL, E. DUPOUX. *The acquisition of allophonic rules: statistical learning with linguistic constraints*, in "Cognition", 2006, vol. 101, n^{o} 3, p. B31–B41.

[91] 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.

[92] S. PETROV, D. KLEIN. *Improved Inference for Unlexicalized Parsing*, in "Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference", Rochester, New York, Association for Computational Linguistics, April 2007, p. 404–411, http://www.aclweb.org/anthology/N/N07/N07-1051.

[93] S. PETROV, R. T. MCDONALD. *Overview of the 2012 Shared Task on Parsing the Web*, in "Proceedings of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL), a NAACL-HLT 2012 workshop", Montréal, Canada, 2012.

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

[95] 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.

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

[97] B. SAGOT, P. BOULLIER. *SxPipe 2: architecture pour le traitement pré-syntaxique de corpus bruts*, in "Traitement Automatique des Langues", 2008, vol. 49, n^{o} 2, p. 155-188, http://hal.inria.fr/inria-00515489/en/.

[98] 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.

[99] B. SAGOT, D. FIŠER. *Building a free French wordnet from multilingual resources*, in "OntoLex", Marrakech, Maroc, 2008, http://hal.inria.fr/inria-00614708/en/.

[100] B. SAGOT, D. FIŠER. *Construction d'un wordnet libre du français à partir de ressources multilingues*, in "Traitement Automatique des Langues Naturelles", Avignon, France, 2008, http://hal.inria.fr/inria-00614707/en/.

[101] 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, 2008, http://hal.archives-ouvertes.fr/hal-00402305.

[102] 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.

[103] 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.

[104] 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.

[105] B. SAGOT. *The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French*, in "7th international conference on Language Resources and Evaluation (LREC 2010)", Malte Valletta, 2010, http://hal.inria.fr/inria-00521242/en.

[106] B. SAGOT, G. WALTHER, P. FAGHIRI, P. SAMVELIAN. *A new morphological lexicon and a POS tagger for the Persian Language*, in "International Conference in Iranian Linguistics", Uppsala, Sweden, 2011, http://hal.inria.fr/inria-00614711/en.

[107] B. SAGOT, G. WALTHER, P. FAGHIRI, P. SAMVELIAN. *Développement de ressources pour le persan : le nouveau lexique morphologique PerLex 2 et l'étiqueteur morphosyntaxique MElt-fa*, in "TALN 2011 - Traitement Automatique des Langues Naturelles", Montpellier, France, June 2011, http://hal.inria.fr/inria-00614710/en.

[108] B. SAGOT, G. WALTHER. *A morphological lexicon for the Persian language*, in "7th international conference on Language Resources and Evaluation (LREC 2010)", Malte Valletta, 2010, http://hal.inria.fr/inria-00521243/en.

[109] B. SAGOT, G. WALTHER. *Non-Canonical Inflection: Data, Formalisation and Complexity Measures*, in "The Second Workshop on Systems and Frameworks for Computational Morphology (SFCM 2011)", Zürich, Suisse, August 2011, http://hal.inria.fr/inria-00615306/en/.

[110] D. SEDDAH, M. CANDITO, B. CRABBÉ. *Adaptation De Parsers Statistiques Lexicalisés Pour Le Français : Une Évaluation Complète Sur Corpus Arborés*, in "Conférence sur le traitement automatique des langues naturelles - TALN'09", Senlis, France, 2009, http://hal.inria.fr/inria-00525749/en/.

[111] D. SEDDAH, M. CANDITO, B. CRABBÉ. *Cross Parser Evaluation and Tagset Variation: A French Treebank Study*, in "Proceedings of the 11th International Conference on Parsing Technologies - IWPT'09", Paris, France, Association for Computational Linguistics, 2009, p. 150-161, http://hal.inria.fr/inria-00525750/en/.

[112] D. SEDDAH, G. CHRUPAŁA, Ö. ÇETINOGLU, J. VAN GENABITH, M. CANDITO. *Lemmatization and Statistical Lexicalized Parsing of Morphologically-Rich Languages*, in "Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages - SPMRL 2010", États-Unis Los Angeles, CA, 2010, http://hal.inria.fr/inria-00525754/en.

[113] D. SEDDAH. *Exploring the Spinal-Stig Model for Parsing French*, in "Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)", Malte Malta, 2010, http://hal.inria.fr/inria-00525753/en.

[114] R. STERN, B. SAGOT. *Resources for Named Entity Recognition and Resolution in News Wires*, in "Entity 2010 Workshop at LREC 2010", Malte Valletta, 2010, http://hal.inria.fr/inria-00521240/en.

[115] J. STRNADOVÁ, B. SAGOT. *Construction d'un lexique des adjectifs dénominaux*, in "Traitement Automatique des Langues Naturelles", Montpellier, France, 2011, http://hal.inria.fr/inria-00617062/en/.

[116] S. TAGLIAMONTE, D. DENIS. *Linguistic ruin? LOL! Instant messaging and teen language*, in "American Speech", 2008, vol. 83, n^{o} 1, 3.

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

[118] 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.

[119] É. VILLEMONTE DE LA CLERGERIE. *Parsing Mildly Context-Sensitive Languages with Thread Automat a*, in "Proc. of COLING'02", August 2002.

[120] 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.

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

[122] É. VILLEMONTE DE LA CLERGERIE. *Building factorized TAGs with meta-grammars*, in "The 10th International Conference on Tree Adjoining Grammars and Related Formalisms - TAG+10", New Haven, CO États-Unis, 2010, p. 111-118, http://hal.inria.fr/inria-00551974/en/.

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

[124] G. WALTHER, L. NICOLAS. *Enriching Morphological Lexica through Unsupervised Derivational Rule Acquisition*, in "WoLeR 2011 at ESSLLI (International Workshop on Lexical Resources)", Ljubljana, Slovénie, August 2011, http://hal.inria.fr/inria-00617064/en/.

[125] G. WALTHER, B. SAGOT, K. FORT. *Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish*, in "International Conference on Lexis and Grammar", Serbie Belgrade, Sep 2010, http://hal.inria.fr/hal-00510999/en.

[126] G. WALTHER, B. SAGOT. *Modélisation et implémentation de phénomènes flexionnels non-canoniques*, in "Traitement Automatique des Langues", 2011, vol. 52, n^{o} 2, http://hal.inria.fr/inria-00614703/en/.

[127] G. WALTHER, B. SAGOT. *Problèmes d'intégration morphologique d'emprunts d'origine anglaise en français*, in "30th International Conference on Lexis and Grammar", Nicosia, Chypre, 2011, http://hal.inria.fr/inria-00616779/en/.

[128] T. WASOW. *Postverbal behavior*, CSLI, 2002.

[129] H. YAMADA, Y. MATSUMOTO. *Statistical Dependency Analysis with Support Vector Machines*, in "The 8th International Workshop of Parsing Technologies (IWPT2003)", 2003.

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