IN PARTNERSHIP WITH:
Université Denis Diderot (Paris 7)

Activity Report 2011

Project-Team ALPAGE

Large-scale 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).

1. Members

Research Scientists

Pierre Boullier [Emeritus Senior Researcher (DR-E) Inria, HdR]
Pascal Denis [Junior Researcher (CR) Inria]
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]
Philippe Muller [delegation from University Paul Sabatier, Toulouse, (September 2009 to August 2011)]
Djamé Seddah [Associate Professor (MC) Univ. Paris 4]

Technical Staff

Mickael Morardo [Inria DTI-funded Engineer in collaboration with Lingua et Machina (since 2011)]
Géraldine Walther [Research Engineer funded by the ANR project EDyLex (March-May 2011)]

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]
Valérie Hanoka [PhD student (CIFRE) at Verbatim Analysis & Univ. Paris 7]
Enrique Henestroza Anguiano [PhD funded by the ANR project SEQUOIA (since November 2009)]
Emmanuel Lassalle [PhD student (ENS stipendium) Univ. Paris 7 (since September 2010)]
Pierre Magistry [PhD student (allocataire) Univ. Paris 7 (since September 2010)]
Charlotte Roze [PhD student (allocataire) Univ. Paris 7 (since October 2009)]
Rosa Stern [PhD student (CIFRE) AFP & Univ. Paris 7 (since November 2009)]
Juliette Thuilier [PhD student (since 2008)]

Post-Doctoral Fellows

Marianna Apidianaki [funded by the ANR project EDyLex (January to September 2011)]
Yayoi Nakamura-Delloye [funded by the ANR project EDyLex (December 2010 to August 2011)]
Sattisvar Tandabany [funded by the ANR project SEQUOIA (from November 2009 to June 2011)]

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.

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 [63], [108], [116]) 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 [128], [125]. 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 [93] or automatic [103], [104] 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 [79], or probabilistic models derived from lexicalized Tree Adjoining grammars, such as Stochastic Tree Insertion Grammars [75].

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 ...) [134], [102]. 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 [96]. 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 [59] and derive the best input for syntagmatic statistical parsing [81]. Benchmarking several PCFG-based learning frameworks [12] 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) [104].
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 [75] 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 [122].
Pursuing our work on PCFG-LA based parsing, we investigated the automatic conversion of the treebank into dependency syntax representations [70], 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 [115]. At the syntactic level, work has been achieved on automatic acquisition of atomic syntactic information and automatic detection of errors in the lexicon [135],[11]. At the semantic level, automatic wordnet development tools have been described [107], [130], [91], [90]. 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* [110],[9], developed within the Alexina framework, as well as a wordnet for French, the WOLF [8], the first freely available resource of the kind.

In the last 2 years, Alpage members have shown how to benefit from other more linguistically-oriented resources, such as the Lexique-Grammaire and DICOVALENCE , in order to improve the coverage and quality of the 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 [115], Polish [117], English, Spanish [98], [97] and Persian [120], 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 [7]. 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 [121].

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 (see section 6.5).

3.5. Discourse structures

Participants: Laurence Danlos, Charlotte Roze, Pascal Denis, Philippe Muller.

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

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

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

- first, some connectors (in particular conjunctions of subordination) introduce pure modifiers and must not be considered as bearing discourse relations,
- second, other elements than connectors can lexicalize discourse relations, in particular verbs like *précéder / to precede* or *causer / to cause*, which have facts or fact eventualities as arguments [82].

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, [83],[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.

3.6. Coreference resolution

Participants: Pascal Denis, Philippe Muller, Laurence Danlos.

An important challenge for the understanding of natural language texts is the correct computation of the *discourse entities* that are mentioned therein —persons, locations, abstract objects, and so on. In addition to identifying individual referential expressions (e.g., *Nicolas Sarkozy*, *Neuilly*, *l'UMP*) and properly typing them (e.g. *Nicolas Sarkozy* is a PERSON, *Neuilly* is a LIEU), the task is also to determine the other mentions with which these expressions are coreferential. Part of the difficulty of this task is that natural languages provide many ways to refer to the same entity (including the use of pronouns such as *il*, *ses* and definite descriptions such as *le président*, making them highly ambiguous. The identification of coreferential links and other anaphoric links (such as "associative anaphora") plays a key role for various applications, such as extraction and retrieval of information, but also the summary or automatic question-answering systems. This central role of coreference resolution has been recognized by the inclusion of this task in different international evaluation campaigns, beginning with the campaigns *Message Understanding Conference* (in particular, MUC-6 and MUC-7)[1], and more recently *Automatic Content Extraction* (ACE)[2] and *Anaphora Resolution Evaluation* (ARE)[3]. The creation and distribution of corpora developped as part of these campaigns have significantly boosted research in automatic coreference resolution. In particular, they have made possible the application of machine learning techniques (mostly supervised ones) to the problem of coreference resolution. This in turn has led to the development of systems that were both more robust and more precise, thus making more realistic their integration within these larger systems. Some of the best systems based on supervised learning methods are described in [123], [99], [95], [100], [94], [88]. Note that a few attemtps were also made at using unsupervised techniques (mostly clustering methods) for the task [74], [101], but these systems are still far from reaching the performance of their supervised counterparts.

[1] See, respectively: http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html and http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.
[2] http://www.nist.gov/speech/tests/ace/
[3] http://clg.wlv.ac.uk/events/ARE/

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), text generation (4.6), spelling correction (4.7) and experimental linguistics (4.8).

4.2. Information extraction and knowledge acquisition

Participants: Éric Villemonte de La Clergerie, Rosa Stern, Yayoi Nakamura-Delloye, Marianna Apidianaki, 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.

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. Shallow processing of e-mails

Participants: Benoît Sagot, Laurence Danlos.

Shallow processing is one of the most important NLP application domains. This includes, in particular, detecting named entities in a broad sense (person names, organization names, locations, addresses, date and time mentions, and others), with many possible purposes, such as text normalization and even anonymization, but more importantly for extracting events and other kinds of structured information from text. This is what the new company Kwaga is trying to do on e-mails, challenging difficulties related to the high level of noise that characterizes e-mail corpora (spelling mistakes, shortenings, inter-e-mail structure...). In 2009-2010, an ARITT contract has been set up to try and study the usability of Alpage's SXPipe shallow processing chain for part of this purpose.

4.5. 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 head by François Brown de Colstoun, a former INRIA researcher, that provides services for developing specialized multilingual terminologies for its clients. It develops the 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.

4.6. Generation of textual reports about statistical data: EASYTEXT

Participant: Laurence Danlos.

Since 2010, the generation system EASYTEXT has been polished up so that it is operational at Kantar Media which sailed it to a bunch of customers. As Kantar Media was pleasantly surprised by the quality of the automatically generated texts, they asked for further extensions of EASYTEXT which are currently worked on, especially an extension to generate English texts.

EASYTEXT has been presented at two international conferences [34], [27].

4.7. 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 digitalization, and two other partners. It has been rerouted to the "Investissements d'avenir" framework, and has been accepted. It will start in early 2012.

4.8. 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 [66], [133], the existence of graded grammaticality judgments [92].

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 [76] gave way on the one hand to more abstract characterizations of the language faculty [76], and on the other hand to the construction of detailed, formally explicit, and often implemented, alternative formulation of the generative approach [65], [106]. For French several grammars have been implemented in this trend, among which the tree adjoining grammars of [68], [80] 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], Sattisvar Tandabany, 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 SEQUOIA ANR project (see section 8.2.1), 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). A partial port for Window Cygwin has been successful but has not yet been integrated and finalized.

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 [127].

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. [125], [126], and very large amount of data (700 millions of words) in the SCRIBO project.

DYALOG is used at LORIA (Nancy), University of Coruña (Spain), Instut Gaspard Monge (Univ. Marne La Vallée), University of Nice, and a few other users.

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 [129]. 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.

5.4. The Bonsai PCFG-LA parser

Participants: Benoît Crabbé [correspondant], Marie Candito, Pascal Denis, Djamé Seddah.

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

Alpage has developped as support of the research papers [81], [70], [71], [12] 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 [61] 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, Pierre Boullier, Éric Villemonte de La Clergerie.

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

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

SXPipe, now in version 2 [109] is a modular and customizable chain aimed to apply to raw corpora a cascade of surface processing steps. It is used

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

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

5.7. MElt

Participants: Pascal Denis [correspondant], Benoît Sagot.

MElt is a part-of-speech tagger, trained for French (on the French TreeBank and coupled with the Le*fff*), English [89], Spanish, Kurmanji Kurdish [131] and Persian [56], [42]. It is state-of-the-art for French. It is distributed freely as a part of the Alpage linguistic workbench.

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 are under development, in particular for Spanish (the Le*ffe*), Polish, Slovak, English, Galician, Persian, Kurdish.

Alexina also hosts *Aleda* [124], an large-scale entity database currently developed for French but under development for English, 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], Marianna Apidianaki.

The WOLF (Wordnet Libre du Français) is a wordnet for French, i.e., a lexical semantic database. The development of WOLF started in 2008 [113], [8]. At this time, we focused on benefitting 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. Since 2008, work specific to each of both resources has been done [114], but since end-2010 the collaboration has been re-activated. This is due among others to the fact that the joint development of WOLF and sloWNet is one of the main objectives of the two-year PROTEUS bilateral PHC project co-headed by Benoît Sagot (2010-2011, see section 8.3.2). Moreover, the EDyLex project also contributed to funding the improvement of the WOLF, in particular through the work of Marianna Apidianaki.

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

Participants: Enrique Henestroza Anguiano [correspondant], Pascal Denis.

FREDISTis a freely-available (LGPL license) Python package that implements methods for the automatic construction of distributional thesauri [31].

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 thresholding, 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

Participants: Laurence Danlos [correspondant], Pascal Denis, Philippe Muller.

Apetite provides a set of tools to handle ISO-TimeML annotations, predict temporal structures from timex/event mark-ups, and different ways of evaluating the results. It is licensed under the Cecill, a GPL-like license http://www.irit.fr/~Philippe.Muller/tools/apetite-0.7.tgz.

In parallel, Alpage developed the *French TimeBank* [22], [21], a freely-available corpus annotated with ISO-TimeML-compliant temporal information (dates, events and relations between events).

5.12. System EasyRef

Participant: Éric Villemonte de La Clergerie [maintainer].

PASSAGE action

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 parsing with DyALog/FRMG

Participant: Éric Villemonte de La Clergerie.

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 [127] and on the Lefff morphological and syntactic lexicon [118]. It relies on the notion of factorized grammar, themselves generated from a representation that lies at a higher level of abstraction, named Meta-Grammars [129]. 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.

Éric de La Clergerie has continued to improve the coverage, quality and efficiency of the French meta-grammar FRMG. On the EasyDev corpus (around 4000 sentences), parsing times have improved over 2011 from an average of 1.03s per sentence to 0.28s, coverage (in terms of sentences with full parses) has improved from 72.5% to 82.60%, and accuray (in terms of f-mesure over relations) from 64.54% to 68.28%.

A part of the accuracy gains comes from the addition of a new output format for FRMG, namely the CONLL format, allowing us to use the CONLL-based dependency version of the French Treebank (around 12K sentences) for training and evaluation. We also used new machine learning techniques to improve FRMG's disambiguation algorithm, allowing us to combine heuristic based disambiguation rules (with manually provided weights) with more standard parsing features associated with automatically learned weights. More precisely, the idea was to study the efficiency of the disambiguation rules over the French treebank and to favor (resp. penalize) well-working (resp. bad working) rules by adjusting their weight, taking into account additional (and more standard) features. Using these techniques, on `ftb6_3` test part, FRMG improved from a base accuracy of 82.31% (in terms of CONLL Labeled Attachment Score) to 84.54%. These gains resulting from a training over the French TreeBank have also been observed (with however a lesser impact) on the EasyDev corpus (using a different format and using a different evaluation metric).

6.2. Task-based evaluation of syntactic lexica: coupling FRMG with various resources

Participants: Éric Villemonte de La Clergerie, Benoît Sagot.

The FRMG symbolic parser was used for comparing the performances of various syntactic lexicons as sources of information for parsing. The idea is to convert syntactic lexica other than the Lefff into the Lefff's format, i.e., turn them into Alexina lexicons, and then use the resulting lexica together with the FRMG grammar for producing several parsers. These parsers only differ by the lexical information they rely on. Preliminary results had already been obtained in 2009 [119], but were restricted to one external lexicon, namely Lexicon-Grammar tables, and only to verbal entries (other entries were gathered from the Lefff when using Lexicon-Grammar-based verbal entries). However, conversion tools for other resources, such as Dicovalence [136], had already been developed, in the context of the development and improvement of the Lefff. Moreover, the development of a new version of the Lefff verbal entries

Task-based evaluation results have been obtained on parsing with FRMG, showing that the Lefff performs better than both Lexique-Grammaire and DICOVALENCE (after conversion to the Alexina formalism) [48], [49]. The new version of the Lefff, mentioned above, leads for now to lower results than the current version, but its results are better than with Lexique-Grammaire or DICOVALENCE data, despite a significant increase of the average amount of entries per lemma. These results are satisfying both because they show that the Lefff is a useful resource for symbolic parsing, but also because they illustrate the relevance of converting other resources into the Alexina formalism, in order to merge the valuable linguistic information they contain — as done in the last years for improving the Lefff [85], [84], [86], [111], [87], [112].

6.3. Information extraction from corpora parsed with FRMG

Participants: Yayoi Nakamura-Delloye, Rosa Stern, Éric Villemonte de La Clergerie, Benoît Sagot.

Following previous experiments, in particular in the context of the FUI-funded project Scribo that ended in 2010 [4], work has been achieved for extracting information from corpora parsed with FRMG.

In the context of the EDyLex project, we have proposed two pattern-based named entity extraction methods for ontology enrichment [36], [35]. The proposed methods are characterized by the use of entity relation patterns obtained by our unsupervised extraction method. These patterns correspond to syntactic paths that connect two named entities in dependency trees produced by FRMG. This work aims to take advantage of parsing benefits and also offers solutions for parsing disadvantage. The proposed methods are characterized by the use of entity relation patterns obtained by our unsupervised extraction method. These patterns correspond to syntactic paths that connect two named entities in dependency trees. This work aims to take advantage of parsing benefits and also offers solutions for parsing disadvantage.

We also developed a mechanism for integrating the results into an domain ontology, namely the ontology under deployment at the Agence France-Presse [37].

6.4. Advances in statistical parsing

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

6.4.1. Improving statistical dependency parsing

Alpage has provided state-of-the art results for French statistical Parsing, adapting existing techniques for French, a richer morphological language than English, either for constituency parsing or dependency parsing. The Bonsai tool (see section 5.4) is available, that gathers preprocessing tools and models for dependency parsing French. We have innovated in the tuning of tagsets and the handling of unknown words. In the last years, Alpage has then contributed on four main points:

- conversion of the French Treebank [59] used as constituency training data into dependencies [72], the resulting treebank being used by several teams for dependency parsing;
- an original method to reduce lexical data sparseness and include coverage and robustness by replacing tokens by unsupervised word clusters or morphological clusters [69], [121], [73]; all of our morphological clustering approaches were integrated into our parsing chains; data driven lemmatization required the adaptation of a state-of-the-art part-of-speech tagger and lemmatizer (Morfette [77]) based on a data-driven joint model benefiting of the inclusion of external lexica such as the Le*fff* [121].
- a parser-agnostic postprocessing step, developed this year, which uses specialized models for dependency parse correction [30]: dependencies in an input parse tree are revised by selecting, for a given dependent, the best governor from within a small set of candidates, using a discriminative linear ranking model that includes a rich feature set that encodes syntactic structure in the input parse tree; the parse correction framework can correct attachments using either a generic model or specialized models tailored to difficult attachment types like coordination and pp-attachment; our experiments have shown that parse correction, combining a generic model with specialized models for difficult attachment types, can successfully improve the quality of predicted parse trees output by several representative state-of-the-art dependency parsers for French.
- an adaptation of the above-mentioned technique of word clustering to the problem of adapting statistical parsers to different text domains [25]. We show that in order to parse texts from a different domain than the one a statistical parser is trained on (namely to parse *target domain* text using a parser trained on *indomain* treebank), word clusters computed over a bridge corpus that couples indomain an target domain raw texts do improve parsing performance on target domain, without degrade performance on indomain texts (contrary to previous domain adaptation techniques). To evaluate these experiments, we use as target domain biomedical texts. We have supervised the manual syntactic annotation of a test corpus from the biomedical domain (European Public Assessment Reports concerning the marketing authorization of medicinal products).

[4] NAKAMURADELLOYE:2010:HAL-00511541:1,NAKAMURADELLOYE:2010:HAL-00511481:1

Besides this line of work, it should be noted that two parsing models built around Stochastic Tree Insertion Grammars are currently under investigation: experiments have been conducted on Spinal TIGs [122]. Moreover, we are still improving the TIG-based dependency parser MICA, developed in collaboration with University of Marseilles, Columbia university and AT&T [61] (see section 5.5).

6.4.2. Functional labelling

Alpage worked towards the improvement of a functional labeller to be used as a post-parsing tool on an unfolded parse forest (as outputted e.g. by the Berkeley parser in the Bonsai architecture) using CRF models of various orders thereby extending the previous maximum entropy labeller designed in the team. The use of CRFs for modelling triggered a collaboration with Isabelle Tellier and JP Prost (LIFO, Orleans). The labeller implementation has been considerably improved and the accurracy of the labeller has improved as well on correct treebank trees. However we found out that the feature engineering work outweights the formal improvements since we were able to show that the use of higher order graphical models were not contributing significantly to improve an unstructured model. Our modest gains come mostly from feature engineering. Moreover we notice that combined with a constituent parser the labeller does not improve at all on constituent parsing output. The reason being that our current architecture for the Bonsai parser is sequential (which is unsatisfactory). Following experiments on n-best parsing outputs, we observe that the labeller can drastically improve on better parses where its input is indeed correct. This suggest investigating formulating constituent parsing and functional labelling as a joint task requiring to address serious efficiency issues. We intend to tackle the two drawbacks of our current architecture (sequential process, parse forest unfolding) by formulating constituent parsing as a joint task with functional labelling in the next few months.

6.4.3. Parsing spontaneous oral text

Alpage also got involved in parsing spontaneous oral text taken from ESTER 3 data (with overlaps) generated in the ANR ETAPE project in collaboration with A. Abeillé (LLF) with the aim of preannotating a seed for a future treebank of oral French which would considerably support work in experimental linguistics led in the Labex. He has also a collaboration set up with A. Abeillé, C. Gardent and C. Cerisara for ensuring interoperability accross ongoing efforts for producing oral treebanks for French. The way to carry out the task was by using a form of preprocessing of oral text to simulate a written entry to the Bonsai parser trained on written text. In the next few months we intend to test semi-supervised learning techniques to speed up the annotation process made by the LLF lab.

6.5. 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.

6.5.1. Improvements of the Aleda entity database

Within the Alexina framework, we develop since 2012 the entity database *Aleda* [124], aimed at constituting such a reference. *Aleda* was first developed for French but is under development for English. *Aleda* is 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.

In 2011, major efforts have been made for improving the coverage, precision and richness of the French *Aleda*: improvements in the tool for creating an XML almost-raw-text version of the wikipedia, new method for identifying and typing entities among wikipedia articles, based on infoboxes and wikipedia categories, richer database structure for storing more detailed information about each entity, and many other improvements. A paper about these advances has been sumbitted to LREC 2012.

6.5.2. 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 [24] 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 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-supervized way thanks to large-scale corpora automatically annotated by NP. This adaptation does not require any additional manual anotation and illustrates the complementarity between numeric and symbolic techniques for tackling linguistic tasks.

6.5.3. Nomos, a statistical named 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 learnt from examples of linkings. 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 detection. 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.

6.6. Extending wordnets

Participants: Benoît Sagot, Marianna Apidianaki, Valérie Hanoka.

The WOLF (see section 5.9) is a freely available, automatically created wordnet for French, the biggest drawback of which has until now been the lack of general concepts that are typically expressed with highly polysemous vocabulary that is on the one hand the most valuable for applications in human language technologies but also the most difficult to add to wordnet accurately with automatic methods on the other. In collaboration with Darja Fišer (University of Ljubljana), we have developed a self-training-like technique for acquiring a classifier that is able to assign appropriate synset ids (i.e., senses) to new words, extracted from non-disambituated multilingual sources of lexical knowledge, such as Wiktionaries and Wikipedia [39], [40]. Automatic and manual evaluation shows high coverage as well as high quality of the resulting lexico-semantic repository. Another important advantage of the approach is that it is fully automatic and language-independent and can therefore be applied to any other language still lacking a wordnet. Indeed, it was applied to Slovene as well.

Other techniques were used as well and are the basis of various submitted conference papers. They rely, among others, on morphological derivation, on graph-based representation of highly multilingual lexicons extracted from numerous wiktionaries, and on automatically induced sense clusters.

6.7. Unsupervized lexical semantics

Participant: Marianna Apidianaki.

6.7.1. Unsupervized word sense induction and disambiguation

Word sense induction (WSI) is the task aimed at automatically identifying the senses of words in texts, without the need for handcrafted resources or annotated data. Up till now, most WSI algorithms extract the different senses of a word 'locally' on a per-word basis, i.e. the different senses for each word are determined separately. In collaboration with Tim van de Cruys, at Alpage in 2010, now at University of Cambridge [19], [50], we have compared the performance of such algorithms to a new algorithm that uses a 'global' approach, i.e. the different senses of a particular word are determined by comparing them to, and demarcating them from, the senses of other words in a full-blown word space model. The induction step and the disambiguation step are based on the same principle: words and contexts are mapped to a limited number of topical dimensions in a latent semantic word space. The intuition is that a particular sense is associated with a particular topic, so that different senses can be discriminated through their association with particular topical dimensions; in a similar vein, a particular instance of a word can be disambiguated by determining its most important topical dimensions. We evaluated our model on the SemEval-2010 word sense induction and disambiguation task. All systems that participated in this task use a local scheme for determining the different senses of a word. We obtain state-of-the-art results.

6.7.2. Unsupervized cross-lingual lexical substitution

Cross-Lingual Lexical Substitution (CLLS) is the task that aims at providing for a target word in context several alternative substitute words in another language. The proposed sets of translations may come from external resources or be extracted from textual data. In 2011, we have introduced a new approach for this task [18], namely the use of an unsupervised cross-lingual word-sense induction method. This method identifies the senses of words by clustering their translations according to their semantic similarity. We evaluated the impact of using clustering information for CLLS on the SemEval-2010 CLLS data set. Our system performs better on the 'out-of-ten' measure than the systems that participated in the SemEval task.

6.8. Unsupervized segmentation: the case for Mandarin Chinese

Participants: Pierre Magistry, Benoît Sagot.

For most languages using the Latin alphabet, tokenizing a text on spaces and punctuation marks is a good approximation of a segmentation into lexical units. Although this approximation hides many difficulties, they do not compare with those arising when dealing with languages that do not use spaces, such as Mandarin Chinese. Many segmentation systems have been proposed, some of them use linguistitically motivated unsupervized algorithms. However, standard evaluation practices fail to account for some properties of such systems. New results [33] have shown that a simple model, based on an entropy-based reformulation of a language-independent hypothesis put forward by Harris in 1955, allows for segmenting a corpus and extracting a lexicon from the results. Tested on the Academia Sinica Corpus, our system allows for inducing a segmentation and a lexicon with good intrinsic properties and whose characteristics are similar to those of the lexicon underlying the manually-segmented corpus. Recent unpublished work using a slightly different model have improved these results. In parallel, preliminary experiments on other languages (Hindi, Singalese, Tamil, French) and original vizualisation techniques have already led to promising results.

6.9. Computational morphology

Participants: Benoît Sagot, Géraldine Walther.

Although computational morphology has been a topic of interest for Alpage for several years now, several new research topics have received attention in 2011, often in collaboration with morphologists from the Laboratoire de Linguistique Formelle (University Paris 7).

6.9.1. Inflectional morphology

Non-canonical inflection (suppletion, deponency, heteroclisis...) is extensively studied in theoretical approaches to morphology. However, these studies often lack practical implementations associated with large-scale lexica. Yet these are precisely the requirements for ob- jective comparative studies on the complexity of morphological descriptions. We have shown [16], [43] how the Parsli model of inflectional morphology [132], which can represent many non-canonical phenomena, as well as a formalisation and an implementation thereof can be used to evaluate the complexity of competing morphological descriptions. After illustrating the properties of the model with data about French, Latin, Italian, Persian and Sorani Kurdish verbs and about noun classes from Croatian and Slovak we have conducted experiments on the complexity of four competing descriptions of French verbal inflection. The complexity is evaluated using the information-theoretic concept of description length. We show that the new concepts introduced in the model by the Parsli model enable reducing the complexity of morphological descriptions w.r.t. both traditional or more recent models.

6.9.2. Derivational morphology

This year, in relation with the ANR project EDyLex (see section 8.2.2), work has started targeted towards the acquisition of lexical information at the level of derivational morphology, both using semi- and non-supervized techniques.

Semi-supervized techniques have been used in a work dedicated to French denominal adjectives, for which we have implemented an automatic technique based on large-scale lexicons and corpora for extracting derivation links between base nouns and derived adjectives based on the same stem [46]. The resulting derivational lexicon, which is freely available, has already been partially manually validated. Future work include a full validation and adding denominal adjectives with a suppletive base.

Unsupervized techniques have been used for extraction of derivational links that appear more systematically, although their definition is less linguistically motivated as such [51].

6.9.3. Morphological issues concerning loan words

Also in the context of the ANR project EDyLex (see section 8.2.2), we have carried out a preliminary study on the morphological issues raised by borrowing phenomena, concerning in particular French nouns and verbs borrowed from English [52]. Using techniques that are similar to those used on derivational morphology, we have extracted a significant amount of loan words from a large raw corpus. We have proposed a model of the borrowing phenomenon, that takes into account graphemic (spelling), phonetic and morphological variability.

6.10. Allophony and word segmentation in language acquisition models

Participants: Luc Boruta, Benoît Crabbé.

Allophonic rules are responsible for the great variety in phoneme realizations. Infants can not reliably infer abstract word representations without knowledge of their native allophonic grammar. We have explored the hypothesis that some properties of infants' input, referred to as indicators, are correlated with allophony. First, we provide an extensive evaluation of individual indicators that rely on distributional or lexical information. This evaluation relies on a phonetically transcribed corpus, generated automatically from a phonemically transcribed English, French and Japanese child-directed corpus. As such corpora do not exist as such, we used automatically extracted allophonic grammars of various sizes leading to various granularity levels, using our own allophonic rule extraction algorithm [57]. Then, we present a first evaluation of the combination of indicators of different types, considering both logical and numerical combinations schemes [23]. Though distributional and lexical indicators are not redundant, straightforward combinations do not outperform individual indicators.

Models of the acquisition of word segmentation are typically evaluated using phonemically transcribed corpora. Accordingly, they implicitly assume that children know how to undo phonetic variation when they learn to extract words from speech. Moreover, whereas models of language acquisition should perform similarly across languages, evaluation is often limited to English samples. Using the phonetically annotated corpora described above, that cover three typologically different languages, we evaluated the performance of state-of-the-art statistical models given inputs where phonetic variation has not been reduced. We have measured segmentation robustness across different levels of segmental variation, simulating systematic allophonic variation or errors in phoneme recognition. We have shown that these models do not resist an increase in such variations and do not generalize to typologically different languages. From the perspective of early language acquisition, the results strengthen the hypothesis according to which phonological knowledge is acquired in large part before the construction of a lexicon.

6.11. Modelling the acquisition of syntactic categories by children

Participant: Benoît Crabbé.

B. Crabbé co-supervised A. Gutman for an M2 thesis (MPRI) in collaboration with A. Christophe (LSCP/ENS) in the domain of psycholinguistic modelling. The topic was 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 had 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.

6.12. Modelling and extracting discourse structures

Participants: Laurence Danlos, Charlotte Roze.

6.12.1. Cross-lingual 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. Depending on its sense, the translation of a connective into another language can vary greatly, either using an equivalent connective, or using a different construction or even no explicit connective at all .

On the basis of data provided by the bilingual concordancer TransSearch which propose statistical word alignment [64], [53] made a semi-manual annotation of the English translation of two French connectives ("en effet" and "alors que"). The results of this annotation show that the translations of these connectives do not correspond to the “transpots” identified by TransSearch and even less to the translations proposed in bilingual dictionaries.

The conclusions of this work were presented at an European workshop organized by the project COMTIS[5], and some members decide to use our technic for other connectives and other aligned corpora (e.g. Europarl).

6.12.2. Discourse relations inference rules

In 2011 we have developed a new methodology for building discourse relations inference rules, to be integrated into an algebra of these relations [54], [38]. The construction of such an algebra has as main objective the improvement of the comparison of discourse structures within the evaluation of discourse annotations and the creation of a gold-standard corpus. The inference rules can also help detecting inconsistencies in discourse structures, in order to improve human or machine annotation. The premises of rules already studied lead to the formulation of inference rules, established by the theoretical definition of discourse relations, manually constructed data and extracted data. By manually annotating discourses, we also compute inference probabilities. We have illustrated the adopted methodology taking as theoretical background the Segmented Discourse Representation Theory [60].

6.12.3. Discourse structure and factivity

Discursive annotations proposed in theories of discourse such as RST (Rhetorical Structure Theory) or SDRT (Segmented Representation Theory Dicourse) have the advatange of building a global discourse structure linking all the information in a text. Discursive annotations proposed in PDTB (Penn Discourse Tree Bank) have the advatange of identifying the "source" of each information – thereby answering to questions such as who says or thinks what?

In collaboration with Owen Rambow (Columbia University), we have proposed [26], [28] a unified approach for discursive annotations combining the strengths of these two streams of research. This unified approach relies crucially on factivity information, as encoded in the English corpus FactBank. We intend to pursue this avenue of research by initiating in 2012 the development of a French FactBank.

6.13. Modelling and extracting temporal structures

Participants: Pascal Denis, Philippe Muller.

Temporal information has been the focus of recent attention in information extraction. An elegant approach to learning temporal orderings from texts is to formulate this problem as a constraint optimization problem, which can be then given an exact solution using Integer Linear Programming. This works well for cases where the number of possible relations between temporal entities is restricted to the mere precedence relation, but becomes impractical when considering all possible interval relations.

We have proposed this year two innovations [29], inspired from work on temporal reasoning, that control this combinatorial blow-up, therefore rendering an exact ILP inference viable in the general case. First, we propose to translate the network of constraints from temporal intervals to their end-points, to handle a drastically smaller set of constraints, while preserving the same temporal information. Second, we have show that additional efficiency is gained by enforcing coherence on particular subsets of the entire temporal graphs. We evaluate these innovations through various experiments on TimeBank 1.2 using standard evaluation metrics, and compare our ILP formulations with various baselines and oracle systems.

[5] http://www.idiap.ch/project/comtis

The evaluation of temporal information extraction, i.e., the comparison of two annotations of a given text, is also a scientific challenge. This is because relations between events in a story are intrinsically interdependent and cannot be evaluated separately. A proper evaluation measure is also crucial in the context of a machine learning approach to the problem. Finding a common comparison referent at the text level is not obvious, and we have argued, in collaboration with Xavier Tannier (LIMSI), in favor of a shift from event based measures to measures on a unique textual object, a minimal underlying temporal graph, or more formally the transitive reduction of the graph of relations between event boundaries [15].

6.14. Automatic meronymy discovery

Participants: Emmanuel Lassalle, Pascal Denis.

Bridging descriptions are a special kind of anaphora whose interpretation requires not only identifying an antecedent, but also inferring a specific relation linking it to the anaphor. The resolution of bridging anaphora represents a very challenging task in discourse processing. It is considerably much harder than standard coreferential anaphora resolution for which shallow predictors (like distance, string matching, or morphosyntactic agreement) have been shown to be rather effective. Part of the challenge is due to an important information bottleneck. Lexical resources like WordNet are still too poor and uneven in coverage to provide a realistic solution. In turn, more recent approaches to bridging resolution have turned to web-based extraction methods. To date, the most complete and best-performing approach combines focus and lexical distance predictors using machine learning techniques [105].

We have focused on mereological bridging anaphora (that is, cases wherein the inferred relation is a part-whole relation).[6] Moreover, we have worked on French, a language for which current lexical resources have a very low coverage. The system, presented in [32] is similar to a system developed for English [105], but it was enriched to integrate meronymic information extracted automatically from both web queries and raw text using syntactic patterns. Through various experiments on the DEDE corpus [78], we show that although still mediocre the performance of our system compare favorably to those obtained for English by the above-mentioned system. In addition, our evaluation indicates that the different meronym extraction methods have a cumulative effect, but that the text pattern-based extraction method is more robust and leads to higher accuracy than the Web-based approach.

6.15. Statistical models of word order in French

Participants: Juliette Thuilier, Benoît Crabbé.

We study the problem of choice in the ordering of French words using statistical models along the lines of [66] and [67]. 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, we mainly studied the problem of the relative ordering of postverbal complements. The focus of this investigation is the relative order of direct object and indirect object of French ditransitive verbs. The first part of this work is based on corpora data that we extracted from two journalistic corpora (French Tree Bank and Est-Républicain) and a radio corpus (ESTER). These data were manually annotated and validated for semantic categories (animacy and semantic class of the ditransitive verb). Based on these data, we built statistical models showing that the relative length of complements and verbal lemmas are the most important factors, and that, differently from English or German, categories as animacy or definiteness seem to play no role in the relative ordering.

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.

[6] An illustrative English example is is the following discourse: *The car will not move. The engine is broken.*

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. In a crosslinguistic perspective, we plan to strengthen the comparison with the constraints observed in other languages such as English or German.

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 framework that is closest to implementing construction grammar is Data Oriented Parsing.

6.16. Asessing the Amazon Mechanical Turk plateform

Participant: Benoît Sagot.

In collaboration with Gilles Adda and Joseph Mariani from LIMSI and with Karën Fort from INIST, we have assessed some crowdsourced microworking systems and especially Amazon Mechanical Turk, the use of which has been steadily growing in language processing in the past few years [41], [17]. According to the mainstream opinion expressed in the articles of the domain, this type of on-line working platforms allows to develop very quickly all sorts of quality language resources, for a very low price, by people doing that as a hobby or wanting some extra cash. We have demonstrated that the situation is far from being that ideal, be it from the point of view of quality, price, workers' status or ethics and bring back to mind already existing or proposed alternatives. Our goal was threefold:

- to inform researchers, so that they can make their own choices with all the elements of the reflection in mind,
- to ask for help from funding agencies and scientific associations, and develop alternatives,
- to propose practical and organizational solutions in order to improve new language resources development, while limiting the risks of ethical and legal issues without letting go price or quality.

6.17. Finite state formalisms for Egyptian Hieroglyphic transliteration

Participant: François Barthélemy.

The task of transliterating an Egyptian Hieroglyphic text into the latin alphabet was studied [20], as a model problem to compare two finite-state formalisms: the first one is a cascade of binary transducers; the second one is a class of multitape transducers expressing simultaneous constraints, implemented using the Karamel language [62]. The two systems were compared regarding their expressivity and readability.

The first system tends to produce smaller machines, but is more tricky. On the other hand, the Karamel language provides a more abstract description of the forms, using an explicit tree structure and separating the different pieces of information on different tapes, according to semantic criteria. But the Karamel machine is much larger. Karamel is a high-level declarative formalism whereas non contextual rewrite rules are an efficient low-level language.

7. 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), Kwaga (ARITT contract, see section 4.4), TNS-Sofres (see section 4.6), Lingua et Machina (DTI-funded engineer, see section 4.5) and soon Viavoo (a joint "CIFRE" PhD is about to start) and Diadeis (the "Investissements d'Avenir" project PACTE will start in early 2012, see section 4.7).

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 head 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 4.8). 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 crosslinguistic perspective", "Computational semantic analysis", "Language Resources").

8.2. National Initiatives

8.2.1. ANR project Sequoia (2009 – 2011)

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

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

8.2.2. ANR project EDyLex (2010 – 2012)

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

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

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

8.2.3. "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 has been accepted, and will start in early 2012

PACTE aims at improving the performace 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 applicative 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.

8.3. European Initiatives

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

Participant: Benoît Sagot.

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

8.3.2. French-Slovene bilateral project "Building Slovene-French Linguistic Ressources" (2010 – 2011)

Participant: Benoît Sagot [principal investigator, jointly with Mojca Schlamberger-Brezar].

The objective of this project, jointly lead by Benoît Sagot (Alpage) and Mojca Schlamberger-Brezar (University of Ljubljana) is the development of multilingual linguistic resources for Slovene and French. The French funding is provided by EGIDE. The project is organized around two main goals: the development of a French-Slovene aligned and morphosyntactically annotated corpus, and the extension using semi-automatic techniques (automatic and manual validation construction) of the WOLF and of SloWNet, the wordnets for both languages. All these resources will be made available to the community by a distribution under a free license (e.g., LGPL-LR).

8.4. International Initiatives

8.4.1. 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 ISO events and has played a role of expert (in particular on morpho-syntactic annotations [MAF], feature structures [FSR & new FSD], and syntactic annotations [SynAF]).

9. Dissemination

9.1. Animation of the scientific community

- 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 regular issue 52/1 (2011). 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 is deeply involved in a forthcoming special issue of the major journal in our field of research, Computational Linguistics. Djamé Seddah is one of the guest editor of this issue devoted to "Parsing of morphologically-rich languages" while Marie Candito is a reviewer for this issue.
- 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 IWPT'11, LGC'11, TALN'11, CLA'11, DepLing'11, TEMA'11, and was a reviewer for IJCNLP'11.
- Djamé Seddah and Benoît Sagot are elected board member of the French NLP society (ATALA); Benoît Sagot is Secretary since September 2010; Djamé Seddah is Program Chair of the "Journées ATALA" (one day long workshops in NLP, 4 or 5 per year).
- 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 is co-chairing 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 expected to be released during this ACL 2012 workshop, and the results will be presented at IWPT 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.
- Laurence Danlos organized CID 2011 (Constraints in Discourse) that was held at Agay-Roches Rouges, France in September 2011.
- Benoît Sagot organized WoLeR 2011, an ESSLLI 2011 workshop on Lexical Resources (http://alpage.inria.fr/~sagot/woler2011/), that was held in Ljubljana, Slovenia in August 2011.
- Marie Candito was workshop chair at EMNLP 2011 (Conference on Empirical Methods in Natural Language Processing) that was held in Edinburgh in july 2011.
- Benoît Crabbé organized the Alpage research seminar
- Benoît Crabbé co-organized the research seminar : lectures in experimental linguistics (Univ P7)
- Éric de La Clergerie was a project reviewer for the ANR

9.2. Participation to workshops, conferences, and invitations

Invited talks and seminars:

- Éric de La Clergerie gave an invited talk at the 1st International "Work-Conference on Linguistics, Biology and Computer Science: Interplays" in Tarragona, Spain
- Benoît Crabbé and Juliette Thuilier gave an invited seminar at the University of Bordeaux (Séminaire ERSSàB)
- Benoît Crabbé gave an invited seminar at the University of Düsseldorf
- Marianna Apidianaki gave an invited seminar at the University of Ljubljana (JOTA talk)

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 2011, Montpellier, France.
- Marianna Apidianaki, Marie Candito and Enrique Henestroza Anguiano participated to EMNLP 2011, Edinburgh, UK.
- Marie Candito participated to IWPT'11 (International Conference in Parsing Technologies) Dublin, Ireland.
- Pascal Denis, Marianna Apidianaki and Luc Boruta participated to ACL-HLT 2011 in Portland, United States.
- Marianna Apidianaki participated to CICLing 2011 in Tokyo, Japan.
- Pascal Denis participated to the International Joint Conference on Artificial Intelligence (IJCAI 2011) in Barcelona, Spain
- Pascal Denis participated to the Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011) in Faro, Portugal.
- Juliette Thuilier participated to the 18th International Conference on HPSG in Seattle, United States.
- Éric de La Clergerie participated to the 1st International "Work-Conference on Linguistics, Biology and Computer Science: Interplays" in Tarragona, Spain.
- Juliette Thuilier participated to the Colloque AFLS in Nancy, France.
- Pierre Magistry participated to ROCLING in Taipei, Taiwan.
- Juliette Thuilier participated to the Conference "Architectures and Mechanisms for Language Process" (AMLAP 2011) in Paris, France.
- Rosa Stern and Benoît Sagot participated to the ATALA workshop on named entities in Paris, France.
- Éric de La Clergerie, Laurence Danlos and Benoît Sagot participated to the Lexis-Grammar Conference in Nicosia, Cyprus.
- Benoît Sagot participated to the 2nd workshop on Systems and Frameworks for Computational Morphology in Zürich, Switzerland.
- Benoît Sagot participated to the 5th Language and Technology Conference in Poznań, Poland.
- Benoît Sagot participated to the European Summer School on Logics, Language and Computation in Ljubljana, Slovenia.
- François Barthelémy participated to the FSMNLP/CIAA (joint events) in Blois, France.
- Laurence Danlos and Charlotte Roze participated to the Muldico Exploratory Workshop "Towards a multilingual database of connectives" in Les Diablerets, Switzerland.
- Laurence Danlos and Charlotte Roze participated to Constraints In Discourse (CID 2011) in Agay, France.
- Laurence Danlos participated to the 13th European Workshop on Natural Language Generation (ENLG) in Nancy, France.

9.3. 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 (half-delegation since September 2011): Information retrieval (2nd year of professional Master, 12h); Clustering and Classification (2nd year of professional Master, 12h); Probabilistic methods for Natural language processing (1st year of Master, 48h); Machine translation (1st year of Master, 48h); Probabilities and statistics for Natural language processing (3rd year of Licence, 24h);

Benoît Crabbé (INRIA delegation until August 2010): 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);

Benoît Sagot: Parsing systems (2nd year of Master, 24h). Introduction to NLP (3rd year of License in Computer Science, 24h).

Pascal Denis: Computational Semantics (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).

Djamé Seddah (half-delegation since September 2011): 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 2011:

PhDs in progress:

- Luc Boruta, *Indicators of Allophony and Phonemehood*, started in September 2009, co-supervized by Emmanuel Dupoux (LSCP/ENS) and Benoît Crabbé
- 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, co-supervized by Pascal Denis
- François-Régis Chaumartin, *Extraction automatisée de connaissances d’une encyclopédie*, started in October 2005, supervized by Sylvain Kahane

- 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 and co-supervized by Pascal Denis
- 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
- Juliette Thuilier, *Contraintes préférentielles et ordre des mots en français*, started in September 2008, supervized by Laurence Danlos and co-supervized by Benoît Crabbé

9.4. PhD committees

+ Benoît Sagot was a member of Ourania Voskaki's PhD Committee (Université Paris-Est Marne-la-Vallée, IGM); the title of her dissertation is "Le lexique-grammaire des verbes du grec moderne — Constructions transitives non locatives à un complément d'objet direct"
+ Laurence Danlos was a referee member and Éric de La Clergerie was a member of Elsa Tolone's PhD Committee (Université Paris-Est Marne-la-Vallée, IGM); the title of her dissertation is "Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français"
+ Laurence Danlos was a referee member of Laurent Kevers' PhD Committee (Université de Louvain La Neuve); the title of his dissertation is "Accès sémantique aux bases de données documentaires: "Techniques symboliques de traitement automatique du langage pour l'indexation the´matique et l'extraction d'informations temporelles"

9.5. Commissions

+ Laurence Danlos was a member of the "Comité de Sélection" for a Full Professor position in Computational Linguistics (CNU 07) at University Paris 3.
+ Benoît Sagot was a member of the "Comités de Sélection" for an Assistant Professor position in Computer Science (CNU 27) at University Paris 13 (LIPN), for an Assistant Professor position in Computer Science or Applied Mathematics (CNU 26/27) at University Paris 4, and 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 at University Paris 13 (LIPN).
+ Benoît Sagot was a member of 4 recruitment committees for PhD, post-doc and engineer positions within the LabEx (excellency cluster) EFL.
+ Benoît Sagot is a member of the Inria's Scientific and Technological Orientation Council (COST-GTRI), and, as such, participated in the evaluation of the Inria Associate Team proposals and of ERCIM post-doc applications.

+ Éric de La Clergerie was a member of to the mid-term Phd defense jury of Karen Fort (LIPN), and participated in the evaluation of the Phd proposal of Silvia Necsulescu (Universitat Pompeu Fabra, Barcelona).
+ Marie Candito was a member of the AERES evaluation committee for the Centre de Recherche en Linguistique et Traitement automatique des Langues–Lucien Tesnière (EA 2283), at Université de Franche-Comté.

9.6. INRIA Evaluation

Alpage was evaluated during the evaluation seminar of the INRIA team "Audio, Speech and Language Processing" which took place at the Mercure Hotel in Rungis, on October 13-14, 2011.

In the "Overall Theme Evaluation" section, that precedes a more detailed analysis of Alpage's achievements, the evaluators consider that Alpage has *continued to publish theoretical contributions in parsing and formal language theory, using statistical and machine learning methods with particular attention to robustness issues in processing unrestricted texts. Alpage is also acting as a developer/providing of open-source NLP resources (lexicons, grammars, etc) for French. These resources are important both for the academic community and for the industrial sector. As a result, ALPAGE enjoys a prominent position in the field, both at the national and international level.*

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 fondé sur les TAG synchrones*, in "Traitement Automatique des Langues", 2009, vol. 50, n^{o} 1.

[6] P. DENIS, B. SAGOT. *Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging*, in "Language Resources and Evaluation", 2012, http://hal.inria.fr/inria-00614819/en.

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

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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

Articles in International Peer-Reviewed Journal

[14] P. DENIS, B. SAGOT. *Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging*, in "Language Resources and Evaluation", 2012, http://hal.inria.fr/inria-00614819/en.

[15] X. TANNIER, P. MULLER. *Evaluating Temporal Graphs Built from Texts via Transitive Reduction*, in "Journal of Artificial Intelligence Research", 2011, n^{o} 40, p. 375–413 [*DOI :* 10.1613/JAIR.3118], http://hal.inria.fr/inria-00602459/en.

[16] 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.

International Conferences with Proceedings

[17] G. ADDA, B. SAGOT, K. FORT, J. MARIANI. *Crowdsourcing for Language Resource Development: Critical Analysis of Amazon Mechanical Turk Overpowering Use*, in "LTC 2011 : Proceedings of the 5th Language and Technology Conference", Poznan, Pologne, November 2011, http://hal.archives-ouvertes.fr/hal-00648187/en/.

[18] M. APIDIANAKI. *Unsupervised Cross-Lingual Lexical Substitution*, in "EMNLP 2011 Workshop on Unsupervised Learning in NLP", Edimbourg, United Kingdom, July 2011, 11 p., http://hal.inria.fr/hal-00607671/en.

[19] M. APIDIANAKI, T. VAN DE CRUYS. *A Quantitative Evaluation of Global Word Sense Induction*, in "CICLing'11 - 12th International Conference on Intelligent Text Processing and Computational Linguistics", Tokyo, Japan, Springer, February 2011, vol. 6608, p. 253–264 [*DOI :* 10.1007/978-3-642-19400-9_20], http://hal.inria.fr/hal-00607673/en.

[20] F. BARTHÉLEMY, S. ROSMORDUC. *Intersection of Multitape Transducers vs. Cascade of Binary Transducers: The Example of Egyptian Hieroglyphs Transliteration*, in "Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing", Blois, France, Association for Computational Linguistics, July 2011, p. 74–82, http://www.aclweb.org/anthology/W11-4410.

[21] A. BITTAR, P. AMSILI, P. DENIS. *French TimeBank : un corpus de référence sur la temporalité en français*, in "TALN 2011 - Traitement Automatique des Langues Naturelles", Montpellier, France, M. LAFOURCADE, V. PRINCE (editors), Laboratoire d'Informatique de Robotique et de Microélectronique, June 2011, vol. 1, p. 259-270, http://hal.inria.fr/inria-00606633/en.

[22] 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.

[23] L. BORUTA. *Combining Indicators of Allophony*, in "ACL-HLT 2011 - The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies", Portland, United States, June 2011, http://hal.inria.fr/inria-00605804/en.

[24] 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/en.

[25] 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/en/.

[26] L. DANLOS. *Analyse discursive et informations de factivité*, in "TALN", Montpellier, France, June 2011, http://hal.inria.fr/inria-00598880/en.

[27] L. DANLOS, F. MEUNIER, V. COMBET. *EasyText: an Operational NLG System*, in "ENLG 2011, 13th European Workshop on Natural Language Generation", Nancy, France, September 2011, http://hal.inria.fr/inria-00614760/en.

[28] L. DANLOS, O. RAMBOW. *Discourse Relations and Propositional Attitudes*, in "CID 2011 - Fourth International workshop on Constraints in Discourse", Agay-Roches Rouges, France, September 2011, http://hal.inria.fr/inria-00614763/en.

[29] P. DENIS, P. MULLER. *Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition*, in "IJCAI-11 - International Joint Conference on Artificial Intelligence", Barcelone, Spain, 2011, http://hal.inria.fr/inria-00614765/en.

[30] 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.

[31] E. HENESTROZA ANGUIANO, P. DENIS. *FreDist: Automatic construction of distributional thesauri for French*, in "TALN - 18ème conférence sur le traitement automatique des langues naturelles", Montpellier, France, France, 2011, p. 119–124, http://hal.inria.fr/hal-00602004/en.

[32] E. LASSALLE, P. DENIS. *Combining meronym discovery methods for bridging resolution in French*, in "Discourse Anaphora and Anaphor Resolution Colloquium", Faro, Portugal, 2011, http://hal.inria.fr/inria-00614785/en.

[33] P. MAGISTRY, B. SAGOT. *Segmentation et induction de lexique non-supervisées du mandarin*, in "TALN'2011 - Traitement Automatique des Langues Naturelles", Montpellier, France, ATALA, June 2011, http://hal.inria.fr/inria-00605899/en.

[34] F. MEUNIER, L. DANLOS, V. COMBET. *EasyText : : un système opérationnel de génération de textes*, in "TALN 2011 - Traitement Automatique des Langues Naturelles 2011", Montpellier, France, June 2011, http://hal.inria.fr/inria-00607708/en.

[35] Y. NAKAMURA-DELLOYE. *Extraction non-supervisée de relations basée sur la dualité de la représentation*, in "TALN 2011 : Traitement Automatique des Langues Naturelles)", Montpellier, France, June 2011, 6, http://hal.inria.fr/hal-00601798/en.

[36] Y. NAKAMURA-DELLOYE. *Named entity extraction for ontology enrichment*, in "IPSJ Special Interest Group - Information Fundamentals and Access Technologies (IFAT)", Japan, August 2011, 1, http://hal.inria.fr/hal-00606077/en.

[37] Y. NAKAMURA-DELLOYE, R. STERN. *Extraction de relations et de patrons de relations entre entités nommées en vue de l'enrichissement d'une ontologie*, in "TOTh 2011 : Terminologie & Ontologie : Théories et Applications", Annecy, France, May 2011, 50, http://hal.inria.fr/hal-00601801/en.

[38] C. ROZE. *Towards a Discourse Relation Algebra for Comparing Discourse Structures*, in "Proceedings of Constraints In Discourse (CID 2011)", Agay, France, September 2011, p. 1-7, http://hal.archives-ouvertes.fr/hal-00655825/en/.

[39] B. SAGOT, D. FIŠER. *Extending wordnets by learning from multiple resources*, in "Language and Technology Conference", Poznań, Pologne, November 2011, http://hal.inria.fr/hal-00655785/en/.

[40] B. SAGOT, D. FIŠER. *Automatic Extension of WOLF*, in "Global Wordnet Conference 2012", Matsue, Japon, Global Wordnet Association + Toyohashi University of Technology + National Institute of Japanese Language and Linguistics, January 2012, PHC PROTEUS 22718UC, http://hal.inria.fr/hal-00655774/en/.

[41] B. SAGOT, K. FORT, G. ADDA, J. MARIANI, B. LANG. *Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé*, in "TALN'2011 - Traitement Automatique des Langues Naturelles", Montpellier, France, 2011, http://hal.inria.fr/inria-00617067/en.

[42] 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.

[43] 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.

[44] P. SAMVELIAN, L. DANLOS, B. SAGOT. *On the predictability of light verbs*, in "30th International Conference on Lexis and Grammar", Nicosia, Cyprus, 2011, http://hal.inria.fr/inria-00617506/en.

[45] D. SEDDAH, J. LE ROUX, B. SAGOT. *Towards Using Data Driven Lemmatization for Statistical Constituent Parsing of Italian*, in "EVALITA 2012", Roma, Italy, December 2011, http://hal.inria.fr/hal-00660495.

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

[47] J. THUILIER. *Case Suffixes and Postpositions in Hungarian*, in "Proceedings of the 18th International Conference on Head-Driven Phrase Structure Grammar", Seattle, États-Unis, C. PUBLICATIONS (editor), November 2011, p. 209-226, http://hal.archives-ouvertes.fr/hal-00643111/en/.

[48] E. TOLONE, B. SAGOT. *Using Lexicon-Grammar Tables for French Verbs in a Large-Coverage Parser*, in "LTC 2009 - 4th Language and Technology Conference", Poznań, Poland, Z. VETULANI (editor), Lecture Notes in Artificial Intelligence, Springer, 2011, vol. 6562, p. 183-191 [*DOI :* 10.1007/978-3-642-20095-3_17], http://hal.inria.fr/inria-00607488/en.

[49] E. TOLONE, É. VILLEMONTE DE LA CLERGERIE, B. SAGOT. *Évaluation de lexiques syntaxiques par leur intégration dans l'analyseur syntaxique FRMG*, in "30th International Conference on Lexis and Grammar", Nicosia, Cyprus, 2011, http://hal.inria.fr/inria-00616694/en.

[50] T. VAN DE CRUYS, M. APIDIANAKI. *Latent Semantic Word Sense Induction and Disambiguation*, in "ACL HLT 2011 - 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies", Portland, Oregon, United States, June 2011, p. 1476–1485, http://hal.inria.fr/hal-00607672/en.

[51] G. WALTHER, L. NICOLAS. *Enriching Morphological Lexica through Unsupervised Derivational Rule Acquisition*, in "WoLeR 2011 at ESSLLI : International Workshop on Lexical Resources", Ljubljana, Slovenia, August 2011, WoLeR 2011 is endorsed by FlaReNet, and supported by the Alpage team and the EDyLex French national grant (ANR-09-CORD-008)., http://hal.inria.fr/inria-00617064/en.

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

National Conferences with Proceeding

[53] L. DANLOS, C. ROZE. *Traduction (automatique) des connecteurs de discours*, in "TALN (Traitement Automatique des Langues Naturelles)", Montpellier, France, June 2011, http://hal.inria.fr/inria-00602555/en.

[54] C. ROZE. *Vers une algèbre des relations de discours pour la comparaison de structures discursives*, in "Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)", Montpellier, France, June 2011, http://hal.inria.fr/inria-00602546/en.

Conferences without Proceedings

[55] Y. NAKAMURA-DELLOYE. *Panorama des grammaires japonaises et défi à la tradition*, in "Colloque " Comment peut-on écrire une grammaire ? "", Montpellier, France, January 2011, 11 pages, http://hal.inria.fr/hal-00606072/en.

[56] 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.

Research Reports

[57] L. BORUTA. *A note on the generation of allophonic rules*, INRIA, January 2011, n^{o} RT-0401, http://hal.inria.fr/inria-00559270/en.

Scientific Popularization

[58] B. SAGOT. *Ressources lexicales libres pour le français*, in "Culture & Recherche", 2011, vol. 124, 53, http://hal.inria.fr/inria-00617066/en.

References in notes

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

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

[61] 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/.

[62] F. BARTHÉLEMY. *A Testing Framework for Finite-State Morphology*, in "Proceedings of the 14th International Conference on Implementation and Application of Automata (CIAA'09)", Sydney, Australia, LNCS, 2009, vol. 5642, p. 75-83.

[63] 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.

[64] J. BOURDAILLET, S. HUET, P. LANGLAIS, G. LAPALME. *TransSearch: from a bilingual concordancer to a translation finder*, in "Machine Translation", 2010, vol. 24 (3-4), p. 241-271.

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

[66] 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.

[67] 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.

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

[69] 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/.

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

[71] 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.

[72] 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.

[73] 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.

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

[75] 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.

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

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

[78] G. CLAIRE, H. MANUÉLIAN. *Création d'un corpus annoté pour le traitement des descriptions définies*, in "Traitement Automatique des Langues", 2005, vol. 46, n^{o} 1.

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

[80] 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.

[81] 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.

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

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

[84] L. DANLOS, B. SAGOT. *Constructions pronominales dans Dicovalence et le lexique-grammaire–intégration dans le Lefff*, in "Proceedings of the 27th Lexicon-Grammar Conference", L'Aquila, Italie, 2008, http://hal.inria.fr/inria-00524741/en/.

[85] L. DANLOS, B. SAGOT. *Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français*, in "Proceedings of the workshop "Lexicographie et informatique : bilan et perspectives"", Nancy, France, 2008, http://hal.inria.fr/inria-00524742/en/.

[86] L. DANLOS, B. SAGOT. *Constructions pronominales dans Dicovalence et le lexique-grammaire – Intégration dans le Lefff*, in "Lingvisticae Investigationes", 2009, vol. 32, n^{o} 2, p. 293-304 [*DOI :* 10.1075/LI.32.2.11DAN], http://hal.inria.fr/inria-00515459/en/.

[87] L. DANLOS, B. SAGOT, R. STERN. *Analyse discursive des incises de citation*, in "2ème Congrès Mondial de Linguistique Française - CMLF 2010", États-Unis La Nouvelle Orléans, Institut de Linguistique Française, 2010, http://hal.inria.fr/inria-00511397/en.

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

[89] 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.

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

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

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

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

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

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

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

[97] 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.

[98] 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.

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

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

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

[102] 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.

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

[104] S. PETROV, 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.

[105] M. POESIO, R. MEHTA, A. MAROUDAS, J. HITZEMAN. *Learning to resolve bridging references*, in "Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics", Stroudsburg, PA, USA, ACL '04, Association for Computational Linguistics, 2004, http://dx.doi.org/10.3115/1218955.1218974.

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

[107] 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.

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

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

[110] 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.

[111] B. SAGOT, L. DANLOS. *Verbes de citation et Tables du Lexique-Grammaire*, in "International Conference on Lexis and Grammar", Serbie Belgrade, Sep 2010, http://hal.inria.fr/inria-00521229/en.

[112] B. SAGOT, L. DANLOS, R. STERN. *A Lexicon of French Quotation Verbs for Automatic Quotation Extraction*, in "7th international conference on Language Resources and Evaluation - LREC 2010", Malte Valetta, 2010, http://hal.inria.fr/inria-00515461/en.

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

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

[115] 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.

[116] 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.

[117] 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.

[118] 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.

[119] B. SAGOT, E. TOLONE. *Intégrer les tables du Lexique-Grammaire à un analyseur syntaxique robuste à grande échelle*, in "Actes de la session poster de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)", Senlis, France, June 2009, electronic version (10 pp.), http://hal.archives-ouvertes.fr/hal-00461893/en/.

[120] 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.

[121] 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.

[122] 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.

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

[124] 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.

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

[126] 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.

[127] 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.

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

[129] 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/.

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

[131] 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.

[132] G. WALTHER. *Measuring Morphological Complexity*, in "Linguistica", 2012, vol. 51, Internal and External Boundaries of Morphology. À paraître.

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

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

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

[136] K. VAN DEN EYNDE, P. MERTENS. *Le dictionnaire de valence* DICOVALENCE *: manuel d'utilisation*, 2006, http://bach.arts.kuleuven.be/dicovalence/manuel_061117.pdf.