IN PARTNERSHIP WITH:
Université Denis Diderot (Paris 7)

Activity Report 2013

Project-Team ALPAGE

Large-scale deep linguistic processing

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

RESEARCH CENTER
Paris - Rocquencourt

THEME
Language, Speech and Audio

Table of contents

Project-Team ALPAGE

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

ALPAGE is a joint team with University Paris–Diderot (Paris 7). It was created on July 1st, 2007 as a team, on January 1st, 2008 as an Inria project-team, and became an UMR-I on January 1st, 2009 (UMR-I 001).

Creation of the Project-Team: 2008 January 01.

1. Members

Research Scientists

Pierre Boullier [Inria, Emeritus Senior Researcher, HdR]
Benoît Sagot [Inria, Researcher]
Eric Villemonte de La Clergerie [Inria, Researcher]

Faculty Members

Laurence Danlos [Team leader, Univ. Paris VII, Professor, HdR]
François Barthélemy [CNAM, Associate Professor]
Marie-Hélène Candito [Univ. Paris VII, Associate Professor, partial delegation (50%) at Inria]
Benoît Crabbé [Univ. Paris VII, Associate Professor]
Djamé Seddah [Univ. Paris IV, Associate Professor, partial delegation (50%) at Inria]

External Collaborators

Juliette Thuilier [Univ. Paris IV, until Jun 2013]
Masoud Alamzadeh [until Jan 2013]

Engineers

Paul Bui Quang [Inria, until Oct 2013]
Margot Colinet [Inria, granted by LabEx EFL via Univ. Paris XIII, from Oct 2013]
Vanessa Combet [Inria, granted by LabEx EFL via Univ. Paris XIII, until Oct 2013]
Mickaël Morardo [Inria]
Virginie Mouilleron [Inria, granted by LabEx EFL via Univ. Paris XIII, until Dec 2013]

PhD Students

Enrique Henestroza Anguiano [until Feb 2013]
Marion Baranes [viavoo, R&D Engineer]
Chloé Braud [Univ. Paris VII]
Marianne Djemaa [Inria, granted by ANR ASFALDA project]
Valérie Hanoka [Verbatim Analysis, supported by a CIFRE grant, until Dec 2013]
Emmanuel Lassalle [ENS Lyon, until Dec 2013]
Pierre Magistry [Univ. Paris VII]
Corentin Ribeyre [Univ. Paris VII]
Raphaël Salmon [Yseop, supported by a CIFRE grant, from Feb 2013]

Post-Doctoral Fellows

Margaret Grant [Univ. Paris VII, granted by LabEx EFL, co-affiliated to the Laboratoire de Linguistique Formelle, until Nov 2013]
Kata Gábor [Inria, granted by "Investissements d'avenir" project PACTE, from Jun 2013]
Julie Hunter [Inria, granted by ANR Polymnie project, from Sep 2013]
Yves Scherrer [Univ. Paris VII, granted by LabEx EFL, until Sep 2013]
Damien Nouvel [Inria, granted by ANR EDyLex project]
Rosa-Devi Stern [Inria, granted by ANR EDyLex project, until Jun 2013]

Administrative Assistant

Christelle Guiziou [Inria]

Others
Sarah Beniamine [Inria, M2 student, May-Jun 2013]
Maximin Coavoux [Inria, M2 student, May-Jun 2013]

2. Overall Objectives

2.1. Overall Objectives

The Alpage team is specialized in **Language modeling**, **Computational linguistics** and **Natural Language Processing (NLP)**. These fields are considered central in the new Inria strategic plan, and are indeed of crucial importance for the new information society. Applications of this domain of research include the numerous technologies grouped under the term of "language engineering". This includes domains such as machine translation, question answering, information retrieval, information extraction, text simplification, automatic or computer-aided translation, automatic summarization, foreign language reading and writing aid. From a more research-oriented point of view, experimental linguistics can be also viewed as an "application" of NLP.

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

Natural language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate. Natural language generation systems convert information from computer databases into human language. Alpage focuses on *text* understanding and generation (by opposition to *speech* processing and generation).

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

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

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

- text normalization, automatic and semi-automatic spelling correction;
- text mining;
- automatic generation;
- with a more long-term perspective, automatic or computer-aided translation.

2.2. Highlights of the Year

2.2.1. Nomination at the Institut Universitaire de France

Laurence Danlos is a Senior Member of the Institut Universitaire de France since October 2013

2.2.2. Statistical Parsing of Morphologically Rich Languages

Since several years, Djamé Seddah, together with Marie-Hélène Candito and more generally the whole Alpage team, has played a major role in setting up and animating an international network of researchers focusing on parsing morphologically rich languages (MRLs).

This year, Djamé Seddah has led the organization of the **first shared task on parsing MRLs**, hosted by the fourth SPMRL workshop [29]. Its primary goal was to bring forward work on parsing morphologically ambiguous input in both dependency and constituency parsing, and to show the state of the art for MRLs. We compiled data for as many as 9 languages, which represents an immense scientific and technical challenge.

Alpage participated to this shared task with two systems. The first one, applied to French only, belongs to the Bonsai series of parsers, adapted in collaboration with the LIGM in order to better deal with multi-word units [19]. It was **ranked first**, and is therefore the best known parser for French to date.

The other Alpage system which took part to this shared task is Éric Villemonte De La Clergerie's new DyALog-based shift-reduced parser [30], which was applied to all 9 languages. It is the **second best system overall**.

3. Research Program

3.1. From programming languages to linguistic grammars

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

Historically, several members of Alpage were originally specialists in the domain of modeling and parsing for programming languages, and have been 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 [49], [79], [86]) 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 [102], [99]. 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

Participants: Djamé Seddah, Marie-Hélène Candito, Benoît Crabbé, Éric Villemonte de La Clergerie, Benoît Sagot, Corentin Ribeyre, Enrique Henestroza Anguiano, Pierre Boullier, Maximin Coavoux.

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 [68] or automatic [74], [75] 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 [58], or probabilistic models derived from lexicalized Tree Adjoining grammars, such as Stochastic Tree Insertion Grammars [56].

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 ...) [108], [73]. 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 [70]. 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. Alpage is the first French team to have turned the French TreeBank into a resource usable for training statistical parsers, to distribute a dependency version of this treebank, and to make freely available various state-of-the art statistical POS-taggers and parsers for French. 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 [46], [45] and derive the best input for syntagmatic statistical parsing [60]. Benchmarking several PCFG-based learning frameworks [11] has led to state-of-the-art results for French, the best performance being obtained with the split-merge Berkeley parser (PCFG with latent annotations) [75].
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 [56] 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 [97].
Pursuing our work on PCFG-LA based parsing, we investigated the automatic conversion of the treebank into dependency syntax representations [53], 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.

3.3. Dynamic wide coverage lexical resources

Participants: Benoît Sagot, Laurence Danlos, Rosa Stern, Valérie Hanoka, 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 [85]. At the syntactic level, work has been achieved on automatic acquisition of atomic syntactic information and automatic detection of errors in the lexicon [109],[10]. At the semantic level, automatic wordnet development tools have been described [77], [103], [65], [64]. 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* [81],[8], developed within the Alexina framework, as well as a wordnet for French, the WOLF [7], the first freely available resource of the kind.

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

Moreover, an increasing effort has been made towards multilingual aspects. In particular, Alexina lexicons developed in 2010 or before exist for Slovak [85], Polish [87], English, Spanish [72], [71] and Persian [91], 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 the key efforts towards deep large-scale linguistic processing, these components need to be included inside a full and robust processing chain, able to handle any text from any source. The development of such linguistic chains, such as SxPipe , is not a trivial task [6]. Moreover, when used as a preliminary step before parsers, the quality of parsers' results strongly depends on the quality of such chains. In that regard, less-standard pre-processings such as word clustering have led to promising results [93].

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

3.5. Discourse structures

Participants: Laurence Danlos, Charlotte Roze.

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

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

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

- first, some connectors (in particular conjunctions of subordination) introduce pure modifiers and must not be considered as bearing discourse relations,
- second, other elements than connectors can lexicalize discourse relations, in particular verbs like *précéder / to precede* or *causer / to cause*, which have facts or fact eventualities as arguments [61].

There are three main frameworks used to model discourse structures: RST, SDRT , and, more recently, the TAG-based formalism D-LTAG. Inside Alpage, Laurence Danlos has introduced D-STAG (Discourse Synchronous TAGs, [62],[5]), which subsumes in an elegant way both SDRT and RST, to the extent that SDRT and RST structures can be obtained by two different partial projections of D-STAG structures. As done in D-LTAG, D-STAG extends a lexicalized TAG analysis so as to deal with the level of discourse. D-STAG has been fully formalized, and is hence possible to implement (thanks to Synchronous TAG, or even TAG parsers), provided one develops linguistic descriptions in this formalism.

4. Application Domains

4.1. Overview

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

4.2. Information extraction and knowledge acquisition

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

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

Knowledge acquisition for a given restricted domain is something that has already been studied by some Alpage members for several years. 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 wires 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 was a CIFRE PhD student at Alpage and at AFP, and worked in 2013 within the ANR project EDyLex).

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* have been co-developed by Verbatim Analysis and by Inria.

In 2013, Verbatim Analysis has bought Inria's part of the intellectual property of the first version of *vera*. A second version has been released, which is co-owned by Verbatim Analysis and Inria.

4.4. Multilingual terminologies and lexical resources for companies

Participants: Éric Villemonte de La Clergerie, Mickaël Morardo.

Lingua et Machina is a small company now headed by François Brown de Colstoun, a former Inria researcher, that provides services for developing specialized multilingual terminologies for its clients. It develops the WEB framework Libellex for validating such terminologies. A formal collaboration with ALPAGE has been set up, with the recruitment of Mikael Morardo in 2012 as an engineer, funded by Inria's DTI. He pursued his work on the extension of the web platform *Libellex* for the visualization and validation of new types of lexical resources. In particular, he has integrated a new interface for handling monolingual terminologies, lexical networks, and bilingual wordnet-like structures, including the WOLF.

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

Participants: Benoît Sagot, Kata Gábor, Éric Villemonte de La Clergerie.

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. Alpage is working within the "Investissements d'avenir" project PACTE, headed by Numen, a company specialized in text digitalization, and three other partners. Kata Gábor is doing a post-doc funded by PACTE (see 6.7)

4.6. Experimental and quantitative linguistics

Participants: Benoît Crabbé, Margaret Grant, Juliette Thuilier, Benoît Sagot.

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 [51], [107], the existence of graded grammaticality judgments [67].

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 [57] gave way on the one hand to more abstract characterizations of the language faculty [57], and on the other hand to the construction of detailed, formally explicit, and often implemented, alternative formulation of the generative approach [50], [76]. For French several grammars have

been implemented in this trend, such as the tree adjoining grammars of [52], [59] 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. In the perspective of better understanding which factors influence the relative ordering of post verbal complements across languages, Meg Grant (post-doc funded by the LabEx EFL), Juliette Thuilier (former PhD at Alpage), Anne Abeillé (LLF) and Benoit Crabbé designed psycholinguistic experiments (questionnaires and recall tasks) with a specific focus on French and on the influence of the animacy factor.

On the other hand we are also collaborating 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.

In parallel, quantitative methods are applied to computational morphology, in collaboration with formal linguists from LLF (CNRS & U. Paris Diderot; Géraldine Walther, Olivier Bonami) and descriptive linguists from CRLAO (CNRS and Inalco; Guillaume Jacques) and HTL (CNRS, U. Paris Diderot and U. Sorbonne Nouvelle; Aimée Lahaussois) — see 6.5.

5. Software and Platforms

5.1. Syntax

Participants: Pierre Boullier [correspondant], Benoît Sagot.

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

The (currently beta) version 6.0 of the SYNTAX system (freely available on Inria GForge) includes various deterministic and non-deterministic CFG parser generators. It includes in particular an efficient implementation of the Earley algorithm, with many original optimizations, that is used in several of Alpage's NLP tools, including the pre-processing chain SXPipe and the LFG deep parser SXLFG. This implementation of the Earley algorithm has been recently extended to handle probabilistic CFG (PCFG), by taking into account probabilities both during parsing (beam) and after parsing (n-best computation). SYNTAX 6.0 also includes parsers for various contextual formalisms, including a parser for Range Concatenation Grammars (RCG) that can be used among others for TAG and MC-TAG parsing.

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

5.2. DyALog

Participant: Éric Villemonte de La Clergerie [maintainer].

DYALOG *on Inria GForge: http://dyalog.gforge.inria.fr/*

DYALOG provides an environment to compile and execute grammars and logic programs. It is essentially based on the notion of tabulation, i.e. of sharing computations by tabulating traces of them. DYALOG is mainly used to build parsers for Natural Language Processing (NLP). It may nevertheless be used as a replacement for traditional PROLOG systems in the context of highly ambiguous applications where sub-computations can be shared.

The current release **1.13.0** of DYALOG is freely available by FTP under an open source license and runs on Linux platforms for x86 and architectures and on Mac OS intel (both 32 and 64bits architectures).

The current release handles logic programs, DCGs (*Definite Clause Grammars*), FTAGs (*Feature Tree Adjoining Grammars*), FTIGs (*Feature Tree Insertion Grammars*) and XRCGs (*Range Concatenation Grammars* with logic arguments). Several extensions have been added to most of these formalisms such as intersection, Kleene star, and interleave operators. Typed Feature Structures (TFS) as well as finite domains may be used for writing more compact and declarative grammars [101].

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 FRMG, 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. [99], [100], and very large amount of data (700 millions of words) in the SCRIBO project. New results concerning FRMG are described in 6.11.

A new statistical dependency parser, based on a shift-reduce algorithm, was also developed in 2013 within the DYALOG system (see 6.12).

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 [12]. 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 Villemonte 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.

The various tools on Metagrammars are available on Inria GForge. FRMG is used directly or indirectly (through a Web service or by requiring parsed corpora) by several people and actions (ANR Rhapsodie, ANR Chronoline, ...)

5.4. The Bonsai PCFG-LA parser

Participants: Marie-Hélène Candito [correspondant], Djamé Seddah, Benoît Crabbé.

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

Alpage has developed as support of the research papers [60], [53], [54], [11] a statistical parser for French, named Bonsai, trained on the French Treebank. This parser provides both a phrase structure and a projective dependency structure specified in [4] as output. This parser operates sequentially: (1) it first outputs a phrase structure analysis of sentences reusing the Berkeley implementation of a PCFG-LA trained on French by Alpage (2) it applies on the resulting phrase structure trees a process of conversion to dependency parses using a combination of heuristics and classifiers trained on the French treebank. The parser currently outputs several well known formats such as Penn treebank phrase structure trees, Xerox like triples and CONLL-like format for dependencies. The parsers also comes with basic preprocessing facilities allowing to perform elementary sentence segmentation and word tokenisation, allowing in theory to process unrestricted text. However it is believed to perform better on newspaper-like text. See 6.12 for recent work and results involving Bonsai.

The parser is available under a GPL license.

5.5. The MICA parser

Participants: Benoît Sagot [correspondant], Pierre Boullier.

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

MICA (Marseille-Inria-Columbia- AT&T) is a freely available dependency parser [48] currently trained on English and Arabic data, developed in collaboration with Owen Rambow and Daniel Bauer (Columbia University) and Srinivas Bangalore (AT&T). MICA has several key characteristics that make it appealing to researchers in NLP who need an off-the-shelf parser, based on Probabilistic Tree Insertion Grammars and on the SYNTAX system. MICA is fast (450 words per second plus 6 seconds initialization on a standard high-end machine) and has close to state-of-the-art performance (87.6% unlabeled dependency accuracy on the Penn Treebank).

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

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

5.6. Alpage's linguistic workbench, including SxPipe

Participants: Benoît Sagot [correspondant], Rosa Stern, Marion Baranes, Damien Nouvel, Virginie Mouilleron, Pierre Boullier, Éric Villemonte de La Clergerie.

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

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

SXPipe [80] is a modular and customizable chain aimed to apply to raw corpora a cascade of surface processing steps. It is used

- as a preliminary step before Alpage's parsers (e.g., FRMG);
- for surface processing (named entities recognition, text normalization, unknown word extraction and processing...).

Developed for French and for other languages, SXPipe includes, among others, various named entities recognition modules in raw text, a sentence segmenter and tokenizer, a spelling corrector and compound words recognizer, and an original context-free patterns recognizer, used by several specialized grammars (numbers, impersonal constructions, quotations...). In 2012, SXPipe has received a renewed attention in four directions:

- Support of new languages, and most notably German (although this is still at a very preliminary stage of development;
- Analysis of unknown words, in particular in the context of the ANR project EDyLex and of the collaboration with *viavoo*; this involves in particular (i) new tools for the automatic pre-classification of unknown words (acronyms, loan words...) (ii) new morphological analysis tools, most notably automatic tools for constructional morphology (both derivational and compositional), following the results of dedicated corpus-based studies (see 6.2 for new results);
- Development of new local grammars for detecting new types of entities and improvement of existing ones, in the context of the PACTE project (see 6.7 for new results).

5.7. MElt

Participants: Benoît Sagot [correspondant], Pierre Magistry.

MElt is a part-of-speech tagger, initially developed in collaboration with Pascal Denis (Magnet, Inria — then at Alpage), which was trained for French (on the French TreeBank and coupled with the Le*fff*), also trained on English [63], Spanish [69], Italian [94], German, Dutch, Polish, Kurmanji Kurdish [104] and Persian [89], [90]. It is state-of-the-art for French.

It is now able to handle noisy corpora (French and English only).

MElt also includes a lemmatization post-processing step.

A specific effort has been made towards the usability of MElt by linguists. In particular, a training session has been organized, and a user guide has been written.

Moreover, a preliminary version of MElt which accepts input DAGs has been developed.

MElt 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 exist, at various stages of development, in particular for Spanish (the Le*ffe*), Polish, Slovak, English, Galician, Persian, Kurdish, Italian,German, as well as for Latin verbs and a subset of Maltese and Khaling verbs. These lexicons are used in various tools, including instances of the MElt POS-tagger, and for studies in quantitative morphology.

Alexina also hosts *Aleda* [98], [88] a large-scale entity database currently developed for French but under development for English, Spanish and German, extracted automatically from Wikipedia and Geonames. It is used among others in the SXPipe processing chain and its NP named entity recognition, as well as in the NOMOS named entity linking system.

5.9. The free French wordnet WOLF

Participants: Benoît Sagot [correspondant], Sarah Beniamine.

The WOLF (Wordnet Libre du Français) is a wordnet for French, i.e., a lexical semantic database. The development of WOLF started in 2008 [82], [83]. At this time, we focused on benefiting from available resources of three different types: general and domain-specific bilingual dictionaries, multilingual parallel corpora and Wiki resources (Wikipedia and Wiktionaries). This work was achieved in a large part in collaboration with Darja Fišer (University of Ljubljana, Slovenia), in parallel with the development of a free Slovene wordnet, sloWNet. However, it was also impacted by specific collaborations, e.g., on adverbial synsets [84].

In 2013, a beta version of a new version of WOLF (version 1.0b2) was published, which integrates and extends the various efforts performed and published somewhat independently in 2012.

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

5.10. OGRE (Optimized Graph Rewriting Engine)

Participants: Corentin Ribeyre [correspondant], Djamé Seddah, Éric Villemonte de La Clergerie, Marie-Hélène Candito.

OGRE (Optimized Graph Rewriting Engine) is a graph rewriting system specifically designed for manipulating linguistic trees and graphs [78]. It relies on a rule specification language for expressing graph rewriting patterns. The transformation is performed in two steps:

1. First, the system performs simple transformations following the rewriting patterns;
2. Second, constraints can be applied on edges, which applies transformations depending on their environment that are propagated while all constraints are satisfied.

The system has been designd for the analysis and manipulation of attributed oriented and multi-relational graphs.

Web site: http://www.corentinribeyre.fr/projects/view/OGRE

5.11. Automatic construction of distributional thesauri

Participants: Marie-Hélène Candito [correspondant], Enrique Henestroza Anguiano.

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

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

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

5.12. Tools and resources for time processing

Participant: Laurence Danlos [correspondant].

Alpage developed the *French TimeBank*, a freely-available corpus annotated with ISO-TimeML-compliant temporal information (dates, events and relations between events) [1].

5.13. LexViz

Participants: Mikael Morardo [maintainer], Éric Villemonte de La Clergerie.

In the context of the industrial collaboration of ALPAGE with the company Lingua & Machina, we have extended their WEB plateform Libellex with a new component used to visualize and collaboratively validate lexical resources. In particular, this extension is used to manage terminological lists and lexical networks. The implemented graph-based representation has proved to be intuitive and quite useful for navigating in such large lexical resources (on the order to 10K to 100K entries).

5.14. Mgwiki

Participants: Paul Bui Quang [maintainer], Éric Villemonte de La Clergerie.

In the context of Inria ADT Mgwiki, Paul Bui Quang has developed a linguistic wiki that may used to discuss linguistic phenomena with the possibility to add annotated illustrative sentences. The work is essentially devoted to the construction of an instance for documenting and discussing FRMG, with the annotations of the sentences automatically provided by parsing them with FRMG. This instance also offers the possibility to parse small corpora with FRMG and an interface of visualization of the results.

Another instance was deployed for managing the annotation guide for the Deep version of the Sequoia treebank, confirming the potential of the notion of linguistic wiki

5.15. NewsProcess

Participants: Éric Villemonte de La Clergerie [maintainer], Damien Nouvel.

NewsProcess is an HTTP-based service that may be used to process AFP news through the Alpage Processing Chain, in order to extract information, in particular citations. The chain has been completed to track the emergence of new words in the news.

In the context on ANR EdyLex, a new version of NewsProcess has been designed for processing AFP news wires and extracting information about unknown words (see 6.2)

5.16. System EasyRef

Participant: Éric Villemonte de La Clergerie [maintainer].

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 the former ANR projectPASSAGE.

EASYREF is maintained under Inria GForge.

6. New Results

6.1. Unsupervised segmentation of Mandarin Chinese

Participants: Pierre Magistry, Benoît Sagot.

In Chinese script, very few symbols can be considered as word boundary markers. The only easily identifiable boundaries are sentence beginnings and endings, as well as positions before and after punctuation marks. Although the script doesn't rely on typography to define (orthographic) "words", a word-level segmentation is often re- quired for further natural language processing, which is a highly non-trivial task.

A great variety of methods have been proposed in the literature, mostly in supervised machine learning settings. Our work addresses the question of unsupervised segmentation, i.e., without any manually segmented training data. Although supervised learning typically performs better than unsupervised learning, we believe that unsuper- vised systems are worth investigating as they require less human labour and are likely to be more easily adaptable to various genres, domains and time periods. They can also provide more valuable insight for linguistic studies.

Amongst the unsupervised segmentation systems described in the literature, two paradigms are often used: Branching Entropy (BE) and Minimum Description Length (MDL). The system we have developed relies on both. We have introduced a new algorithm [22] which searches in a larger hypothesis space using the MDL criterion, thus leading to lower Description Lengths than other previously published systems. Still, this improvement concerning the Description Length does not come with better results on the Chinese word segmentation task, which raises interesting issues. However, it turns out that it is possible to add very simple constraints to our algorithm in order to adapt it to the specificities of Mandarin Chinese in a way that leads to results better than the state-of-the-art on the Chinese word segmentation task.

Moreover, an important part of discrepancies between the various segmentation guidelines concerns the so-called "factoids." This term covers a variety of language phenomena that include: numbers, dates, addresses, email addresses, proper names, and others. We have shown that specific treatment of a subset of such expressions is both sound (as factoids to not resort to general language, which we try and capture with our segmentation model, both rather to conventions that are easy to encode as rules). By augmenting the local grammars of SXPipe to deal with the aforementioned expressionsin Chinese, and use them as a pre-processing for our task, we can discard the matched expressions from the training data and segment them accordingly to the guidelines as a post-processing step. Our results show a significant improvement over previous results.

6.2. Dynamic extension of a French lexical resources based a text stream

Participants: Damien Nouvel, Benoît Sagot, Rosa Stern, Virginie Mouilleron, Marion Baranes.

Lexical incompleteness is a recurring problem when dealing with natural language and its variability. It seems indeed necessary today to regularly validate and extend lexica used by tools processing large amounts of textual data. This is even more true when processing real-time text flows. In this context, we have introduced two series of techniques for addressing words unknown to lexical resources, and applied them to French within the context of the EDyLex ANR project:

- **Extending a morphological lexicon** We have studied neology (from a theoretic and corpus-based point of view) and developed modules for detecting neologisms in AFP news wires in real time and inferring information about them (lemma, category, inflectional class) [24]. We have shown that we are able, using among others modules for analyzing derived and compound neologisms, to generate lexical entries candidates in real time and with a good precision, to be added in the Le*fff* lexicon.
- **Extending an entity database** We have also extended our previous work on named entities detection and linking in order to be able to extract new named entities from AFP news wires and create candidate entries for the *Aleda* entity database.

6.3. Transferring lexical knowledge from a resourced language to a closely-related resource-free language

Participants: Yves Scherrer, Benoît Sagot.

We have developed a generic approach for the transfer of part-of-speech (POS) annotations from a resourced language (RL) towards an etymologically closely related non-resourced language (NRL), without using any bilingual (i.e., parallel) data. We rely on two hypotheses. First, on the lexical level, the two languages share a lot of cognates, i.e., word pairs that are formally similar and that are translations of each other. Second, on the structural level, we admit that the word order of both languages is similar, and that the set of POS tags is identical. Thus, we suppose that the POS tag of one word can be transferred to its translational equivalent in the other language.

The proposed approach consists of two main steps. In the first step, we induce a translation lexicon from monolingual corpora. This step relies on several methods, including a character-based statistical machine translation model to infer cognate pairs, and 3-gram and 4-gram contexts to infer additional word pairs on the basis of their contextual similarity. This step yields a list of $< w_{\mathrm{NRL}}, w_{\mathrm{RL}} >$ pairs. In the second step, the RL lexicon entries are annotated with POS tags with the help of an existing resource, and these annotations are transferred onto the corresponding NRL lexicon entries. We complete the resulting tag dictionary with heuristics based on suffix analogy. This results in a list of $< w_{\mathrm{NRL}}, t >$ pairs, covering the whole NRL corpus.

We have evaluated our methods on several language pairs. We have worked among others onfive language pairs of the Iberic peninsula, where Spanish and Portuguese play the role of RLs: Aragonese–Spanish, Asturian–Spanish, Catalan–Spanish, Galician– Spanish and Galician–Portuguese [27]. We have also conducted experiments on germanic [28] and slavic languages. We have also applied it in a slightly different context, in collaboration with Tomaž Erjavec (IJS, Slovenia), namely that of inducing resources for historical Slovene based on existing resources for contemporary Slovene [26]. Although no direct comparison can be performed, because of the novelty of the task, our results are very satisfying in so far that they are almost as high as published result on a related but simpler task, that of unsupervized part-of-speech tagging — which, contrarily to our work, relies on an existing morphological lexicon for the language at hand.

6.4. Building a large-scale translation graph

Participants: Valérie Hanoka, Benoît Sagot.

Large-scale general-purpose multilingual translation databases are useful in a wide range of Natural Languages Processing (NLP) tasks. This is especially true concerning researches tackling problems specific to under-resourced languages, as translation databases can be used for adapting existing resources in other languages. This has been applied for example for the development of wordnets in languages other than English. There is thus a real need in NLP for *open-source* multilingual lexical databases that compiles as many translations as can be found on any freely available resource in any language.

We have developed, and are about to release, a new open-source heavily multilingual (over 590 languages) translation database built using several sources, namely various wiktionaries and the OPUS parallel corpora.

Our graph was built in several steps. We first extracted a preliminary set of translation and synonym pairs, which we stored in a large translation and synonym graph. We then applied filtering techniques for increasing the accuracy of this graph. We have evaluated the accuracy of our graph as being as high as 98% for translations extracted from wiktionaries.

6.5. Computational morphology

Participant: Benoît Sagot.

In 2013, following previous collaborative work [92], [105], we have designed and developed Alexina$_{\text{PARSLI}}$ in collaboration with Géraldine Walther (LLF and DDL), a formalism for encoding inflectional descriptions (lexicon and grammar) that aims at filling the gap between morphologically and typologically motivated approaches on the one hand and implemented approaches on the other hand, as will be discussed in the remainder of this section. Indeed, Alexina$_{\text{PARSLI}}$ is both:

- an **implementation formalism for PARSLI**, a formal model of inflectional morphology [106] that accounts for concepts underlying the canonical approach of morphological typology;

- an **extension of the Alexina lexical framework** developed at Alpage for modeling lexical information and developing lexical resources. The Alexina framework now supports both morphological grammars that use the original Alexina morphological formalism as well as new grammars developed in Alexina$_{\text{PARSLI}}$.

The Alexina$_{\text{PARSLI}}$ formalism and tools have been proven greatly beneficial to works both in descriptive and formal morphology, in particular in studies about Latin passivisation and Maltese verbal inflection [106] and in studies comparing the compacity of morphological descriptions [106], [92], [105], as well as in NLP, for the efficient development of a large-scale and linguistically sound morphological lexicon for German (a paper describing this new lexicon is to be presented at the LREC 2014 conference).

In collaboration with Géraldine Walther and Guillaume Jacques (CRLAO, CNRS), within operation LR4.11 from strand 6 of the LabEx EFL, we have also developed two Alexina$_{\text{PARSLI}}$ descriptions of (part of) the Khaling (Kiranti, Sino-Tibetan) verbal inflectional system, together with a medium-scale lexicon. Our study shows that an explicit account for the so-called direct-inverse marking, based on concepts developed within PARSLI, allows for a more compact account of this inflectional system [42].

6.6. Extracting Derivational Relations from an Inflectional Lexicon

Participants: Marion Baranes, Benoît Sagot.

Derivational morphological can provide useful information for natural language processing tasks. Indeed, it can improve any application which has to deal with unknown words such as information extraction, spell-checking and others.

We define a morphological family as a set of semantically related lexical entries which differ by their prefix and/or suffix, thus limiting ourselves to concatenative derivational morphology. We shall denote as derivationally related two morphological lexical entries that belong to the same morphological family.

We have developed a system which performs an analogy-based unsupervized extraction of weighted transformation rules that relate derivationally related lexical entries, and use these rules for extracting derivational relations within an existing inflectional lexicon. Our transformation rules can also be used to infer morphological information (both inflectional and derivational) for wordforms unknown to the inflectional lexicon. Our system is language-independent, although restricted to concatenative derivational morphology. We have evaluated it on four languages: English, French, German and Spanish. Our results will be published at the LREC 2014 conference.

6.7. Improving post-OCR correction with shallow linguistic processing

Participants: Kata Gábor, Benoît Sagot.

Providing wider access to national cultural heritage by massive digitalization confronts the actors of the field to a set of new challenges. State of the art optical character recognition (OCR) software currently achieve an error rate of around 1 to 10% depending on the age and the layout of the text. While this quality may be adequate for indexing, documents inteded for reading need to meet higher standards. A reduction of the error rate by a factor of 10 to 100 becomes necessary for the diffusion of digitalized books and journals through emerging technologies such as e-books.

Within the PACTE project, an “Investissements d’avenir” project led by the Numen company, we have worked on the automatic post-processing of digitalized documents in the aim of reducing the OCR error rate by using contextual information and linguistic processing, by and large absent from current OCR engines. At the current stage of the project, we are focusing on French texts coming from the archives of the French National Library (Bibliothèque Nationale de France).

We adopted a hybrid approach, making use of both statistical classification techniques and linguistically motivated modules to detect OCR errors and generate correction candidates. The technology is based on the noisy chanel model, widely used in the field of machine translation and spelling correction and subsequently in OCR post-correction. As to linguistically enhanced models, POS tagging was succesfully applied to spelling correction. However, to our knowledge, little work has been done to exploit linguistic analysis for post-OCR correction.

We have proposed to integrate a shallow processing module to detect certain types of named entities, and a POS tagger trained specifically to deal with NE-tagged input. Our studies demonstrate that linguistically informed processing can efficiently contribute to reduce the error rate by 1) detecting false corrections proposed by the statistical correction module, 2) detecting a certain amount of OCR errors not detected by the statistical correction module.

6.8. Named Entity Linking

Participants: Rosa Stern, Benoît Sagot.

The Ph.D. research work started in 2009 lead in 2013 to the development of a joint entity recognition and linking system for the processing of textual data at the Agence France Presse (AFP).

This system, Nomos, allows to use any existing named entity recognition system, as well as combinations of such systems; their results are passed to a linking module in charge of the association between each detected mention and a unique reference within an existing data inventory. The two tasks (recognition and linking) are jointly operated: the recognition module presents a set of possible detections, which are further disambiguated by the linking module concurrently to the search for the best linking solution to each mention. This joint approach is justified by the need to limit the error propagation between two such modules in a pipeline system.

Experiments were achieved in order to evaluate the performance of Nomos over AFP news wires. They showed that the joint approach, relatively to a purely sequential one, improves the system's global precision, i.e. the linking accuracy as well as the named entity recognition task itself. A gain of 3 points (87,6) is observed for the recognition precision with a low recall loss, while a gain of 8 points (92,9) is observed when several recognition systems are combined - although with a more significant loss of recall.

The Nomos system also allows to anchor of the AFP's textual production in the Linked Data network and the Semantic Web paradigm, since the annotations derived from the entity linking associate each entity to an identified resource in repositories such as Wikipedia, DBPedia, Geonames or the New York Times Linked Data.

6.9. Treebanking at Alpage

Participants: Djamé Seddah, Benoît Sagot, Marie-Hélène Candito, Corentin Ribeyre, Benoît Crabbé, Éric Villemonte de La Clergerie, Virginie Mouilleron, Vanessa Combet.

Since the advents of supervized methods for building accurate statistical parsing models, treebank engineering has become of crucial importance. In fact building a treebank, namely a set of carefully annotated syntactic parses with possibly different annotation layers and covering potentially different text domains, can be seen as providing a parser with both a grammar and a set of probabilities used for disambiguation. The main problem of such approaches lies in the nature of the lexical probabilities: they force the parsing model to be extremely sensitive to its training data and hence limit its performance to some low upper-bound when applied in out-of-domain scenario.

6.9.1. Written French Treebanks

Originating from the merging of two NLP teams specialized in grammar engineering and in which the creation of the first treebank for French was initiated [46], it is no wonder that we decided to increase the coverage of our French Treebank-based parsers by building out-of-domain treebanks: the Sequoia Corpus, [55], [18], made from Europarl, biomedical and wikipedia data, and the French Social Media Bank (outside English, the first data set covering Facebook, Twitter and other social media noisy text data) [95], [96]. We built those two corpus for two purposes: first, we wanted to evaluate the performance of our nlp chains (tokenization, tagging, parsing) on out-of-domain data, being noisy or not ; then we increased the coverage of our French treebank based models by simply adding those new data set to the canonical training set (using of-course many lexical variation, morphological clustering, brown clustering, etc.). We're also on the process of finalizing a new 2600 sentence data set, made essentially of questions, which are strikingly absent from all the treebanks we've been using and developing. So far, only one such data set exist and only for English: the Question-Bank [66]. Our very preliminary results show that simply adding a third of that corpus to the French Treebank greatly improve our parser performance.

Finally, Alpage is leading, in collaboration with the Nancy-based team Calligrame, a project to annotate the Sequoia corpus and the French Treebank with a richer, "deeper" syntactic layer, at the interface between syntax and semantics. A paper describing this effort is to appear at the LREC 2014 conference.

6.9.2. Spoken French Treebank

In collaboration with Anne Abeillé (LLF, CNRS), we have also contributed to the deign of a spoken treebank for French based on data produced in the ANR ETAPE. Contrary to other languages such as English, where spoken treebanks such as the Switchboard corpus treebank (Meteer, 1995), there is no sizable spoken corpus for French annotated for syntactic constituents and grammatical functions. Our project is to build such a resource which will be a natural extension of the Paris 7 treebank (Abeillé et al. 2003) for written French, in order to be able to compare with similar annotations written and spoken French. We have reused and adapted the parser (Petrov et al., 2006) which has been trained on the written treebank, with manual correction and validation. The first results are promising [32].

6.10. Linear time constituent parser

Participant: Benoît Crabbé.

We have designed an efficient and accurate lexicalized LR inspired discriminative parsing algorithm that recasts some current advances in dependency parsing to the constituency setting. We specifically designed and evaluated a Graph Structured Stack-based parser (Huang et al. 2010) using some additional specific approximate inference techniques such as the max violation update for the perceptron (Huang et al. 2012) . By contrast with dependency parsing however, lexicalized constituent parsing raises some additional correctness issues that motivate the explicit use of an LR automata instead of a simpler shift reduce framework.

The parsing model is linear in time and has been evaluated on French data, where it turns out to be state of the art on SPMRL 2013 datasets [29] both in time and in accuracy. The parsing framework has been designed to be further extended with compositional semantic representations and allows in principle an easy integration of ressources — such as those developped in the team — considered to be important for parsing morphologically rich languages.

6.11. Improving FRMG through partially supervised learning

Participant: Éric Villemonte de La Clergerie.

Since the emergence of several statistical parsers for French developed on the French TreeBank (FTB), including those developed at Alpage, it was important to be able to compare the symbolic meta-grammar-based parser FRMG with these statistical parsers on their native treebank, but also possibly to extend the comparison for other treebanks.

A first necessary step in this direction was a conversion from FRMG's native dependency scheme into FTB's dependency scheme, a tedious task highlighting the differences in design at all levels (segmentation, parts of speech, representation of the syntactic phenomena, etc.). A preliminary evaluation has shown that accuracy is good, but largely below the scores reached by the statistical parsers.

A challenge was then to explore if training on the FTB could be used to improve the accuracy of a symbolic parser like FRMG. However, the main difficulty arises from the fact that FTB's dependency scheme has little in common with FRMG's underlying grammar, and that no reverse conversion from FTB to FRMG structures is available. Such a conversion could be investigated but would surely be difficult to develop. Instead, we tried to exploit directly FTB data, using only very minimal assumptions, nevertheless leading to important gains and results close to those obtained by the statistical parsers [31]: it was possible to tune the disambiguisation process of FRMG and strongly increase its accuracy, from 83% up to 87.17% (in terms of CONLL Labeled Attachment Score), a level comparable to those reached by statistical parsers trained on the FTB. Preliminary experiments show that (a) disambiguisation tuning also improve the performances on other corpora and (b) that FRMG seems to be more stable than statistical parsers on corpora other than the FTB. Finer-grained comparison of FRMG wrt statistical parsers have been done that provide some insight for further improvements of FRMG.

The interest is that the technique should be easily adaptable for training data with different annotation schemes. Furthermore, our motivation was not just to improve the performances on the FTB and for the annotation scheme of FTB, for instance by training a reranker (as often done for domain adaptation), but to exploit the FTB to achieve global improvement over all kinds of corpora and for FRMG native annotation scheme.

6.12. Statistical parsing of Morphologically Rich Languages

Participants: Djamé Seddah, Marie-Hélène Candito, Éric Villemonte de La Clergerie, Benoît Sagot.

6.12.1. The SPMRL shared task

Since several years, Djamé Seddah, together with Marie-Hélène Candito and more generally the whole Alpage team, has played a major role in setting up and animating an international network of researchers focusing on parsing morphologically rich languages (MRLs).

In 2013, Djamé Seddah led the organization of the first shared task on parsing MRLs, hosted by the fourth SPMRL workshop and described in a 36-page overview paper that constitutes an in-depth state-of-the-art analysis and review of the domain [29]. The primary goal of this shared task was to bring forward work on parsing morphologically ambiguous input in both dependency and constituency parsing, and to show the state of the art for MRLs. We compiled data for as many as 9 languages, which represents an immense scientific and technical challenge.

6.12.2. DyALog-SR

The SPMRL 2013 shared task was the opportunity to develop and test, with promising results, a simple beam-based shift-reduce dependency parser on top of the tabular logic programming system DYALOG. We used (Huang and Sagae, 2010) as the starting point for this work, in particular using the same simple arc-standard strategy for building projective dependency trees. The parser was also extended to handle ambiguous word lattices, with almost no loss w.r.t. disambiguated input, thanks to specific training, use of oracle segmentation, and large beams. We believe that this result is an interesting new one for shift-reduce parsing.

The current implementation scales correctly w.r.t. sentence length and, to a lesser extent, beam size. Nevertheless, for efficiency reasons, we plan to implement a simple C module for beam management to avoid the manipulation in DyALog of sorted lists. Interestingly, such a module, plus the already implemented model manager, should also be usable to speed up the disambiguation process of DyALog-based TAG parser FRMG (de La Clergerie, 2005a). Actually, these components could be integrated in a slow but on-going effort to add first-class probabilities (or weights) in DyALog, following the ideas of (Eisner and Filardo, 2011) or (Sato, 2008).

6.12.3. The Alpage-LIGM French parser

The second Alpage system that participated to the SPMRL shared task, although on French language only, was developed in collaboration with Mathieu Constant (LIGM), based on the Bonsai architecture. This system is made of several single statistical dependency parsing systems whose outputs are combined into a reparser. We use two types of single parsing architecture: (a) pipeline systems; (b)“joint” systems.

The pipeline systems first perform multi-word expression (MWE) analysis before parsing. The MWE analyzer merges recognized MWEs into single tokens and the parser is then applied on the sentences with this new tokenization. The parsing model is learned on a gold training set where all marked MWEs have been merged into single tokens. For evaluation, the merged MWEs appearing in the resulting parses are expanded, so that the tokens are exactly the same in gold and predicted parses.

The “joint” systems directly output dependency trees whose structure comply with the French dataset annotation scheme. Such trees contain not only syntactic dependencies, but also the grouping of tokens into MWEs, since the first component of an MWE bears dependencies to the subsequent components of the MWE with a specific label. At that stage, the only missing information is the POS of the MWEs, which we predict by applying a MWE tagger in a post-processing step.

This parsing system obtains the best results for French, both for overall parsing and for MWE recognition, using a reparsing architecture that combines several parsers, with both pipeline architecture (MWE recognition followed by parsing), and joint architecture (MWE recognition performed by the parser).

6.13. Towards a French FrameNet

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

The ASFALDA project [1] is a three-year project which started in October 2012, with the objective of building semantic resources (generalizations over predicates and over the semantic arguments of predicates) and a corresponding semantic analyzer for French. We chose to build on the work resulting from the FrameNet project [47], [2] which provides a structured set of prototypical situations, called *frames*, along with a semantic characterization of the participants of these situations (called *frame elements*, FEs). The resulting resources will consist of :

1. a French lexicon in which lexical units are associated to FrameNet frames,
2. a semantic annotation layer added on top of existing syntactic French treebanks
3. and a frame-based semantic analyzer, focused on joint models for syntactic and semantic analysis.

In the first year of the project, we focused on the first of these objectives. A team of 10 active members, from Alpage, the Laboratoire de Linguistique Formelle (LLF), the MELODI team (IRIT - Toulouse) and the CEA-List partners achieved :

- the delimitation and adaptation to French of a set of FrameNet frames, in order to cover a set of specific notional domains (commercial transaction, communication, cognitive positions, judgment/evaluation, temporal relations, spatial position, causality).
- and the semi-automatic construction of a French lexicon in which French lexical units are associated with frames

The current resource contains 110 frames, and roughly 2500 lexical units / frame pairs. The next phase consists in automatic pre-annotation of semantic annotations, that will serve as basis for the manual validation phase.

Note that a publication describing the project and these first achievements shall be presented at the LREC 2014 conference.

6.14. Modelisation of discourse structures with DSTAG

Participant: Laurence Danlos.

This work was done within the ANR Polymnie, in collaboration with Sylvain Pogodalla and Philippe de Groote from LORIA.

Neg-Raising (NR) verbs form a class of verbs with a clausal complement that show the following behavior: when a negation syntactically attaches to the matrix predicate, it can semantically attach to the embedded predicate. Such an implication does not always hold. Some contexts make it impossible to consider the negation as having scope over the embedded predicate only. This corresponds to the non-NR reading of the predicate.

We have developed and published [20] an account of NR predicates within Tree Adjoining Grammars (TAG) that relies on a Montague-like semantics for TAG. The different properties of NR predicates are rendered at different levels: the ambiguity of the readings is modeled by lexical ambiguity; the scoping and cyclicity properties are modeled through the lexical semantics and the higher-order interpretation of adjunction nodes; spurious am- biguities are avoided using fine-grained types for terms representing derivation trees. This provides us with a base layer where to account for interactions with discourse connectives and discourse representation represented in DSTAG.

[1] https://sites.google.com/site/anrasfalda/
[2] https://framenet.icsi.berkeley.edu/

6.15. Annotation of discourse structures on the FTB

Participants: Laurence Danlos, Margot Colinet.

With the aim of annotating the French TreeBank (FTB, already annotated for syntax) with discourse information, we have been working on the first step of the project, namely identify all the occurrences of discourse connectives in the FTB. This raises problems for lexemes which are ambiguous with a discourse usage and other uses. In collaboration with Mathilde Dargnat (ATILF) and Grégoire Winterstein, we have been working on the preposition *pour* (around 1500 occurrences) and the adverb *alors* (300 occurrences). This work is the basis for a future annotation manual.

In parallel, we have been working on adverbial discourse connectives and published on the topic [17]. This paper focuses on the following question: does the only syntactic argument of an adverbial discourse connective correspond to its second semantic argument? It shows that this is not always the case, which is a problem for the syntax-semantics interface. This interface brings us to distinguish two classes of adverbial connectives we sketch the study of.

6.16. Pairwise coreference models

Participant: Emmanuel Lassalle.

In collaboration with Pascal Denis (Magnet, Inria), we have proposed a new method for significantly improving the performance of pairwise coreference models [34]. Given a set of indicators, our method learns how to best separate types of mention pairs into equivalence classes for which we construct distinct classification models. In effect, our approach finds an optimal fea- ture space (derived from a base feature set and indicator set) for discriminating coreferential mention pairs. Although our approach explores a very large space of possible feature spaces, it remains tractable by exploiting the structure of the hierarchies built from the indicators.

In the framework of decision trees, this method can be seen as a pruning procedure and thus can be combined with different methods for expanding a decision tree. It can also be compared to polynomial kernels, but has the advantage of a lower computational complexity [21]. Our experiments on the CoNLL-2012 Shared Task English datasets (gold mentions) indicate that our method is robust relative to different clustering strategies and evaluation metrics, showing large and consistent improvements over a single pairwise model using the same base features. Our best system obtains a competitive 67.2 of average F1 over MUC, B3, and CEAF which, despite its simplicity, places it above the mean score of other systems on these datasets.

6.17. Identification of implicit discourse relations

Participant: Chloé Braud.

In collaboration with Pascal Denis (Magnet, Inria), we have developed a system for identifying "implicit" discourse relations (that is, relations that are not marked by a discourse connective) [33]. Given the little amount of available annotated data for this task, our system also resorts to additional automatically labeled data wherein unambiguous connectives have been suppressed and used as relation labels, a method introduced by Marcu and Echihabi (2002). As shown by Sporleder and Lascarides (2008) for English, this approach doesn't generalize well to implicit relations as annotated by humans. We have shown that the same conclusion applies to French due to important distribution differences between the two types of data. In consequence, we propose various simple methods, all inspired from work on domain adaptation, with the aim of better combining annotated data and artificial data. We have evaluated these methods through various experiments carried out on the ANNODIS corpus: our best system reaches a labeling accuracy of 45.6%, corresponding to a 5.9% significant gain over a system solely trained on manually labeled data.

7. Bilateral Contracts and Grants with Industry

7.1. Contracts with Industry

Alpage has developed several collaborations with industrial partners. Apart from grants described in the next section, specific collaboration agreements have been set up with the following companies:

- Verbatim Analysis (license agreement, transfer agreement, "CIFRE" PhD, see section 4.3),
- Lingua et Machina (DTI-funded engineer, see section 4.4), Viavoo,
- Yseop ("CIFRE" PhD of Raphael Salmon which started in 2012 on automatic text generation)
- CEA-List ("CIFRE" PhD of Quentin Pradet on the development of lexical resources which help annotating semantic roles; e.g., development of a French VerbNet)

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-Hélène Candito, Benoît Crabbé, Pascal Denis, Charlotte Roze, Pierre Magistry, Djamé Seddah, Juliette Thuilier, Éric Villemonte de La Clergerie.

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

Alpage is a very active member of the LabEx EFL together with other linguistic teams we have been increasingly collaborating with: LLF (University Paris 7 & CNRS) for formal linguistics, LIPN (University Paris 13 & CNRS) for NLP, LPNCog (University Paris 5 & CNRS) LSCP (ENS, EHESS & CNRS) for psycholinguistics, MII (University Paris 4 & CNRS) for Iranian and Indian studies. Alpage resources and tools have already proven relevant for research at the junction of all these areas of linguistics, thus drawing a preview of what the LabEx is about: experimental linguistics (see Section 4.6). Moreover, the LabEx provides Alpage with opportunities for collaborating with new teams, e.g., on language resource development with descriptive linguists (see 6.5 for example). In 2013, two post-docs funded by Labex EFL have worked at Alpage (Yves Scherrer) or jointly at Alpage and LLF (Margaret Grant).

Benoît Sagot is the head one of the 7 autonomous scientific "strands" of the LabEx EFL, namely the strand 6 on "Language Resources". Marie-Hélène Candito and Benoît Crabbé are respectively deputy-head of strands 5 on "Computational semantic analysis" and 2 on "Experimental grammar from a cross-linguistic perspective". Several project members are in charge of research operations within these 3 strands.

8.2. National Initiatives

8.2.1. ANR

8.2.1.1. ANR project ASFALDA (2012 – 2015)

Participants: Marie-Hélène Candito [principal investigator], Marianne Djemaa, Benoît Sagot, Éric Villemonte de La Clergerie, Laurence Danlos.

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

The project aims to provide both a French corpus with semantic annotations and automatic tools for shallow semantic analysis, using machine learning techniques to train analyzers on this corpus. The target semantic annotations are structured following the FrameNet framework [47] and can be characterized roughly as an explicitation of "who does what when and where", that abstracts away from word order / syntactic variation, and to some of the lexical variation found in natural language.

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

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

The project is structured as follows:

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

The scientific key aspects of the project are:

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

8.2.1.2. ANR project EDyLex (2010 – 2013)

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

EDyLex was an ANR project (STIC/CONTINT) headed by Benoît Sagot, which came to an end on June 30, 2013. The focus of the project was the dynamic acquisition of new entries in existing lexical resources that are used in syntactic and semantic parsing systems: how to detect and qualify an unknown word or a new named entity in a text? How to associate it with phonetic, morphosyntactic, syntactic, semantic properties and information? Various complementary techniques will be explored and crossed (probabilistic and symbolic, corpus-based and rule-based...). Their application to the contents produced by the AFP news agency (Agence France-Presse) constitutes a context that is representative for the problems of incompleteness and lexical creativity: indexing, creation and maintenance of ontologies (location and person names, topics), both necessary for handling and organizing a massive information flow (over 4,000 news wires per day).

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

In 2013, several important developments have been achieved:

- Finalization of a beta version of the first non-alpha release of the WOLF (Free French WordNet)
- Improvement or development of modules for automatic detection, classification and morphological analysis of unknown words (neologisms, new named entities) in French corpora and integration within a full-featured processing pipeline (see 6.2);
- Collaboration with Vocapia for interfacing the results of this pipeline with Vocapia's language models, in order to improve speech recognition systems used at AFP;
- Use of an EDyLex-specific version of the NewsProcess architecture, previously developed at Alpage, for meeting the expectations of the EDyLex project in terms of lexicon extension from dynamic corpora, here AFP news wires.

8.2.1.3. ANR project Polymnie (2012-2015)

Participants: Laurence Danlos, Éric Villemonte de La Clergerie.

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

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

The role of Alpage in this project is to develop sentential or discursive grammars written in TAG so as to study their conversion in ACG. First results achieved in 2013 are described in 6.14.

8.2.2. Other national initiatives

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

Participants: Benoît Sagot, Kata Gábor.

PACTE (*Projet d'Amélioration de la Capture TExtuelle*) is an "Investissements d'Avenir" project sumbitted within the call "Technologies de numérisation et de valorisation des contenus culturels, scientifiques et éducatifs". It started in November 2012, although the associated fundings only arrived at Alpage in July 2013.

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

PACTE involves 3 companies (Numen, formerly 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.

The results obtained at Alpage in 2013 within PACTE are described in 6.7

8.2.3. Consortium Corpus Écrits within the TGIR Huma-Num

Participants: Benoît Sagot, Djamé Seddah.

Huma-Num is a TGIR (Very Large Research Infrastructure) dedicated to digital humanities. Among Huma-Num initiatives are a dozen of consortia, which bring together most members of various research communities. Among them is the *Corpus Écrits* consortium, which is dedicated to all aspects related to written corpora, from NLP to corpus development, corpus specification, standardization, and others. All types of written corpora are covered (French, other languages, contemprorary language, medieval language, specialized text, non-standard text, etc.). The consortium Corpus Écrits is managed by the Institut de Linguistique Française, a CNRS federation of which Alpage is a member since June 2013, under the supervision of Franck Neveu.

Alpage is involved in various projects within this consortium, and especially in the development of corpora for CMC texts (blogs, forum posts, SMSs, textchat...) and shallow corpus annotation, especially with MElt.

8.3. European Initiatives

8.3.1. Collaborations in European Programs, except FP7

Program: COST

Project acronym: PARSEME

Project title: Parsing and multi-word expressions. Towards linguistic precision and computational efficiency in natural language processing

Duration: 03/2013- 03/2017

Coordinator: Agata SAVARY

Other partners: 24 participating countries

Abstract: This Action aims at increasing and enhancing the support of the European multilingual heritage from Information and Communication Technologies (ICT). This general aim is addressed through improving linguistic representativeness, precision and computational efficiency of Natural Language Processing (NLP) applications. The Action focuses on the major bottleneck of these applications: Multi-Word Expressions (MWEs), i.e. sequences of words with unpredictable properties such as “to count somebody in” or “to take a haircut.” A breakthrough in their modelling and processing can only result from a coordinated effort of multidisciplinary experts in different languages. COST is the most adequate framework answering this need. Fourteen European languages will be addressed from a cross-theoretical and cross-methodological perspective, necessary for coping with current fragmentation issues. Expected deliverables include enhanced language resources and tools, as well as recommendations of best practices for cutting-edge MWE-aware language models. The Action will lead to a better understanding of the nature of MWEs. It will establish a long-lasting collaboration within a multilingual network of MWE specialists. It will pave the way towards competitive next generation text processing tools which will pay greater attention to language phenomena.

8.4. International Initiatives

8.4.1. Inria International Partners

8.4.1.1. Informal International Partners

Alpage has active collaborations with several international teams. The most active in 2013 have been:

- collaboration with Columbia University (United States), in particular on discourse modeling (Laurence Danlos, with Owen Rambow) and on computational morphology (Benoît Sagot, with Owen Rambow and Nizar Habash)
- collaboration with the Weizmann Institute of Science (Israel) on parsing morphologically rich languages (Djamé Seddah, with Reut Tsarfaty)
- collaboration with the Indiana University (United States) on parsing morphologically rich languages (Djamé Seddah, with Sandra Kubler)
- collaboration with the Uppsala University (Sweden) on statistical parsing (Marie-Hélène Candito and sDjamé Seddah, with Joakim Nivre)

9. Dissemination

9.1. Scientific Animation

- Alpage members were involved in many Program, Scientific or Reviewing Committees for other journals and conferences. For example, Éric Villemonte De La Clergerie participated to the program committees of IWPT'13, LGC'13, HMGE'13, DepLing'13, EPIA'13 TeMA, AAAI'13 Student workshop, plus reviewing for ACL'13, NAACL'13, and TALN'13; Djamé Seddah was a PC member of SPMRL'13, EMNLP, IJCNLP, CICLing, TALN, IWPT; Laurence Danlos was a reviewer for ACL, the Generative Lexicon Workshop, TALN, and the TAL journal; Benoît Crabbé was a reviewer for Formal Grammar, CSSP, TALN, RECITAL, Dependency Linguistics, Journal of Language Modelling, Language Ressources and Evaluation; Benoît Sagot was a Program Committee member or reviewer for the LRE and for the TAL journal (vol. 54:1 and 54:2), as well as for the conferences TALN, ACL, SFCM, IJCNLP and SPMRL
- Participation of Éric Villemonte De La Clergerie to the editorial board of French journal T.A.L. as "Redacteur en chef" (chief editor)
- Benoît Sagot is elected board member of the French NLP society (ATALA) and was its Secretary until June 2013.
- Benoît Sagot is a member of the scientific board of the consortium Corpus Écrits, which belongs to the TGIR Huma-Num
- Benoît Sagot represents Alpage at the board of the Institut de Linguistique Française, a CNRS federation (Alpage is a member of the Institut de Linguistique Française since June 2013)
- 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 UFR of Linguistics of University Paris Diderot.
- Benoît Crabbé and Laurence Danlos are members of the Administrative board of the UFR of Linguistics of University Paris Diderot.
- Benoît Crabbé co-organized the research seminar : lectures in experimental linguistics (Univ P7)
- Benoît Crabbé is responsible for the L3 computational linguistics (Univ Paris Diderot)
- Benoît Sagot is a member of the Governing Board and of the Scientific Board of the **LabEx EFL**, as head of the research strand on language resources; he is also in charge of several reserach operations; Benoît Crabbé is deputy-head of the research strand on experimental grammar;Marie-Hélène Candito and Laurence Danlos are in charge of one research operation each. Laurence Danlos is a member of the Scientific Board, representing Alpage.

- Djamé Seddah is one of the founders of the statistical parsing of morphologically rich language initiative (**SPMRL**) that started during IWPT'09. He was the program co-chair of the successful SPMRL 2010 NAACL-HLT Workshop and of its 2011 (at IWPT) and 2012 (at ACL) editions. In 2013, he was chair of the Shared Task that was organized together with this year's edition of SPMRL, for which he, Marie-Hélène Candito, Benoît Crabbé and Benoît Sagot were member of the program committee. Finally, Alpage is a major sponsor of this series of workshops.
- Marie-Hélène Candito organized the Alpage research seminar.
- Benoît Sagot served as a project reviewer for the National Research Agency (ANR), calls "bilatéraux SHS" and "STIC".
- Djamé Seddah is the "Responsable pédagogique" for the C2I courses for Université Paris-Sorbonne
- Benoît Crabbé tought an invited course on Experimental Grammar at the Summer School on Empirical Linguistics at the Berder island (France), organized by the LabEx EFL (international audience)
- Benoît Sagot was an invited speaker at the workshop on Computational Approaches to Morphological Complexity organized by the Surrey Morphology Group in Paris
- many members of the team were invited to give talks in France and aborad, as for example in Geneva (Marie-Hélène Candito), Düsseldorf (Djamé Seddah, Benoît Sagot), Montreal (Djamé Seddah, Rosa Stern), Columbia University (Benoît Sagot).

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Laurence Danlos, Introduction to NLP, 28h, L3, Université Paris-Diderot, France

Laurence Danlos, Discourse, NLU and NLG, 28h, M2, Université Paris-Diderot, France

Marie-Hélène Candito, Information retrieval, 12h, M2, Université Paris-Diderot, France

Marie-Hélène Candito, Clustering and Classification, 12h, M2, Université Paris-Diderot, France

Marie-Hélène Candito, Automatic semantic analysis, 12h, M2, Université Paris-Diderot, France

Marie-Hélène Candito, Machine translation, 48h, M1, Université Paris-Diderot, France

Benoît Crabbé, Linguistic data analysis, 24h, M2, Université Paris-Diderot, France

Benoît Crabbé, Introduction to computer science, 24h, L3, Université Paris-Diderot, France

Benoît Crabbé, Introduction to Statistics and Probability, 24h, L3, Université Paris-Diderot, France

Benoît Crabbé, Probabilistic methods for NLP, 48h, M1, Université Paris-Diderot, France

Benoît Crabbé, Introduction to Computational Linguistics, 24h, L3, Université Paris-Diderot, France

Benoît Sagot, Natural Language Parsing, 28h, M2, Université Paris-Diderot, France

Djamé Seddah, Linguistic Models and Computational Linguistics, 38h, M1, Université Paris-Sorbonne, France

Djamé Seddah, Machine Translation, 30h, M2, Université Paris-Sorbonne, France

Djamé Seddah, C2I, 30h, L2, Université Paris-Sorbonne, France

9.2.2. Supervision

Charlotte Roze, *Vers une algèbre des relations de discours*, Université Paris-Diderot, May 22, PhD thesis supervized by Laurence Danlos, co-supervized by Philippe Muller

Enrique Henestroza Anguiano, *Efficient Large-Context Dependency Parsing and Correction with Distributional Lexical Resources*, Université Paris-Diderot, June 27, PhD thesis supervized by Laurence Danlos, co-supervized by Alexis Nasr and Marie-Hélène Candito

Rosa Stern, *Identification automatique d'entités pour l'enrichissement de contenus textuels*, Université Paris-Diderot, June 28, PhD thesis supervized by Laurence Danlos, co-supervized by Benoît Sagot

Pierre Magistry, *Unsupervised Word Segmentation and Wordhood Assessment: the case for Mandarin Chinese*, Université Paris-Diderot, December 19, PhD thesis supervized by Sylvain Kahane, co-supervized by Benoît Sagot and Marie-Claude Paris

Corentin Ribeyre, *vers la syntaxe profonde pour l'interface syntaxe-sémantique*, started in November 2012, supervized by Laurence Danlos, co-supervized by Djamé Seddah and Éric Villemonte De La Clergerie.

Isabelle Dautriche, *Exploring early syntactic acquisition: a experimental and computational approach*, started in September 2012, co-supervized by Benoît Crabbé.

Marion Baranes, *Correction orthographique contextuelle de corpus multilingues et multicanaux*, started in January 2012, supervized by Laurence Danlos, co-supervized by Benoît Sagot, in collaboration with the *viavoo* company.

Marianne Djemaa, *Création semi-automatique d'un FrameNet du français, via interface syntaxe-sémantique*, started in October 2012, supervized by Laurence Danlos, co-supervised by Marie-Hélène Candito

Chloé Braud, *Développement d'un système complet d'analyse automatique du discours à partir de corpus annotés, bruts et bruités*, started in September 2011, supervized by Laurence Danlos, co-supervized by Pascal Denis

Valérie Hanoka, *Construction semi-automatique de réseaux lexicaux spécialisés multilingues*, started in January 2011, supervized by Laurence Danlos, co-supervized by Benoît Sagot

Emmanuel Lassalle, *Résolution automatique des anaphores associatives*, started in September 2010, supervised by Laurence Danlos, co-supervized by Pascal Denis

9.2.3. Juries

We do not mention in this section participations to PhD defense committees as supervisor or co-supervisor.

- Laurence Danlos served in the PhD defense committe of Noémie-Fleur Sandillon-Rezer. Field: computer science. Title: *Acquisition de grammaires catégorielles à partir de corpus annotés*. Other members of the committee: Christian Rétoré (U. Bordeaux, supervisor), Richard Moot (CNRS, co-supervisor), Mark Steedman (University of Edinburgh, reviewer), Annie Forêt (U. Rennes, reviewer), Alexis Nasr (U. de la Méditerrannée), Géraud Sénizergues (U. Bordeaux), Tim van de Cruys (CNRS).
- Marie-Hélène Candito served in the PhD defense committe of Assaf Urieli. Field: linguistics. Title: *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Other members of the committee: Ludovic Tanguy (CNRS, supervisor), Alexis Nasr (U. de la Méditerrannée, reviewer), Éric Wehrli (LATL, reviewer), Nabil Hathout (CNRS).
- Djamé Seddah served in the PhD defense committe of Anthony Sigogne. Field: computer science. Title: *Intégration de ressources lexicales riches dans un analyseur syntaxique probabiliste*. Other members of the committee: Eric Laporte (U. Paris-Est, supervisor), Matthieu Constant (U. Paris-Est, co-supervisor), Alexis Nasr (U. de la Méditerrannée, reviewer), Thierry Poibeau (CNRS, reviewer), Isabelle Tellier (U. Sorbonne-Nouvelle).
- Benoît Sagot and Laurence Danlos served in the PhD defense committee of Rosa Stern, as PhD supervisors. Field: linguistics. Title: *Identification automatique d'entités pour l'enrichissement de contenus textuels*. Other members of the committe: Adeline Nazarenko (LIPN, reviewer), Frédéric Béchet (U. de la Méditerrannée, reviewer), Éric Charton (U. Polytechnique de Montréal), Denis Teyssou (AFP, CIFRE industrial supervisor)

- Laurence Danlos served in the PhD defense committe of Charlotte Roze, as PhD supervisor. Field: linguistics. Title: *Vers une algèbre des relations de discours*. Other members of the committee: Philippe Muller (IRIT, co-supervisor), Francis Corblin (U. Paris Sorbonne, reviewer), Owen Rambow (Columbia University, reviewer), Liesbeth Degand (U. Catholique de Louvain), Laure Vieu (CNRS).
- Laurence Danlos (supervisor) and Marie-Hélène Candito (co-supervisor) served in the PhD defense committe of Enrique Henestroza Anguiano. Field: computer science. Title: *Efficient Large-Context Dependency Parsing and Correction with Distributional Lexical Resources*. Other members of the committe: Alexis Nasr (U. de la Méditerrannée, co-supervisor), Matthieu Constant (U. Paris-Est, reviewer), Joakim Nivre (Uppsala University, reviewer), Bernd Bohnet (University of Birmingham).
- Sylvain Kahane (supervisor) and Benoît Sagot (co-supervisor) served in the PhD defense committe of Pierre Magistry. Field: linguistics. Title: *Unsupervised Word Segmentation and Wordhood Assessment: the case for Mandarin Chinese*. Other members of the jury: Marie-Claude Paris (U. Paris-Diderot, co-supervisor), Hsieh Shukai (National Taiwan University, reviewer), Yves Lepage (Waseda University, reviewer), Pierre Zweigenbaum (CNRS).

9.3. Popularization

- Benoît Sagot has given in December 2013 a tutorial about the MElt POS-tagger for approx. 20 linguists, as part of the activities of the Corpus Écrits consortium, which belongs to the TGIR Huma-Num.
- General NLP presentation at the ISN day (for Math teachers, computer science option, December 4th)
- Interventions at Digilinguo (organized by DGLFLF) on April 15th and June 5th, in particular for the promotion of Mgwiki
- Intervention at the Inria I-Match day, organized during the "Futur en Seine" event (June 14th)

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, pp. 269–289

[3] P. BOULLIER, B. SAGOT. *Are Very Large Context-Free Grammars Tractable?*, in "Trends in Parsing Technology", H. BUNT, P. MERLO, J. NIVRE (editors), Text, Speech and Language Technology, Springer, Oct 2010, vol. 43, http://hal.inria.fr/inria-00516341/en

[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, pp. 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] 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

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

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

[9] B. SAGOT, G. WALTHER. *Implementing a formal model of inflectional morphology*, in "Third International Workshop on Systems and Frameworks for Computational Morphology", Berlin, Germany, C. MAHLOW, M. PIOTROWSKI (editors), Communications in Computer and Information Science, Springer, September 2013, vol. 380, pp. 115-134 [*DOI :* 10.1007/978-3-642-40486-3_7], http://hal.inria.fr/hal-00927277

[10] B. SAGOT, É. VILLEMONTE DE LA CLERGERIE. *Error Mining in Parsing Results*, in "Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics", Sydney, Australia, Association for Computational Linguistics, July 2006, pp. 329–336

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

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

Publications of the year

Doctoral Dissertations and Habilitation Theses

[13] E. HENESTROZA ANGUIANO. , *Analyse syntaxique probabiliste en dependances : approches efficaces a` large contexte avec ressources lexicales distributionnelles*, Université Paris-Diderot - Paris VII, June 2013, http://hal.inria.fr/tel-00860720

[14] C. ROZE. , *Vers une algèbre des relations de discours*, Université Paris-Diderot - Paris VII, May 2013, http://hal.inria.fr/tel-00881243

[15] R. STERN. , *Identification automatique d'entités pour l'enrichissement de contenus textuels*, Université Paris-Diderot - Paris VII, June 2013, http://hal.inria.fr/tel-00939420

Articles in International Peer-Reviewed Journals

[16] B. CRABBÉ, D. DUCHIER, C. GARDENT, J. LE ROUX, Y. PARMENTIER. *XMG : eXtensible MetaGrammar*, in "Computational Linguistics", September 2013, vol. 39, n^{o} 3, pp. 591-629, http://hal.inria.fr/hal-00768224

[17] L. DANLOS. *Connecteurs de discours adverbiaux: Problèmes à l'interface syntaxe-sémantique*, in "Linguisticae Investigationes", December 2013, vol. 36, n^{o} 2, pp. 261-275, http://hal.inria.fr/hal-00932184

[18] D. SEDDAH, M. CANDITO, E. HENESTROZA ANGUIANO, H. ANGUIANO ENRIQUE. *A Word Clustering Approach to Domain Adaptation: Robust parsing of source and target domains*, in "Journal of Logic and Computation", February 2013 [*DOI :* 10.1093/LOGCOM/EXS082], http://hal.inria.fr/hal-00940224

International Conferences with Proceedings

[19] M. CONSTANT, M. CANDITO, D. SEDDAH. *The LIGM-Alpage Architecture for the SPMRL 2013 Shared Task: Multiword Expression Analysis and Dependency Parsing*, in "Fourth Workshop on Statistical Parsing of Morphologically Rich Languages", Seattle, United States, October 2013, pp. 46-52, http://hal.inria.fr/hal-00932372

[20] L. DANLOS, P. DE GROOTE, S. POGODALLA. *A Type-Theoretic Account of Neg-Raising Predicates in Tree Adjoining Grammars*, in "LENLS 10 - Logic and Engineering of Natural Language Semantics 10", Kanagawa, Japan, S. YATABE, D. BEKK, E. MCCREADY (editors), October 2013, http://hal.inria.fr/hal-00868382

[21] E. LASSALLE, P. DENIS. *Improving pairwise coreference models through feature space hierarchy learning*, in "ACL 2013 - Annual meeting of the Association for Computational Linguistics", Sofia, Bulgaria, Association for Computational Linguistics, 2013, http://hal.inria.fr/hal-00838192

[22] P. MAGISTRY, B. SAGOT. *Can MDL Improve Unsupervised Chinese Word Segmentation?*, in "Sixth International Joint Conference on Natural Language Processing: Sighan workshop", Nagoya, Japan, October 2013, 2 p. , http://hal.inria.fr/hal-00876389

[23] S. A. MIRROSHANDEL, A. NASR, B. SAGOT. *Enforcing Subcategorization Constraints in a Parser Using Sub-parses Recombining*, in "NAACL 2013 - Conference of the North American Chapter of the Association for Computational Linguistics", Atlanta, United States, June 2013, http://hal.inria.fr/hal-00936492

[24] B. SAGOT, D. NOUVEL, V. MOUILLERON, M. BARANES. *Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel*, in "TALN - Traitement Automatique du Langage Naturel", Les sables d'Olonne, France, June 2013, pp. 407-420, http://hal.inria.fr/hal-00832078

[25] B. SAGOT, G. WALTHER. *Implementing a formal model of inflectional morphology*, in "Third International Workshop on Systems and Frameworks for Computational Morphology", Berlin, Germany, C. MAHLOW, M. PIOTROWSKI (editors), Communications in Computer and Information Science, Springer, September 2013, vol. 380, pp. 115-134 [*DOI :* 10.1007/978-3-642-40486-3_7], http://hal.inria.fr/hal-00927277

[26] Y. SCHERRER, T. ERJAVEC. *Modernizing historical Slovene words with character-based SMT*, in "BSNLP 2013 - 4th Biennial Workshop on Balto-Slavic Natural Language Processing", Sofia, Bulgaria, July 2013, http://hal.inria.fr/hal-00838575

[27] Y. SCHERRER, B. SAGOT. *Lexicon induction and part-of-speech tagging of non-resourced languages without any bilingual resources*, in "RANLP Workshop on Adaptation of language resources and tools for closely related languages and language variants", Hissar, Bulgaria, September 2013, http://hal.inria.fr/hal-00862693

[28] Y. SCHERRER, B. SAGOT. *Étiquetage morphosyntaxique de langues non dotées à partir de ressources pour une langue étymologiquement proche*, in "Atelier TALARE, TALN 2013", Les Sables d'Olonne, France, ATALA, June 2013, http://hal.inria.fr/hal-00838569

[29] D. SEDDAH, R. TSARFATY, S. KÜBLER, M. CANDITO, J. D. CHOI, R. FARKAS, J. FOSTER, I. GOENAGA, K. GOJENOLA GALLETEBEITIA, Y. GOLDBERG, S. GREEN, N. HABASH, M. KUHLMANN, W. MAIER, J. NIVRE, A. PRZEPIÓRKOWSKI, R. ROTH, W. SEEKER, Y. VERSLEY, V. VINCZE, M. WOLIŃSK, A. WRÓBLEWSKA, É. VILLEMONTE DE LA CLERGERIE. *Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages*, in "Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages", Seattle, Washington, United States, Association for Computational Linguistics, 2013, pp. 146–182, http://hal.inria.fr/hal-00877096

[30] É. VILLEMONTE DE LA CLERGERIE. *Exploring beam-based shift-reduce dependency parsing with DyALog: Results from the SPMRL 2013 shared task*, in "4th Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL'2013)", Seattle, United States, 2013, http://hal.inria.fr/hal-00879129

[31] É. VILLEMONTE DE LA CLERGERIE. *Improving a symbolic parser through partially supervised learning*, in "The 13th International Conference on Parsing Technologies (IWPT)", Naria, Japan, 2013, http://hal.inria.fr/hal-00879358

National Conferences with Proceedings

[32] A. ABEILLÉ, B. CRABBE. *Vers un treebank du français parlé*, in "TALN 2013 - 20ème conférence du Traitement Automatique du Langage Naturel", Sables d'Olonne, France, June 2013, http://hal.inria.fr/hal-00936490

[33] C. BRAUD, P. DENIS. *Identification automatique des relations discursives "implicites" à partir de données annotées et de corpus bruts*, in "TALN - 20ème conférence du Traitement Automatique du Langage Naturel 2013", Sables d'Olonne, France, June 2013, vol. 1, pp. 104-117, http://hal.inria.fr/hal-00830983

[34] E. LASSALLE, P. DENIS. *Apprentissage d'une hiérarchie de modèles à paires spécialisés pour la résolution de la coréférence*, in "TALN 2013 - 20ème conférence du Traitement Automatique du Langage Naturel 2013", Les Sables-d'Olonne, France, June 2013, http://hal.inria.fr/hal-00825617

[35] M. MORARDO, É. VILLEMONTE DE LA CLERGERIE. *Vers un environnement de production et de validation de ressources lexicales semantiques*, in "Atelier TALN 2013 SemDIS", Les Sables d'Olonne, France, ATALA, 2013, http://hal.inria.fr/hal-00879127

[36] Q. PRADET, J. BAGUENIER DESORMEAUX, G. DE CHALENDAR, L. DANLOS. *WoNeF : amélioration, extension et évaluation d'une traduction française automatique de WordNet*, in "TALN 2013 - 20ème conférence du Traitement Automatique du Langage Naturel", Les Sables d'Olonne, France, June 2013, pp. 76-89, http://hal.inria.fr/cea-00932340

[37] C. RIBEYRE. *Vers un système générique de réécriture de graphes pour l'enrichissement de structures syntaxiques*, in "TALN 2013 - 20ème conférence du Traitement Automatique du Langage", Les Sables d'Olonne, France, Université de Nantes, June 2013, pp. 178-191, http://hal.inria.fr/hal-00830967

Conferences without Proceedings

[38] A. ABEILLÉ, B. CRABBE, D. GODARD, J.-M. MARANDIN. *French Polar Questions : a corpus study*, in "Congrès international des linguistes", Genève, Switzerland, July 2013, http://hal.inria.fr/hal-00936493

[39] B. SAGOT. *Comparing Complexity Measures*, in "Computational approaches to morphological complexity", Paris, France, Surrey Morphology Group, February 2013, http://hal.inria.fr/hal-00927276

[40] B. SAGOT. *Les catégories prédicatives dans le Lefff*, in "Journée d'étude " CAtégories Prédicatives et Traitement Automatique des Langues " (CAPTAL)", Lille, France, February 2014, http://hal.inria.fr/hal-00943675

[41] Y. SCHERRER. *Continuous variation in computational morphology - the example of Swiss German*, in "TheoreticAl and Computational MOrphology: New Trends and Synergies (TACMO)", Genève, Switzerland, 19th International Congress of Linguists, July 2013, http://hal.inria.fr/hal-00851251

[42] G. WALTHER, G. JACQUES, B. SAGOT. *Uncovering the inner architecture of Khaling verbal morphology*, in "3rd Workshop on Sino-Tibetan Languages of Sichuan", Paris, France, September 2013, http://hal.inria.fr/hal-00927278

Scientific Books (or Scientific Book chapters)

[43] H. GOEBL, Y. SCHERRER, P. SMEČKA. *Kurzbericht über die Dialektometrisierung des Gesamtnetzes des "Sprachatlasses der deutschen Schweiz" (SDS)*, in "Vielfalt, Variation und Stellung der deutschen Sprache", K. SCHNEIDER-WIEJOWSKI, B. KELLERMEIER-REHBEIN, J. HASELHUBER (editors), De Gruyter, 2013, pp. 153-176, http://hal.inria.fr/hal-00932593

[44] B. SAGOT. *Construction de ressources lexicales pour le traitement automatique des langues*, in "Ressources Lexicales – Contenu, construction, utilisation, évaluation", N. GALA, M. ZOCK (editors), Lingvisticæ Investigationes Supplementa, John Benjamins, 2013, vol. 30, pp. 217-254, http://hal.inria.fr/hal-00927281

References in notes

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

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

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

[48] 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/

[49] 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, pp. 269–289

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

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

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

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

[54] 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

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

[56] D. CHIANG. *Statistical parsing with an automatically-extracted Tree Adjoining Grammar*, in "Proceedings of the 38th Annual Meeting on Association for Computational Linguistics", 2000, pp. 456–463

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

[58] M. COLLINS. , *Head Driven Statistical Models for Natural Language Parsing*, University of PennsylvaniaPhiladelphia, 1999

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

[60] 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, pp. 45–54

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

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

[63] 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

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

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

[66] J. JUDGE, A. CAHILL, J. VAN GENABITH. *Questionbank: Creating a corpus of parse-annotated questions*, in "Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics", Association for Computational Linguistics, 2006, pp. 497–504

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

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

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

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

[71] 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

[72] 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

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

[74] 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

[75] 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, pp. 404–411, http://www.aclweb.org/anthology/N/N07/N07-1051

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

[77] 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

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

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

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

[81] 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

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

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

[84] B. SAGOT, K. FORT, F. VENANT. *Extending the Adverbial Coverage of a French WordNet*, in "Proceedings of the NODALIDA 2009 workshop on WordNets and other Lexical Semantic Resources", Odense, Danemark, 2008, http://hal.archives-ouvertes.fr/hal-00402305

[85] 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, pp. 156–163

[86] 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, pp. 271–286

[87] 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

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

[89] 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

[90] 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

[91] 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

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

[93] 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

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

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

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

[97] 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

[98] 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

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

[100] 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

[101] 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

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

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

[104] 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

[105] 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/

[106] G. WALTHER. , *Sur la canonicité en morphologie — Perspective empirique, formelle et computationnelle*, Université Paris-Diderot, 2013

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

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

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