IN PARTNERSHIP WITH:
Université Denis Diderot (Paris 7)

Activity Report 2014

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, Machine Learning

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 (EPI), and became an UMR-I on January 1st, 2009 (UMR-I 001). Since January 1st, 2014, Benoît Sagot has taken over form Laurence Danlos as the director of the UMR-I Alpage and as the (acting) team leader of the EPI Alpage.

Creation of the Project-Team: 2008 January 01.

1. Members

Research Scientists
Benoît Sagot [Team leader, Inria, Researcher]
Pierre Boullier [Inria, Senior Researcher (Emeritus)]
Éric Villemonte de La Clergerie [Inria, Researcher]

Faculty Members
Marie-Hélène Candito [Univ. Paris VII, Associate Professor]
Benoit Crabbé [Univ. Paris VII, Associate Professor]
Laurence Danlos [Univ. Paris VII, Professor, HdR]

Engineers
Paul Bui Quang [Inria, until Oct 2014]
Margot Colinet [Inria, granted by Univ. Paris XIII]
Vanessa Combet [Inria, granted by LabEx EFL]
Pierre Magistry [Inria, granted by Caisse des Dépôts et Consignations, from Aug 2014]
Mikaël Morardo [Inria, until Apr 2014]
Virginie Mouilleron [Inria, granted by LabEx EFL]
Jacques Steinlin [Inria, from Jun 2014]

Post-Doctoral Fellows
Kata Gábor [Inria, granted by Caisse des Dépôts et Consignations]
Julie Hunter [Inria, until Aug 2014, granted by ANR Polymnie project]
Alexandra Simonenko [Univ. Paris VII, granted by LabEx EFL]

Visiting Scientists
Clement Beysson [ENS Cachan, until Aug 2014]
Kristina Gulordava [Univ. of Geneva, since Oct 2014]
James Pustejovsky [Brandeis University, from Mar 2014 until Apr 2014]

Administrative Assistant
Christelle Guiziou [Inria]

Others
Lucie Barque [Univ. Paris XIII, Associate Professor, from Apr 2014]
Djamé Seddah [Univ. Paris IV, Associate Professor]
Marion Baranes [viavoo and Univ. Paris VII]
Sarah Beniamine [Univ. Paris VII, granted by LabEx EFL]
Chloé Braud [Min. Ens. Sup. Recherche and Univ. Paris VII]
Maximin Coavoux [Univ. Paris VII]
Marianne Djemaa [Inria, granted by ANR ASFALDA project]
Valérie Hanoka [Univ. Paris VII]
Emmanuel Lassalle [Univ. Paris VII and ENS]

Quentin Pradet [CEA and Univ. Paris VII]
Corentin Ribeyre [Min. Ens. Sup. Recherche and Univ. Paris VII]
Raphaël Salmon [Yseop and Univ. Paris VII]

2. Overall Objectives

2.1. Overall Objectives

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

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

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

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

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

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

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

3. Research Program

3.1. From programming languages to linguistic grammars

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

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 (e.g., grammar size [1]) 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 [59], [108], [116]) are also parsable in polynomial time.

Unification-based formalisms They combine a context-free backbone with logic arguments as decoration on non-terminals. Most famous representatives are Definite Clause Grammars (DCGs) where PROLOG powerful unification is used to compute and propagate these logic arguments. More recent formalisms, like Lexical Functional Grammars (LFGs) and Head-Driven Phrasal Structure Grammars (HPSGs) rely on more expressive Typed Feature Structures (TFS) or constraints.

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

An efficient way to develop large-coverage hand-crafted symbolic grammars is to use adequate tools and adequate levels of representation, and in particular Meta-Grammars, one of Alpage's areas of expertise, especially with the FRMG grammar and parser for French based on the DyALog logic programming environment [136], [130]. Meta-Grammars (MGs) 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, Benoit Crabbé, Éric Villemonte de La Clergerie, Benoît Sagot, Corentin Ribeyre, Pierre Boullier, Maximin Coavoux.

[1] boullier:2010:inria-00516341:1

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 [87] or automatic [100], [101] 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 [73], or probabilistic models derived from lexicalized Tree Adjoining grammars, such as Stochastic Tree Insertion Grammars [71].

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 ...) [140], [98]. 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 [92]. 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 [55], [54] and derive the best input for syntagmatic statistical parsing [75]. Benchmarking several PCFG-based learning frameworks [122] 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) [101].

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 [71] 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 [126].

Pursuing our work on PCFG-LA based parsing, we investigated the automatic conversion of the treebank into dependency syntax representations [67], 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. Robust linguistic processing

Participants: Djamé Seddah, Benoît Sagot, Éric Villemonte de La Clergerie, Marie-Hélène Candito, Kata Gábor, Pierre Magistry, Marion Baranes.

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, especially out-of-domain text genres. Such texts that exhibit properties (e.g., lexical and syntactic properties) that are different or differently distributed than what is found on standard data (e.g., training corpora for statistical parsers). The development of shallow processing chains, such as SxPipe (see 5.5), is not a trivial task [110]. Obviously, they are often used as such, and not only as pre-processing tools before parsing, since they perform the basic tasks that produce immediately usable results for many applications, such as tokenization, sentence segmentation, spelling correction (e.g., for improving the output of OCR systems), named entity detection, disambiguation and resolution, as well as morphosyntactic tagging.

Still, when used as a preliminary step before parsers, the quality of parsers' results strongly depends on the quality of such chains. This is especially the case, beyond the standard out-of-domain corpora mentioned above, for user-generated content. Indeed, until very recently out-of-domain text genres that have been prioritized have not been Web 2.0 sources, but rather biomedical texts, child language and general fiction (Brown corpus). Adaptation to user-generated content is a particularly difficult instance of the domain adaptation problem since Web 2.0 is not really a domain: it consists of utterances that are often ungrammatical. It even shares some similarities with spoken language [129]. The poor overall quality of texts found on such media lead to weak parsing and even POS-tagging results. This is because user-generated content exhibits both the same issues as other out-of-domain data, but also tremendous issues related to tokenization, typographic and spelling issues that go far beyond what statistical tools can learn from standard corpora. Even lexical specificities are often more challenging than on edited out-of-domain text, as neologisms built using productive morphological derivation, for example, are less frequent, contrarily to slang, abbreviations or technical jargon that are harder to analyse and interpret automatically.

In order to fully prepare a shift toward more robustness, we developed a first version of a richly annotated corpus of user-generated French text, the French Social Media Bank [7], which includes not only POS, constituency and functional information, but also a layer of "normalized" text. This corpus is fully available and constitutes the first data set on Facebook data to date and the first instance of user generated content for a morphologically-rich language. Thanks to the support of the Labex EFL through, we are currently the finalizing the second release of this data set, extending toward a full treebank of over 4,000 sentences (see section 6.9).

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

However, this raises issues concerning the normalization step. A good sandbox for working on this challenging task is that of POS-tagging. For this purpose, we did leverage Alpage's work on MElt, a state-of-the art POS tagging system [80] (see 5.5). A first round of experiments on English have already led to promising results during the shared task on parsing user-generated content organized by Google in May 2012 [102], as Alpage was ranked second and third [125]. For achieving this result, we brought together a preliminary implementation of a normalization wrapper around the MElt POS tagger followed by a state-of-the art statistical parser improved by several domain adaptation techniques we originally developed for parsing edited out-of-domain texts. Those techniques are based on the unsupervized learning of word clusters *a la* Brown and benefit

from morphological treatments (such as lemmatization or desinflection) [123]. More recent developments are sketched in section 4.2

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

3.4. Dynamic wide coverage lexical resources

Participants: Benoît Sagot, Laurence Danlos, Éric Villemonte de La Clergerie, Marie-Hélène Candito, Lucie Barque, Valérie Hanoka, Marianne Djemaa, Quentin Pradet.

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

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

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

Successful experiments have been conduced by Alpage members with different languages for the automatic acquisition of morphological knowledge from raw corpora [115]. At the syntactic level, work has been achieved on automatic acquisition of atomic syntactic information and automatic detection of errors in the lexicon [142],[6]. At the semantic level, automatic wordnet development tools have been described [104], [137], [84], [81]. 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* [111], [117], developed within the Alexina framework. At the semantic level, Alpage members have developed or are developing various syntactico-semantic or semantic resources, including:

- a wordnet for French, the WOLF [112], the first freely available resource of the kind (see 5.7);
- a French FrameNet lexicon (together with an annotated corpus) within the ASFALDA ANR project (see sections 8.1.2.1 and 6.10);
- and a French VerbNet, Verb∋net (see 6.12).

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* , the WOLF, the French FrameNet lexicon and the French VerbNet. 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 2014 or before exist for German [38], Slovak, Polish, English, Spanish, Persian, Latin (verbs only), Kurmanji Kurdish, Maltese (verbs only, restricted to the so-called first *binyan*) and Khaling, not including freely-available lexicons adapted to the Alexina framework.

3.5. Discourse structures

Participants: Laurence Danlos, James Pustejovsky, Jacques Steinlin, Chloé Braud, Julie Hunter, Raphaël Salmon.

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 [77].

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, [78],[4]), which subsumes in an elegant way both SDRT and RST, to the extent that SDRT and RST structures can be obtained by two different partial projections of D-STAG structures. As done in D-LTAG, D-STAG extends a lexicalized TAG analysis so as to deal with the level of discourse. D-STAG has been fully formalized, and is hence possible to implement (thanks to Synchronous TAG, or even TAG parsers), provided one develops linguistic descriptions in this formalism.

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 four 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; related to this group lies the automatic correction of the output of OCR systems;

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, 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 5.5) and Alexina morphological lexicons. Several other parts of *vera* have been co-developed by Verbatim Analysis and by Inria.

4.4. Multilingual terminologies and lexical resources for companies

Participant: Éric Villemonte de La Clergerie.

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 Mikaël 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: Kata Gábor, Pierre Magistry, Benoît Sagot, 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 and Pierre Magistry are doing post-docs funded by PACTE (see 6.3)

4.6. Experimental and quantitative linguistics

Participants: Benoit Crabbé, Benoît Sagot, Alexandra Simonenko, Sarah Beniamine, Kristina Gulordava.

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 [62], [139], the existence of graded grammaticality judgments [86].

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 [72] gave way on the one hand to more abstract characterizations of the language faculty [72], and on the other hand to the construction of detailed, formally explicit, and often implemented, alternative formulation of the generative approach [61], [103]. For French several grammars have been implemented in this trend, such as the tree adjoining grammars of [65], [76] 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 [121]. 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 and through language evolution.

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 particlar in the context of Sarah Beniamine's PhD, co-supervized by Benoît Sagot (Alpage) and Olivier Bonami (LLF, CNRS, U. Paris Diderot and U. Paris Sorbonne). Collaborative work in this area is also conducted in collaboration with descriptive linguists from CRLAO (CNRS and Inalco; Guillaume Jacques) and HTL (CNRS, U. Paris Diderot and U. Sorbonne Nouvelle; Aimée Lahaussois) and formal linguists from DDL (CNRS and Université Lyon 2; Géraldine Walther).

5. New 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.

In 2014, an in-depth rewriting of the RCG parser has started, in order for RCG parsers produced by SYNTAX to handle input DAGs while remaining efficient [60], although parsing time complexity might, on such inputs, become exponential w.r.t. their length, whereas RCGs exactly cover the set of languages that are parsable in polynomial time (if the input is a string).

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 of DYALOG (version 1.14.0) 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). In particular, it has been ported for the CLANG/LLVM compiler used in recent Mac OS systems (Mavericks).

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 [135]. Version 1.14.0 now includes an efficient handler for feature-based statistical models, derived from the work on DYALOG-SR and now used in FRMG parser.

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) [130], [134], and very large amount of data (700 millions of words) in the SCRIBO project. New results concerning FRMG are described in 6.5.

DYALOG is also used to run DYALOG-SR, a transition-based dependency parser (see new results in 6.5)

DYALOG and other companion modules (including DYALOG-SR) 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 [132]. 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, Benoit Crabbé.

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

Alpage has developed as support of the research papers [75], [67], [68], [122] 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 [66] as output. This parser operates sequentially: (1) it first outputs a phrase structure analysis of sentences reusing the Berkeley implementation of a PCFG-LA trained on French by Alpage (2) it applies on the resulting phrase structure trees a process of conversion to dependency parses using a combination of heuristics and classifiers trained on the French treebank. The parser currently outputs several well known formats such as Penn treebank phrase structure trees, Xerox like triples and CONLL-like format for dependencies. The parsers also comes with basic preprocessing facilities allowing to perform elementary sentence segmentation and word tokenisation, allowing in theory to process unrestricted text. However it is believed to perform better on newspaper-like text.

The parser is available under a GPL license.

5.5. Alpage's linguistic workbench, including SxPipe and MElt

Participants: Benoît Sagot [correspondant], Kata Gábor, Marion Baranes, Pierre Magistry, Pierre Boullier, Éric Villemonte de La Clergerie, Djamé Seddah.

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, two packages are of particular importance: the SXPipe pre-processing chain, and the MElt part-of-speech tagger.

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

- as a preliminary step before Alpage's parsers (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...). It can now be augmented with modules developed during the former ANR EDyLex project for analysing unknown words; 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. New local grammars for detecting new types of entities and improvement of existing ones, developed in the context of the PACTE project, will soon be integrated within the standard configuration.

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 [79], Spanish [88], Italian [124], German [38], Dutch, Polish, Kurmanji Kurdish [138] and Persian [119], [120]. It is state-of-the-art for French. It is now able to handle noisy corpora (French and English only; see below). MElt also includes a lemmatization post-processing step. A preliminary version of MElt which accepts input DAGs has been developed in 2013, and is currently under heavy rewriting and improvement in the context of the PACTE project (see 6.3).

MElt is distributed freely as a part of the Alpage linguistic workbench.

In 2014, additional efforts have been achieved for a better pre-processing of noisy input text. This covers two different scenarios:

- user-generated content (see 6.2); two sets of tools are available for processing user-generated content: (i) very noisy computer-mediated content, such as found on social media, forums or blogs, are adressed within the MElt part-of-speech tagger via a three-step procedure (normalisation, tagging, de-normalisation with tag redistribution); this work is performed in relation with the CoMeRe project, funded by the Institut de Linguistique Française [14]; (ii) less noisy customer data, for preparing shallow semantic analysis; this work is performed in collaboration with the viavoo company [17].
- output of OCR systems, in the context of the PACTE project (see 6.3).

5.6. 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* [128], [118] 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.

5.7. The free French wordnet WOLF

Participants: Benoît Sagot [correspondant], Valérie Hanoka.

The WOLF (Wordnet Libre du Français) is a wordnet for French, i.e., a lexical semantic database. The development of WOLF started in 2008 [112], [113]. 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 [114].

In 2014, updated betas of the new version of the WOLF have been published (now version 1.0b4), which integrates and extends the various efforts performed and published somewhat independently in 2012, together with the result of additional filtering, both manual and semi-automatic.

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

5.8. 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 [105]. 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

5.9. LexViz

Participants: Mikaël 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).

5.10. 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. Large parsed corpora (like French Wikipedia or Wikisource) are also available. The parsed corpora can also be queried through the use of the DPath language. The resulting wiki has been officially opened in 2014 on http://alpage.inria.fr/frmgwiki.

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

The source code of the wiki is available on the GForge.

6. New Results

6.1. Highlights of the Year

Benoit Crabbé is a Junior Member of the Institut Universitaire de France (IUF) since October 2014. Two out of the five academic staff at Alpage are now member of the IUF, Laurence Danlos being a Senior Member since October 2013.

6.2. Automatic text normalisation

Participants: Benoît Sagot, Marion Baranes.

Since the emergence of the web, one of the goals of natural language processing (NLP) tools has been analysing raw noisy text documents such as blogs, review sites or social networks. These texts commonly contain misspellings, redundant punctuation, smileys, etc. Consequently they require specific preprocessing before being used in different NLP applications. That is why, we worked at Alpage on the development of a new corpora and the implementation of an automatic system for normalisation of such texts:

- **Corpus crap** In 2014, a large-scale extension of the number of normalisation rules used by the MElt part-of-speech tagger for processing noisy computer-generated content has been achieved. This work was carried out in the context of and based on corpora developed within the CoMeRe project, funded by the Institut de Linguistique Française and lead by Thierry Chanier [14].
- **Normalisation system** We have implemented a modular system which follows SxPipe [109]. This system detects if an unknown word to a reference lexicon corresponds to a non-word error (and is not a neologisme or a borrowing). Then, it attempts to normalize non-word errors and grammatical errors. In 2014, we focused on these two latter tasks. First, we have implemented a system which suggests one or several normalization candidates for these non-word errors. As described in [17], to do that, we use an analogy-based approach for acquiring normalisation rules and use them in the same way as lexical spelling correction rules. Secondly, we propose to normalize grammatical errors. To do that, we check for each word if it has common homophones. If this is the case, we consider these homophones as possible candidates for normalization. Finally, we filter all these candidates in order to keep only the one which is the most probable. This filtration is done using a probabilistic

model based on a n-gram system. Moreover, the implementation of this system of normalisation motivated a side task. We developed an unsupervised method for acquiring pairs of lexical entries belonging to the same morphological family, i.e., derivationally related words, starting from a purely inflectional lexicon. This work, detailed in [16], allows us to create new linguistic resources for English, French, German and Spanish which contains derivational relations.

6.3. The impact of morphosyntactic processing on post-OCR error correction

Participants: Kata Gábor, Benoît Sagot, Pierre Magistry.

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. To our knowledge, very little work has been done to exploit linguistic analysis for post-OCR error correction. Within the PACTE project we are conducting research on reducing the OCR error rate by using contextual information and linguistic processing.

In 2014 we continued our investigations on how named entity recognition can benefit OCR error detection by applying context-aware error correction rules directly to the OCR output. Several grammars have been created or improved to adress OCR problems occurring within different types of named entities. As a result, the SxPipe-PACTE toolchain was created to correct named entities in a noisy input [45], [31].

While the symbolic error correction method works with a very high precision, its limitation lies in its relatively low coverage. In order to deal with the errors occurring outside the recognized entitites, we studied the possibility of using lattice-based part of speech tagging to select the best correction hypothesis in context. Different methods were investigated to generate correction hypotheses, using word alignment software or by observing frequently occurring error types. The initial results confirm that a significant number of the remaining OCR errors can be corrected via lattice-based tagging, as long as the noise introduced by correction hypotheses is controlled.

6.4. Linear-time discriminant syntactico-semantic parsing

Participants: Benoit Crabbé, Maximin Coavoux, Djamé Seddah.

In this module we study efficient and accurate models of statistical phrase structure parsing. We focus on linear time lexicalized parsing algorithms (shift reduce, left corner) with approximations entailing linear time processing. The existing prototype involves a global discriminant parsing model of the large margin family (Perceptron,Mira, SVM avatars) able to parse user defined structured input tokens [23]. Thus the model can take into account various sources of information for taking decisions such as word form, part of speech, morphology or semantic classes inter alia.

Our participation to the SPRML 2014 shared task on parsing morphologically rich languages has been a first step towards testing our model in a multilingual setting where we were among the state of the art systems and state of the art on some languages such as Polish. To our knowledge the parser is one of the fastest existing multilingual parser worldwide (4000 8000 tokens/sec.). In order to ease model design for multilingual settings, we currently study efficient feature selection procedures for automating model adaptation to new languages.

The ongoing investigation aims to integrate continuous semantic representations into the model such as word embeddings in order to leverage data sparsity and estimation issues recurrent in lexicalized parsing. To this end we study neural-network-based architectures for structured phrase structure parsing.

6.5. Playing with DyALog-based parsers

Participant: Éric Villemonte de La Clergerie.

Éric de la Clergerie has continued the development of DyALog-SR, a transition-based dependency parser running on top of DyALog and initiated in 2013 to participate to SPMRL'2013. Thanks to DyALog's tabulation functionalities, this parser implements a dynamic programming algorithm to explore larger search space through the use of beams.

In order to participate to SemEval'14 Task on "broad coverage semantic dependency parsing", DYALOG-SR was extended to handle non-connected dependency graphs rather than standard dependency trees. This was achieved by considering a richer set of transitions, besides the usual Shift and Reduce transitions. However, while working, this extended set of transitions was not ensuring the expected gains when using beams. The issue was finally solved after long investigations, with the identification of multiple causes. One of them was related to the fact that transition paths of various lengths may lead to a final state. In consequence, a noop transition was added to compensate on shorter paths.

A second axe of work was a thorough use of DYALOG-SR over the French TreeBank (FTB) to compare its performances to those published for other parsers. By enriching its set of features and improving the update strategy of the perceptron-based statistical model of DYALOG-SR, we were able to reach state-of-the-art results.

However, the best results were obtained by coupling DYALOG-SR with FRMG, our large-coverage French grammar (derived from a meta-grammar). The results from FRMG were used as features to guide the statistical DYALOG-SR parser. This innovative step proved to provide us with the best results published so far for the FTB (over 90% of Labeled Attachement Score [LAS] over the test part of the FTB) [41].

The improvements of FRMG was pursued in 2014, at the level of the underlying meta-grammar (to extend its coverage over 96% on the FTB) but also by adapting the statistical models developed for DYALOG-SR (in replacement of older and slower SQLite-based models).

6.6. Multiword expressions and statistical parsing

Participants: Sarah Beniamine, Marie-Hélène Candito, Benoît Sagot, Djamé Seddah.

Multi-word expressions recognition (MWE recognition) and syntactic parsing are two tasks that have been extensively investigated. Yet, systems combining both tasks have been rather rare. In particulat, works on parsing have tended to use training and test data with gold MWEs (generally with each MWE) merged into one token. In 2013, Djamé Seddah led the organization of the first shared task on statistical parsing Morphologically Rich Languages (SPMRL) [127], hosted by the fourth SPMRL workshop. 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. The shared task proposed a data set for 9 languages. The French part of this data set is particular, in that it uses a representation combining MWEs and syntax, which allows to investigate techniques for performing parsing and MWE recognition. A first system was proposed for the dependency parsing track of the Shared Task, in collaboration with Matthieu Constant (LIGM, Université Marne-la-Vallée) [74]. This work investigates pipeline and joint architecture for both tasks. In 2014, Marie Candito and Matthieu Constant continued that line of work [2], focusing on using an alternative representation of syntactically regular MWEs, which captures their syntactic internal structure. The objective of such representation was two fold. First, it is well-known that the MWE status is not clear-cut, and that MWE status can hold due to syntactic and/or semantic criteria. In particular, syntactically regular MWEs exhibit various degrees of semantic non-compositionality. For such MWEs, an atomic representation fails to capture internal partial semantic composition, and also fails to take advantage of the internal syntactic regularity. Indeed, one hypothesis of this work was that augmenting the regularity of the syntactic representations could help parsing. The results of this work is that while this hypothesis could not be verified, the resulting system has comparable performance to that of previous works on this dataset, but it has the advantage of predicting both syntactic dependencies and the internal structure of MWEs, a crucial feature to capture the various degrees of semantic compositionality of MWEs.

In the same time, Sarah Beniamine and Benoît Sagot also investigated the use of internal regular structures for MWEs, yet for *syntagmatic* syntactic parsing. The objective is to guide a parser with predicted MWEs, while keeping a regular syntactic representation.

6.7. Graph-based approaches for deep-syntatic and semantic parsing

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

With most state-of-the-art statistical parsers routinely crossing a ninety percent performance plateau in capturing tree structures, the question of *what next* crucially arises. Most of the structures used to train current parsing models are degraded versions of a more informative data set: the Wall Street journal section of the Penn treebank ([91]) which is often stripped from its richer set of annotations (i.e. traces and functional labels are removed), while, for reasons of efficiency and availability, projective dependency trees are often given preference over richer graph structures [96], [107]. This led to the emergence of *surface* syntax-based parsers [70], [97], [100] whose output cannot by itself be used to extract full-fledged predicate argument-structures. For example, control verb constructions, it-cleft structures, argument sharing in ellipsis coordination, etc. are among the phenomena requiring a graph to be properly accounted for. The dichotomy between what can usually be parsed with high accuracy and what lies in the deeper syntactic description has initiated a line of research devoted to closing the gap between surface syntax and richer structures.

At Alpage, we built our work on the widely known transition-based parsing approach [95], which is state-of-the-art to parse surfacic syntatic trees [141]. Shift-reduce transition-based parsers essentially rely on *configurations* formed of a stack and a buffer, with stack transitions used to move from a configuration to the next one, until reaching a final configuration.

6.8. English Broad-coverage Semantic Dependency Parsing

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

We successfully tested our graph-based approach described in Section 6.7 on a shared task on broad-coverage semantic dependency parsing part of the International Workshop on Semantic Evaluation (SemEval 2014, [99]). We were given three resources, which constitute parallel semantic annotations over the same common text (the Penn Treebank (PTB), [91]). The first one is part of the tectogrammatical layer of the Prague Czech-English Dependency Treebank, the second one is the reduction of the Minimal Recursion Semantics, available through the HPSG annotation of the PTB, into bi-lexical dependencies [82]. Finally, the third one is the predicate-argument structures extracted from the Enju Parser [131]. The shared task consisted of two tracks: a closed one where we needed to use these three resources only and an open one, where we could use whatever we needed to produce the best semantic representations.

At Alpage, we developed two semantic parsers: The first one is based on a previous work on DAG parsing [107] and the second one on the FRMG surfacic syntactic parser [133]. We use two parsers to assess the validity of our approach. The top performing models we submitted used a mix of syntactic features (tree fragments from a constituent syntactic parser [100], dependencies from a syntactic parser [58], elementary spinal trees using a spine grammar [126], etc.) to improve our results. Our intuition is that syntax and semantic are not independent of each other and using syntax could improve semantic parsing. Our systems performs well and were able to compete with the top performers. Those systems, as well as the software needed to parse these new data sets, are already available.

6.9. Development of syntactic and deep-syntactic treebanks: Extending our Coverage

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

Taking its roots in the teams that initiated the first syntactically annotated the French Treebank, the first metagrammar compiler and one of the best wide coverage grammars, Alpage has a strong tendency to focus on creating pioneer resources that serve both to extend our linguistics knowledge and to nurture accurate parsing models. Recently, we focused on extending the lexical coverage of our parsers using semi-supervized techniques (see above) built on edited texts. In order to evaluate these models, we built the first free out-domain treebank for French (the Sequoia treebank, [69]) covering various domains such as Wikipedia, Europarl and bio medical texts on which we established the state-of-the-art. Exploring other kind of texts (speech, user generated content), we faced however various issues inherently tied to the nature of these productions. Syntactic divergences from the norm are actually prominent and are a severe bottleneck for any data driven

parsing model. Simply because a structure not present in a training set cannot be reproduced. This analysis naturally occurred as a side effect of our experiments in parsing social media texts. Actually, the first version of the French Social Media Bank (FSMB) was conceived as a stress test for our tool chains (tokenization, tagging, parsing). Our recent experiments showed that to reach a decent performance plateau, we need to include some of the target data into our training set. Focusing on processing direct questions and social media texts, we built two treebanks of about 2,500 sentences each: one devoted to questions and one built to extend the FSMB [2]. These initatives are funded by the Labex EFL.

- The French Social Media Bank 2.0: We are about to release the second part of the FSMB, 2600 sentences from Twitter, Facebook and other sources, with an extended annotation scheme able to describe more precisely the various phenomena at stakes in the social media text streams. To do so we extended our pre-processing chain (included and available in the MeLT tagger) to include a much more robust normalizer and tokenizer than the one we used to build the first version of the FSMB. The building phase being over, publications on this topics are on preparation.
- The French Question Bank: The building of a treebank made solely of questions comes from the simple fact that in both the FTB and the Sequoia treebank, there's only 150 direct questions. Making the parsing of such constructions extremely difficult for our data driven parsers. Following our now classical methodology, we selected more than 3200 sentences coming from governmental sources, from the TREC ressources – allowing to have a strong set of aligned sentences with the English ressources – and from social media sources as well. In the case of the TREC part, those are the questions used by [85], which allows some potentially interesting cross-language experiments. Unlike in the English Question Bank, phrasal-movement are annotated with functional paths and not traces. This allows to maintain a strong compatibility with the FTB annotation scheme. Our Question bank is the only resources of its kind for any other languages than English.

Both ressources are available in constituency and dependency. The later being still verified for the FSMB 2.0.

Note that we just started another annotation campaign aiming at adding a deep syntax layer to these two data sets, following the Deep Sequoia as presented above. These resources will prove invaluable to building a robust data driven syntax to semantic interface.

In the same time, Alpage collaborated with the Nancy-based Inria team Sémagramme in the domain of deep syntax analysis. Deep Syntax is intended as an intermediary level of representation, at the interface between syntax and semantics, which partly abtracts away from syntactic variation, and aims at providing the canonical grammatical functions of predicates. This means for instance neutralizing diathesis alternation and making explicit argument sharing, such as occurring for infinitival verbs. The advantage of a deep syntactic representation is to provide a more regular representation to serve as basis for semantic analysis. Note though it is computationally more complex, as we switch from surface syntactic trees to deep syntactic graphs, since shared arguments are made explicit.

We collaboratively defined a deep syntactic representation scheme for French and built a gold deep syntactic treebank [21], [43]. More precisely, each team used an automatic surface-to-deep syntax converter module, applied it on the Sequoia corpus (already annotated for surface syntax), and manually corrected it. Remaining differences were collaboratively adjudicated. The surface-to-deep syntax converter tool used by Alpage is built around the OGRE Graph Rewriting Engine built by Corentin Ribeyre [105].

The Deep Sequoia Treebank is too small to train a deep syntactic analyzer directly. In order to obtain more annotated data, we further used the surface-to-deep syntax converter to obtain predicted (non validated) deep syntactic representations for the French Treebank [36], which is much bigger than the Sequoia treebank (more than 18.000 sentences compared to $3,000$ sentences). We performed an evaluation of a small subset of the resulting deep syntactic graphs. The high level of performance we obtained (more than 98% of F-score in labeled dependencies recovery task) which suggests that the deep syntax version of the French Treebank can be used as pseudo-gold data to train deep syntactic parsers, or to extract syntactic lexicons augmented with quantitative information.

[2] Let us note that the ever evolving nature of user generated content makes this a necessity.

6.10. Towards a French FrameNet

Participants: Marie-Hélène Candito, Marianne Djemaa, Benoît Sagot.

The ASFALDA project [3] is an ANR project coordinated by Marie Candito. 5 partners collaborate on the project, on top of Alpage : the Laboratoire d'Informatique Fondamentale de Marseille(LIF), the Laboratoire de Linguistique Formelle (LLF), the MELODI team (IRIT - Toulouse) and the CEA-List. It 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 [57], [4] which provides a structured set of prototypical situations, called *frames*, along with a semantic characterization of the participants of these situations (called *frame elements*). 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 2014, we first finished the work on the lexicon, which was started in 2013 [19]. The step 2 (semantic annotations on top of syntactic representations) is ongoing :

- We wrote the annotation guide. In particular Marianne Djemaa focused on how to annotate phenomenon known to exhibit syntax/semantic divergences [42].
- We designed the annotation workflow and built an automatic pre-annotator, which proposes candidate semantic annotations that must be disambiguated manually.
- We started in july 2014 to manage six annotators, who were hired to perform the manual annotation phase.

6.11. Towards a morpho-semantic resource for French designed for Word Sense Disambiguation

Participant: Lucie Barque.

The most promising WSD methods are those relying on external knowledge resources [93] but semantic resources for French are scarse. Moreover, existing resources offer fine grained sense dictinctions that do not fit to WSD. Our aim is to provide the NLP community with a broad-coverage morpho-semantic lexicon for French that relies on coarse-grained sense distinctions for polysemic units. Preliminary results concern nouns, on which we have first focused because their semantic description, compared to verbs, crucially lacks (for information retrieval, for instance) and because the regular polysemy phenomenon (recurring cases of polysemy within semantic classes) mainly occurs in nominal semantic classes:

- We proposed a linguistically motivated description of general semantic labels for nouns, that will allow for coarse-grained sense distinctions [40]
- Regular polysemy of nouns that can denote an event or a participant of this event has also been described for a large number of French nouns in [12]

6.12. Development of Verb∋net

Participants: Laurence Danlos, Quentin Pradet.

[3] https://sites.google.com/site/anrasfalda/
[4] https://framenet.icsi.berkeley.edu/

VerbNet is an English lexical resources for verbs, which is internationally known and widely used in numerous NLP applications [89]. Verb∋net is a French adaptation of this resource. It is semi-automatically developed thanks to the use of two French existing resources created in the 70's: LG, Lexique-Grammaire developed at LADL under the supervision of Maurice Gross, and LVF, Lexique des verbs du français by Dubois and Dubois-Charlier. The idea is to map English classes, which gather verbs with a common syntactic and semantic behavior, into classes of LG and LVF, then to manually adapt the syntactic frames according to French grammar while keeping the thematic roles and the semantic information, [35], [28]. This work is currently under progress in collaboration with Takuya Nakamura (Institut Gaspard Monge) and the resource should be freely available in 2015.

6.13. Development of FDTB1

Participants: Laurence Danlos, Margot Colinet, Jacques Steinlin.

FDTB1 is the first step towards the creation of the French Discourse Tree Bank (FDTB) with a discourse layer on top of the syntactic one which is available in the French Tree Bank (FTB). In this first step, we have identified all the words or phrases in the corpus that are used as "discourse connectives". The methodology was the following: first, we highlighted all the items in the corpus that are recorded in LexConn [106], a lexicon of French connectives with 350 items, next we eliminated some of these items with the following criteria:

1. first, we filtered out the LexConn items that are annotated in FTB with parts of speech incompatible with a connective use, e.g. *bref* annotated as *Adj* instead of *Adv*, *en fait* annotated as *Pro V* instead of (compound) *Adv*;
2. second, as we lay down for theoretical and pratical reasons that elementary arguments of connectives must be clauses or VPs, we filtered out e.g. LexConn prepositions that introduce NPs;
3. last, we filtered out LexConn prepositions and adverbials with a non-discursive function.

The last criterion requires a manual work contrarily to the two others. For example the preposition *pour (to)*, is ambiguous between a connective use *(Fred s'est dépeché pour être à la gare à 17h (Fred hurried to be at the station at 17h))* and a preposition introducing a complement *(Fred s'est dépeché pour aller à la gare (Fred hurried to go to the station))*, and the disambiguation between the two uses is subtle and so the topic of a long paper [22], whose results have been used to enhance Le*fff*, [44].

The FDTB corpus contains 18 535 sentences and FDTB1 identifies 9 833 discourse connectives. This ressource is freely available.

6.14. Discourse Parsing

Participants: Laurence Danlos, Chloé Braud.

Discourse parsing goal is to reflect the rhetorical structure of a document, how pieces of text are linked in order to form a coherent document. Understanding such links could benefits to several other natural language applications (summarization, language generation, information extraction...). A discourse parser corresponds to two major subtasks: a segmentation step wherein discourse units (DUs) are extracted, and a parsing step wherein these DUs are (recursively) related through "discourse (rhetorical) relations". The more difficult task in discourse parsing is the labeling of the relations between DUs, especially when no so-called connective overtly marks the relation (we then talk about implicit relations as opposed to explicit ones). In her PhD work, Chloé Braud develops a discourse relation classifier, carrying experiments on French and English. Focusing on the problem on implicit relation identification, this work tries to tackle the lack of manually annotated data, a discourse specific difficulty, by exploiting the similarities between explicit and implicit relations. In 2014, this work lead to systems based on domain adaptation methods [18], [13], demonstrating improvements on the French corpus Annodis [56].

6.15. Multilingual and cross-lingual terminology extraction

Participants: Valérie Hanoka, Benoît Sagot.

Language diversity spans more than 7000 languages. Among them, 24 macrolanguages [5] have at least 50 million first-language speakers. Traditional terminology techniques, which are mostly based on language-dependent linguistic tools (part of speech tagging, phrase chunking) requires a considerable effort to be developed for a new language. This effort is likely to be even more critical if the term extraction is to be based on noisy text (i.e. displaying linguistic creativity, spelling errors and ungrammatical sentences). In this context, the need has arisen to examine the issue of a less language-specific method for term extraction.

To that end, our approach take advantage of existing language typologies in order to alleviate for the lack of language-dependent linguistic processing. We based our reflexions and experiments on a sample of 7 typologically different language: Arabic, Chinese, English, French, German, Polish and Turkish.
As a starting point, we considered the minimal textual preprocessing (character normalization, segmentation) needed to allow for a comprehensive multilingual approach to automatic term extraction. In order to gain further insight on the influence of the morphology for term extraction, we examined the impact of the deletion of selected morphological information on words of morphologically rich languages.
For the different settings, models based on Conditional Random Fields (CRF) have been trained on existing gold data. We proposed an adapted version of the evaluation algorithm of [94] able to issue terminological scores for all the language of our sample. The scores thus obtained allowed to identify the best experimental setting for each language tested.

The results were surprising in two ways: First, the cross-lingual [6] application of models works well (the best cross-lingual models' accuracies range from 0.8% to 0.97%). Secondly, the languages which makes the overall best cross-lingual models are those who have the richest morphology (i.e: Turkish).

Finally, we developped and used a multilingual translation graph [32] to extend the multilingual terminology obtained using two methods: those presented in [83] and a more formal one, based on a simulated annealing clustering algorithm.

6.16. Word order variation in Old French

Participants: Benoit Crabbé, Alexandra Simonenko, Benoît Sagot.

As participant of the strand *Experimental Grammar* of the Labex EFL project *Empirical Foundations of Linguistics* [7] we study word order issues on Old French and more specifically the relative ordering of complements of ditransitive verbs. The inquiry seeks to identify several factors influencing the ordering of Old French complementation in different texts (varying in dates and genres) by carrying quantitative and statistical work from annotated Old French data. [8]

The quantitative results will be compared with what is known from corpus studies on the relative ordering of subject and complement in Old French [90]. It will also be compared to the quantitative results obtained on the relative ordering of complements of ditransitive verbs in Modern French [8] and modern English [64]. This comparative perspective is expected to provide new insights on French language evolution.

6.17. Cross linguistic factors governing word order

Participants: Benoit Crabbé, Kristina Gulordava.

In many languages, flexible word order often has a pragmatic role and marks the introduction of new information, a focus or a topic shift. Other cases of language- internal word order variation are alternations between two options such as *Mary gave John a book* and *Mary gave a book to John*, which are conditioned on syntactic and semantic factors such as the complexity of the constituents (as in *Mary gave John a book she had read ten times*), their animacy or the meaning of the verb [63].

[5] A macrolanguage is defined as "multiple, closely related individual languages that are deemed in some usage contexts to be a single language" in the ISO 639-3 standard.
[6] A model trained on data of one language and applied to data of another language
[7] www.labex-efl.org
[8] SRCMF corpus: http://srcmf.org/; MCVF: http://www.voies.uottawa.ca

One of the goals of this module is to investigate the connection between the quantitative aspects of word order variation across languages and the quantitative aspects of word order variation within a language. We study the corresponding patterns in language-internal variation by looking at the syntactically annotated corpora of various languages. Focusing on the variation of the internal word order of the noun-phrase as a case study, we explore to which extent a computational corpus-based analysis can provide new evidence not only for empirical, but also for theoretical linguistic research.

6.18. Anaphoricity detection and coreference resolution

Participant: Emmanuel Lassalle.

Resolving coreference in a text, that is, partitioning mentions (noun phrases, verbs, etc) into referential entities, is a challenging task in NLP leading to many different approaches. Anaphoricity detection, on the other hand, consists in deciding whether a mention is anaphoric (aka discourse-old) or non-anaphoric (discourse-new). This task is strongly related to coreference resolution and has been mainly addressed as a preliminary task to solve, leading to pipeline architectures.

A first line of work compares several methods for learning latent structures encoding coreference clusters that optionally take into account very accurate constraints on mention pairs. We study the relationship between standard decoding strategies used with pairwise models and those used with structured learning of latent structures, providing both topological and empirical comparisons. We also show that further gains can be obtained by the addition of pairwise constraints. Our experiments on the CoNLL-2012 dataset show that our best system obtains state-of-the-art results, and significant gains compared to standard locally-trained models.

Our second line of work introduces a new structured model for learning anaphoricity detection and coreference resolution in a jointly. Specifically, we use a latent tree to represent the full coreference and anaphoric structure of a document at a global level, and we jointly learn the parameters of the two models using a version of the structured perceptron algorithm. This model is refined by the use of pairwise constraints, and our experiments on the CoNLL-2012 English datasets show large improvements in both coreference resolution and anaphoricity detection, compared to various competing architectures. Our best coreference system obtains a CoNLL score of 81.97 on gold mentions, which is to date the best score reported on this setting.

This work has been achieved in collaboration with Pascal Denis, a former Alpage member, now at Inria Lille-Nord-Europe (EPI Magnet).

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 (PhD of Marion Baranes, employed at viavoo, which started in 2012 on automatic normalisation of noisy texts),
- Yseop (“CIFRE” PhD of Raphael Salmon which started in 2012 on automatic text generation)
- CEA-List (PhD of Quentin Pradet on the annotation of semantic roles in specific domains. The thesis has finished on the 12/31/2015 (defense on the 02/06/2015).
- Proxem (consulting)

8. Partnerships and Cooperations

8.1. National Initiatives

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

Participants: Laurence Danlos, Benoît Sagot, Chloé Braud, Marie-Hélène Candito, Benoit Crabbé, Pierre Magistry, Djamé Seddah, Sarah Beniamine, Maximin Coavoux, 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.

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 Benoit 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.

8.1.2. ANR

8.1.2.1. ANR project ASFALDA (2012 – 2015)

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

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 [57] 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.

8.1.2.2. ANR project Polymnie (2012-2016)

Participants: Laurence Danlos, Éric Villemonte de La Clergerie, Julie Hunter.

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. ACG allows for the encoding of a large variety of grammatical formalisms, in particular Tree Adjoining grammars (TAG).

The role of Alpage in this project is to develop sentential or discursive grammars written in TAG and to participate in their conversion in ACG. Results were first achieved in 2014 concerning text generation: GTAG formalism created by Laurence Danlos in the 90's has been rewritten in ACG [25], [26], [27]. As regards discursive analysis, D-STAG formalism created by Laurence Danlos in the 00's is currently been rewritten in ACG and enhanced to cover attributions with some preliminary linguistic work on attributions [33].

8.1.3. Other national initiatives

8.1.3.1. "Investissements d'Avenir" project PACTE (2012 – 2015)

Participants: Benoît Sagot, Kata Gábor, Pierre Magistry.

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 2014 within PACTE are described in 6.3

8.1.3.2. FUI project COMBI (2014-2016)

Participants: Laurence Danlos, Vanessa Combet, Jacques Steinlin.

COMBI is is an "FUI 16" project. It started in February 2014 for a two year duration. It groups 5 industrial partners (Temis, Isthma, Kwaga, Yseop and Qunb) and Alpage. Temis and Istma work on data mining from texts and big data. Kwaga works on the interpretation and inferences that can be drawn from the data retrieved in the analysis module. Alpage and Qunb work, under the supervision of Yseop, on the production of respectively texts and graphics describing the results of the interpretation module. Currently, COMBI aims at creating the full chain for a user case concerning the weekly activity of an on-line service.

Alpage works on text generation, with the adaptation of TextElaborator, a generation system developed in the 10's by WatchAssistance and based on G-TAG. Alpage also works on the opportunity to describe pieces of information by texts, graphics or both.

8.1.3.3. 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.

8.2. European Initiatives

8.2.1. Collaborations in European Programs, except FP7 & H2020

Program: **IC1207 COST**

Project acronym: PARSEME

Project title: PARSing and Multi-word Expressions

Duration: March 2013 - March 2017

Coordinator: Agata Savary

Other partners: interdisciplinary experts (linguists, computational linguists, computer scientists, psycholinguists, and industrials) from 30 countries

Abstract: The general aim of PARSEME is increasing and enhancing the ICT support of the European multilingual heritage. This aim is pursued via more detailed objectives: (1) to put multilingualism in focus of linguistic and technological studies; (2) to establish a long-lasting cross-lingual, cross-theoretical and cross-methodological research network in natural language processing (NLP); (3) to bridge the gap between linguistic precision and computational efficiency in NLP applications.

Program: **ISCH COST Action IS1312**

Project acronym: TextLink

Project title: Structuring Discourse in Multilingual Europe

Duration: April 2014 - April 2018

Coordinator: Liesbeth Degand

Other partners: experts in computational linguistics and discourse from 24 countries

France MC members: Laurence Danlos and Philippe Muller (IRIT)

Abstract: With partners from across Europe, TextLink will unify numerous but scattered linguistic resources on discourse structure. With its resources searchable by form and/or meaning and a source of valuable correspondences, TextLink will enhance the experience and performance of human translators, lexicographers, language technology and language learners alike.

8.3. International Initiatives

8.3.1. Inria International Partners

8.3.1.1. Informal International Partners

Alpage has active collaborations with several international teams. The most active in 2014 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)
- collaboration with the Emory University (USA) on broad coverage parsing of unlabeled and noisy Korean data set (Djamé Seddah, with Jinho D. Choi).
- collaboration with the Indiana University (United States) on parsing morphologically rich languages (Djamé Seddah, with Sandra Kubler)
- collaboration with the University of Ljubljana (Slovenia) on wordnet development (Benoît Sagot, with Darja Fišer)
- collaboration with the Uppsala University (Sweden) on statistical parsing (Marie-Hélène Candito and sDjamé Seddah, with Joakim Nivre)
- collaboration with the Weizmann Institute of Science (Israel) on parsing morphologically rich languages (Djamé Seddah, with Reut Tsarfaty)

8.4. International Research Visitors

8.4.1. Visits of International Scientists

James Pustejovsky from Brandeis University (Boston, USA) was invited Professor at Alpage in April 2014. His stay was funded by Inria, his travel by Alpage. He is specialist in computational semantics and the creator of the “Generative Lexicon”. During his stay in Paris, he gave two lectures with a large audience. The topic was on the computational model of events. The notion of event has long been central for both modeling the semantics of natural language as well as reasoning in goal-driven tasks in artificial intelligence. James outlined a unified theory of event structure. James has also been working with Alpage members. First on the French lexical resources developed at Alpage, namely Framenet (Marie Candito) and Verbenet (Laurence Danlos). Second on the role of attributions in discourse structure within the linguistic work made at Alpage for the ANR Polymnie (Laurence Danlos and Julie Hunter).

8.4.1.1. Internships

Kristina Gulordava is a visiting research student from the University of Geneva (LATL) supervised by Paola Merlo, visting ALPAGE from September 2014 to January 2015. Her Phd thesis is dedicated to the study of generic cross linguistic constraints across languages. Her goal is to investigate the connection between the quantitative aspects of word order variation across languages and the quantitative aspects of word order variation within a language. She explores to which extent a computational corpus-based analysis can provide new evidence not only for empirical, but also for theoretical linguistic research.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events selection

9.1.1.1. Member of the conference program committee

- Benoît Sagot has served as member of the program committee for the following events: TALN 2014, SPMRL-SANCL 2014 (Workshop at COLING 2014)
- Marie Candito has served as member of the program committee for the following events: SPMRL-SANCL 2014 (Workshop at COLING 2014)
- Pierre Magistry has served as a member of the program committee for the following event: “Questionnements sur la didactique des sinogrammes”, workshop organised by the CERLOM, Inalco
- Éric Villemonte de La Clergerie has served as PC member for the following events: COLING 2014 (area Resources), AAAI SA-14 (AAAI-14 Student Abstract and Poster Program), TALN’14 SemDis Workshop, STAIRS-2014 (the 7th Starting AI Researcher Symposium)

9.1.1.2. Reviewer

- Benoit Crabbé has served as a reviewer for the following events : TALN 2014, RECITAL, COLING, Formal Grammar.
- Benoît Sagot has served as reviewer for the following events: ACL 2014, LREC 2014, WoLE 2014 (LREC workshop), PolTAL 2014, Coldoc 2014, RECITAL 2014, NLP4CMC, SHESL-HTL 2015
- Chloé Braud has served as a reviewer for the following event: ConSOLE XXIII
- Corentin Ribeyre has served as a reviewer for the following event: ConSOLE XXIII
- Kata Gábor served as a reviewer for the following event(s)/journal(s): LREC 2014, Cicling 2014, ConSOLE XXIII
- Laurence Danlos has served as a reviewer for the following conferences: LREC 2014, TALN 2014, CMLF 2014
- Marianne Djemaa has served as a reviewer for the following event: ConSOLE XXIII
- Marie Candito has served as reviewer for the following events: ACL 2014, EMNLP 2014, COLING 2014, SEMDIS 2014 (workshop at TALN 2014)
- Éric Villemonte de La Clergerie has served as a reviewer for the following events: LREC 2014, LATIN 2014

9.1.2. Journal

9.1.2.1. Member of the editorial board

- Laurence Danlos served as a member of the editorial board of *Traitement Automatique des Langues (TAL)*

- Éric Villemonte de La Clergerie has served as Chief Editor of *Traitement Automatique des Langues (TAL)*

9.1.2.2. Reviewer

- Benoit Crabbé has served as a reviewer for the following journals: Journal of Language Modelling, Language Ressources Evaluation, Linguisticae Investigationes
- Benoît Sagot has served as a reviewer for the following journal: Natural Language Engineering
- Lucie Barque has served as a reviewer for the following journal: Neologica
- Éric Villemonte de La Clergerie has served as reviewer for the following journal: Journal of Language Modelling

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence: Benoit Crabbé, Introduction à la programmation, 24 heures en équivalent TD, niveau L3, Université Paris Diderot, France.

Licence: Benoit Crabbé, Introduction à la linguistique informatique, 24 heures en équivalent TD, niveau L3, Université Paris Diderot, France.

Licence: Benoit Crabbé, Introduction aux probabilités pour la linguistique, 24 heures en équivalent TD, niveau L3, Université Paris Diderot, France.

Licence: Chloé Braud, Programmation 2, 28 heures en équivalent TD, niveau L3, Université Paris 7, France

Licence: Corentin Ribeyre, TD d'Algorithmique, 24 heures en équivalent TD, niveau L3, Université Paris 7 Diderot, France

Licence: Laurence Danlos, Introduction au TAL, 32 heures en équivalent TD, niveau L3, Université Paris-Diderot, France

Licence: Lucie Barque, Sémantique lexicale, 45 heures en équivalent TD, niveau L2, Université Paris 13, France

Licence: Lucie Barque, Morphologie lexicale, 22,5 heures en équivalent TD, Université Paris 13, France

Licence: Lucie Barque, Fondamentaux de la grammaire, 67,5 heures en équivalent TD, niveau L1, Université Paris 13, France

Licence: Lucie Barque, Introduction aux sciences du langage, 22,5 heures en équivalent TD, niveau L2, Université Paris 13, France

Licence: Marianne Djemaa, Programmation 1 (TD), 24 heures en équivalent TD, niveau L3, Université Paris Diderot, France

Licence: Marie Candito, Linguistique de corpus, 28 heures en équivalent TD, niveau L3, Université Paris Diderot, France

Master: Benoit Crabbé, Méthodes probabilistes pour le TAL, 48 heures en équivalent TD, niveau M1, Université Paris Diderot, France.

Master: Benoit Crabbé, Linguistique empirique et expérimentale, 24 heures en équivalent TD, niveau M2, Université Paris Diderot, France.

Master: Benoît Sagot, Analyse syntaxique du langage naturel, 28 heures en équivalent TD, niveau M1, Université Paris 7, France

Master: Chloé Braud, Sémantique Computationnelle, 24 heures en équivalent TD, niveau M1, Université Paris 7, France

Master: Corentin Ribeyre, TD de Langages Formels, 24 heures en équivalent TD, niveau M1, Université Paris 7 Diderot, France

Master: Éric Villemonte de La Clergerie, "Structures Informatiques et Logiques pour la Modélisation Linguistique", 18 heures en équivalent TD, niveau M2, MPRI, France

Master: Laurence Danlos, Discours: Analyse et génération de textes, 32 heures en équivalent TD, niveau M2, Université Paris-Diderot, France

Master: Lucie Barque, Ressources lexicales pour le TAL, 27 heures en équivalent TD, niveau M2, Université Paris 13, France

Master: Marie Candito, Analyse sémantique automatique du langage naturel, 28 heures en équivalent TD, niveau M2, Université Paris Diderot, France

Master: Marie Candito, Traduction automatique, 51 heures en équivalent TD, niveau M1, Université Paris Diderot, France

Master: Maximin Coavoux, Approches probabilistes du TAL (TD), 24 heures en équivalent TD, niveau M1, Université Paris Diderot, France

9.2.2. Supervision

PhD in progress: Marion Baranes, “Normalisation de textes bruités”, started in January 2012, supervised by Laurence Danlos (superviser) and Benoît Sagot (co-superviser)

PhD in progress: Valérie Hanoka, “Extraction et Structuration de Terminologie Multilingue”, started in January 2011, supervised by Laurence Danlos (superviser) and Benoît Sagot (co-superviser)

PhD in progress: Emmanuel Lassale, “Structured learning for coreference resolution: a joint approach”, Université Paris-Diderot, started in October 2010, supervised by Laurence Danlos (superviser) and Pascal Denis (co-superviser)

PhD in progress: Chloé Braud, “Analyse discursive : les relations implicites”, Université Paris-Diderot, started in October 2011, supervised by Laurence Danlos (superviser) and Pascal Denis (co-superviser)

PhD in progress: Quentin Pradet, “Annotations en rôles sémantiques du français en domaine spécifique”, Université Paris-Diderot, started in October 2011, supervised by Laurence Danlos (superviser) and Gael de Calendar (co-superviser)

PhD in progress: Raphael Salmon, “Implémentation d’un système de génération à base de contraintes”, Université Paris-Diderot, started in October 2013, supervised by Laurence Danlos (superviser) and Alain Kaeser (co-superviser)

PhD in progress: Jacques Steinlin, “Les relations implicites et les AltLex dans le FDTB”, Université Paris-Diderot, started in Novembre 2014, supervised by Laurence Danlos

PhD in progress: Marianne Djemaa, "Création semi-automatique d’un FrameNet du français", started in October 2012, supervised by Marie Candito

PhD in progress: Corentin Ribeyre, “Vers la syntaxe profonde pour l’interface syntaxe-sémantique ”, started in November 2012, supervised by Laurence Danlos (superviser), Djamé Seddah (co-superviser) and Éric Villemonte de La Clergerie (co-superviser). from data driven approaches to graph based approaches.

PhD in progress: Isabelle Dautriche, “Exploring early syntactic acquisition: a experimental and computational approach”, started in September 2012, supervised by Anne Christophe (LSCP, superviser) and Benoit Crabbé (co-superviser)

PhD in progress: Maximin Coavoux, “Représentations continues pour l’analyse syntaxique et sémantique automatique”, started in September 2015, supervised by Benoit Crabbé.

9.2.3. Juries

- Marie Candito served as a member in the PhD defense committee of Ophélie Lacroix. Title: De l'étiquetage syntaxique pour les grammaires catégorielles de dépendances à l'analyse par transition dans le domaine de l'analyse en dépendances non-projective. University: Université de Nantes Angers Le Mans. PhD supervisors: Colin de la Higuera and Denis Béchet. Defense date: December 8th, 2014.
- Marie Candito served as a member in the PhD defense committee of Sandrine Ollinger. Title: Le raisonnement analogique en lexicographie, son informatisation et son application au Réseau Lexical du Français. University: Université de Lorraine. PhD supervisor: Alain Polguère. Defense date: December 15th, 2014.
- Laurence Danlos served as a member in the PhD defense committee of Sai Qian. Title: Accessibility of Referents in Discourse Semantics University : Université de Lorraine PhD supervisors : Philippe de Groote and Maxime Amblard. Defense date : July 11th, 2014.
- Éric Villemonte de La Clergerie as served as a member in the PhD defense of Simon Petijean. Title: "Génération modulaire de Grammaires Formelles". PhD supervisors: Denys Duchier and Yannick Parmentier. University: Université d'Orléans. Defense date: December 11th 2014
- Benoît Sagot served as a member in the PhD defense committee of Édouard Grave. Title: A Markovian approach to distributional semantics. University: Universite´ Paris VI - Pierre et Marie Curie. PhD supervisers: Francis Bach and Guillaume Obozinski. Defense date: January 20th, 2014.

9.2.4. Other activities

- 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)
- Benoît Sagot is a member of the Permanent Committee of the TALN conference organized by ATALA.
- Laurence Danlos is the deputy chair of the Doctoral School for Linguistic Sciences (École Doctorale de Sciences du Langage).
- Laurence Danlos is the chair of the Scientific Committee of the UFR of Linguistics of University Paris Diderot.
- Marie-Hélène Candito is the deputy chair of the UFR of Linguistics of University Paris Diderot.
- Djamé Seddah is the head teacher (*directeur pédagogique*) of the transversal Computer Science class (C2I) at the University Paris-Sorbonne.
- Benoit Crabbé and Laurence Danlos are members of the Administrative board of the UFR of Linguistics of University Paris Diderot.
- Benoit Crabbé co-organized the research seminar : lectures in experimental linguistics (Univ P7)
- Benoit Crabbé is responsible for the L3 computational linguistics (Univ Paris Diderot)
- Laurence Danlos is responsible for the M2 computational linguistics (Univ Paris Diderot)
- Éric Villemonte de La Clergerie has served as SIGParse secretary
- 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; Benoit 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.
- Marie-Hélène Candito organized the Alpage research seminar.

9.3. Popularization

- Éric Villemonte de La Clergerie has presented some simple NLP tasks at the ISN day (for Math teachers, computer science option, April 8th, http://euler.ac-versailles.fr/webMathematica/reflexionpro/2013_2014/Seminaire/eric_de_la_clergerie.pdf)
- Éric Villemonte de La Clergerie has presented a survey on the use on NLP tools for MOOCs at the MoocLab day on “Moocs : de la correction automatique à la personnalisation des cursus” (Paris, January 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] M. CANDITO, M. CONSTANT. *Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing*, in "ACL 14 - The 52nd Annual Meeting of the Association for Computational Linguistics", Baltimore, United States, ACL, June 2014, https://hal.inria.fr/hal-01022415

[3] B. CRABBÉ. *An LR-inspired generalized lexicalized phrase structure parser*, in "COLING", Dublin, Ireland, 2014, https://hal.inria.fr/hal-01105142

[4] 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

[5] 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, https://hal.inria.fr/hal-00927281

[6] 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

[7] 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

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

[9] R. TSARFATY, D. SEDDAH, Y. GOLDBERG, S. KÜBLER, Y. VERSLEY, M. CANDITO, J. FOSTER, I. REHBEIN, L. TOUNSI. *Statistical Parsing of Morphologically Rich Languages (SPMRL) What, How and Whither*, in "Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages", États-Unis Los Angeles, Association for Computational Linguistics, 2010, pp. 1–12

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

Publications of the year

Articles in International Peer-Reviewed Journals

[11] M. APIDIANAKI, B. SAGOT. *Data-driven Synset Induction and Disambiguation for Wordnet Development*, in "Language Resources and Evaluation", November 2014, vol. 48, n° 4, pp. 655-677 [*DOI :* 10.1007/s10579-014-9291-2], https://hal.inria.fr/hal-01088000

[12] L. BARQUE, P. HAAS, R. HUYGHE. *La polysémie nominale événement / objet : quels objets pour quels événements ?*, in "Neophilologica", 2014, pp. 170-187, https://hal.archives-ouvertes.fr/hal-01104652

[13] C. BRAUD, P. DENIS. *Identifier les relations discursives implicites en combinant données naturelles et données artificielles*, in "Traitement Automatique des Langues", December 2014, vol. 55, n° 1, 31 p. , https://hal.inria.fr/hal-01094346

[14] T. CHANIER, C. POUDAT, B. SAGOT, G. ANTONIADIS, C. R. WIGHAM, L. HRIBA, J. LONGHI, D. SEDDAH. *The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres*, in "JLCL - Journal for Language Technology and Computational Linguistics", 2014, vol. 29, n° 2, pp. 1-30, Final version to Special Issue of JLCL (Journal of Language Technology and Computational Linguistics (JLCL, http://jlcl.org/): Building And Annotating Corpora Of Computer-Mediated Discourse : Issues and Challenges at the Interface of Corpus and Computational Linguistics (ed. by Michael Beißwenger, Nelleke Oostdijk, Angelika Storrer & Henk van den Heuvel), https://halshs.archives-ouvertes.fr/halshs-00953507

[15] A. GUTMAN, D. ISABELLE, B. CRABBÉ, A. CHRISTOPHE. *Bootstrapping the Syntactic Bootstrapper: Probabilistic Labeling of Prosodic Phrases*, in "Language Acquisition", September 2014, 25 p. [*DOI :* 10.1080/10489223.2014.971956], https://hal.inria.fr/hal-01105141

International Conferences with Proceedings

[16] M. BARANES, B. SAGOT. *A Language-Independent Approach to Extracting Derivational Relations from an Inflectional Lexicon*, in "Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)", Reykjavik, Iceland, May 2014, https://hal.inria.fr/hal-01002723

[17] M. BARANES, B. SAGOT. *Normalisation de textes par analogie: le cas des mots inconnus*, in "TALN - Traitement Automatique du Langage Naturel", Marseille, France, July 2014, pp. 137-148, https://hal.inria.fr/hal-01019998

[18] C. BRAUD, P. DENIS. *Combining Natural and Artificial Examples to Improve Implicit Discourse Relation Identification*, in "Coling", Dublin, Ireland, August 2014, https://hal.inria.fr/hal-01017151

[19] M. CANDITO, P. AMSILI, L. BARQUE, F. BENAMARA, G. DE CHALENDAR, M. DJEMAA, P. HAAS, R. HUYGHE, Y. Y. MATHIEU, P. MULLER, B. SAGOT, L. VIEU. *Developing a French FrameNet: Methodology and First results*, in "LREC - The 9th edition of the Language Resources and Evaluation Conference", Reykjavik, Iceland, May 2014, https://hal.inria.fr/hal-01022385

[20] M. CANDITO, M. CONSTANT. *Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing*, in "ACL 14 - The 52nd Annual Meeting of the Association for Computational Linguistics", Baltimore, United States, ACL, June 2014, https://hal.inria.fr/hal-01022415

[21] M. CANDITO, G. PERRIER, B. GUILLAUME, C. RIBEYRE, K. FORT, D. SEDDAH, É. VILLEMONTE DE LA CLERGERIE. *Deep Syntax Annotation of the Sequoia French Treebank*, in "International Conference on Language Resources and Evaluation (LREC)", Reykjavik, Iceland, May 2014, https://hal.inria.fr/hal-00969191

[22] M. COLINET, L. DANLOS, M. DARGNAT, G. WINTERSTEIN. *Uses of the prepostion <<pour>> introducing an infinitival clause: description, formal criteria and corpus annotation* , in "4ème Congrès Mondial de Linguistique Française", Berlin, Germany, F. NEVEU, P. BLUMENTHAL, L. HRIBA, A. GERSTENBERG, J. MEINSCHAEFER, S. PRÉVOST (editors), SHS Web of Conferences, EDP Sciences, July 2014, vol. 8, pp. 3041 - 3058 [*DOI :* 10.1051/SHSCONF/20140801071], https://hal.inria.fr/hal-01084546

[23] B. CRABBÉ. *An LR-inspired generalized lexicalized phrase structure parser*, in "COLING", Dublin, Ireland, 2014, https://hal.inria.fr/hal-01105142

[24] B. CRABBÉ. *Un analyseur discriminant de la famille LR pour l'analyse en constituants*, in "TALN 2014", Marseille, France, 2014, https://hal.inria.fr/hal-01105143

[25] L. DANLOS, A. MASKHARASHVILI, S. POGODALLA. *An ACG Analysis of the G-TAG Generation Process*, in "INLG 2014 - 8th International Natural Language Generation Conference", Philadelphia, PA, United States, M. MITCHELL, K. MCCOY, D. MCDONALD, A. CAHILL (editors), Association for Computational Linguistics, June 2014, pp. 35-44, https://hal.inria.fr/hal-00999595

[26] L. DANLOS, A. MASKHARASHVILI, S. POGODALLA. *An ACG View on G-TAG and Its g-Derivation*, in "LACL 2014 - Eight International Conference on Logical Aspects of Computational Linguistics", Toulouse, France, N. ASHER, S. SOLOVIEV (editors), Springer, June 2014, vol. 8535, pp. 70-82 [*DOI :* 10.1007/978-3-662-43742-1_6], https://hal.inria.fr/hal-00999633

[27] L. DANLOS, A. MASKHARASHVILI, S. POGODALLA. *Génération de textes : G-TAG revisité avec les Grammaires Catégorielles Abstraites*, in "TALN 2014 - 21ème conférence sur le Traitement Automatique des Langues Naturelles", Marseille, France, Actes de TALN 2014, Association pour le Traitement Automatique des Langues, July 2014, vol. 1, pp. 161-172, https://hal.inria.fr/hal-00999589

[28] L. DANLOS, T. NAKAMURA, Q. PRADET. *Vers la création d'un Verbnet du français*, in "TALN - 21ème conférence sur le Traitement Automatique des Langues Naturelles, Atelier FondamenTAL", Marseille, France, July 2014, https://hal.inria.fr/hal-01084681

[29] D. FIŠER, B. SAGOT. *Automatic extension and cleaning of sloWNet*, in "Devete konference Jezikovne Tehnologije / Ninth Language Technologies Conference", Ljubljana, Slovenia, October 2014, https://hal.inria.fr/hal-01078839

[30] K. GÁBOR. *Le syste`me WoDiS - WOlf & DIStributions pour la substitution lexicale*, in "Sémantique Distributionnelle - Atelier TALN 2014", Marseille, France, July 2014, https://hal.inria.fr/hal-01022406

[31] K. GÁBOR, B. SAGOT. *Automated Error Detection in Digitized Cultural Heritage Documents*, in "EACL 2014 Workshop on Language Technology for Cultural Heritage", Göteborg, Sweden, April 2014, https://hal.inria.fr/hal-01022402

[32] V. HANOKA, B. SAGOT. *YaMTG: An Open-Source Heavily Multilingual Translation Graph Extracted from Wiktionaries and Parallel Corpora*, in "Language Resources and Evaluation Conference", Reykjavik, Iceland, European Language Resources Association, May 2014, https://hal.inria.fr/hal-01022306

[33] J. HUNTER, L. DANLOS. *Because we say so*, in "EACL - 14th Conference of the European Chapter of the Association for Computational Linguistics, Workshop on Computational Approaches to Causality in Language", Gothenburg, Sweden, April 2014, https://hal.inria.fr/hal-00985486

[34] M. MORARDO, É. VILLEMONTE DE LA CLERGERIE. *Towards an environment for the production and the validation of lexical semantic resources*, in "The 9th edition of the Language Resources and Evaluation Conference (LREC)", Reykjavik, Iceland, ELRA, May 2014, https://hal.inria.fr/hal-01005464

[35] Q. PRADET, L. DANLOS, G. DE CHALENDAR. *Adapting VerbNet to French using existing resources*, in "LREC'14 - Ninth International Conference on Language Resources and Evaluation", Reykjavík, Iceland, May 2014, https://hal.inria.fr/hal-01084560

[36] C. RIBEYRE, M. CANDITO, D. SEDDAH. *Semi-Automatic Deep Syntactic Annotations of the French Treebank*, in "The 13th International Workshop on Treebanks and Linguistic Theories (TLT13)", Tübingen, Germany, Proceedings of TLT 13, Tübingen Universität, December 2014, https://hal.inria.fr/hal-01089198

[37] C. RIBEYRE, É. VILLEMONTE DE LA CLERGERIE, D. SEDDAH. *Alpage: Transition-based semantic graph parsing with syntactic features*, in "International Workshop on Semantic Evaluation", Dublin, Ireland, August 2014, https://hal.inria.fr/hal-01052485

[38] B. SAGOT. *DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German*, in "Language Resources and Evaluation Conference", Reykjavik, Iceland, European Language Resources Association, May 2014, https://hal.inria.fr/hal-01022288

[39] Y. SCHERRER, B. SAGOT. *A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages*, in "Language Resources and Evaluation Conference", Reykjavik, Iceland, European Language Resources Association, May 2014, https://hal.inria.fr/hal-01022298

[40] D. TRIBOUT, L. BARQUE, P. HAAS, R. HUYGHE. *De la simplicité en morphologie*, in "Congrès Mondial de Linguistique Française (CMLF 2014)", Berlin, Germany, 2014 [*DOI :* 10.1051/SHSCONF/20140801182], https://hal.archives-ouvertes.fr/hal-01091007

[41] É. VILLEMONTE DE LA CLERGERIE. *Jouer avec des analyseurs syntaxiques*, in "TALN 2014", Marseilles, France, ATALA, July 2014, https://hal.inria.fr/hal-01005477

National Conferences with Proceedings

[42] M. DJEMAA. *Traitement FrameNet des constructions à attribut de l'objet*, in "RECITAL'2014 - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues", Marseille, France, Aix-Marseille Université, July 2014, pp. 13 - 25, https://hal.archives-ouvertes.fr/hal-01075232

[43] G. PERRIER, M. CANDITO, B. GUILLAUME, C. RIBEYRE, K. FORT, D. SEDDAH. *Un schéma d'annotation en dépendances syntaxiques profondes pour le français*, in "TALN - Traitement Automatique des Langues Naturelles", Marseille, France, July 2014, pp. 574-579, https://hal.inria.fr/hal-01054407

[44] B. SAGOT, L. DANLOS, M. COLINET. *Sous-catégorisation en pour et syntaxe lexicale*, in "Traitement Automatique du Langage Naturel 2014", Marseille, France, July 2014, https://hal.inria.fr/hal-01022351

[45] B. SAGOT, K. GÁBOR. *Détection et correction automatique d'entités nommées dans des corpus OCRisés*, in "Traitement Automatique du Langage Naturel 2014", Marseille, France, July 2014, https://hal.inria.fr/hal-01022378

Conferences without Proceedings

[46] M. GRANT, J. THUILIER, B. CRABBÉ, A. ABEILLÉ. *The role of animacy in sentence production: Evidence from French*, in "Annual Conference of the Canadian Linguistics Association (ACL)", Toronto, Canada, 2014, https://hal.inria.fr/hal-01105148

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

Scientific Books (or Scientific Book chapters)

[48] B. CRABBÉ, D. DUCHIER, Y. PARMENTIER, S. PETITJEAN. *Constraint-driven Grammar Description*, in "Constraints and Language", P. BLACHE, H. CHRISTIANSEN, V. DAHL, D. DUCHIER, J. VILLADSEN (editors), Cambridge Scholar Publishing, October 2014, pp. 93-121, https://hal.archives-ouvertes.fr/hal-01059206

[49] L. DANLOS, P. DE GROOTE, S. POGODALLA. *A Type-Theoretic Account of Neg-Raising Predicates in Tree Adjoining Grammars*, in "New Frontiers in Artificial Intelligence. JSAI-isAI 2013 Workshops, LENLS, JURISIN, MiMI, AAA, and DDS, Kanagawa, Japan, October 27-28, 2013, Revised Selected Papers", Y. NAKANO, K. SATOH, D. BEKKI (editors), Lecture Notes in Computer Science, Springer International Publishing, November 2014, vol. 8417, pp. 3-16 [*DOI :* 10.1007/978-3-319-10061-6_1], https://hal.inria.fr/hal-00868382

[50] K. FORT, G. ADDA, B. SAGOT, J. MARIANI, A. COUILLAULT. *Crowdsourcing for Language Resource Development: Criticisms About Amazon Mechanical Turk Overpowering Use*, in "Human Language Technology Challenges for Computer Science and Linguistics", Z. VETULANI, J. MARIANI (editors), Springer International Publishing, July 2014, pp. 303-314 [*DOI :* 10.1007/978-3-319-08958-4_25], https://hal.inria.fr/hal-01053047

[51] J. THUILIER, A. ABEILLÉ, B. CRABBÉ. *Ordering preferences for postverbal complements in French*, in "Ecological and Data-Driven Perspectives in French Language Studies", Cambridge Scholars Publishing, 2014, https://hal.inria.fr/hal-01105151

Other Publications

[52] B. CRABBÉ, D. SEDDAH. *Multilingual discriminative shift reduce phrase structure parsing for the SPMRL 2014 shared task*, 2014, First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages, https://hal.inria.fr/hal-01105345

[53] M. GRANT, J. THUILIER, B. CRABBÉ, A. ABEILLÉ. *The role of conceptual accessibility on word order alternations in French: Evidence from sentence recall*, 2014, International Workshop on Language Production, https://hal.inria.fr/hal-01105144

References in notes

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

[55] 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

[56] S. AFANTENOS, N. ASHER, F. BENAMARA, M. BRAS, C. FABRE, L.-M. HO-DAC, A. LE DRAOULEC, P. MULLER, M.-P. PERY-WOODLEY, L. PRÉVOT, J. REBEYROLLES, L. TANGUY, M. VERGEZ-COURET, L. VIEU. *An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus*, in "Proceedings of LREC", 2012

[57] C. F. BAKER, C. J. FILLMORE, J. B. 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", Montreal, Canada, 1998, pp. 86-90

[58] B. BOHNET. *Very High Accuracy and Fast Dependency Parsing is Not a Contradiction*, in "Proceedings of the 23rd International Conference on Computational Linguistics", Stroudsburg, PA, USA, COLING '10, Association for Computational Linguistics, 2010, pp. 89–97

[59] 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

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

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

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

[63] J. BRESNAN, A. CUENI, T. NIKITINA, R. H. BAAYEN. *Predicting the Dative Alternation*, in "Cognitive Foundations of Interpretation", G. BOUME, I. KRAEMER, J. ZWARTS (editors), Royal Netherlands Academy of Science, 2007

[64] J. BRESNAN, M. FORD. *Predicting syntax: Processing dative constructions in American and Australian varieties of English*, in "Language", 2010, vol. 86, pp. 168–213, http://dx.doi.org/10.1353/lan.0.0189

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

[66] 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

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

[68] 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

[69] 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

[70] E. CHARNIAK. *A Maximum-Entropy-Inspired Parser*, in "Proc. of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL)", Seattle, 2000, pp. 132–139

[71] 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

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

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

[74] 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, https://hal.archives-ouvertes.fr/hal-00932372

[75] B. CRABBÉ, M. CANDITO. *Expériences D'Analyse Syntaxique Statistique Du Français*, in "Actes de la 15ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'08)", Avignon, France, 2008, pp. 45–54

[76] 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

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

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

[79] 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

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

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

[82] D. FLICKINGER, Y. ZHANG, V. KORDONI. *DeepBank: A dynamically annotated treebank of the Wall Street Journal*, in "Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories", 2012, pp. 85–96, ftp://59.108.48.12/parsing/reading/hpsg/lingo/DeepBank_tlt11.pdf

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

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

[85] 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

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

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

[88] 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

[89] B. LEVIN. *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago PressChicago, IL, 1993

[90] C. MARCHELLO-NIZIA. *L'évolution du français: ordre des mots, démonstratifs, accent tonique*, Collection linguistique, A. Colin, 1995, http://books.google.fr/books?id=bzRiQgAACAAJ

[91] M. MARCUS, B. SANTORINI, M. A. MARCINKIEWICZ. *Building a large annotated corpus of English: The Penn Treebank*, in "Computational Linguistics", 1993, vol. 19, n^{o} 2, pp. 313–330, http://acl.ldc.upenn.edu/J/J93/J93-2004.pdf

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

[93] R. NAVIGLI. *Word Sense Disambiguation: A Survey*, in "ACM Comput. Surv.", February 2009, vol. 41, n^{o} 2, pp. 10:1–10:69, http://doi.acm.org/10.1145/1459352.1459355

[94] A. NAZARENKO, H. ZARGAYOUNA. *Evaluating term extraction*, in "Proceegings of International Conference Recent Advances in Natural Language Processing (RANLP'09)", Borovets, Bulgarie, 2009, pp. 299-304, http://hal.archives-ouvertes.fr/hal-00517090

[95] J. NIVRE, J. HALL, J. NILSSON, A. CHANEV, G. ERYIĞIT, S. KÜBLER, S. MARINOV, E. MARSI. *MaltParser: A Language-Independent System for Data-Driven Dependency Parsing*, in "Natural Language Engineering", 2007, vol. 13, n^{o} 2, pp. 95–135

[96] J. NIVRE, J. NILSSON. *Pseudo-projective dependency parsing*, in "Proc. of the 43rd Annual Meeting on Association for Computational Linguistics", Association for Computational Linguistics, 2005, pp. 99–106

[97] J. NIVRE. *An efficient algorithm for projective dependency parsing*, in "Proc. of the 8th International Workshop on Parsing Technologies (IWPT", Citeseer, 2003

[98] 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

[99] S. OEPEN, M. KUHLMANN, Y. MIYAO, D. ZEMAN, D. FLICKINGER, J. HAJIC, A. IVANOVA, Y. ZHANG. *SemEval 2014 Task 8: Broad-Coverage Semantic Dependency Parsing*, in "Proc. of the 8th International Workshop on Semantic Evaluation", 2014, pp. 63–72

[100] 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

[101] 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://aclweb.org/anthology/N07-1051

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

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

[104] 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

[105] 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

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

[107] K. SAGAE, J. TSUJII. *Shift-Reduce Dependency DAG Parsing*, in "Proc. of the 22nd International Conference on Computational Linguistics (Coling 2008)", 2008, pp. 753–760

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

[109] 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/

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

[111] 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

[112] B. SAGOT, D. FIŠER. *Building a free French wordnet from multilingual resources*, in "OntoLex", Marrakech, Morocco, May 2008, https://hal.inria.fr/inria-00614708

[113] 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/

[114] 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

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

[116] B. SAGOT. *Linguistic facts as predicates over ranges of the sentence*, in "Lecture Notes in Computer Science 3492 (© Springer-Verlag), Proceedings of LACL'05", Bordeaux, France, April 2005, pp. 271–286

[117] 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

[118] 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

[119] 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

[120] 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

[121] H. SCHUTZE. *Ambiguity in Language Learning: computational and Cognitive Models*, Stanford, 1995

[122] 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, October 2009, pp. 150-161

[123] 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

[124] 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

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

[126] 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

[127] 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, October 2013, pp. 146–182, https://hal.archives-ouvertes.fr/hal-00877096

[128] R. STERN, B. SAGOT. *Resources for Named Entity Recognition and Resolution in News Wires*, in "Entity 2010 Workshop at LREC 2010", Malte Valletta, 2010

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

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

[131] Y. TSURUOKA, Y. MIYAO, J. TSUJII. *Towards efficient probabilistic HPSG parsing: integrating semantic and syntactic preference to guide the parsing*, in "Proceedings of the IJCNLP-04 Workshop on Beyond Shallow Analyses", Citeseer, 2004

[132] 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, United States, June 2010, pp. 111-118, https://hal.inria.fr/inria-00551974

[133] 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, October 2013, https://hal.inria.fr/hal-00879129

[134] 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

[135] 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

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

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

[138] 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

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

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

[141] Y. ZHANG, J. NIVRE. *Transition-based dependency parsing with rich non-local features*, in "Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2", Association for Computational Linguistics, 2011, pp. 188–193

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