IN PARTNERSHIP WITH:
Université Denis Diderot (Paris 7)

Activity Report 2015

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

Creation of the Project-Team: 2008 January 01

Keywords:

Computer Science and Digital Science:

3.1.1. - Modeling, representation
3.1.7. - Open data
3.2.2. - Knowledge extraction, cleaning
3.2.4. - Semantic Web
3.3.2. - Data mining
5.8. - Natural language processing
8.2. - Machine learning
8.4. - Natural language processing

Other Research Topics and Application Domains:

1.3.2. - Cognitive science
9.5.10. - Digital humanities
9.5.8. - Linguistics
9.7.1. - Open access
9.7.2. - Open data

1. Members

Research Scientists

Benoît Sagot [Team leader, Inria, Researcher]
Pierre Boullier [Inria, Senior Researcher, Emeritus]
Laurent Romary [Inria, Senior Researcher, HdR]
Éric Villemonte de La Clergerie [Inria, Researcher]

Faculty Members

Lucie Barque [Univ. Paris XIII, Associate Professor]
Mathieu Constant [Univ. Paris Est, Associate Professor “en délégation”, from Sep 2015]
Benoit Crabbé [Univ. Paris VII, Associate Professor]
Laurence Danlos [Univ. Paris VII, Professor, HdR]
Marie Candito [Univ. Paris VII, Associate Professor]
Djamé Seddah [Univ. Paris IV, Associate Professor]

Engineers

Margot Colinet [Inria, granted by ANR Polymnie project until May 2015 and in Jul 2015, and by Labex EFL in June 2015]
Vanessa Combet Meunier [Inria, granted by Labex EFL and by Caisse des Dépôts et Consignations]
Noemie Faivre [Inria, until Jun 2015, granted by ANR ASFALDA project]
Pierre Magistry [Inria, granted by Caisse des Dépôts et Consignations until March 2014, then by Caisse des Dépôts et Consignations project from Oct 2015 to Nov 2015, then from LabEx EFL since Dec 2015]
Virginie Mouilleron [Inria, until Mar 2015, granted by Labex EFL and ANR ASFALDA project]
Charles Riondet [Inria, granted by H2020 PARTHENOS, from Nov 2015]
Stéphane Riou [CNRS, granted (and hosted most of the time) by Institut de Linguistique Française, from May 2015]

Jacques Steinlin [Inria, granted by Caisse des Dépôts et Consignations, until May 2015]

PhD Students

Quentin Pradet [Inria, granted by LabEx EFL, until Fev 2015]
Marion Baranes [viavoo and Univ. Paris VII, until Oct 2015]
Sarah Beniamine [(member of Alpage until August 2015) Univ. Paris VII, granted by Labex EFL]
Timothée Bernard [ENS Lyon]
Chloé Braud [Univ. Paris VII, until Dec 2015]
Maximin Coavoux [Univ. Paris VII]
Marianne Djemaa [Univ. Paris VII, granted by ANR ASFALDA- project]
Valérie Hanoka [Univ. Paris VII, until July 2015]
Corentin Ribeyre [Min. Ens. Sup. Recherche and Univ. Paris VII, then Inria (Relais Thèse)]
Raphaël Salmon [Yseop and Univ. Paris VII, granted by CIFRE]

Post-Doctoral Fellows

Kata Gábor [Inria, until Feb 2015, granted by Caisse des Dépôts et Consignations]
Alexandra Simonenko [Univ. Paris VII, granted by LabEx EFL]

Administrative Assistant

Christelle Guiziou [Inria]

Others

Emmanuel Lassalle [Univ. Paris VII, until March 2015]
Rachel Bawden [Inria, M2 intern, from Feb 2015 until Jul 2015]
Laurine Lamy [Inria, intern, from Jul 2015 until Sep 2015]
Nicholas Parslow [Inria, M2 intern, from Feb 2015 until Jul 2015]

2. Overall Objectives

2.1. Overall Objectives

The Alpage team is specialised in **Language modelling**, **Computational linguistics** and **Natural Language Processing (NLP)**. These fields are 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, data mining, text simplification, automatic or computer-aided translation, automatic summarisation, 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 **transdisciplinary** domain: 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) and in applied mathematics (to automatically acquire linguistic or general knowledge). It is one of the specificities of Alpage to put together both researchers with a background in computer science (Inria members) and researchers with a background more oriented towards linguistics, all of them working on a single topic: simulation on computers of human understanding and production of language.

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 mostly on French. One of the main objectives of the team is to develop **generic** linguistically relevant *and* computationally efficient tools and resources for French which are freely distributed. These products are dedicated to the francophone community so as to help French to be part of the new information society. However, Alpage 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., English, Spanish, Polish, Persian and others). This is of course of high relevance, among others, for language-independant modelling and multi-lingual tools and applications.

Alpage covers all linguistics domains, although not at the same level. At the creation of the team, the morphological and syntactic levels was the most developed and led to a number of applications, especially with industrial partners. However, the importance of the semantic and discourse levels has increased during the evaluation period and the interface between syntax and semantics has been better worked on. Our goal is also to apply our knowledge, tools and resources in various contexts such as research in experimental linguistics, operational applications and prototypes as well as standardisation of linguistic resources and annotations.

Our four main objectives, as reworded and updated while writing the 2015 Inria evaluation report, are the following:

- **Objective i: Towards large scale natural language understanding at the sentence level** This objective covers all the work carried out on shallow processing, tagging, syntactic parsing, deep-syntactic parsing and shallow semantic parsing.
- **Objective ii : Language resource development, evaluation and use** This objective covers all language resource development efforts that range from morphology to semantics including syntax, but not including supra-sentential (discourse) resources.
- **Objective iii : Modelling and parsing supra-sentential phenomena** This objectives covers all efforts, including language resource development efforts, regarding discourse and other phenomena that cross sentence boundaries (e.g. anaphora).
- **Objective iv : Application domains** This objectives regroups the three main application domains for Alpage: empirical linguistics, academic downstream NLP applications and industrial applications.

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

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

than MCS formalisms, Range Concatenation Grammars (RCGs, introduced and used by Alpage members, such as Pierre Boullier and Benoît Sagot [49], [90], [96]) 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 [110], [106]. 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 Candito, Benoit Crabbé, Éric Villemonte de La Clergerie, Benoît Sagot, Corentin Ribeyre, Pierre Boullier, Maximin Coavoux.

Contrary to symbolic approaches to parsing, in statistical parsing, the grammar is extracted from a corpus of syntactic trees : a treebank. The main advantage of the statistical approach is to encode within the same framework the parsing and disambiguating tasks. The extracted grammar rules are associated with probabilities that allow to score and rank the output parse trees of an input sentence. This obvious advantage of probabilistic context-free grammars has long been counterbalanced by two main shortcomings that resulted in poor performance for plain PCFG parsers: (i) the generalization encoded in non terminal symbols that stand for syntagmatic phrases is too coarse (so probabilistic independence between rules is too strong an assertion) and (ii) lexical items are underused. In the last decade though, effective solutions to these shortcomings have been proposed. Symbol annotation, either manual [73] or automatic [80], [81] 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 [59], or probabilistic models derived from lexicalized Tree Adjoining grammars, such as Stochastic Tree Insertion Grammars [56].

In the same period, totally different parsing architectures have been proposed, to obtain dependency-based syntactic representations. The properties of dependency structures, in which each word is related to exactly one other word, make it possible to define dependency parsing as a sequence of simple actions (such as read buffer and store word on top of a stack, attach read word as dependent of stack top word, attach read word as governor of stack top word ...) [113], [79]. 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 [77]. 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 [44], [43] and derive the best input for syntagmatic statistical parsing [60]. Benchmarking several PCFG-based learning frameworks [99] 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) [81].

In parallel to the work on dependency based representation, presented in the next paragraph, we also conducted a preliminary set of experiments on richer parsing models based on Stochastic Tree Insertion Grammars as used in [56] and which, besides their inferior performance compared to PCFG-LA based parser, raise promising results with respect to dependencies that can be extracted from derivation trees. One variation we explored, that uses a specific TIG grammar instance, a *vertical* grammar called *spinal* grammars, exhibits interesting properties wrt the grammar size typically extracted from treebanks (a few hundred unlexicalized trees, compared to 14 000 CFG rules). These models are currently being investigated in our team [102].

Pursuing our work on PCFG-LA based parsing, we investigated the automatic conversion of the treebank into dependency syntax representations [55], 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 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 , is not a trivial task [91]. 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 [105]. 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.

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 [69]. 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 [82], as Alpage was ranked second and third [101]. 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) [100].

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 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 [95]. At the syntactic level, work has been achieved on automatic acquisition of atomic syntactic information and automatic detection of errors in the lexicon [114],[6]. At the semantic level, automatic wordnet development tools have been described [85], [111], [71], [70]. 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* [92], [97], 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 [94], [16], the first freely available resource of the kind;
- a French FrameNet lexicon (together with an annotated corpus) within the ASFALDA ANR project;
- and a French VerbNet.

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 exist for German, 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, Jacques Steinlin, Chloé Braud, Timothée Bernard, 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 [63].

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, [64],[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.
- Empirical 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, text mining, spelling correction and empirical linguistics.

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 and Alexina morphological lexicons. Several other parts of *vera* have been co-developed by Verbatim Analysis and 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 have worked as PACTE-funded post-docs until the end of the project in March 2015.

4.6. Empirical linguistics

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

Alpage is a team that dedicates efforts in producing ressources and algorithms for processing large amounts of textual materials. These ressources can be applied not only for purely NLP purposes but also for linguistic purposes. Indeed, the specific needs of NLP applications led to the development of electronic linguistic resources (in particular lexica, annotated corpora, and treebanks) that are sufficiently large for carrying statistical analysis on linguistic issues. In the last 10 years, pioneering work has started to use these new data sources to the study of English grammar, leading to important new results in such areas as the study of syntactic preferences [51], [112], the existence of graded grammaticality judgments [72].

The reasons for getting interested for statistical modelling of language can be traced back by looking at the recent history of grammatical works in linguistics. In the 1980s and 1990s, theoretical grammarians have been mostly concerned with improving the conceptual underpinnings of their respective subfields, in particular through the construction and refinement of formal models. In syntax, the relative consensus on a generative-transformational approach [57] gave way on the one hand to more abstract characterizations of the language faculty [57], and on the other hand to the construction of detailed, formally explicit, and often implemented, alternative formulation of the generative approach [50], [83]. For French several grammars have been implemented in this trend, such as the tree adjoining grammars of [54], [61] 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 [98]. 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 has been supported in the past years by one PhD project, whose defense has now taken place.

In parallel, quantitative methods are applied to computational morphology, in particlar in relation with Sarah Beniamine's PhD supervised by Olivier Bonami (LLF, CNRS, U. Paris Diderot and U. Paris Sorbonne) [31], [20], [32]. 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. Highlights of the Year

5.1. Highlights of the Year

In 2015, Alpage has obtained three new national fundings: the team is a partner of two new ANR projects (PARSEME-FR and SoSweet) and an industrial contract ("RAPID" project VeRDI).

5.1.1. Awards

Best Paper Award at the TALN 2015 conference .

BEST PAPER AWARD:

[22]
M. COAVOUX, B. CRABBÉ. *Comparaison d'architectures neuronales pour l'analyse syntaxique en constituants*, in "TALN 2015", Caen, France, 2015, https://hal.inria.fr/hal-01174613

6. New Software and Platforms

6.1. Alexina

Atelier pour les LEXiques INformatiques et leur Acquisition
FUNCTIONAL DESCRIPTION

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 Lefff, a morphological and syntactic lexicon for French.

- Participants: Benoît Sagot and Laurence Danlos
- Contact: Benoît Sagot
- URL: http://gforge.inria.fr/projects/alexina/

6.2. Bonsai

FUNCTIONAL DESCRIPTION

Alpage has developed 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.

- Participants: Marie-Hélène Candito, Djamé Seddah and Benoit Crabbé
- Contact: Marie-Hélène Candito
- URL: http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

6.3. Crapbank

French Social Media Bank
FUNCTIONAL DESCRIPTION

The French Social Media Bank is a treebank of French sentences coming from various social media sources (Twitter(c), Facebook(c)) and web forums (JeuxVidéos.com(c), Doctissimo.fr(c)). It contains different kind of linguistic annotations: - part-of-speech tags - surface syntactic representations (phrase-based representations) as well as normalized form whenever necessary.

- Contact: Djamé Seddah

6.4. DyALog

FUNCTIONAL DESCRIPTION

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.

- Participant: Éric Villemonte de La Clergerie
- Contact: Éric Villemonte de La Clergerie
- URL: http://dyalog.gforge.inria.fr/

6.5. FDTB1

- Contact: Laurence Danlos

6.6. FQB

French QuestionBank
FUNCTIONAL DESCRIPTION

The French QuestionBanks is a corpus of around 2000 questions coming from various domains (TREC data set, French governmental organisation, NGOs, etc..) it contains different kind of annotations - morpho-syntactic ones (POS, lemmas) - surface syntaxe (phrase based and dependency structures) with long-distance dependency annotations.

The TREC part is aligned with the English QuestionBank (Judge et al, 2006).

- Contact: Djamé Seddah

6.7. FRMG

- Participant: Éric Villemonte de La Clergerie
- Contact: Éric de La Clergerie
- URL: http://mgkit.gforge.inria.fr/

6.8. French Question Bank

- Contact: Djamé Seddah

6.9. LexConn

- Contact: Laurence Danlos

6.10. LexViz

FUNCTIONAL DESCRIPTION

In the context of the industrial collaboration of ALPAGE with the company Lingua et 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).

- Participants: Éric Villemonte de La Clergerie and Mickael Morardo
- Contact: Éric Villemonte de La Clergerie

6.11. MElt

Maximum-Entropy lexicon-aware tagger
KEYWORD: Part-of-speech tagger
FUNCTIONAL DESCRIPTION

MElt is a freely available (LGPL) state-of-the-art sequence labeller that is meant to be trained on both an annotated corpus and an external lexicon. It was developed by Pascal Denis and Benoît Sagot within the Alpage team, a joint Inria and Université Paris-Diderot team in Paris, France. MElt allows for using multiclass Maximum-Entropy Markov models (MEMMs) or multiclass perceptrons (multitrons) as underlying statistical devices. Its output is in the Brown format (one sentence per line, each sentence being a space-separated sequence of annotated words in the word/tag format).

MElt has been trained on various annotated corpora, using Alexina lexicons as source of lexical information. As a result, models for French, English, Spanish and Italian are included in the MElt package.

MElt also includes a normalization wrapper aimed at helping processing noisy text, such as user-generated data retrieved on the web. This wrapper is only available for French and English. It was used for parsing web data for both English and French, respectively during the SANCL shared task (Google Web Bank) and for developing the French Social Media Bank (Facebook, twitter and blog data).

- Contact: Benoît Sagot
- URL: https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=MElt

6.12. Mgwiki

FUNCTIONAL DESCRIPTION

Mgwiki is 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.

- Participants: Éric Villemonte de La Clergerie and Paul Bui–quang
- Contact: Éric Villemonte de La Clergerie
- URL: http://alpage.inria.fr/frmgwiki/

6.13. OGRE

Optimized Graph Rewriting Engine
FUNCTIONAL DESCRIPTION

OGRE is a graph rewriting system specifically designed for manipulating linguistic trees and graphs, It relies on a rule specification language for expressing graph rewriting patterns. The transformation is performed in two steps:

First, the system performs simple transformations following the rewriting patterns,

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.

- Participants: Corentin Ribeyre, Djamé Seddah, Éric Villemonte de La Clergerie and Marie-Hélène Candito
- Contact: Corentin Ribeyre
- URL: http://www.corentinribeyre.fr/projects/view/OGRE

6.14. SYNTAX

FUNCTIONAL DESCRIPTION

Syntax system 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 Sx Pipe 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).

- Participants: Pierre Boullier, Philippe Deschamps and Benoît Sagot
- Contact: Pierre Boullier
- URL: http://syntax.gforge.inria.fr/

6.15. Sequoia corpus

FUNCTIONAL DESCRIPTION

The Sequoia corpus contains French sentences, annotated with various linguistic information: - parts-of-speech - surface syntactic representations (both constituency trees and dependency trees) - deep syntactic representations (which are deep syntactic dependency graphs)

- Contact: Djamé Seddah

6.16. SxPipe

SCIENTIFIC DESCRIPTION

Developed for French and for other languages, Sx Pipe 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.

FUNCTIONAL DESCRIPTION

SxPipe 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) and for surface processing (named entities recognition, text normalization, unknown word extraction and processing...).

- Participants: Pierre Boullier, Benoît Sagot, Kata Gábor, Marion Baranes, Pierre Magistry, Éric Villemonte de La Clergerie and Djamé Seddah
- Contact: Benoît Sagot
- URL: http://lingwb.gforge.inria.fr/

6.17. VerbeNet

- Contact: Laurence Danlos

6.18. hyparse

Alpage Hybrid Parser

KEYWORDS: Parsing - NLP

FUNCTIONAL DESCRIPTION

Multilingual Phrase Structure Parser

- Contact: Benoit Crabbé
- URL: http://hyparse.gforge.inria.fr

6.19. DyALog-sr

DYALOG-sr

KEYWORDS: Parsing - NLP

FUNCTIONAL DESCRIPTION

DyALog-SR is a transition-based dependency parser, built on top of DyALog system. Parsing relies on dynamic programming techniques to handle beams. Supervised learning exploit a perceptron and aggressive early updates. DyALog-SR can handle word lattice and produce dependency graphs (instead of basic trees). It was tested during several shared tasks (SPMRL'2013 and SEMEVAL'2014). It achieves very good accuracy on French TreeBank, alone or by coupling with FRMG parser.

- Contact: Éric de La Clergerie

7. New Results

7.1. Playing with DyALog-based parsers

Participants: Éric Villemonte de La Clergerie, Nicholas Parslow.

Éric de la Clergerie has continued the development of two DyALog-based parsers, namely DyALog-SR, a transition-based dependency parser, and FRMG, a wide-coverage French TAG based on an underlying meta-grammar.

The coverage of FRMG has been extended to cover more (rare) syntactic phenomena. A new conversion scheme has been added for the French version of the Universal Dependency Scheme. Preliminary evaluation experiments have been conducted on the French UD corpus, with both FRMG and DyALog-SR. FRMG has also been evaluated on the French SPMRL corpus, alone and with coupling with DyALog-sr

A new notion of secondary edges has been investigated in FRMG metagrammar and parser to provide additional dependency edges, helpful for understanding parsing outputs. In particular, secondary edges are used to denote controls between a verb and its hidden subject.

FRMG's disambiguation tuning is learned from CONLL-like treebanks using supervised learning method. We have conducted preliminary experiments to use unsupervised learning methods with observed accuracy gains between 1 to 1.5 points w.r.t. the no tuning case. However, trying to mix supervised and unsupervised methods have shown no significant gain w.r.t. the supervised case.

The hybridation of FRMG and DyALog-SR have been tried on a larger spectrum of treebanks.

FRMG has also been exploited during the Master internship of Nicholas Parslow about the use of NLP tools to provide feedback information and correlations on essays written by non-native French learners. In particular, the correction mechanism of FRMG has been extended to cover more cases of frequent errors and provide more explicit messages.

7.2. Linear-time discriminant syntactico-semantic parsing

Participants: Benoit Crabbé, Maximin Coavoux, Rachel Bawden.

In this module we study efficient and accurate models of statistical phrase structure parsing. We focus on linear time lexicalized parsing algorithms (shift reduce) with approximations entailing linear time processing. The existing prototype involves a global discriminant parsing model of the large margin family (Perceptron,Mira, SVM) able to parse user defined structured input tokens [62]. 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 model has been generalized in a multilingual setting where we are among the state of the art systems and state of the art on some languages [23]. To our knowledge the parser is one of the fastest existing multilingual phrase structure parser. In order to ease model design for multilingual settings, we currently study efficient feature selection procedures for automating model adaptation to new languages.

We have also extended our model to continuous representations by means of deep learning methods. We currently have a neural network based decision procedure for parsing [22]. It involves both greedy search and beam based search techniques. Current work focuses on the design of dynamic oracles for improving greedy search procedures. This framework is currently tested in the multilingual setting too.

Further work involves to tackle the knowledge acquisition bottleneck problem by integrating either symbolic knowledge such as dictionaries or semi-supervised procedures for improving the formal representation of lexical dependencies in order to leverage data sparsity and estimation issues recurrent in lexicalized parsing.

7.3. French Deep Syntactic Dependency Parsing

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

At Alpage, we used two distinct but complementary approaches to parse and produce deep syntactic dependency graphs from the DeepSequoia and the DeepFTB (crossref here). The first one was developped by using OGRE [87], [86], a graph rewriting system (crossref here). We developed a set of rewriting rules to transform surfacic syntactic dependency trees into deep syntactic dependency graphs, then we applied this set of rules on previously parsed surfacic trees. Those trees were produced using up to three different surfacic syntactic parsers: FRMG [109], DyALog-SR [109] and Mate [47]. The results were convincing and on par with what we got on English.

The second approach was based on the work made last year regarding the English broad-coverage semantic dependency parsing. We reused our two graph parsers (the first one is based on a previous work on DAG parsing [89] and the second one on the FRMG surfacic syntactic parser [109]) to parse the same graphs. As we previously have shown on English, the use of a mix of syntactic features (tree fragments from a constituent syntactic parser [80], dependencies from a syntactic parser [47], elementary spinal trees using a spine grammar [102], etc.) improve our results. Our intuition is that syntax and semantic are not independent of each other and using syntax could improve semantic parsing. Finally, we extended a dual-decomposition third-order graph parser [76] to incorporate our syntactic feature set and we were able to reach the best performances to this day on the task for both English [28] and French (Ribeyre et al, to appear).

7.4. Towards a French FrameNet

Participants: Marie Candito, Marianne Djemaa, Benoît Sagot.

The ASFALDA project [2] 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. The project started in October 2012, and will end in march 2016. Its objective is to build 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 [45], [3] 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 2015, we continued the corpus annotation phase, which started in 2014. We currently have about 90 frames and 790 lexical units with at least one annotated occurrence, totalizing about $11,000$ annotated occurrences. We also set up :

- procedures for checking the coherence of the annotations
- a procedure for extracting the "annotated lexicon", namely extract quantitative information about the annotated lexical units, and syntax/semantics interface information (in terms of the probabilistic distributions of the syntactic paths used to express a given semantic role)
- the graphical vizualization of the annotated corpus

We also just started a collaboration with the LIF laboratory for using deep syntactic representations for predicting semantic frames and roles.

7.5. Development of Verb∋net

Participants: Laurence Danlos, Quentin Pradet, Lucie Barque.

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

VerbNet is an English lexical resources for verbs, which is internationally known and widely used in numerous NLP applications [74]. 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, [84], [68] [14]. A first version of this work has been achieved in June 2015 in collaboration with Takuya Nakamura (Institut Gaspard Monge) [33].

The next step was to verify the coherence of the resource. A particular focus has been to check the way alternations have been encoded and to document this encoding. A journal article extracted from this documentation has been submitted to the *TAL* journal and Verb∋net will be released after getting the feedback of the editorial board.

7.6. Development of the French Discourse TreeBank (FDTB)

Participants: Laurence Danlos, Margot Colinet, Jacques Steinlin, Pierre Magistry.

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 [88], 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 [58], whose results have been used to enhance Le*fff* [93].

FDTB1 identifies 9 833 discourse connectives (among 18 535 sentences). This ressource is freely available and has been released in May 2015 [36].

FDTB2 is the next step in the creation of the FDTB. It consists in annotating the arguments of the discourse connectives identified in FDTB1 as well as the senses of these connectives (senses expressed through a set of discourse relations). This resource is still worked on.

7.7. Discourse Parsing

Participants: Chloé Braud, Laurence Danlos.

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 most 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, defended in December 2015, Chloé Braud develops a discourse relation classifier, carrying experiments on French and English. Focusing on the problem on implicit relation identification, this work explores ways of using raw data in combination with the available manually annotated data: this work led to systems based on domain adaptation methods exploiting automatically annotated explicit relations – demonstrating improvements on the French corpus Annodis and on the English corpus PDTB –, and to systems using word embeddings built from raw text to efficiently transform a word based representation of the data – leading to state-of-the art performance or above on the English corpus PDTB without the need of hand-crafted resources [21].

7.8. 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 [78] 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 [107]
- 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 [46]
- From a morphological point of view, nouns denoting events in French are mostly deverbal nouns (eg. *conversation* 'conversation', *promenade* 'stroll'), but there are also underived event nouns (eg. *guerre* 'war', *séisme* 'earthquake'). We compared their semantic properties in [35].
- Some lexical meanings are not easily captured by ontological semantic classes and a closer look has to be taken at them. Relational meanings in relational nouns are one of them [15].

7.9. Development of the Corpus de Référence du Français

Participants: Stéphane Riou, Benoît Sagot.

The 'Initiative Corpus de Référence du Français' (ICRF) is a project of Institut de Linguistique Française (ILF-FR2393 CNRS), coordinated by its director Franck Neveu and by Benoît Sagot.

The purpose of the ICRF is the development of a first prototype of the future French Reference Corpus, so as to assess the feasibility of this project and evaluate its potential impact. ICRF reuses existing freely-available French corpora, supplemented by additional data in an opportunistic fashion (e.g. a French media critic corpus and the corpus of talks given at an workshop on ethics and neurodegenerative diseases). ICRF preserves copyright and authorship of all corpora used. These corpora have been or will be part-of-speech tagged with MElt, converted to XML (TEI-P5-compliant) and made accessible via a web interface. The aim of ICRF is not to replace individual corpora and the interface will therefore allow, whenever possible, to easily recover access to each individual corpus. ICRF adds 5 metadata tags to categorize each individual corpus: spoken/written, text type and genre, linguistic competence level, date and linguistic area.

In 2015, the normalisation, tagging and conversion to XML of individual corpora has started, following the design of format specifications. The development of the web interface has already started, and a prototype is now available. Users can perform queries (search by tokens and/or POS) and use basic linguistic tools on the corpora (e.g. a concordancer). It is therefore more than a simple search interface or a download site: it improves research and selection of corpus.

7.10. Word order variation in Old French

Participants: Benoit Crabbé, Alexandra Simonenko.

As participant of the strand *Experimental Grammar* of the Labex EFL project *Empirical Foundations of Linguistics* [4] 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. [5]

The first quantitative results [29] will be compared with what is known from corpus studies on the relative ordering of subject and complement in Old French [75]. 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 [53]. This comparative perspective is expected to provide new insights on French language evolution.

7.11. Cross linguistic factors governing word order

Participant: Benoit Crabbé.

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 [52].

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 [25], we explore, in collaboration with Kristina Gulordava (PhD at the University of Geneva, former international visitor at Alpage), to which extent a computational corpus-based analysis can provide new evidence not only for empirical, but also for theoretical linguistic research.

8. Bilateral Contracts and Grants with Industry

8.1. 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 (contract ended in Dec 2014), see section 4.3),
- Lingua et Machina (DTI-funded engineer, see section 4.4),
- viavoo (PhD of Marion Baranes, employed at viavoo, started in 2012 and defended in Oct 2015 about the automatic normalisation of noisy texts),
- Yseop (“CIFRE” PhD of Raphael Salmon started in 2012 about automatic text generation)
- CEA-List (PhD of Quentin Pradet on the annotation of semantic roles in specific domains (defense in Feb 2015).
- Proxem (consulting)

[4] www.labex-efl.org
[5] SRCMF corpus: http://srcmf.org/; MCVF: http://www.voies.uottawa.ca

9. Partnerships and Cooperations

9.1. National Initiatives

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

Participants: Laurence Danlos, Benoît Sagot, Chloé Braud, Marie 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, both before the start of the LabEx EFL and within several EFL "scientific operations" (see Section 4.6). Moreover, the LabEx provides Alpage with opportunities for collaborating with new teams, e.g., on language resource development and empirical studies in collaboration with descriptive linguists.

The LabEx EFL's scientific activities are spread accross 7 autonomous scientific "strands". In 2015,Benoît Sagot, Marie Candito and Benoit Crabbé were respectively deputy-head of strand 6 on "Language Resources", strands 5 on "Computational semantic analysis" and strand 2 on "Experimental grammar from a cross-linguistic perspective". Several project members are in charge of research operations within these 3 strands.

9.1.2. ANR

9.1.2.1. ANR project PARSEME-FR (2016 - 2019)

Participants: Marie Candito, Mathieu Constant [principal investigator], Benoit Crabbé, Laurence Danlos, Éric Villemonte de La Clergerie, Djamé Seddah.

PARSEME-FR is a 4-year ANR research project headed by Mathieu Constant (LIGM, Université Paris-Est Marne-la-Vallée, currently in "délégation" at Alpage). PARSEME-FR partners are LIGM, Alpage, LI (Université de Tours), LIF (Aix-Marseille Université) and LIFO (Université d'Orléans). This project aims at improving linguistic representativeness, precision and computational efficiency of Natural Language Processing (NLP) applications, notably parsing. The project focuses on the major bottleneck of these applications: Multi-Word Expressions (MWEs), i.e. groups of words with a certain degree of idiomaticity such as "hot dog", "to kick the bucket", "San Francisco 49ers" or "to take a haircut". In particular, it aims at investigating the syntactic and semantic representation of MWEs in language resources, the integration of MWE analysis in (deep) syntactic parsing and its links to semantic processing. Expected deliverables include enhanced language resources (lexicons, grammars and annotated corpora) for French, MWE-aware (deep) parsers and tools linking predicted MWEs to knowledge bases. This proposal is a spin-off of the European IC1207 COST action PARSEME on the same topic.

Alpage will be participate mainly to two tasks: (i) the production of an evaluation corpus annotated with MWE and (ii) the production of MWE-aware statistical parsers, both for surface syntax and deep syntax. MWE recognition can be viewed as part of a more ambitious task of recovering the semantic units of a sentence. Combining it to deep syntactic parsing will provide a further step towards semantic parsing.

9.1.2.2. ANR project ASFALDA (2012 – 2016)

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

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

The project aims to provide both a French corpus with semantic annotations and automatic tools for shallow semantic analysis, using machine learning techniques to train analyzers on this corpus. The target semantic annotations are structured following the FrameNet framework [45] 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.

9.1.2.3. ANR project Polymnie (2012-2016)

Participants: Laurence Danlos, Éric Villemonte de La Clergerie, Timothée Bernard.

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 [65], [66], [67]. As regards discursive analysis, D-STAG formalism created by Laurence Danlos in the 00's has also been rewritten in ACG in 2015 [24] and enhanced with some preliminary linguistic work on attributions [39].

9.1.3. Other national initiatives

9.1.3.1. "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's aims was the improvement of 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 was more specifically targeted to 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 involved 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.

9.1.3.2. FUI project COMBI (2014-2016)

Participants: Laurence Danlos, Vanessa Combet Meunier, 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.

9.1.3.3. Institut de Linguistique Française and Consortium Corpus Écrits within the TGIR Huma-Num

Participants: Benoît Sagot, Stéphane Riou, 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, and in the development of a preliminary version of the future Corpus de Référence du Français (French Reference Corpus).

9.2. European Initiatives

9.2.1. H2020 PARTHENOS

Participants: Laurent Romary, Charles Riondet.

This EU project Parthenos of the H2020 INFRADEV program aims et strengthening the cohesion of research in the broad sector of Linguistic Studies, Humanities, Cultural Heritage, History, Archaeology and related fields through a thematic cluster of European Research Infrastructures, integrating initiatives, e-infrastructures and other world-class infrastructures, and building bridges between different, although tightly interrelated, fields. Within this project started in May 2015, Alpage has the leadership over the work package dedicated to the promotion and development of standards in the humanities.

In 2015, Laurent Romary and Charles Riondet have identified digital humanities use cases where standards play a central role and specified an architecture for organising standards related information (specification, software, bibliography, reference material, experts) at the service of scholars in the humanities.

9.2.2. H2020 EHRI

Participants: Laurent Romary, Charles Riondet.

The EHRI 2 (European Holocaust Research Infrastructure), also in the INFRADEV program of H2020, seeks to transform archival research on the Holocaust, by providing methods and tools to integrate and provide access to a wide variety of archival content. The project has started in June 2015 and will led us to work on both standards for the representation of archival content and develop data mining components for archival textual data.

In 2015, we have focused on the identification of available data sources resulting from the first phase of the project in the previous years and compile specifications for the description of authorities according to the EAC (Encoded Archival Context) standard.

9.2.3. H2020 Iperion

Participant: Laurent Romary.

The H2020 Iperion project aims at coordinating infrastructural activities in the cultural heritage domain. Our team has a small participation in relation to the definition of data management and representation issues. This will directly contribute to increase our experience in curating the kind of heterogeneous linguistic data that we gathered over the years.

In 2015, we have designed a questionnaire for all data producers in the project in order to gather feedback on their existing practices (data flows, licences, formats) concerning the creation, management and dissemination of cultural heritage data. On this basis, we have produced a first version of the data management plan for the project.

9.2.4. 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 aim of this project is to improve linguistic representativeness, precision and computational efficiency of Natural Language Processing (NLP) applications, focusing on the major bottleneck of these applications: Multi-Word Expressions (MWEs), i.e., sequences of words with unpredictable properties such as "to count somebody in" or "to take a haircut". A breakthrough in their modelling and processing is targeted, as the result of a coordinated effort of multidisciplinary experts working on fourteen different languages.

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: This action will facilitate European multilingualism by (1) identifying and creating a portal into discourse-level resources within Europe - including annotation tools, search tools, and discourse-annotated corpora; (2) delineating the dimensions and properties of discourse annotation across corpora; (3) organising these properties into a sharable taxonomy; (4) encouraging the use of this taxonomy in subsequent discourse annotation and in cross-lingual search and studies of devices that relate and structure discourse; and (5) promoting use of the portal, its resources and sharable taxonomy.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific events organisation

10.1.1.1. General chair, scientific chair

- Laurence Danlos is co-chair of the TALN conference which will be held in Paris in July 2016.

10.1.1.2. Member of organizing committees

- Marie Candito has co-organised (with Jihno Choi and Yannick Versley) the 6th Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2015) in Bilbao, Spain
- Benoît Sagot is the co-organiser (with Olivier Bonami) of the workshop on Computational methods for descriptive and theoretical morphology (CMDTM) to be held during the 17th International Morphology Meeting (IMM 2016) in Vienna, Austria

10.1.2. Scientific events selection

10.1.2.1. Member of conference program committees or reviewer

- Marie Candito has served as reviewer or member of the program committee for the following events: ACL 2015, EMNLP 2015, LAW 2015, DEPLING 2015, RECITAL 2015
- Mathieu Constant has served as member of the program committee for the following events: PARSEME 6, LREC 2016
- Benoit Crabbé has served as a reviewer or member of the program committee for the following events: TALN 2015, RECITAL, DepLing, EMNLP.
- Laurence Danlos has served as a reviewer or member of the program committee for the following events: LREC, TALN

- Corentin Ribeyre has served as a reviewer or member of the program committee for the following event: EMNLP 2015
- Laurent Romary has served as member of the program committee for the following events: LT4VarDial, CMLC-3 (co-located with Corpus Linguistics 2015), SLSP 2015, LREC 2016, DH Conference
- Benoît Sagot has served as a reviewer or member of the program committee for the following events: TALN 2015, LAW 2015, LREC 2016, CMDTM workshop of IMM 2016
- Djamé Seddah has served as a reviewer or member of the program committee for the following events: ACL 2015, CONLL 2015, TLT 2015, LAW 2015, SPMRL 2015
- Éric Villemonte de La Clergerie has served as a reviewer or member of the program committee for the following events: EPIA'15, IJCAI'15, Depling'15, LREC 2016

10.1.3. Journals

10.1.3.1. Member of the editorial boards

- Laurence Danlos has served as a member of the editorial board of the journal *TAL*.
- Laurent Romary is co-editor of the overlay Journal of Data Mining and Digital Humanities and for the DH Commons journa

10.1.3.2. Reviewer - Reviewing activities

- Marie Candito has served as a reviewer for the following journals: Language Resources and Evaluation, Traitement Automatique des Langues
- Benoit Crabbé has served as a reviewer for the following journal: Journal of Language Modelling.
- Laurence Danlos has served as a reviewer for the following journal: Traitement Automatique des Langues
- Laurent Romary has been reviewing for the Language Resources and Evaluation Journal Special Issue on Replicability and Reproducibility of Language Technology Experiments, the Journal on Computing and Cultural Heritage
- Benoît Sagot has served as a reviewer for the following journals: Journal of Language Modeling, Language Resources and Evaluation, Traitement Automatique des Langues
- Djamé Seddah has served as a reviewer for the following journals: Journal of Linguist Modeling, TALLIP (Transactions on Asian and Low-Resource Language Information Processing)
- Éric Villemonte de La Clergerie has served as reviewer for the following journals: Journal of Language Modelling, Language Resource and Evaluation

10.1.4. Invited talks

- Chloé Braud has given an invited talk on 09/03/2015 at Copenhagen University, Lowlands team (Danemark)
- Benoit Crabbé has given two invited talk on 20/04/2015 and 21/04/2015 at Université de Genève, CLCL group and department of Linguistics (Switzerland)
- Corentin Ribeyre has given an invited talk on 31/08/2015 at Université de Genève, CLCL group (Switzerland)
- Laurent Romary gave an introductory talk for the launch of the Maltese chapter of DARIAH in Valetta, Malta on 03/09/2015
- Laurent Romary gave an introductory talk for the launch of the Serbian chapter of DARIAH in Belgrade, Serbia on 10/12/2015, in the presence of the Ministry of Culture and Ministry of Higher Education and Research of Serbia
- Laurent Romary has given a keynote at the DiXiT Convention week on 17/09/2015 in The Hague, Netherlands (hal-01254365)

- Laurent Romary has given a keynote at the Spanish DH conference in Madrid, Spain on 06/10/2015
- Laurent Romary has given an invited talk on 24/09/2015 at the 1st Workshop on Digital Humanities in Prag, Czech Republic
- Benoit Crabbé has given an invited talk on 5/11/2015 at the Institut fuer Sprache und Information at Heinrich-Heine Universitaet Duesseldorf
- Laurent Romary has given an invited talk on 16/11/2015 at the Intangible Cultural Heritage and Innovation Symposium in Berlin (ITN DCH network)
- Laurent Romary gave the welcoming talk for the launch of the Digital Humanities Laboratory at the University of Warsaw on 10/12/2015
- Laurent Romary was invited to the final panel of the 10 year anniversary of the HAL-SHS portal on 18/12/2015
- Éric Villemonte de La Clergerie has given an invited talk on 09/07/2015 at Montpelier (LIRMM).
- Éric Villemonte de La Clergerie has given an invited talk and participated to a panel during the colloquium "Technologies pour les Langues Régionales de France" organized by DGLFLF (Meudon, February 19-20, 2015).
- Éric Villemonte de La Clergerie has given a talk and participated to a panel on "Humains, Algorithmes et data-sciences" during the GFII forum (Paris, December 7-9th 2015)

10.1.5. Leadership within the scientific community

10.1.5.1. Involvement in international initiatives

- Djamé Seddah is playing a key role in the SPMRL initiative that was initiated during the IWPT'09 conference organised in Paris by Alpage, which involves several leading international teams. As a result, Alpage members have served as programme co-chair or member of all editions of the successful SPMRL Workshop and Shared Task series hosted successively by NAACL-HLT (2010), IWPT (2011), ACL (2012) [103], EMNLP (2013) [104], CoLing (2014) and IWPT (2015). Djamé Seddah also served as a co-editor of a special issue of *Computational Linguistics* on this topic [108].
- Alpage is involved in the ISO subcommittee TC 37/SC 4 on "Language Resource Management". Éric Villemonte de La Clergerie has participated in various ISO meetings as an expert, in particular on morpho-syntactic annotations (MAF), feature structures (FSR & new FSD), and syntactic annotations (SynAF) [48]. Within the same subcommittee, Laurent Romary is the convenor of the working group on lexical resources (WG4).
- Éric de La Clergerie is the Secretary of SIGParse, the Association for Computational Linguistics (ACL) Special Interest Group in natural language parsing.
- Mathieu Constant is a member of the International Advisory Board of SIGFSM, the Association for Computational Linguistics (ACL) Special Interest Group in Finite-State Methods.
- Laurent Romary heads the Board of Directors of the European Research Infrastructure Consortium DARIAH established by the European Commission to coordinate Digital Humanities infrastrcuture activities in Europe.
- Laurent Romary is member of the TEI Archiving, Publishing, and Access Service (TAPAS) project advisory board
- Laurent Romary is member of the International Advisory board of the Belgrade Center for Digital Humanities
- Mathieu Constant serves as a member of the Management Committee and as a substitute member of the Steering Committee of the European COST action PARSEME

10.1.5.2. Involvement in national initiatives

- Alpage has many responsabilities wihtin the LabEx EFL. Until February 2015, Benoît Sagot served as head of the research strand on language resources, and was therefore a member of the Governing Board and of the Scientific Board; since then, he is deputy head of this research strand, and therefore deputy member of both boards; Marie Candito is deputy head of the research strand on computational semantics; Benoit Crabbé is deputy head of the research strand on experimental grammar; Laurence Danlos is a member of the Scientific Board of the LabEx EFL, representing Alpage;
- Benoît Sagot is in charge of the scientific and technical aspects of the development of the future Corpus de Référence du Français (French Reference Corpus), a project initiated and funded by the Institut de Linguistique Française;
- Benoît Sagot is a member of the scientific board of the consortium Corpus Écrits, which belongs to the TGIR Huma-Num;
- Benoît Sagot is a member of the executive board of the Institut de Linguistique Française, a CNRS federation of research teams involved in French linguistics;
- Laurent Romary is the leader of the scientific committe of the EquipEx Ortolang, of which Benoît Sagot is also a member.
- Laurent Romary is chairman of the scientific council of ABES (Agence Bibliographique de l'Enseignement Supérieur)
- Laurent Romary is also member of several scientific committee: Labex 'Les passes dans le présent' (PasP), ABES (Agence Bibliographique de l'Enseignement Supérieur, chair), OpenAIRE 2020 Advisory Board, OpenEdition (UMS Cleo)
- Laurent Romary is the Inria scientific advisor for Scientific and Technical Information, in charge in particular of the Open Access strategy.

10.1.5.3. Other activities for the scientific community

- Éric de La Clergerie and Benoît Sagot have served as project reviewers for ANR
- Benoît Sagot and Djamé Seddah are elected board members of the French NLP society (ATALA)
- Laurence Danlos and Benoît Sagot are members of the Permanent Committee of the TALN conference organised by ATALA

10.1.6. Scientific expertise

- Laurent Romary has advised the European Patent Office for the specification of their model for representing the patent information, based on recommandations by the TEI (Text Encoding Initiative). This model is now used for representing a database containing 200 million documents associated with over 2 billion of individual annotations.
- Éric Villemonte de La Clergerie has participated to several AFNOR meetings in relation with ISO TC37SC4 "Language Resource Management"

10.1.7. Research administration

10.1.7.1. University duties

- Lucie Barque is deputy director of the Linguistic department at Université Paris Nord
- Laurence Danlos is a member of the Scientific Committee of the Linguistics UFR of University Paris Diderot, which she chairs since 2014.
- Laurence Danlos is the deputy chair of the Doctoral School for Linguistic Sciences (École Doctorale de Sciences du Langage).
- Benoit Crabbé and Laurence Danlos are members of the Administrative board of the UFR of Linguistics of University Paris Diderot.
- Marie Candito is deputy director of the UFR of Linguistics.

- Within the Computational Linguistic curriculum at Université Paris–Diderot (L3 to M2), Laurence Danlos is in charge of the M2 and coordinates the whole curriculum, and Benoit Crabbé is in charge of the L3 year.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Licence: Lucie Barque, Sémantique lexicale, 22,5 heures en équivalent TD, niveau L2, Université Paris 13, France

Licence: Lucie Barque, Phonétique, 22,5 heures en équivalent TD, niveau L2, Université Paris 13, France

Licence: Lucie Barque, Dictionnaires électroniques, 22,5 heures en équivalent TD, niveau L2, Université Paris 13, France

Licence: Lucie Barque, Syntaxe et sémantique, 22,5 heures en équivalent TD, niveau L3, Université Paris 13, France

Licence: Lucie Barque, Introduction aux sciences du langage, 22,5 heures en équivalent TD, niveau L3, Université Paris 13, France

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

Licence: Marie Candito, Probabilités et statistiques pour le TAL, 28 heures en équivalent TD, niveau L3, Université Paris Diderot, France

Licence: Maximin Coavoux, Programmation 2, 28 heures en équivalent TD, niveau L3, Université Paris Diderot, France

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

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

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

Master: Lucie Barque, Ressources lexicales pour le TAL, 24 heures en équivalent TD, niveau M2, Université Paris 13, France

Master: Timothée Bernard, Phonétique (TD), 12 heures en équivalent TD, niveau M1, Université Paris Diderot, France

Master: Timothée Bernard, Langages formels (TD), 24 heures en équivalent TD, niveau M1, Université Paris Diderot, France

Master: Marie Candito, Analyse sémantique automatique du langage naturel, 14 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: Marie Candito, Apprentissage automatique pour le TAL, 60 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

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

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

Master: Marianne Djemaa, Sémantique Computationnelle (TD), 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: Benoît Sagot, Analyse syntaxique du langage naturel, 28 heures en équivalent TD, niveau M1, Université Paris 7, France

10.2.2. Supervision

PhD in progress: Timothée Bernard, "Analyse discursive et factualité", started in September 2015, supervised by Laurence Danlos (superviser) and Philippe de Groote (co-superviser)

PhD: Marion Baranes, "Normalisation de textes bruités", started in January 2012, supervised by Laurence Danlos (superviser) and Benoît Sagot (co-superviser), defended on 23 October 2015

PhD: Valérie Hanoka, "Extraction et Structuration de Terminologie Multilingue", started in January 2011, supervised by Laurence Danlos (superviser) and Benoît Sagot (co-superviser), defended on 6 July 2015

PhD: Emmanuel Lassalle, "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), defended on 11 May 2015

PhD: Chloé Braud, "Analyse discursive : les relations implicites", Université Paris-Diderot, started in October 2011, supervised by Laurence Danlos (superviser) and Pascal Denis (co-superviser), defended on 18 December 2015

PhD: 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), defended on 6 February 2015

PhD: 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), defended on 18/09/2015

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: 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: Maximin Coavoux, "Représentations continues pour l'analyse syntaxique et sémantique automatique", started in September 2015, supervised by Benoit Crabbé.

10.2.3. Juries

- Mathieu Constant served as a reviewer (*rapporteur*) in the PhD defense committee of Seyed Abolghasem Mirroshandel. Title: Towards Less-Supervision in Dependency Parsing. University: Aix-Marseille Université. PhD supervisor: Alexis Nasr. Defense date: 10 December 2015.
- Laurence Danlos and Benoît Sagot served as a member in the PhD defense committee of Valerie Hanoka, respectively as supervisor and co-supervisor. Title: Extraction et complétion de terminologie multilingue. University: Université Paris-Diderot. Defense date: 6 July 2015.
- Laurence Danlos and Benoît Sagot served as a member in the PhD defense committee of Marion Baranes, respectively as supervisor and co-supervisor. Title: Normalisation automatique de textes bruités. University: Université Paris-Diderot. Defense date: 23 October 2015.

- Éric Villemonte de La Clergerie served as a member in the PhD defense committee of Jérôme Kirman. Title: Mise au point d'un formalisme syntaxique de haut niveau pour le traitement automatique des langues. PhD supervisors: Sylvain Salvati and Lionel Clément. Defense date: November 4th 2015 (Bordeaux)
- Éric Villemonte de La Clergerie served as a member in the mid-term PhD defense committee of Suzanne Mpouli. Title: Automatic Detection and Analysis of Figurative Similes in Literary Texts. PhD Supervisor: Jean-Gabriel Ganascia. Defense date: July 9th 2015 (Paris)
- Éric Villemonte de La Clergerie served as a reviewer for the PhD dissertation of Daniel Fernández-González. Title: Improvements to the performance and applicability of dependency parsing. PhD supervisors: Manuel Vilares Ferro and Carlos Gómez-Rodríguez. University of Coruña (Spain)

11. 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

Doctoral Dissertations and Habilitation Theses

[11] M. BARANES. *Spelling Normalisation of Noisy Text*, Université Paris-Diderot - Paris VII, October 2015, https://hal.inria.fr/tel-01226159

[12] C. BRAUD. *Automatically Identifying Implicit Discourse Relations using Annotated Corpora and Raw Data*, Universite Paris Diderot-Paris VII, December 2015, https://hal.inria.fr/tel-01256884

[13] V. HANOKA. *Extraction and Extension of Multilingual Terminologies*, Université Paris Diderot (Paris 7), July 2015, https://hal.archives-ouvertes.fr/tel-01257201

[14] Q. PRADET. *Domain-specific French Semantic Role Labeling*, Université Paris Diderot (Paris 7), February 2015, https://hal.inria.fr/tel-01182711

Articles in International Peer-Reviewed Journals

[15] L. BARQUE. *Les noms relationnels de type humain*, in "Langue Française", 2015, vol. 1, n^{o} 185, pp. 29-41, https://halshs.archives-ouvertes.fr/halshs-01175853

[16] D. FIŠER, B. SAGOT. *Constructing a poor man's wordnet in a resource-rich world*, in "Language Resources and Evaluation", 2015, 35 p. [*DOI :* 10.1007/s10579-015-9295-6], https://hal.inria.fr/hal-01174492

[17] L. PREVOT, P. MAGISTRY, C.-R. HUANG. *Un état des lieux du traitement automatique du Chinois*, in "Faits de langues", 2016, forthcoming, https://hal.archives-ouvertes.fr/hal-01231880

[18] L. ROMARY. *Standards for language resources in ISO – Looking back at 13 fruitful years*, in "edition - die Terminologiefachzeitschrift", December 2015, n^{o} 2, https://hal.inria.fr/hal-01220925

[19] L. ROMARY. *TEI and LMF crosswalks*, in "JLCL - Journal for Language Technology and Computational Linguistics", 2015, vol. 30, n^{o} 1, https://hal.inria.fr/hal-00762664

International Conferences with Proceedings

[20] O. BONAMI, S. BENIAMINE. *Implicative structure and joint predictiveness*, in "NetWordS Final Conference", Pise, Italy, V. PIRELLI, C. MARZI, M. FERRO (editors), 2015, https://hal.inria.fr/hal-01178211

[21] C. BRAUD, P. DENIS. *Comparing Word Representations for Implicit Discourse Relation Classification*, in "Empirical Methods in Natural Language Processing (EMNLP 2015)", Lisbonne, Portugal, September 2015, https://hal.inria.fr/hal-01185927

[22] *Best Paper*
M. COAVOUX, B. CRABBÉ. *Comparaison d'architectures neuronales pour l'analyse syntaxique en constituants*, in "TALN 2015", Caen, France, 2015, https://hal.inria.fr/hal-01174613.

[23] B. CRABBÉ. *Multilingual discriminative lexicalized parsing*, in "Empirical Methods in Natural Language Processing", Lisbon, Portugal, 2015, https://hal.inria.fr/hal-01186018

[24] L. DANLOS, A. MASKHARASHVILI, S. POGODALLA. *Grammaires phrastiques et discursives fondées sur les TAG : une approche de D-STAG avec les ACG*, in "TALN 2015 - 22e conférence sur le Traitement Automatique des Langues Naturelles", Caen, France, Actes de TALN 2015, Association pour le Traitement Automatique des Langues, June 2015, pp. 158-169, https://hal.inria.fr/hal-01145994

[25] K. GULORDAVA, P. MERLO, B. CRABBÉ. *Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases*, in "ACL 2015 - the 53rd annual meeting of the Association for Computational Linguistics", Beijing, China, July 2015, https://hal.inria.fr/hal-01174617

[26] R. KALJAHI, J. FOSTER, J. ROTURIER, C. RIBEYRE, T. LYNN, J. LE ROUX. *Foreebank: Syntactic Analysis of Customer Support Forums*, in "Conference on Empirical Methods in Natural Language Processing (EMNLP)", Lisboa, Portugal, September 2015, https://hal.inria.fr/hal-01188170

[27] E. LASSALLE, P. DENIS. *Joint Anaphoricity Detection and Coreference Resolution with Constrained Latent Structures*, in "AAAI Conference on Artificial Intelligence (AAAI 2015)", Austin, Texas, United States, January 2015, https://hal.inria.fr/hal-01205189

[28] C. RIBEYRE, É. VILLEMONTE DE LA CLERGERIE, D. SEDDAH. *Because Syntax does Matter: Improving Predicate-Argument Structures Parsing Using Syntactic Features*, in "Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", Denver, USA, United States, June 2015, https://hal.archives-ouvertes.fr/hal-01174533

[29] A. SIMONENKO, B. CRABBÉ, S. PRÉVOST. *Morphological triggers of syntactic changes: Treebank-based Information Theoretic approach*, in "Treebanks and Linguistic Theories (TLT)", Warsaw, Poland, 2015, https://hal.inria.fr/hal-01240245

Conferences without Proceedings

[30] C. AMMAR, K. HADDAR, L. ROMARY. *Automatic construction of a TMF Terminological Database using a transducer cascade*, in "RANLP-2015", Hissar, Bulgaria, Proceedings of the International Conference "Recent Advances in Natural Language Processing", September 2015, https://hal.inria.fr/hal-01276816

[31] S. BENIAMINE, O. BONAMI, B. SAGOT. *Information-theoretic inflectional classification*, in "1st International Quantitative Morphology Meeting", Belgrade, Serbia, July 2015, https://hal.inria.fr/hal-01178209

[32] S. BENIAMINE, B. SAGOT. *Segmentation strategies for inflection class inference*, in "Décembrettes 9, Colloque international de morphologie", Toulouse, France, Université de Toulouse, December 2015, https://hal.inria.fr/hal-01190524

[33] L. DANLOS, T. NAKAMURA, Q. PRADET. *Traduction de VerbNet vers le français*, in "Congrès ACFAS", Rimouski, Canada, ACFAS, May 2015, 1 p. , https://hal.inria.fr/hal-01179175

[34] R. GARNIER, B. SAGOT. *Could Greek and Italic share a same Indo-European substratum?*, in "22nd International Conference on Historical Linguistics", Naples, Italy, July 2015, https://hal.inria.fr/hal-01256310

[35] R. HUYGHE, L. BARQUE, P. HAAS, D. TRIBOUT. *Underived event nouns in French*, in "JENom6 - 6th Workshop on Nominalizations", Verona, Italy, June 2015, https://halshs.archives-ouvertes.fr/halshs-01175856

[36] J. STEINLIN, M. COLINET, L. DANLOS. *FDTB1 : Identification of discourse connectives in a French corpus*, in "Traitement automatique du langage naturel", Caen, France, June 2015, https://hal.inria.fr/hal-01178382

Scientific Books (or Scientific Book chapters)

[37] T. BLANKE, C. KRISTEL, L. ROMARY. *Crowds for Clouds: Recent Trends in Humanities Research Infrastructures*, in "Cultural Heritage Digital Tools and Infrastructures", A. BENARDOU, E. CHAMPION, C. DALLAS, L. HUGHES (editors), 2016, https://hal.inria.fr/hal-01248562

Scientific Popularization

[38] L. ROMARY. *DARIAH in Motion – Tangible infrastructure for intangible information*, November 2015, Intangible Cultural Heritage and Innovation, https://hal.inria.fr/hal-01260042

Other Publications

[39] T. BERNARD. *Verbes d'attitude propositionnelle et analyse discursive*, Université Paris Diderot - Paris 7, June 2015, https://hal.inria.fr/hal-01256344

[40] M. CANDITO, G. PERRIER. *Guide d'annotation en dépendances profondes pour le français*, 2015, 129 p. , Ce guide décrit le schéma d'annotation en dépendances profondes utilisé pour l'annotation en syntaxe profonde du corpus Sequoia initialement annoté en syntaxe de surface, https://hal.inria.fr/hal-01249907

[41] L. CAPELLI, L. FARHI, L. ROMARY. *A TEI conformant pivot format for the HAL back-office*, October 2015, Text Encoding Initiative Conference and members meeting 2015, Poster, https://hal.archives-ouvertes.fr/hal-01221774

[42] L. ROMARY. *TEI challenges in an accelerating digital world*, September 2015, DiXiT Convention week, https://hal.inria.fr/hal-01254365

References in notes

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

[44] 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

[45] 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

[46] 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

[47] 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

[48] S. BOSCH, S. K. CHOI, É. VILLEMONTE DE LA CLERGERIE, A. CHENGYU FANG, G. FAASS, K. LEE, A. PAREJA-LORA, L. ROMARY, A. WITT, A. ZELDES, F. ZIPSER. *[tiger2] As a standardized serialisation for ISO 24615 - SynAF*, in "TLT11 - 11th international workshop on Treebanks and Linguistic Theories - 2012", Lisbon, Portugal, I. HENDRICKX, S. KÜBLER, K. SIMOV (editors), Ediçoes Colibri, November 2012, pp. 37-60, https://hal.inria.fr/hal-00765413

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

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

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

[52] 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

[53] 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

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

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

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

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

[58] M. 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

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

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

[61] 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

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

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

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

[65] 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), Proceedings of the 8th International Natural Language Generation Conference (INLG), Association for Computational Linguistics, June 2014, pp. 35-44, https://hal.inria.fr/hal-00999595

[66] 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

[67] 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

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

[69] 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

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

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

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

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

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

[75] 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

[76] A. F. T. MARTINS, M. S. C. ALMEIDA. *Priberam: A Turbo Semantic Parser with Second Order Features*, in "Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)", Dublin, Ireland, Association for Computational Linguistics and Dublin City University, August 2014, pp. 471–476

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

[78] 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

[79] 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

[80] 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

[81] 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

[82] 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

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

[84] 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

[85] 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

[86] C. RIBEYRE. *Vers un système générique de réécriture de graphes pour l'enrichissement de structures syntaxiques.*, in "RECITAL 2013 - 15ème Rencontre des Etudiants Chercheurs en Informatique pour le

Traitement Automatique des Langues", Les Sables d'Olonne, France, Université de Nantes, June 2013, pp. 178-191

[87] 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

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

[89] 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

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

[91] 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

[92] 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

[93] 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

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

[95] 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

[96] 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

[97] 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

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

[99] 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

[100] 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

[101] 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

[102] 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

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

[104] 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

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

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

[107] 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

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

[109] 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

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

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

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

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

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