IN PARTNERSHIP WITH:
Université Denis Diderot (Paris 7)

Activity Report 2016

Project-Team ALPAGE

Large-scale deep linguistic processing

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

RESEARCH CENTER
Paris

THEME
Language, Speech and Audio

Table of contents

Project-Team ALPAGE

Creation of the Project-Team: 2008 January 01, end of the Project-Team: 2016 December 31

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
9.5.8. - Linguistics
9.5.10. - Digital humanities
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]
Marie-Hélène Candito [Univ. Paris VII, Associate Professor]
Mathieu Constant [Univ. Paris Est, Associate Professor, until Aug 2016]
Benoît Crabbé [Univ. Paris VII, Associate Professor]
Laurence Danlos [Univ. Paris VII, Professor, HDR]
Djamé Seddah [Univ. Paris IV, Associate Professor]

Engineers

Noemie Faivre [Inria, until May 2016]
Luca Foppiano [Inria]
Pierre Magistry [Inria, until Jan 2016, granted by Caisse des Dépôts et Consignations]
Héctor Martínez Alonso [Inria, from Feb 2016]
Marie Puren [Inria]
Charles Riondet [Inria]
Stéphane Riou [Institut de Linguistique Française (CNRS), until Oct 2016]
Dorian Seillier [Inria, from May 2016]

PhD Students

Timothée Bernard [ENS Lyon]

Maximin Coavoux [Univ. Paris VII]
Daniel Dakota [Inria, Visiting PhD Student, from Oct 2016]
Marianne Djemaa [Univ. Paris VII, until Sep 2016, granted by ANR ASFALDA- project]
Mohamed Khemakhem [Inria, from Jun 2016]
Corentin Ribeyre [Univ. Paris VII, until Jan 2016]
Raphaël Salmon [Yseop and Univ. Paris VII, until Jan 2016, granted by CIFRE]

Administrative Assistant
Christelle Guiziou [Inria]

Others
Graziella Pastore [Inria, from Apr 2016 until Jul 2016]
Laura Ramirez Sanchez [Inria, Intern, from Jul 2016 until Sep 2016]
Adrien Roux [Inria, Intern, from Mar 2016 until Jul 2016]
Vincent Segonne [Inria, Intern, from Feb 2016 until Jul 2016]
De Zhao [Inria, Intern, from Jun 2016 until Jul 2016]

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 than MCS formalisms, Range Concatenation Grammars (RCGs, introduced and used by Alpage members, such as Pierre Boullier and Benoît Sagot [56], [79], [84]) 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.

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

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 [92], [91]. Meta-Grammars (MGs) allows the linguist to focus on a modular description of the linguistic aspects of a grammar, rather than focusing on the specific aspects of a given grammatical formalism. Translation from MGs to grammatical formalisms such as TAG or LFG may be automatically handled. Graphical environments can be used to design MGs and their modularity provides a promising way for sharing the description of common linguistic phenomena across human languages.

3.2. Statistical Parsing

Participants: Djamé Seddah, Marie-Hélène Candito, Benoît 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 [72] or automatic [75], [76] 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 [58].

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 ...) [94], [74]. 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 [73]. 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 [53], [52] and derive the best input for syntagmatic statistical parsing [60]. Benchmarking several PCFG-based learning frameworks [86] 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) [76].

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 [58] 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 [89].

Pursuing our work on PCFG-LA based parsing, we investigated the automatic conversion of the treebank into dependency syntax representations [57], that are easier to use for various NLP applications such as question-answering or information extraction, and that are a better ground for further semantic analysis. This conversion can be applied on the treebank, before training a dependency-based parser, or on PCFG-LA parsed trees. This gives the possibility to evaluate and compare on the same gold data, both syntagmatic- and dependency-based statistical parsing. This also paved the way for studies on the influence of various types of lexical information.

3.3. Robust linguistic processing

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

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 [80]. 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 [90]. 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 [68]. 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 [77], as Alpage was ranked second and third [88]. 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) [87].

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

3.4. Dynamic wide coverage lexical resources

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

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 [83]. At the syntactic level, work has been achieved on automatic acquisition of atomic syntactic information and automatic detection of errors in the lexicon [95],[6]. At the semantic level, automatic wordnet development tools have been described [78], [93], [71], [69]. All such techniques need of course to be followed by manual validation, so as to ensure high-quality results.

For French, these techniques, and others, have lead some Alpage members to develop one of the main syntactic resources for French, the Le*fff* [81], [85], 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 [82], [70], 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, 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 [62].

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, [63],[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. Highlights of the Year

4.1. Highlights of the Year

In 2016, Alpage has obtained several new national fundings: the team is the leader of a new ANR project (Parsiti), and a partner of a new ANR project (Profiterole) and of a new ANR-NSF project (MCM-NL).

5. New Software and Platforms

5.1. Alexina

Atelier pour les LEXiques INformatiques et leur Acquisition
FUNCTIONAL DESCRIPTION

Alexina is Alpage's 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/

5.2. 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 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, Djame Seddah and Benoît Crabbe
- Contact: Marie-Hélène Candito
- URL: http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

5.3. 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: Djame Seddah

5.4. 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: Eric Villemonte De La Clergerie
- Contact: Eric Villemonte De La Clergerie
- URL: http://dyalog.gforge.inria.fr/

5.5. FDTB1

- Contact: Laurence Danlos

5.6. 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: Djame Seddah

5.7. FRMG

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

5.8. Extreme UGC corpus

FUNCTIONAL DESCRIPTION

The Extreme UGC corpus is French three-domain data set focusing on user-generated content, made up of noisy question headlines from a cooking forum, live game chat logs and associated forums from two popular online games (MINECRAFT and LEAGUE OF LEGENDS). Building such an out of domain corpus, allowed us to consider the limits of our current normalization approaches. Currently annotated with part-of-speech, we plan to add other annotations layers.

- Contact: Djame Seddah

5.9. LexConn

- Contact: Laurence Danlos

5.10. LexViz

FUNCTIONAL DESCRIPTION

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

- Participants: Eric Villemonte De La Clergerie and Mickaël Morardo
- Contact: Eric Villemonte De La Clergerie

5.11. 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

5.12. 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: Eric Villemonte De La Clergerie and Paul Bui-Quang
- Contact: Eric Villemonte De La Clergerie
- URL: http://alpage.inria.fr/frmgwiki/

5.13. 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 designed for the analysis and manipulation of attributed oriented and multi-relational graphs. Itis currently being used to convert existing universal dependencies for French to the upcoming 2.0 scheme to be used for the next “big” CoNLL parsing Shared Task of 2017.

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

5.14. 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/

5.15. 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: Djame Seddah

5.16. 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, Eric Villemonte De La Clergerie and Djame Seddah
- Contact: Benoît Sagot
- URL: http://lingwb.gforge.inria.fr/

5.17. Verb∋net

- Contact: Laurence Danlos

5.18. dyalog-sr

KEYWORD: Parsing

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

5.19. hyparse

Alpage Hybrid Parser
KEYWORDS: Parsing - NLP
FUNCTIONAL DESCRIPTION

Multilingual Phrase Structure Parser

- Contact: Benoît Crabbe
- URL: http://hyparse.gforge.inria.fr

5.20. vera

- Participants: Benoît Sagot and Dimitri Tcherniak
- Partner: Verbatim Analysis
- Contact: Benoît Sagot

6. New Results

6.1. Deep syntactic parsing

Participants: Corentin Ribeyre, Marie-Hélène Candito.

Syntax plays an important role in the task of predicting the semantic structure of a sentence. But syntactic phenomena such as alternations, control and raising tend to obfuscate the relation between syntax and semantics. We have investigated how to predict the semantic structure of a sentence, encoded using the FrameNet model, taking advantage of deeper syntactic information than what is usually used. This deep syntactic representation abstracts away from purely syntactic phenomena and proposes a structural organization of the sentence that is closer to the semantic representation, by normalising the syntactic paths between a verb and its arguments. This reduces the variety of the syntactic realization of semantic roles, as shown by a decrease of the entropy of the syntactic paths of a given role.

Experiments conducted on a French corpus annotated with semantic frames showed that a FrameNet semantic parser reaches better performances with such a deep syntactic information [31]. For instance, switching from surface to deep syntactic information leads to a significant gain in FrameNet role identification, especially when this information is predicted (rather than reference information): $+5.1$ points (56.7 to 61.7) on all triggers [2] and $+6.7$ points (61.3 to 68.0) on verbal triggers only. These results clearly show the benefit of using deep syntactic features.

6.2. Multilingual POS-tagging

Participant: Benoît Sagot.

[2] In the sense of FrameNet, i.e. predicative lexial units, which should be assigned a frame.

Morphosyntactic lexicons and word vector representations have both proven useful for improving the accuracy of statistical part-of-speech taggers. We compare the performances of four systems on datasets covering 16 languages, two of these systems being feature-based (MEMMs —in the case of our own system MElt— and CRFs) and two of them being neural-based (bi-LSTMs). We show that, on average, all four approaches perform similarly and reach state-of-the-art results. Yet we obtained better performances with feature-based models on lexically richer datasets (e.g. for morphologically rich languages), whereas neural-based results are higher on datasets with less lexical variability (e.g. for English). These conclusions hold in particular for the MEMM models relying on our system MElt, which benefited from newly designed features [32], [44]. Thus we have shown that, under certain conditions, feature-based approaches enriched with morphosyntactic lexicons are competitive with respect to neural methods.

6.3. Transition-based constituency parsing with HyParse

Participants: Benoît Crabbé, Maximin Coavoux.

Transition-based parsing reduces the parsing task to predict a sequence of atomic decisions. These decisions are taken while sequentially reading words from a buffer and combining them incrementally into syntactic structures. The resulting structures are often dependency structures but can also be constituents, as is the case for our parser HyParse. Such an approach is therefore linear in the length of the input sentence, making transition-based parsing computationally efficient relative to other approaches. The challenge in transition-based parsing is modelling which action should be taken in each state it encounters as it progresses in a sentence provided as an input.

Training of a transition-based parser therefore consists in training a function that maps each of the unboundedly many states the parser might encounter to the best possible action, or transition, it should take. This function generally relies on a huge set of features, often conveniently grouped in the form of more abstract feature templates. Yet selecting the optimal subset of feature(template)s remains a challenge.

The training procedure therefore requires the help of an “oracle”, that is a function that returns the action that the parser should take in a given parser state given the gold parse. If the oracle assumes that the next action is necessarily the one given in the gold parse, it is said to be "static” and the oracle is deteminist. In order to train the parser to take relevant decisions when in an erroneous state, we can introduce some non determinism in the oracle in order to explore not only gold transition sequences but also near-gold transition sequences. This is the purpose of a dynamic oracle. Dynamic oracle training has shown substantial improvements for dependency parsing in various settings, but had not previously been explored for constituent parsing.

The two research directions we have investigated reflect the two above-mentioned challenges.

First, in collaboration with Rachel Bawden, now PhD student at LIMSI, we resumed our work on developing an efficient, language-independent model selection method for our parser HyParse [61]. It is designed for model selection when faced with a large number of possible feature templates, which is typically the case for morphologically rich languages, for which we want to exploit morphological information. The method we proposed uses multi-class boosting for iterative selection in constant time, using virtually no *a priori* constraints on the search space. We did however use a pre-ranking step before selection in order to guide the selection process. Our experiments have illustrated the feasibility of the method for our working language, French and resulted in high-performing, compact models much more efficiently than naive methods [22].

Second, we developed a dynamic oracle for HyParse. First, we replaced the traditional feature-based approach used in the above-described experiments by a neural approach. This is a way to overcome the feature selection issue addressed in the above-described work. The neural network weighting function we developed uses a non-linear hidden layer to automatically capture interactions between variables, and embeds morphological information in a vector space, as is usual for words and other symbols. Then, we developed our dynamic oracle based on this neural function and conducted experiments on the 9 languages of the SPMRL dataset in order to assess the impact of this oracle [25]. The experiments have shown that a neural greedy parser with morphological features, trained with a dynamic oracle, leads to accuracies comparable with the best currently available non-reranking and non-ensemble parsers.

6.4. French FrameNet

Participants: Marie-Hélène Candito, Marianne Djemaa.

In 2016 we have continued the development of a French FrameNet, within the ASFALDA project. While the first phase of the project focused on the development of a French set of frames and corresponding lexicon (Candito et al., 2014), we have focused this year on the subsequent corpus annotation phase, which targeted four notional domains (commercial transactions, cognitive stances, causality and verbal communication). Given full coverage is not reachable for a relatively "new" FrameNet project such as ours, focusing on specific notional domains allowed us to obtain full lexical coverage for the frames of these domains, while partially reflecting word sense ambiguities. Furthermore, as frames and roles were annotated on two main French Treebanks (the French Treebank and the Sequoia Treebank), we were able to extract a syntactico-semantic lexicon from the annotated frames. In the resource's current status [28], there are 98 frames, 662 frame-evoking words or "triggers", 872 senses, and about 13,000 annotated frames, with their semantic roles assigned to portions of text [3]

During this year's resource development efforts, we have put a specific emphasis on the causality domain (about 4000 instances of causal lexical items with their corresponding semantic frames are included in our resource). In the process of building the French lexicon and preparing the annotation of the corpus, we had to remodel some of the frames proposed in FrameNet based on English data, with hopefully more precise frame definitions to facilitate human annotation. This includes semantic clarifications of frames and frame elements, redundancy elimination, and added coverage. The result is arguably a significant improvement of the treatment of causality in FrameNet itself [34].

6.5. Verb∋net

Participants: Lucie Barque, Laurence Danlos.

VerbNet is a lexical resource for English verbs in which verbs are grouped together based on their ability to appear in similar sets of syntactic frames that correspond as well to alternations exhibited by verbs as to alternative syntactic realizations (Kipper et al. 2004). A French Verbnet, named Verb∋net, was first automatically derived from English VerbNet (Pradet et al., 2014) and is still under development. [13] details how Verb∋net was developed from the English VerbNet while using as far as possible the available lexical resources for French and how the various French alternations are coded, focusing on differences with English (e.g. existence of pronominal forms). One difficulty encountered in the development of Verb∋net springs from the fact that the list of (potentially numerous) frames has no internal organization in VerbNet. [26] proposes a type system for frames that shows whether two frames are variants of a given alternation. Frame typing facilitates coherence checking of the resource in a "virtuous circle".

6.6. French FrameNet

Participant: Benoît Crabbé.

Elaborating on our previous work on Medieval French in collaboration with Sasha Simonenko (McGill) and Sophie Prévost (LATTICE), we have conducted the first large-scale quantitative investigation of the syncretisation of verbal subject agreement in this language and test a classic analysis which relates non-syncretic agreement and null subjects as parts of the same grammar. We have shown that agreement syncretisation and the emergence of overt pronominal subjects proceeded at the same rate. Under the Constant Rate Hypothesis of Kroch (1989), which states that a grammatical change has the same rate in different contexts, these results are compatible with the traditional analysis [40], [39], [33]. However, we show that this analysis also generates a number of predictions which are not borne out by the quantitative data. We conclude that a more complex model of interaction of subject and inflection parameters is needed. Such a model may for instance be one where the type of an ending (non-syncretic vs. syncretic), presumably dependent on some unrelated phonological mechanism, presents a parsing difficulty for a null subject-licensing grammar and thus lowers its probability to be chosen by the speaker, which eventually drives it to extinction, similarly to the grammar competition model proposed in Yang (2010).

[3] The French FrameNet is freely available at http://asfalda.linguist.univ-paris-diderot.fr/frameIndex.xml.

We have also investigated the effects of the text form (prose vs. verse) on diachronic grammatical changes in Medieval French using parsed treebanks and (1 million words with PTB-like annotations). Despite the common intuition that the prose is somehow more "advanced" than the verse contemporary to it with respect to grammatical changes, the magnitude of the difference has remained unknown in the absence of quantificational evaluations. At the same time, the prevalence of verse in the earliest periods of documented French (i.e. X–XII c.) results in a strong and unavoidable correlation between time and form, which potentially undermines the results of the studies attempting to formally model Medieval French evolution. We have compared two historical changes across text forms (namely the loss of pro-drop and that of OV_{finite} order), and shown that verse and prose behave differently, at least regarding the OV_{finite} order, thus contradicting Kroch's (1989) Constant Rate Hypothesis [38].

6.7. Modelling discourse-level information

Participants: Laurence Danlos, Timothée Bernard.

We have continued our work on the formalisation of discourse-level information. First, we have proposed in [24] a new model in STAG syntax and semantics for subordinate conjunctions (SubConjs) and attributing phrases —attitude/reporting verbs (AVs; *believe*, *say*) and attributing prepositional phrase (APPs; *according to*). This discourse-oriented model is based on the observation that SubConjs and AVs are not homogeneous categories. Indeed, previous work has shown that SubConjs can be divided into two classes according to their syntactic and semantic properties. Similarly, AVs have two different uses in discourse: evidential and intentional. While evidential AVs and APPs have strong semantic similarities, they do not appear in the same contexts when SubConjs are at play. Our proposition aims at representing these distinctions and capturing these various discourse-related interactions.

We have also investigated how sentential and discourse TAG-based grammars can be interfaced, in collaboration with Aleksandre Maskharashvili and Sylvain Pogodalla (LORIA). Tree-Adjoining Grammars (TAG) have been used both for syntactic parsing, with sentential grammars, and for discourse parsing, with discourse grammars (see for example our D-STAG model or the D-LTAG model). Yet the modelling of discourse connectives (coordinate conjunctions, subordinate conjunctions, adverbs...) in TAG-based formalisms for discourse differ from their modelling in sentential grammars. Because of this mismatch, an intermediate processing step is required between the sentential and the discourse processes, both in parsing and in generation [27]. We have developed a method to smoothly interface sentential and discourse TAG grammars, without using such an intermediate processing step. This method, based on Abstract Categorial Grammars (ACG), allows for building D-STAG discourse structures that are direct acyclic graphs (DAG) and not only trees.

6.8. Detecting omissions in journalistic texts

Participants: Héctor Martínez Alonso, Benoît Sagot.

In the journalistic genre that is characteristic of online news, editors make frequent use of citations as prominent information; yet these citations are not always in full. The reasons for leaving information out are often motivated by the political leaning of the news platform.

Existing approaches to the detection of political bias rely on bag-of-words models that examine the words present in the writings. In the context of the VerDI project (see below), we have initiated work aimed at going beyond such approaches, which focus on what is said, by instead focusing on what is *ommited*. Thus, this method requires a pair of statements; an original one, and a shortened version with some deleted words or spans. The task is then to determine whether the information left out in the second statement conveys *substantial* additional information. If so, we consider that a certain statement pair presents an omission. To tackle this question, we used a supervised classification framework, for which we require a dataset of sentence pairs, each pair manually annotated for omission.

We have developed a small reference corpus for evaluation purposes, using and comparing both crowd and expert annotation. This corpus has allowed us to examine which features help automaticallly identify cases of omission. In addition to straightforward measures of word overlap (the Dice coefficient), we also determined that there is a good deal of lexical information that determines whether there is an omission. This work is, to the best of our knowledge, the first empirical study on omission identification in statement pairs. We shall make all data and annotations freely available upon publication.

6.9. Models for interoperable lexical data

Participants: Mohamed Khemakhem, Laurent Romary.

Lexical data play an essential role in computational linguistic in two complementary ways:

- They serve as basic resources with which computational linguistic process can be parameterized. Such lexical resources are usually automatically or semi-automatically produced, are highly structured and may cover various levels of linguistic description from basic morpho-syntactic content to semantic representations;
- When created manually either for the purpose of describing a language (mono- or multilingual dictionary) or as a by product other language based activities (e.g. technical writing, translation), they may serve as a primary source of observation to analyse the way the lexicon of a language is organized, is used in domain oriented content, or how languages vary across time, space and usage.

The Alpage team has a specific expertise in the domain of lexical data, having been involved in the recent years in the creation of reference resources for the French language in particular, but also as driving force in the definition of international standards for the modelling and representation of both semasiological (word to sense) and onomasiological (concept to term) lexical information:

- ISO 16642 (TMF, Terminological Markup framework) and ISO 30042 (TBX, TermBase eXchange) as reference standards for the interchange of terminological data, for instance between translators' workbenches, but also for the modelling of dialectal information in linguistics;
- ISO 24613 (LMF, Lexical Markup Framework), a modular modelling framework for the representation of both machine and human semasiological resources;
- The Text Encoding Initiative (TEI), which since its inception has provided an XML based format for human readable dictionaries, widely used in most last scale dictionary projects worldwide.

One of the difficulties in lexical modelling is to identify the proper modelling framework for a given lexical resource but also to ensure maximal interoperability across heterogeneous lexical content. In the recent period, we have been working on the following aspects:

- Participation in the on going revision of ISO 30046, and planning of a possible integration of a TBX dialect in the TEI guidelines;
- Setting up the revision of ISO 24613 as a multi-part standard. Alpage is now involved in the provision of a reference TEI based serialisation of LMF and the part dedicated to etymological/diachronical information;
- Proposing an extension to the TEI guidelines for the representation of etymological information in dictionaries thus offering a formal basis for the study of diachronical phenomena across dictionaries [46];
- Organising a workshop in the context of the COST action eNEL that brought together the most relevant experts in the field in order to provide a set of constraints to apply the TEI guidelines in a more interoperable way across dictionary projects;
- Starting working on a machine learning based process to extract lexical content and structure automatically from digitized legacy dictionaries, This activity, base don the architecture of the Grobid library, is the basis of the PhD work by Mohamed Khemakhem.

6.10. Open data in the arts and humanities

Participants: Luca Foppiano, Marie Puren, Charles Riondet, Laurent Romary, Dorian Seillier.

The issue of open data has become increasingly important in various scholarly domains for it impacts on the visibility of the corresponding works, the capacity to provide evidence for reported facts and results, but also let other scholars build up new research on existing data sets. This is particularly acute in the humanities where primary sources play an essential role in providing the core material of scholarly results and for which the digital turn has offered a unique perspective of building up a wealth of structure information about human traces at large.

Based upon the experience gained in the definition of the open access policy at Inria [42], [50], [43], we have pursued various activities leading to a better understanding of the technical, editorial and political factors that may improve the wide dissemination of scholarly data sets in the humanities:

- Carry out a large scale questionnaire on data re-use within the partnership of the Iperion projects, which showed the lack of a coherent data management policy across cultural heritage laboratories in Europe from the points of view of documentation, archiving, licencing and re-use [49];
- Design a concept [16], [41] to improve the general fluidity of research results in the humanities based on data quality assessment, data journals and above all the setting of of a data re-use charter between scholars and cultural research institutions in the humanities. This action, carried out in the context of the Parthenos project has started with the organisation of two high level workshops in Berlin and Paris with representatives of major cultural research institutions;
- Coordinate as leader of WP 4 (Standards) in the Parthenos project a major overview of the needs and possible deployment of standards in the humanities based of an in depth survey of possible research scenario and associated practices in the domain of standards (Deliverable 4.1 published in October 2016). This has been accompanied by specific technical developments such as the proposition of an extension to the TEI guidelines for the representation of embedded stand-off annotations [45], [51];
- Develop specific modules for mining digital sources in the humanities, in particular in the domain of named entity recognition as an improvement of the NERD software initially developed in the European Cendari project.

7. Bilateral Contracts and Grants with Industry

7.1. Contracts with Industry

Alpage has developed several collaborations with industrial partners. Apart from grants described in the next section and informal discussions, 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 5.20),
- Yseop (“CIFRE” PhD of Raphael Salmon started in 2012 about automatic text generation)
- Agence France-Presse (on-going discussions aimed at a renewal of a long-lasting collaborations, involving several joint projects and a CIFRE PhD)

8. Partnerships and Cooperations

8.1. National Initiatives

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

Participants: Laurence Danlos, Benoît Sagot, Marie-Hélène Candito, Benoît Crabbé, Pierre Magistry, Djamé Seddah, 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". 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 2016,Benoît Sagot, Marie Candito and Benoît Crabbé were respectively deputy-head of strand 6 on "Language Resources", strand 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.

8.1.2. ANR

8.1.2.1. ANR project Profiterole (2017 - 2020)

Participants: Benoît Crabbé, Éric Villemonte de La Clergerie, Benoît Sagot.

PROFITEROLE is a 4-year ANR research project led by Sophie Prévost (LATTICE) that involves computational linguists and specialists of Medieval French from LATTICE (Univ. Paris 3, CNRS, ENS), ALPAGE and ICAR (Univ. Lyon, ENS).

PROFITEROLE has three closely correlated main goals that fall within the fields of linguistics and Natural Language Processing (NLP): (1) formal and computational modeling phonological, morphological and syntactic aspects of the diachronic evolution of French; (2) targeting the development of a methodology to explore and annotate heterogeneous linguistic data while providing automatic analysers for various stages of the French language; (3) expanding linguistic resources for French, by building a large annotated corpus (1 million words) of Medieval French (9th-15th centuries) and morphological lexicons (plus NLP tools) covering several stages of French. Alpage members will essentially be involved on the computational and formal modeling aspects of the project and on the design of automated processing tools for lexicon and syntax.

8.1.2.2. ANR project PARSITI (2016 - 2020)

Participants: Marie-Hélène Candito, Djamé Seddah [principal investigator], Benoît Crabbé, Éric Villemonte de La Clergerie, Benoît Sagot.

Exploiting multilingual user-generated content (UGC), for applications such as information extraction, text mining or summarization, and facilitate their access to a wider audience implies a qualitative step-ahead in Natural Language Understanding. This is because UGC differs from better-studied edited data in many ways, including by non-canonical syntax, highly contextualised nature and rich lexical variability. The ParSiTi ANR project focuses on three critical aspects: (1) Robust Parsing Technologies, (2) Accurate Machine Translation Engines and (3) Context-aware Methods, all backed by State-of-the-Art Morphological Analysers and Normalization tools. To showcase the different models and algorithms designed during the project, a Machine Translation System will be developed that will be able to translate UGC between French, Arabic and English.

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

Participants: Marie-Hélène Candito, Mathieu Constant [principal investigator], Benoît 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 is participating 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.

8.1.2.4. ANR project SoSweet (2015 - 2019)

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

Led by Jean-Phillipe Magué (ENS Lyon), the SoSweet project focuses on the synchronic variation and the diachronic evolution of the variety of French language used on Twitter. Its goal is to provide a state-of-the-art socio-linguistic description of half a billion tweets collected over 5 years.

Alpage, specialized in natural language processing, takes care of the linguistics enrichment part, which provides the other partners with normalized and structurally enriched forms of text. Alpage is also responsible of providing distributional analysis of our corpus, by the means of various forms of word clustering in order to define sociolinguistic variants in the tweets.

8.1.2.5. ANR project ASFALDA (2012 – 2016)

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

Alpage is principal investigator team for the ANR project ASFALDA, lead by Marie 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 [54] and can be characterized roughly as an explicitation of "who does what when and where", that abstracts away from word order / syntactic variation, and to some of the lexical variation found in natural language.

The project relies on an existing standard for semantic annotation of predicates and roles (FrameNet), and on existing previous effort of linguistic annotation for French (the French Treebank). The original FrameNet project provides a structured set of prototypical situations, called frames, along with a semantic characterization of the participants of these situations (called *roles*). We propose to take advantage of this semantic database, which has proved largely portable across languages, to build a French FrameNet, meaning both a lexicon listing which French lexemes can express which frames, and an annotated corpus in which occurrences of frames and roles played by participants are made explicit. The addition of semantic annotations to the French Treebank, which already contains morphological and syntactic annotations, will

boost its usefulness both for linguistic studies and for machine-learning-based Natural Language Processing applications for French, such as content semantic annotation, text mining or information extraction.

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

The project is structured as follows:

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

The scientific key aspects of the project are:

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

8.1.2.6. 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 [64], [65], [66]. As regards discursive analysis, D-STAG formalism created by Laurence Danlos in the 00's has also been rewritten in ACG in 2015 [67] (see also [27]).

8.1.3. Other national initiatives

8.1.3.1. "RAPID" project VerDI (2016 – 2019)

Participants: Benoît Sagot, Héctor Martínez Alonso.

The ANR "RAPID" project VerDI focuses on the automatic identification of information dissimulation on the Internet and on social networks. Such dissimulations can be produced by omiting crucial pieces of information within documents or during written online discussions, by hiding them within a massive information flow, or using other techniques. VerDI aims at extending an existing journalistic fact-checking tool developed by Trooclick, the company that leads the project.

8.1.3.2. FUI project COMBI (2014-2016)

Participant: Laurence Danlos.

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

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

8.1.3.3. Institut de Linguistique Française and Consortium CORLI 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 CORLI consortium (following, among other, the *Corpus Écrits* consortium in which previously participating), which is dedicated, among other topics, 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 CORLI 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).

8.2. European Initiatives

8.2.1. FP7 & H2020 Projects

8.2.1.1. H2020 PARTHENOS

Participants: Laurent Romary, Luca Foppiano, Mohamed Khemakhem, Marie Puren, Charles Riondet, Dorian Seillier.

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.

8.2.1.2. H2020 EHRI

Participants: Laurent Romary, Luca Foppiano, 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.

8.2.1.3. H2020 Iperion

Participants: Laurent Romary, Luca Foppiano, Marie Puren.

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.

8.2.2. 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.

Program: **ISCH COST Action IS1305**

Project acronym: ENeL

Project title: European Network of e-Lexicography

Duration: October 2013 - October 2017

Coordinator: Prof Martin EVERAERT (NL)

Other partners: interdisciplinary experts (linguists, computational linguists, computer scientists, lexicographers, and industrials) from 31 countries

Abstract: The proposed Action aims to establish a European network of lexicographers in order to deal with the following issues: give easier access to scholarly dictionaries, establish a systematic exchange of expertise on common standards and solutions, develop a common approach to e-lexicography that forms the basis for a new type of lexicography that fully embraces the pan-European nature of much of the vocabularies of the languages spoken in Europe

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Selection

9.1.1.1. Member of the Conference Program Committees or Reviewer

- Marie Candito was a reviewer or member of the program committee for the following events: NAACL 2016, CONLL 2016, RECITAL 2016, COLING 2016
- Benoît Crabbé was a reviewer or member of the program committee for the following events: NAACL 2016, ACL 2016, COLING 2016, EMNLP 2016, CONLL 2016.
- Maximin Coavoux was a sub-reviewer or member of the program committee for the following events: ACL 2016, EMNLP 2016.
- Héctor Martínez Alonso was a reviewer or member of the program committee for the following events: ACL 2016, COLING 2016, CONLL 2016, EMNLP 2016, EACL 2017, NAACL 2016 , MWE 2016 and WNUT 2016.
- Laurent Romary was a member of the program committee for the following events: CMLC-4, Digital Humanities and Iconography, The case of alpine mural painting, Datech 2016, CILA 2016, TOTh (Terminology & Ontology: Theories and applications), LDL-2016, LG-LP 2016, ISA-12
- Benoît Sagot was a member of the program committe for the following events: NAACL 2016, EACL 2017, EMNLP 2016, ACL 2016
- Djamé Seddah was a reviewer or member of the program committee for the following events: ACL 2016, COLING 2016, EMNLP 2016, CONLL 2016, TLT 2016, LAW 2016, TALN 2016 and others workshops.
- Éric Villemonte de La Clergerie was a reviewer or member of the program committee for the following events: COLING'16 (tracks Parsing, and "Resources, Software and Tools"), LREC'16, NAACL'16, DiscoNLP'16, Toth'16

9.1.2. Journal

9.1.2.1. Member of the Editorial Boards

- Laurent Romary is co-editor of the overlay Journal of Data Mining and Digital Humanities and for the DH Commons journal
- Laurent Romary is member of the advisory board of the Journal of the Text Encoding Initiative (jTEI)

9.1.2.2. Reviewer - Reviewing Activities

- Héctor Martínez Alonso was a reviewer for the Artificial Intelligence journal
- Laurent Romary was a reviewer for the jTEI journal

- Benoît Sagot was a reviewer for the following journals: Journal of Language Modelling, Language Resources and Evaluation, Northern European Journal of Language
- Djamé Seddah was a reviewer for the following journals: Asian Languages and Information Processing, Language Ressources and Evaluation

9.1.3. Invited Talks

- Marie Candito gave a talk on 24/03/2016 at the "3ème journée TAL et IA ", Paris
- Marie Candito gave a talk on 27/10/2016 at the DIGICOSME Labex, LIMSI, Orsay
- Héctor Martínez Alonso gave an invited talk on 05/04/2016 at PARSEME/ENeL workshop on MWE e-lexicons, Skopje (Republic of Macedonia)
- Héctor Martínez Alonso gave an invited talk on 01/06/2016 at Darmstadt Technical University (Germany).
- Marie Puren gave an invited talk on 26/10/2016 for theDARIAH's Humanities At Scale Winter School at Charles University in Prague.
- Marie Puren gave an invited talk on 22/11/2016 during the colloquim "DHNord" at the "Maison Européenne des Sciences de l'Homme et de la Société" in Lille.
- Charles Riondet gave an invited talk on 26/10/2016 at the DARIAH's Humanities at Scale Winter School, Charles University, in Prag (Czech Republic)
- Laurent Romary gave a speech at Ready to Reach Out, Conference on Digitization of Cultural Heritage (NL EU presidency) "Data and Dissertations", Amsterdam, The Netherlands on 29/06/2016
- Laurent Romary gave a keynote at ETD 2016 "Data and Dissertations", Lille, France on 11/07/2016 [50]
- Laurent Romary gave a keynote at Language Technologies and & Digital Humanities 2016 on 30/09/2016 in Ljubljana, Slovenia [51]
- Benoît Sagot was an invited panelist during the AnaMorphoSys conference on 20-22/06/2016 in Lyon
- Benoît Sagot gave an invited talk at the EPHE workshop on Digital Humanities on 12/10/2016 in Paris
- Djamé Seddah gave a keynote talk at the workshop "Data Driven Approach to Networks and Language" (Lyon)
- Djamé Seddah gave a keynote talk at the workshop "Challenges for Data-driven Natural Language Analysis beyond Standard Data (Lyon)
- Djamé Seddah gave a invited talk in Dusseldorf (Formal Linguistic Departement) on deep syntax based graph-parsing
- Djamé Seddah gave a invited talk in Lyon (ENS seminar series) on morpho-syntactic analysis in challenging environments
- Djamé Seddah gave a invited talk at the 1st NLP Paris Meetup on graph parsing for French
- Éric Villemonte de La Clergerie gave an invited talk at the meeting on "Information, Me´dias et Informatique" (IRISA Rennes, 15/03/2016)
- Éric Villemonte de La Clergerie gave an invited tutorial at the COST Parseme meeting (Dubrovnik, 27/09/2016)
- Éric Villemonte de La Clergerie gave an invited talk at the AIM-WEST & PARSEME-FR Workshop (IMAG Grenoble, 4/10/2016)
- Éric Villemonte de La Clergerie gave an invited talk at the NLP meetup (Paris, 23/11/2016)

9.1.4. Leadership within the Scientific Community

9.1.4.1. Involvement in international initiatives

- 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) [55]. Within the same subcommittee, Laurent Romary is the convenor of the working group on lexical resources (WG4).
- Laurent Romary is chairman of ISO committee TC 37 “Terminology and other language and content resources”
- Laurent Romary chairs the Board of Directors of the European Research Infrastructure Consortium DARIAH established by the European Commission to coordinate Digital Humanities infrastructure 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

9.1.4.2. Involvement in national initiatives

- Alpage has many responsabilities wihtin the LabEx EFL. Until February 2015, Benoît Sagot is deputy head of this research strand; Marie Candito is deputy head of the research strand on computational semantics; Benoît Crabbé is deputy head of the research strand on experimental grammar; all three are therefore deputy members of the Scientific and Governing Boards of the LabEx; Laurence Danlos is a member of the Scientific Board of the LabEx EFL, representing Alpage;
- 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 or advisory board: Labex ‘Les passes dans le présent’ (PasP), OpenAIRE 2020, OpenEdition (UMS Cleo)
- Laurent Romary is the Inria scientific advisor for Scientific and Technical Information, in charge in particular of the Open Access strategy.

9.1.4.3. Other activities for the scientific community

- Laurence Danlos is member of the Permanent Committee of the TALN conference (CPerm) organised by ATALA.

9.1.5. Scientific Expertise

- Marie Candito was member of the committee for the "prix de thèse ATALA"
- Benoît Crabbé was reviewer for a ERC starting grant
- Laurent Romary was project reviewer for Fonds National Suisse, Switzerland
- Laurent Romary was project reviewer for the Canada Foundation for Innovation, Canada
- Laurent Romary was project reviewer for LabEx PATRIMA, Switzerland
- Laurent Romary has advised the European Patent Office for the specification of their model for representing the non patent literature, based on the TEI (Text Encoding Initiative) guidelines.
- Laurent Romary has been a referee for the selection of a professor at the University of Cologne, Germany
- Laurent Romary has part of the selection committee for a professorship at the Université de Lorraine, France
- Benoît Sagot was project reviewer for the ANR
- Benoît Sagot was an elected board member of the French NLP society (ATALA) until July 2016

- Benoît Sagot was a member of the HCERES evaluation committee for the UMR ATILF (december 2016)
- Djamé Seddah was a reviewer for the Luxemrbourg research funding agency.
- Djamé Seddah was an external reviewer for a Machine Learning PhD thesis (University of Barcelona, sup: Xavier Carreras)
- Éric Villemonte de La Clergerie has participated to several AFNOR meetings in relation with ISO TC37SC4 “Language Resource Management”. He was also a member of the French delegation at the annual ISO TC37 meeting (Copenhagen, June 2016)
- Éric Villemonte de La Clergerie was one of the animators of the working group of the GFII on Knowledge Technologies, and was also involved in the discussions of the GFII about the "regulation of algorithms"

9.1.6. Research Administration

9.1.6.1. University duties

- Lucie Barque was until june 2016 deputy director of the Linguistic department at Université Paris Nord
- Marie Candito is deputy director of the UFR of Linguistics.
- Maximin Coavoux is a student member of the Administrative board of the UFR of Linguistics of University Paris Diderot.
- Benoît Crabbé is a member of the Administrative board of the UFR of Linguistics of University Paris Diderot.
- Laurence Danlos is the chair of the Scientific Committee of the Linguistics UFR of University Paris Diderot.
- Laurence Danlos is the deputy chair of the Doctoral School for Linguistic Sciences (École Doctorale de Sciences du Langage).

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

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, Corpus 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, Pragmatique et Énonciation, 11 heures en équivalent TD, niveau L3, Université Paris 13, France

Licence: Timothée Bernard, TD d’Algorithmique, 24 heures en équivalent TD, niveau L3, Université Paris 7 Diderot, 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: Benoît 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

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

Master: Lucie Barque, La langue et son enseignement, 18 heures en équivalent TD, niveau M1, Université Paris 13, France

Master: Lucie Barque, Problématiques de la néologie, 36 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: Benoît 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: Laurent Romary, Basic encoding and annotation of textual sources in TEI, 24 hours, Master of Library and Information Science, Fach Hochschule Potsdam, Germany

Continuous training: Laurent Romary, Codage de document scientifique en XML TEI, 16 hours, INIST-CNRS, France

9.2.2. 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 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 (tutor in the company Yseop)

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

PhD: 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), defended on January 27, 2016

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

PhD in progress: Loïc Grobol, “Reconnaissance automatique de chaînes de coréférences en français par combinaison d’apprentissage automatique et de connaissances linguistiques”, started in October 2016, supervised by Isabelle Tellier (supervisor), Marco Dinarelli (co-supervisor), and Éric Villemonte de La Clergerie (co-supervisor)

PhD in progress: Axel Herold, “Automatic identification and modeling of etymons in retro-digitized dictionaries”, started in November 2016, supervised by Laurent Romary

PhD in progress: Jack T. Bowers, “Technology, description and theory in language documentation: creating a comprehensive body of multi-media resources for Mixtepec-Mixtec using standards, ontology and Cognitive Linguistics”, started in November 2016, supervised by Laurent Romary

PhD in progress: Mohamed Khemakhem, “Structuration automatique de dictionnaires à partir de modèles lexicaux standardisés”, started in September 2016, supervised by Laurent Romary

9.2.3. Juries

Laurence Danlos was an examinator in the PhD defense committee of Aleksandre Maskharashvili. Title: Discourse Modeling with Abstract Categorial Grammars. University: Université de Lorraine. PhD supervisor: P. de Groote, Sylvain Pogodalla. Defense date: 1st December 2016.

Laurent Romary was a reviewer (*rapporteur*) in the PhD defense committee of Sirine Boukedi Troudi. Title: Outils d’analyse des structures non-têtes. University: Université de Sfax. PhD supervisor: Kais Haddar. Defense date: 11 May 2016.

Laurent Romary was a reviewer (*rapporteur*) in the PhD defense committee of Daouda Sawadogo. Title: Architectures logicielles et mécanismes pour la gestion adaptative et consolidée de ressources numériques dans une application interactive scénarisée. University: Université de La Rochelle. PhD supervisor: Pascal Estraillier and Ronan Champagnat. Defense date: 28 June 2016.

Djamé Seddah and Éric Villemonte de La Clergerie was a member in the PhD defense committee of Corentin Ribeyre as co-supervisor, together with Laurence Danlos (as PhD director). Title: Méthodes d’Analyse Supervisée pour l’Interface Syntaxe-Sémantique. University: Université Paris-Diderot.Defense date: 27 January 2016.

Éric Villemonte de La Clergerie was a member in the PhD defense committee of Julie Belião. Title: How prosody and syntax are mapping: A study of synchronization and congruence. University: Université Paris Ouest. Defense date: 5 December 2016

9.3. Popularization

- Interview of Laurent Romary on new scientific publication models for APS news, American Physical Society (http://www.aps.org/publications/apsnews/201602/arxiv.cfm).
- Laurent Romary and Marie Puren have given an interview on 24/11/2016 on the European project Parthenos for the Inria website (https://www.inria.fr/centre/paris/actualites/epi-alpage-historiens-et-informaticiens-reunis-dans-le-projet-europeen-parthenos).
- Timothée Bernard has welcomed two high school students at Alpage during 4 days, in the framework of the “Science Académie” of the “Paris-Montagne” association.

10. Bibliography

Major publications by the team in recent years

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

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

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

[4] L. DANLOS. *D-STAG : un formalisme d'analyse automatique de discours fondé sur les TAG synchrones*, in "Traitement Automatique des Langues", 2009, vol. 50, n° 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° 1, https://hal.inria.fr/hal-00698896

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

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

Publications of the year

Articles in International Peer-Reviewed Journals

[11] Z. AGIC, A. JOHANNSEN, B. PLANK, H. MARTÍNEZ ALONSO, N. SCHLUTER, A. SØGAARD. *Multilingual Projection for Parsing Truly Low-Resource Languageš*, in " Transactions of the Association for Computational Linguistics", August 2016, https://hal.inria.fr/hal-01426754

[12] M. COAVOUX, B. CRABBÉ. *Prédiction structurée pour l'analyse syntaxique en constituants par transitions : modèles denses et modèles creux*, in "Traitement Automatique des Langues", 2016, vol. 57, n° 1, https://hal.inria.fr/hal-01365252

[13] L. DANLOS, Q. PRADET, L. BARQUE, T. NAKAMURA, M. CONSTANT. *Un Verbenet du français*, in "Traitement Automatique des Langues", September 2016, vol. 57, n° 1, 25 p. , https://hal.inria.fr/hal-01392817

[14] B. GAUME, K. DUVIGNAU, E. NAVARRO, Y. DESALLE, H. CHEUNG, S. HSIEH, P. MAGISTRY, L. PREVOT. *Skillex: a graph-based lexical score for measuring the semantic efficiency of used verbs by human subjects describing actions*, in "Revue TAL", 2016, vol. 55, n° 3, https://hal.archives-ouvertes.fr/hal-01320416

[15] H. MARTINEZ ALONSO, D. ZEMAN. *Universal Dependencies for the AnCora treebanks* , in "Procesamiento del Lenguaje Natural", September 2016, n° 57, https://hal.inria.fr/hal-01426751

[16] L. ROMARY, M. MERTENS, A. BAILLOT. *Datenfluss an der Schnittstelle von Forschung, Infrastruktur und Kulturerbeinstitutionen: der DARIAH-Fahrplan 2016*, in "BIBLIOTHEK Forschung und Praxis", 2016, vol. 39, n° 3, pp. 350–357, https://hal.inria.fr/hal-01285917

Articles in National Peer-Reviewed Journals

[17] C. RIONDET. *Journaux Intimes de Clandestinité : Le cas de Léo Hamon (19040-1944)*, in "Vingtième siècle", October 2016, https://hal.inria.fr/hal-01416988

Invited Conferences

[18] M. PUREN. *A l'épreuve de l'hétérogénéité : données de recherche et interdisciplinarité : L'exemple du projet européen IPERION-CH*, in "DHnord 2016 - Humanités numériques: théories, débats, approches critiques", Lille, France, Maison Européenne des Sciences de l'Homme et de la Société , November 2016, https://hal.archives-ouvertes.fr/hal-01408951

[19] M. PUREN, C. RIONDET. *Research data management, a chance for Open Science. Methods and tutorials to create a Data Management Plan (DMP)*, in "DARIAH's Humanities at Scale Winter School", Prague, Czech Republic, Dariah and Humanities at Scale, October 2016, https://hal.inria.fr/hal-01416978

[20] C. RIONDET. *De Gaulle et l'organisation de la résistance à Paris*, in "De Gaulle et Paris", Paris, France, Comité d'Histoire de la ville de Paris and Fondation Charles de Gaulle, April 2016, https://halshs.archives-ouvertes.fr/halshs-01301278

International Conferences with Proceedings

[21] L. BARQUE. *A survey on semantic productivity*, in "Workshop Expanding the lexicon", Trier, Unknown or Invalid Region, D. GRAS (editor), 2016, https://halshs.archives-ouvertes.fr/halshs-01428252

[22] R. BAWDEN, B. CRABBÉ. *Boosting for Efficient Model Selection for Syntactic Parsing*, in "COLING 2016 - 26th International Conference on Computational Linguistics", Osaka, Japan, December 2016, pp. 1-11, https://hal.inria.fr/hal-01391743

[23] T. BERNARD. *Modelling Subordinate Conjunctions in STAG: A Discourse Perspective*, in "28th European Summer School in Logic, Language & Information", Bozen-Bolzano, Italy, Proceedings of the ESSLLI 2016 Student Session, August 2016, 13 p. , https://hal.inria.fr/hal-01363201

[24] T. BERNARD, L. DANLOS. *Modelling Discourse in STAG: Subordinate Conjunctions and Attributing Phrases*, in "12th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+12)", Dûsseldorf, Germany, Proceedings of the 12th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+12), June 2016, pp. 38-47, https://hal.archives-ouvertes.fr/hal-01329539

[25] M. COAVOUX, B. CRABBÉ. *Neural Greedy Constituent Parsing with Dynamic Oracles*, in "Association for Computational Linguistics (ACL)", Berlin, Germany, 2016, https://hal.inria.fr/hal-01353734

[26] L. DANLOS, M. CONSTANT, L. BARQUE. *Improvement of VerbNet-like resources by frame typing*, in "Workshop on Grammar and Lexicon: interactions and interfaces (GramLex)", Osaka, Japan, Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex), The COLING 2016 Organizing Committee, December 2016, https://hal.inria.fr/hal-01392822

[27] L. DANLOS, A. MASKHARASHVILI, S. POGODALLA. *Interfacing Sentential and Discourse TAG-based Grammars*, in "The 12th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+12)", Düsseldorf, Germany, Proceedings of the 12th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+12), June 2016, https://hal.inria.fr/hal-01328697

[28] M. DJEMAA, M. CANDITO, P. MULLER, L. VIEU. *Corpus annotation within the French FrameNet: a domain-by-domain methodology*, in "Tenth International Conference on Language Resources and Evaluation (LREC 2016)", Portorož, Slovenia, May 2016, https://hal.archives-ouvertes.fr/hal-01391526

[29] M. LHIOUI, K. HADDAR, L. ROMARY. *A new method for interoperability between lexical resources using MDA approach*, in "AISI 2016 The 2nd International Conference on Advanced Intelligent Systems and Informatics", Cairo, Egypt, October 2016, https://hal.inria.fr/hal-01350524

[30] H. MARTINEZ ALONSO, A. JOHANNSEN, B. PLANK. *Supersense tagging with inter-annotator disagreement*, in "Linguistic Annotation Workshop 2016", Berlin, Germany, August 2016, pp. 43 - 48, https://hal.inria.fr/hal-01426747

[31] O. MICHALON, C. RIBEYRE, M. CANDITO, A. NASR. *Deeper syntax for better semantic parsing*, in "Coling 2016 - 26th International Conference on Computational Linguistics", Osaka, Japan, December 2016, https://hal.archives-ouvertes.fr/hal-01391678

[32] B. SAGOT. *Multilingual part-of-speech tagging with MElt*, in "23ème Conférence sur le Traitement Automatique des Langues Naturelles", Paris, France, July 2016, https://hal.inria.fr/hal-01352243

[33] A. SIMONENKO, B. CRABBÉ, S. PRÉVOST. *Taraldsen's Generalization in Diachrony: Evidence from a Diachronic Corpus*, in "West Coast Conference on Formal Linguistics", Salt Lake City, United States, 2016, https://hal.inria.fr/hal-01353741

[34] L. VIEU, P. MULLER, M. CANDITO, M. DJEMAA. *A general framework for the annotation of causality based on FrameNet*, in "Tenth International Conference on Language Resources and Evaluation (LREC 2016)", Portorož, Slovenia, May 2016, https://hal.archives-ouvertes.fr/hal-01391542

[35] S. M. YIMAM, H. MARTÍNEZ ALONSO, M. RIEDL, C. BIEMANN. *Learning Paraphrasing for Multi-word Expressions*, in "MWE 2016 - Multiword Expression Workshop 2016", Berlin, Germany, August 2016, https://hal.inria.fr/hal-01426749

National Conferences with Proceedings

[36] T. BERNARD. *Conjonctions de subordination, verbes de dire et d'attitude propositionnelle : une modélisation STAG pour le discours*, in "18ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement

Automatique des Langues", Paris, France, Actes de la conférence conjointe JEP-TALN-RECITAL 2016, July 2016, vol. volume 3 : RECITAL, pp. 27-39, https://hal.archives-ouvertes.fr/hal-01357125

Conferences without Proceedings

[37] A. BAILLOT. *A certification model for digital scholarly editions: Towards peer review-based data journals in the humanities*, in "Digital Scholarly Editing: Theory, Practice, Methods", Anvers, Belgium, Université d'Anvers, October 2016, https://halshs.archives-ouvertes.fr/halshs-01392880

[38] A. SIMONENKO, B. CRABBÉ, S. PREVOST. *Effects of literary form on grammatical changes: A treebank study*, in "49th Annual Meeting of the Societas Linguistica Europaea (SLE 2016)", Naples, Italy, 2016, https://hal.inria.fr/hal-01365263

[39] A. SIMONENKO, B. CRABBÉ, S. PRÉVOST. *Quantificational dimension of Taraldsen's Generalisation*, in "New Ways of Analyzing Syntactic Variation 2 (NWASV 2)", Ghent, Belgium, 2016, https://hal.inria.fr/hal-01353738

[40] A. SIMONENKO, B. CRABBÉ, S. PRÉVOST. *Taraldsen's Generalisation in Medieval French*, in "Diachronic Generative Syntax conference (DIGS)", Ghent, Belgium, 2016, https://hal.inria.fr/hal-01353736

Scientific Books (or Scientific Book chapters)

[41] 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), Taylor & Francis Group, 2016, https://hal.inria.fr/hal-01248562

[42] L. ROMARY. *Eléments d'une communication scientifique ouverte et publique*, in "Publier, éditer, éditorialiser. Nouveaux enjeux de la production numérique", L. CALDERAN, P. LAURENT, H. LOWINGER, J. MILLET (editors), Information & Stratégie, De Boek, 2016, https://hal.inria.fr/hal-01328192

Research Reports

[43] J. BAETEN, P. ESTRAILLIER, C. KIRCHNER, A. MOATTI, L. ROMARY. *Open Access in Japan – a multi-institutional perspective*, Ambassade de France au Japon, March 2016, https://hal.archives-ouvertes.fr/hal-01290936

[44] B. SAGOT. *External Lexical Information for Multilingual Part-of-Speech Tagging*, Inria Paris, June 2016, n° RR-8924, https://hal.inria.fr/hal-01330301

Other Publications

[45] P. BANSKI, B. GAIFFE, P. LOPEZ, S. MEONI, L. ROMARY, T. SCHMIDT, P. STADLER, A. WITT. *Wake up, standOff!*, September 2016, TEI Conference 2016, https://hal.inria.fr/hal-01374102

[46] J. BOWERS, L. ROMARY. *Deep encoding of etymological information in TEI*, November 2016, working paper or preprint, https://hal.inria.fr/hal-01296498

[47] L. DANLOS, B. CRABBÉ. *Natural Language Processing, 60 years after the Chomsky-Schützenberger hierarchy*, March 2016, Marie Paule Schützenberger 20 ans après, https://hal.inria.fr/hal-01392829

[48] E. NIVAULT, A. MONTEIL, L. FARHI, L. ROMARY. *Implementation of the IFIP Digital Library in the HAL open publication repository*, June 2016, Libraries Opening Paths to Knowledge: LIBER Annual Conference 2016, Poster, https://hal.inria.fr/hal-01327170

[49] L. ROMARY, M. PUREN. *Datasets of IPERION CH*, March 2016, Atelier interdisciplinaire « Matériaux du patrimoine et patrimoine matériel », https://hal.inria.fr/hal-01289058

[50] L. ROMARY. *Elements of a scientific communication policy*, July 2016, ETD 2016 "Data and Dissertations", https://hal.inria.fr/hal-01345623

[51] L. ROMARY. *The Text Encoding Initiative: 30 years of accumulated wisdom and its potential for a bright future*, September 2016, Language Technologies & Digital Humanities 2016, https://hal.inria.fr/hal-01374597

References in notes

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

[53] 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

[54] 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

[55] 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

[56] 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

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

[58] 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

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

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

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

[64] 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

[65] 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

[66] 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

[67] 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

[68] 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

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

[70] 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

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

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

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

[74] 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

[75] 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

[76] 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

[77] 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

[78] 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

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

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

[81] B. SAGOT, L. CLÉMENT, É. VILLEMONTE DE LA CLERGERIE, P. BOULLIER. *The Lefff 2 syntactic lexicon for French: architecture, acquisition, use*, in "Proc. of LREC'06", 2006, http://hal.archives-ouvertes.fr/docs/00/41/30/71/PDF/LREC06b.pdf

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

[83] 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

[84] 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

[85] 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

[86] 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

[87] 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

[88] 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

[89] 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

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

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

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

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

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

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