INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team alpage

Large-scale deep linguistic processing

Paris - Rocquencourt

Theme : Audio, Speech, and Language Processing

Table of contents

1. Team

Research Scientists

Pierre Boullier [Emeritus Senior Researcher (DR-E) Inria, HdR]
Pascal Denis [Junior Researcher (CR) Inria]
Éric Villemonte de La Clergerie [Junior Researcher (CR) Inria]
Benoît Sagot [Junior Researcher (CR) Inria]

Faculty Members

François Barthélemy [Associate Professor (MC) CNAM]
Marie Candito [Associate Professor (MC) Univ. Paris 7]
Benoît Crabbé [Associate Professor (MC) Univ. Paris 7]
Laurence Danlos [Full Professor (PR) Univ. Paris 7, Member of IUF, Team leader, HdR]
Sylvain Kahane [Full Professor (PR) Univ. Paris X, Associate member, HdR]
Philippe Muller [delegation from Université Paul Sabatier, Toulouse, since September 2009]
Djamé Seddah [Associate Professor (MC) Univ. Paris 4]

PhD Students

André Bittar [PhD student Univ. Paris 7 (since 2007), now ATER at Université Paris-Est Marne-la-Vallée]
Luc Boruta [PhD student (allocataire) (since October 2009)]
François-Régis Chaumartin [PhD student Univ. Paris 7]
Elżbieta Gryglicka [PhD student (CIFRE) Thales & Univ. Paris 7]
Enrique Henestroza Anguiano [PhD funded by the ANR project SEQUOIA (since November 2009)]
Emmanuel Lassale [PhD student (ENS stipendium) Univ. Paris 7 (since September 2010)]
Pierre Magistry [PhD student (allocataire) Univ. Paris 7 (since September 2010)]
Charlotte Roze [PhD student (allocataire) Univ. Paris 7 (since October 2009)]
Rosa Stern [PhD student (CIFRE) AFP & Univ. Paris 7 (since November 2009)]
Juliette Thuilier [PhD student (allocataire) Univ. Paris 7 (since 2008)]

Post-Doctoral Fellows

Marianna Apidianaki [funded by the System@tic project Scribo (May to December 2010)]
Kata Gábor [funded by the System@tic project Scribo (May to December 2010)]
Yayoi Nakamura-Delloye [funded by the System@tic project Scribo (March to November 2010), then by the ANR project EDyLex (from December 2010)]
Sattisvar Tandabany [funded by the ANR project SEQUOIA (from November 2009)]
Tim Van de Cruys [funded by the System@tic project Scribo (May to December 2010)]

Administrative Assistant

Christelle Guiziou [Secretary (SAR) Inria]

Other

Gaëlle Recourcé [funded by the Scribo project, part-time, until June 2010]

2. Overall Objectives

2.1. Overall Objectives

The Alpage team is specialized in **Language modeling**, **Computational linguistics** and **Natural Language Processing (NLP)**. These fields are considered central in the new Inria strategic plan, and are indeed of crucial importance for the new information society. Applications of this domain of research include the numerous technologies grouped under the term of "language engineering" (information retrieval, information extraction, spelling, grammatical and semantic correction, automatic summarizing, machine translation, man machine communication, etc).

NLP, the domain of Alpage, is a subfield of both artificial intelligence, linguistics, and cognition. It studies the problems of automated understanding and generation of natural human languages. 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 text generation (by opposition to speech processing and generation).

NLP applications are numerous, and include domains such as machine translation, question answering, information retrieval, information extraction, text simplification, automatic or computer-aided translation, automatic symmetrization, 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 is a multidisciplinary domain. Indeed, it requires an expertise in formal and descriptive linguistics (to develop linguistic models of human languages), in computer science and algorithmics (to design and develop efficient programs that can deal with such models), in applied mathematics (to acquire automatically linguistic or general knowledge) and in other related fields. It is one of the specificities of Alpage to put together NLP specialists with a strong background in all these fields (in particular, linguistics for Paris 7 Alpage members, previously in the Lattice UMR, computer science and algorithmics for Inria members).

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

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

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

2.2. Highlights

2010 is a very important year for Alpage as the team's visibility, dissemination opportunities and involvement in projects and projects proposal has broaden out.

First, Alpage's international connections and collaborations are now numerous and lead to concrete results, which shows that the team's work is recognized internationaly. Alpage members are involved in a European (FP7) project proposal, have organized and organizing several workshops (SPMRL at NAACL 2010, WoLeR at ESSLLI 2011, SSSST at ACL 2011), are in charge of special issues in several journals (Computational Linguistics, TAL).

Second, the number and impact of Alpage's industrial collaborations have increased significantly. Two members of Alpage also work for companies that exploit Alpage's results and technologies, including the INRIA spin-off *Verbatim Analysis*. Two CIFRE PhDs (i.e., PhDs in collaboration with a company), one with AFP and one with Verbatim Analysis (starting Jan 1st, 2011) also contribute to transfering Alpage tools and resources to the industry. Thanks to the Iliatech day organized by INRIA and Alpage, which took place Oct 20, several other industrial contracts have been set up, some of them already accepted (with Lingua et Machina and Dated), an other still under review by the Pôle de Compétitivité System@tic (with Diadeis).

3. Scientific Foundations

3.1. From programming languages to linguistic grammars

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

Historically, several members of Alpage were originally specialists in the domain of modeling and parsing for programming languages, and are working for more than 15 years on the generalization and extension of the techniques involved to the domain of natural language. The shift from programming language grammars to NLP grammars seriously increases complexity and requires ways to handle the ambiguities inherent in every human language. It is well known that these ambiguities are the sources of many badly handled combinatorial explosions.

Furthermore, while most programming languages are expressed by (subclasses) of well-understood context-free grammars (CFGs), no consensual grammatical formalism has yet been accepted by the whole linguistic community for the description of human languages. On the contrary, new formalisms (or variants of older ones) appear constantly. Many of them may be classified into the three following large families:

Mildly Context-Sensitive (MCS) formalisms They manipulate possibly complex elementary structures with enough restrictions to ensure the possibility of parsing with polynomial time complexities. They include, for instance, Tree Adjoining Grammars (TAGs) and Multi-component TAGs with trees as elementary structures, Linear Indexed Grammars (LIGs). Although they are strictly more powerful than MCS formalisms, Range Concatenation Grammars (RCGs, introduced and used by Alpage members, such as Pierre Boullier and Benoît Sagot [59], [92], [97]) are also parsable in polynomial time.

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

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

An efficient way to develop large-coverage hand-crafted symbolic grammars is to use adequate tools and adequate levels of representation, and in particular Meta-Grammars, one of Alpage's areas of expertise [108], [102]. Meta-Grammars allows the linguist to focus on a modular description of the linguistic aspects of a grammar, rather than focusing on the specific aspects of a given grammatical formalism. Translation from MGs to grammatical formalisms such as TAG or LFG may be automatically handled. Graphical environments can be used to design MGs and their modularity provides a promising way for sharing the description of common linguistic phenomena across human languages.

3.2. Statistical Parsing

Contrary to symbolic approaches to parsing, in statistical parsing, the grammar is extracted from a corpus of syntactic trees : a treebank. The main advantage of the statistical approach is to encode within the same framework the parsing and disambiguating tasks. The extracted grammar rules are associated with probabilities that allow to score and rank the output parse trees of an input sentence. This obvious advantage of probabilistic context-free grammars has long been counterbalanced by two main shortcomings that resulted in poor performance for plain PCFG parsers: (i) the generalization encoded in non terminal symbols that stand for syntagmatic phrases is too coarse (so probabilistic independence between rules is too strong an assertion) and (ii) lexical items are underused. In the last decade though, effective solutions to these shortcomings have been proposed. Symbol annotation, either manual [78] or automatic [88], [89] 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 [68], or probabilistic models derived from lexicalized Tree Adjoining grammars, such as Stochastic Tree Insertion Grammars [65].

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 ...) [112], [86]. 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 [82]. These two kinds of statistical dependency parsing allow to benefit from discriminative learning, and its ability to easily integrate various kinds of features, which is typically needed in a complex task such as parsing.

Statistical parsing is now effective, both for syntagmatic representations and dependency-based syntactic representations. Alpage has obtained state-of-the-art parsing results for French, by adapting various parser learners for French, and works on the current challenges in statistical parsing, namely (1) robustness and portability across domains and (2) the ability to incorporate exogenous data to improve parsing attachment decisions. We review below the approaches that Alpage has tested and adapted, and the techniques that we plan to investigate to answer these challenges.

In order to investigate statistical parsers for French, we have first worked how to use the French Treebank [58] and derive the best input for syntagmatic statistical parsing [70]. Benchmarking several PCFG-based learning frameworks [9] 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) [89].
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 [65] 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 [40].

Pursuing our work on PCFG-LA based parsing, we investigated the automatic conversion of the treebank into dependency syntax representations [63], that are easier to use for various NLP applications such as question-answering or information extraction, and that are a better ground for further semantic analysis. This conversion can be applied on the treebank, before training a dependency-based parser, or on PCFG-LA parsed trees. This gives the possibility to evaluate and compare on the same gold data, both syntagmatic- and dependency-based statistical parsing. This also paved the way for studies on the influence of various types of lexical information. Results are described in sections 6.6 and 6.5.

3.3. Dynamic wide coverage lexical resources

Participants: Benoît Sagot, Laurence Danlos, Rosa Stern, Éric Villemonte de La Clergerie.

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

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

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

Successful experiments have been conduced by Alpage members with different languages for the automatic acquisition of morphological knowledge from raw corpora [96]. At the syntactic level, work has been achieved on automatic acquisition of atomic syntactic information and automatic detection of errors in the lexicon [113],[8]. At the semantic level, automatic wordnet development tools have been described [91], [109], [75], [74]. 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* [94],[34], developed within the Alexina framework, as well as a wordnet for French, the WOLF [95], the first freely available resource of the kind.

In the last 2 years, Alpage members have shown how to benefit from other more linguistically-oriented resources, such as the Lexique-Grammaire and DICOVALENCE , in order to improve the coverage and quality of the Le*fff* and the WOLF. This work is a good example of how Inria and Paris 7 members of Alpage fruitful collaborate: this collaboration between NLP computer scientists and NLP linguists have resulted in significant advances which would have not been possible otherwise.

Moreover, an increasing effort has been made towards multilingual aspects. In particular, Alexina lexicons developed in 2009 or before exist for Slovak [96], Polish [98], English, Spanish and Persian (although very preliminarily before 2010, see 6.13), not including freely-available lexicons adapted to the Alexina framework.

3.4. Shallow processing

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

The constitution of resources such as lexica or grammars raises the issues of the evaluation of these resources to assess their quality and coverage. For this reason, Alpage is the leader of the PASSAGE ANR project (ended in June 2010), which is the follow-up of the EASy parsing evaluation campaign held in 2004 and conducted by team LIR at LIMSI.

However, although developing parsing techniques, grammars (symbolic or probabilistic), and lexica constitute obviously the key efforts towards deep large-scale linguistic processing, these components need to be included inside a full and robust processing chain, able to handle any text from any source. The development of such linguistic chains, such as SXPipe , is not a trivial task [7]. Moreover, when used as a preliminary step before parsers, the quality of parsers' results strongly depends on the quality of such chains. In that regard, less-standard pre-processings such as word clustering show promising results (see, e.g., section 6.5)

I fact, such processing chains are mostly used as such, and not only as pre-processing tools before parsing. They aim at performing the basic tasks that produce immediately usable results for many applications, such as tokenization, sentence segmentation, spelling correction, and, most importantly, named entity detection, disambiguation and resolution (see section 6.12).

3.5. Discourse structures

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

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

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

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

- first, some connectors (in particular conjunctions of subordination) introduce pure modifiers and must not be considered as bearing discourse relations,
- second, other elements than connectors can lexicalize discourse relations, in particular verbs like *précéder / to precede* or *causer / to cause*, which have facts or fact eventualities as arguments [71].

There are three main frameworks used to model discourse structures: RST, SDRT , and, more recently, D-LTAG. Inside Alpage, Laurence Danlos has introduced D-STAG (Discourse Synchronous TAGs, [72],[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.

3.6. Coreference resolution

Participants: Pascal Denis, Philippe Muller, Elżbieta Gryglicka, Laurence Danlos.

An important challenge for the understanding of natural language texts is the correct computation of the *discourse entities* that are mentioned therein —persons, locations, abstract objects, and so on. In addition to identifying individual referential expressions (e.g., *Nicolas Sarkozy*, *Neuilly*, *l'UMP*) and properly typing them (e.g. *Nicolas Sarkozy* is a PERSON, *Neuilly* is a LIEU), the task is also to determine the other mentions with which these expressions are coreferential. Part of the difficulty of this task is that natural languages provide many ways to refer to the same entity (including the use of pronouns such as *il*, *ses* and definite descriptions such as *le président*, making them highly ambiguous. The identification of coreferential links and other anaphoric links (such as "associative anaphora") plays a key role for various applications, such as extraction and retrieval of information, but also the summary or automatic question-answering systems. This central role of coreference resolution has been recognized by the inclusion of this task in different international evaluation campaigns, beginning with the campaigns *Message Understanding Conference* (in particular, MUC-6 and MUC-7)[1], and more recently *Automatic Content Extraction* (ACE)[2] and *Anaphora Resolution Evaluation* (ARE)[3]. The creation and distribution of corpora developped as part of these campaigns have significantly boosted research in automatic coreference resolution. In particular, they have made possible the application of

[1] See, respectively: http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html and http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.

[2] http://www.nist.gov/speech/tests/ace/

[3] http://clg.wlv.ac.uk/events/ARE/

machine learning techniques (mostly supervised ones) to the problem of coreference resolution. This in turn has led to the development of systems that were both more robust and more precise, thus making more realistic their integration within these larger systems. Some of the best systems based on supervised learning methods are described in [101], [83], [81], [84], [79];[6]. Note that a few attemtps were also made at using unsupervised techniques (mostly clustering methods) for the task [64], [85], but these systems are still far from reaching the performance of their supervised counterparts.

4. Application Domains

4.1. Panorama

NLP tools and methods have many possible domains of application. Some of then are already mature enough to be commercialized. They can be roughly classified in three groups:

Human-computer interaction : mostly speech processing and text-to-speech, often in a dialogue context; today, commercial offers are limited to restricted domains (train tickets reservation...);

Language writing aid : spelling, grammatical and stylistic correctors for text editors, controlled-language writing aids (e.g., for technical documents), memory-based translation aid, foreign language learning tools, as well as vocal dictation;

Access to information : tools to enable a better access to information present in huge collections of texts (e.g., the Internet): automatic document classification, automatic document structuring, automatic summarizing, information acquisition and extraction, text mining, question-answering systems, as well as surface machine translation. Information access to speech archives through transcriptions is also an emerging field.

Experimental linguistics : tools to explore language in an objective way (this is related, but not limited to corpus linguistics).

Alpage focuses on some applications included in the three last points, such as information extraction and (linguistic and extra-linguistic) knowledge acquisition (4.2), text mining (4.3), text generation (4.6), spelling correction (4.7) and experimental linguistics (4.8).

4.2. Information extraction and knowledge acquisition

Participants: Éric Villemonte de La Clergerie, François-Régis Chaumartin, Rosa Stern, Kata Gábor, Tim van de Cruys, Yayoi Nakamura-Delloye, Marianna Apidianaki, Benoît Sagot.

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

Knowledge acquisition for a given restricted domain is something that has already been studied by some Alpage members for several years (ACI Biotim, biographic information extraction from the Maitron corpus, Scribo project). François-Régis Chaumartin, PhD student at Alpage and CEO of Proxem, is working on information extraction from the English Wikipedia. Indeed, chunking or, better, syntactic (and semantic) parsing gives an access, through learning techniques, to useful information present in documents. Obviously, the progressive extension of Alpage parsing systems to a full syntactic *and* semantic parsing will increase the quality of the extracted information, as well as the scope of information that can be extracted. Such knowledge acquisition efforts bring solutions to current problems related to information access and take place into the emerging notion of *Semantic Web*. The transition from a web based on data (textual documents,...) to a web based on knowledge requires linguistic processing tools which are able to provide fine grained pieces of information, in particular by relying on high-quality deep parsing. For a given domain of knowledge (say, tourism), the extraction of a domain ontology that represents its key concepts and the relations between them is a crucial task, which has a lot in common with the extraction of linguistic information.

All these applications in the domain of information extraction raise exciting challenges that require altogether ideas and tools coming from the domains of computational linguistics, machine learning and knowledge representation.

4.3. Processing answers to open-ended questions in surveys: vera

Participant: Benoît Sagot.

Verbatim Analysis is a startup co-created by Benoît Sagot from Alpage and Dimitri Tcherniak from Towers Watson, a world-wide leader in the domain of employee research (opinion mining among the employees of a company or organization). The aim of its first product, *vera*, is to provide an all-in-one environment for editing (i.e., normalizing the spelling and typography), understanding and classifying answers to open-ended questions, and relating them with closed-ended questions, so as to extract as much valuable information as possible from both types of questions. The editing part relies in part on SXPipe (see section 5.5) and Alexina morphological lexicons. Several other parts of *vera* are co-owned by Verbatim Analysis and by INRIA.

4.4. Shallow processing of e-mails

Participants: Benoît Sagot, Laurence Danlos.

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

4.5. Multilingual terminologies and lexical resources for companies

Participants: Éric Villemonte de La Clergerie, Benoît Sagot.

Lingua et Machina is a small company now head by François Brown de Colstoun, a former INRIA researcher, that provides services for developing specialized multilingual terminologies for its clients. It develops the framework Libellex for validating such terminologies. A formal collaboration with ALPAGE is under preparation, centered around the joint extension and improvement of Lingua et Machina's and ALPAGE's lexical resources and associated tools.

4.6. Generation of textual reports about statistical data: EASYTEXT

Participant: Laurence Danlos.

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

A paper describing roughly this system and the economic stakes of Natural Language Generation has been published in *La Recherche* [57]. See also a column in *La Tribune* (9/11/2010).

Another application of NLG we foresee is the automatic production of captions for photos. There is ongoing discussions with AFP (*Agence France-Presse*) on the topic.

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

Participants: Benoît Sagot, Éric Villemonte de La Clergerie, Laurence Danlos.

NLP tools and resources used for spelling correction, such as large n-gram collections, POS taggers and finite-state machinery are now mature and precise. In industrial setting such as post-processing after large-scale OCR, these tools and resources should enable spelling correction tools to work on a much larger scale and with a much better precision than what can be found in different contexts with different constraints (e.g., in text editors). Moreover, such industrial contexts allow for a non-costly manual intervention, in case one is able to identify the most uncertain corrections. An FUI project on this topic has been proposed in collaboration with Diadeis, a company specialized in text digitalization, and two other partners. It is currently under review.

4.8. Experimental linguistics

Participants: Benoît Crabbé, Juliette Thuilier, Luc Boruta.

Alpage is a team that dedicates efforts in producing ressources and algorithms for processing large amounts of textual materials. These ressources can be applied not only for purely NLP purposes but also for linguistic purposes. Indeed, the specific needs of NLP applications led to the development of electronic linguistic resources (in particular lexica, annotated corpora, and treebanks) that are sufficiently large for carrying statistical analysis on linguistic issues. In the last 10 years, pioneering work has started to use these new data sources to the study of English grammar, leading to important new results in such areas as the study of syntactic preferences [61], [110], the existence of graded grammaticality judgments [77].

The reasons for getting interested for statistical modelling of language can be traced back by looking at the recent history of grammatical works in linguistics. In the 1980s and 1990s, theoretical grammarians have been mostly concerned with improving the conceptual underpinnings of their respective subfields, in particular through the construction and refinement of formal models. In syntax, the relative consensus on a generative-transformational approach [66] gave way on the one hand to more abstract characterizations of the language faculty [66], and on the other hand to the construction of detailed, formally explicit, and often implemented, alternative formulation of the generative approach [60], [90]. For French several grammars have been implemented in this trend, among which the tree adjoining grammars of [62], [69] among others. This general movement led to much improved descriptions and understanding of the conceptual underpinnings of both linguistic competence and language use. It was in large part catalyzed by a convergence of interests of logical, linguistic and computational approaches to grammatical phenomena.

However, starting in the 1990s, a growing portion of the community started being frustrated by the paucity and unreliability of the empirical evidence underlying their research. In syntax, data was generally collected impressionistically, either as ad-hoc small samples of language use, or as ill-understood and little-controlled grammaticality judgements (Schütze 1995). This shift towards quantitative methods is also a shift towards new scientific questions and new scientific fields. Using richly annotated data and statistical modelling, we address questions that could not be addressed by previous methodology in linguistics. In this line, at Alpage we have started investigating the question of choice in French syntax with a statistical modelling methodology. Currently two studies are being led on the position of attributive adjectives w.r.t. the noun and the relative position of postverbal complement. This research has contributed to establish new links with the Laboratoire de Linguistique Formelle (LLF, Paris 7) and the Laboratoire de Psychologie et Neuropsychologie Cognitives (LPNCog, Paris 5) with which we intend to submit joint research projects next year.

On the other hand we have also started a collaboration with the Laboratoire de Sciences Cognitives de Paris (LSCP/ENS) where we explore the design of algorithms towards the statistical modelling of language acquisition (phonological acquisition). This is currently supported by one PhD project.

5. Software

5.1. Syntax

Participants: Pierre Boullier [correspondant], Sattisvar Tandabany, Benoît Sagot.

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

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

Direct NLP users of SYNTAX for NLP, outside Alpage, include Alexis Nasr (Marseilles) and other members of the SEQUOIA ANR project (see section 8.2.1), Owen Rambow and co-workers at Columbia University (New York), as well as (indirectly) all SXPipe and/or SXLFG users. The project-team VASY (INRIA Rhône-Alpes) is one of SYNTAX' user for non-NLP applications.

5.2. System DyALog

Participant: Éric Villemonte de La Clergerie [maintainer].

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

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

The current release **1.13.0** of DYALOG is freely available by FTP under an open source license and runs on Linux platforms for x86 and architectures and on Mac OS intel (both 32 and 64bits architectures). A partial port for Window Cygwin has been successful but has not yet been integrated and finalized.

The current release handles logic programs, DCGs (*Definite Clause Grammars*), FTAGs (*Feature Tree Adjoining Grammars*), FTIGs (*Feature Tree Insertion Grammars*) and XRCGs (*Range Concatenation Grammars* with logic arguments). Several extensions have been added to most of these formalisms such as intersection, Kleene star, and interleave operators. Typed Feature Structures (TFS) as well as finite domains may be used for writing more compact and declarative grammars [107].

C libraries can be used from within DYALOG to import APIs (`mysql`, `libxml`, `sqlite`, ...).

DYALOG is largely used within ALPAGE to build parsers but also derivative softwares, such as a compiler of Meta-Grammars (cf. 5.3). It has also been used for building a parser from a large coverage French TIG/TAG grammar derived from a Meta-Grammar. This parser has been used for the Parsing Evaluation campaign EASy, the two Passage campaigns (Dec. 2007 and Nov. 2009), cf. 8.2.4 and [102], [106], and very large amount of data (700 millions of words) for the SCRIBO action, cf. 8.1.1.

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

DYALOG and other companion modules are available on INRIA GForge.

5.3. Tools and resources for Meta-Grammars

Participant: Éric Villemonte de La Clergerie [maintainer].

`mgcomp`, MGTOOLS*, and FRMG on INRIA GForge: http://mgkit.gforge.inria.fr/*

DYALOG (cf. 5.2) has been used to implement `mgcomp`, a compiler of Meta-Grammar (cf. 6.1). Starting from an XML representation of a MG, `mgcomp` produces an XML representation of its TAG expansion.

The current version **1.5.0** is freely available by FTP under an open source license. It is used within ALPAGE and (occasionally) at LORIA (Nancy) and at University of Pennsylvania.

The current version adds the notion of namespace, to get more compact and less error-prone meta-grammars. It also provides other extensions of the standard notion of Meta-Grammar in order to generate very compact TAG grammars. These extensions include the notion of *Guarded nodes*, i.e. nodes whose existence and non-existence depend on the truth value of a guard, and the use of the regular operators provided by DyALog on nodes, namely disjunction, interleaving and Kleene star. The current release provides a dump/restore mechanism for faster compilations on incremental changes of a meta-grammars.

The current version of `mgcomp` has been used to compile a wide coverage Meta-Grammar FRMG (version 2.0.1) to get a grammar of around 200 TAG trees [47]. Without the use of guarded nodes and regular operators, this grammar would have more than several thousand trees and would be almost intractable. FRMG has been packaged and is freely available.

To ease the design of meta-grammars, a set of tools have been implemented, mostly by Éric de La Clergerie, and collected in MgTools (version **2.2.2**). This package includes a converter from a compact format to a XML pivot format, an Emacs mode for the compact and XML formats, a graphical viewer interacting with Emacs and XSLT stylesheets to derive HTML views. A new version is under development to provide an even more compact syntax and some checking mechanisms to avoid frequent typo errors.

The various tools on Metagrammars are available on INRIA GForge.

5.4. The Bonzai PCFG-LA parser

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

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

Alpage has developped as support of the research papers [70], [63],[3], [9] a statistical parser for French, named Bonzai, trained on the French Treebank. This parser provides both a phrase structure and a projective dependency structure specified in [17] as output. This parser operates sequentially: (1) it first outputs a phrase structure analysis of sentences reusing the Berkeley implementation of a PCFG-LA trained on French by Alpage (2) it applies on the resulting phrase structure trees a process of conversion to dependency parses using a combination of heuristics and classifiers trained on the French treebank. The parser currently outputs several well known formats such as Penn treebank phrase structure trees, Xerox like triples and CONLL-like format for dependencies. The parsers also comes with basic preprocessing facilities allowing to perform elementary sentence segmentation and word tokenisation, allowing in theory to process unrestricted text. However it is believed to perform better on newspaper-like text. The parser is available under a GPL license.

5.5. Alpage's linguistic workbench, including SxPipe

Participants: Benoît Sagot [correspondant], Rosa Stern, Pierre Boullier, Éric Villemonte de La Clergerie.

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

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

SxPipe, now in version 2 [93] is a modular and customizable chain aimed to apply to raw corpora a cascade of surface processing steps. It is used

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

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

5.6. MElt

Participants: Pascal Denis [correspondant], Benoît Sagot.

MElt is a part-of-speech tagger, trained for French (on the French TreeBank and coupled with the Le*fff*) and English [73]. It is state-of-the-art for French. It is distributed freely as a part of the Alpage linguistic workbench.

5.7. The syntactic lexicon Lefff and the Alexina framework

Participants: Benoît Sagot [correspondant], Laurence Danlos.

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

Alexina is Alpage's Alexina framework for the acquisition and modeling of morphological and syntactic lexical information. The first and most advanced lexical resource developed in this framework is the Le*fff*, a morphological and syntactic lexicon for French.

Historically, the Le*fff* 1 was a freely available French morphological lexicon for verbs that has been automatically extracted from a very large corpus. Since version 2, the Le*fff* covers all grammatical categories (not just verbs) and includes syntactic information (such as subcategorization frames); Alpage's tools, including Alpage's parsers, rely on the Le*fff*. The version 3 of the Le*fff*, which has been released in 2008, improves the linguistic relevance and the interoperability with other lexical models.

Other Alexina lexicons are under development, in particular for Spanish (the Le*ff*e), Polish, Slovak, English, Galician, Persian (PerLex, see 6.13), Kurdish.

5.8. System EasyRef

Participant: Éric Villemonte de La Clergerie [maintainer].

PASSAGE action

A collaborative WEB service EASYREF has been developed, in the context of ANR action Passage, to handle syntactically annotated corpora. EASYREF may be used to view annotated corpus, in both EASY or PASSAGE formats. The annotations may be created and modified. Bug reports may be emitted. The annotations may be imported and exported. The system provides standard user right management. The interface has been designed with the objectives to be intuitive and to speed edition.

EASYREF relies on an Model View Controller design, implemented with the Perl Catalyst framework. It exploits WEB 2.0 technologies (i.e. AJAX and JavaScript).

Version 2 has been used by ELDA and LIMSI to annotate a new corpus of several thousands words for PASSAGE.

A preliminary version 3 has been developed by François Guérin and revised by Éric de La Clergerie, relying on Berkeley DB XML to handle very large annotated corpora and to provide a complete query language expanded as XQuery expressions. EASYREF is maintained under INRIA GForge.

6. New Results

6.1. Designing efficient parsers using Meta-Grammars and DyALog

Participant: Éric Villemonte de La Clergerie.

Glossary

MG *Meta-Grammars*

Éric de La Clergerie has continued to improve the coverage, quality and efficiency of the French meta-grammar FRMG. This work is made progressively easier, using the various testing scripts developed in 2009 (confusion matrices for instance). A small but very practical shell (FRMG_SHELL) has been written to ease the use of FRMG, in particular for testing sentences.

Preliminary experiments have been tried to integrate, in the post-parsing disambiguation phase, some quantitative information provided by the knowledge acquisition experiments (cf. 6.3) about for instance the strength of a dependency through some preposition between a governor noun and a governee noun. Weights for several kinds of dependencies have been used, exploiting their frequency of a large corpus but also the semantic proximity between the governors or between the governees, as computed by the knowledge acquisition experiments. However, while real, the benefits of these weights for a better disambiguation remain very small. Further experiments have to be tried to investigate the situation, with maybe a better calibration of the imported weights wrt the various weights manually attached to the disambiguation rules.

6.2. Large scale corpus processing

Participant: Éric Villemonte de La Clergerie.

In the context of the PASSAGE and SCRIBO actions, we have continued the explore the use of distributed computing for processing of very large corpora, largely using GRID 5000 and a local cluster at INRIA Rocquencourt. For the SCRIBO action (cf. 8.1.1), GRID 5000 was in particular used to parse with FRMG a 700 million words corpus (including Wikipedia, Wikisource, and 30 months of AFP news). Several problems were again identified and corrected to get more and more robust treatments.

Distributed computing is also used on local clusters to collect and count patterns in the output of FRMG, using an adapted version of the *map-reduce* algorithm. These patterns with counts are then used as input for knowledge acquisition (cf. 6.2).

6.3. Knowledge acquisition and ontologies

Participants: Éric Villemonte de La Clergerie, Benoît Sagot, Rosa Stern, Kata Gábor, Tim van de Cruys, Yayoi Nakamura-Delloye, Marianna Apidianaki.

The parsing results provided by FRMG over a very large corpus covering various style and domains (30 months of AFP news [2007, 2009, and 2010], wikipedia, wikisource, Est Republicain, EuroParlement, JRC) have been used as input for the acquisition of semantic knowledge. This work was part of the SCRIBO action (cf. 8.1.1), in order to explore the techniques and methodologies that may be used to enrich a lexicalized seed ontology in a supervised way, i.e. using some human validation at some point.

The guiding principle for most of the experiments that were tried is the Harris distributional hypothesis, stating that semantically close words occur in similar contexts, syntactic contexts in our case.

The experiments may be classified around 2 main directions, those trying to identify and organize concepts, and those trying to identify and organize classes of "events", an event class being (often) used to establish some relationship between named entities (for instance, to denote that "someone" becomes president of "something"). These event class may properly denote events, but more generally, may denote the attachment of a property to an entity.

6.3.1. Terminology extraction

A first experiment about concepts, done by Éric de La Clergerie, dealt with terminology extraction, using frequency information to select sequences of Passage chunks that follow some patterns (essentially, nominal phrases possibly with adjectives, completed by prepositional phrases). The selection of the candidate terms was refined using mutual information and a notion of autonomy (a candidate term should not always occur within prepositional phrases or been modified by prepositional phrases). A preliminary detection of variants was used to group close terms (at the form level) and to discard some artifacts that have many variants (such as some date or interval constructions). Two ranked lists of terms were extracted on the AFP corpus and on

the whole corpus. The AFP list was communicated to *Lingua et Machina*, an INRIA startup, to be loaded in their LIBELLEX platform, in order to start some human validation. This experiment was the starting point for a collaboration with this company about the managing of lexical resources.

At this stage, this terminology extraction process is not as sophisticated than some existing software such as Acabit and should only seen as a starting point. More element of information could be used such a derivationnal morphology (to group variants) and coordinations (which were discarded). The exploitation of semantic similarity between words (as provided by the other experiments conducted below) could also be used to group variants.

6.3.2. Building word network

Another experiment led by Éric de La Clergerie was to regroup words by semantic similarity, exploiting as much as possible the Harris distributional hypothesis. Having collected the frequencies of dependencies triples (governor, relation, governee) on lemmas, such as (*assoir_v*, *sur_prep*, *chaise_nc*), a graph was built connecting the words (say *assoir*) with their syntactic context (say *sur_chaise*), the weight of an edge being derived from the frequency of the underlying dependency triple. Using ideas from the Markow CLustering (MCL) algorithm, an iterative algorithm was designed and implemented in Perl to identify pairs of words (and dually pairs of contexts) that are related by a dense set of relatively short paths. One of the interest of the algorithm is to provide a set of weighted contexts explaining the (semantic) proximity of two words. Another particularity of the algorithm is the strong duality between words and contexts, the algorithm being able to group words but also contexts. It also takes into account the fact that the syntactic contexts have an internal structure, including in particular a word (if we assume that X and Y are close, then given some relation r, the contexts r_X and r_Y are also close, and conversely).

Several extensions of the algorithm have been implemented. Random Indexing (RI) was for instance used to complement MCL, with the intuition that RI provides a way to quickly compare the similarity of large context sets, when MCL focus on the specific contexts that tend to group two words (and may miss some global differences). The presence of coordinations between two words was concretized by a bonus used to strengthen their proximity. Similarly, proximity was also strengthened by morphological or spelling closeness. Finally, a bonus may also be added for pairs of similar words found in a external wordnet-like seed resource (we tried with the French WordNet and with Wolf, an acquired French wordnet, but the usefulness of such resources was finally not obvious).

Running 10 to 20 iterations on the dependencies extracted from the whole 700 million parsed corpus (around 2 days, using around Go RAM), we get information about around 22000 words for around 44000 pairs of words. For each pair of words is specified how close they are (in a non symmetrical way) and which syntactic contexts are the most specific to explain the their proximity.

The amount of data makes it difficult to evaluate the quality of these pairs. Several evaluation experiments were tried by comparing with wordnet-like resources (French wordnet and WOLF) but the results remain difficult to interpret. Another option was to visualize and browse the pairs as a connecting graph, using TULIP, a software for large graphs. Browsing the graph, while hiding many important pieces of information (the degree of similarity, the explaining contexts) was nevertheless very useful to assess the interest of such a word network (and of the underlying clustering algorithm). Browsing the graph was also useful to point out some characteristic topological structures, for instance bush-like structures generally characterizes strong clusters (such as all month names).

These topological properties should guide further developments to get a better detection of the strongest word clusters, but also to get a better detection of polysemous words. An important evolution would also to shift from word to terms (resulting from the previous extraction).

6.3.3. Tensor based clustering

Starting from the same material (namely the dependencies with their frequencies), Tim van de Cruys has tried other techniques for word clustering. More precisely, he used a matrix-based approach using tensors and their non-negative factorization to identify the dimensions that cluster words. The base clusters were then used to

build an hierarchical clustering. An advantage of the tensors over matrices is also the possibility to add more axes, for instance to cluster words along the syntactic axe given by the dependencies but also along a thematic axe given by the word occurring in a document or in the keyword field of a document (such as an AFP news).

6.3.4. Cluster labeling

A second step in Tim van de Cruys' work was to establish ontological relationships between concepts, for instance to establish the hyperonomy (IS-A) relation between two concepts. This work was partially achieved by being able to assign labels to word clusters. A word cluster of synonymous words may be seen as denoting a concept and the label may be seen as the hyperonym concept. Cluster labeling was conducted by extracting and analyzing the syntactic content of the first meaningful sentence of Wikipedia pages. The first sentence of the page with title X usually mentions the *genus* of X. This genus found (with variations) for several elements of a cluster may be used as a label. It is even often possible to assign a term (rather than a single word). In practice, the labels are not always assigned to cluster of synonymous words but rather clusters of related words, such as *navire de guerre* (*warship*), *élément chimique* (*chemical element*) or related named entities such as *chanteur français* (*French singer*).

6.3.5. Acquisition of event structures

Kata Gábor worked on the acquisition of event classes (or event structures) characterized by verb classes or verb nominalizations. Semantic verb classes (Gross 1975, Levin 1993, Kipper-Schuler 2005) generalise over a set of syntactic and semantic properties: they participate in the same syntactic alternations and share one or more meaning components. Such classifications can be useful in a variety of NLP tasks, including semantic role labeling (Swier & Stevenson 2004) and information extraction. Semantic classes can be obtained automatically from corpora (Schulte im Walde 2006, O Seaghdha & Copestake 2008, Sun & Korhonen 2009, Messiant et al. 2010), in compliance with the distributional hypothesis which states that semantically related words tend to occur in similar contexts.

Kata Gábor applied an unsupervised learning method to obtain verb clusters from corpora, using a set of syntactic and semantic features (complement structure and the semantic profile of the arguments which fill the complement positions).

To overcome the sparse data problem, we expanded the verbal data by distributional information about corresponding deverbal nouns. In order to do so, she developed another algorithm to detect event nominalisations and to map nominal complement structure to that of finite verbs. The detection of event nominalisations is based on three measures:

1. distributional similarity;
2. morphological similarity;
3. the so-called 'event indicator' score which shows how likely it is that the noun refers to an event.

The intuition behind the notion of event indicator is that event nominalisations are characterized by a high proportion of occurrences in verb-like dependent positions (where a clause could also occur). The event indicator is obtained from corpus data by quantifying the proportion of syntactic heads having the noun as a dependent in a position where infinitives are also accepted (e.g. "*accepter de restreindre*" - "*accepter la restriction*").

6.3.6. Dependency paths between named entities

Yayoi Nakamura-Delloye worked on the identification of dependency-based extraction patterns denoting relations between named entities, for instance the membership relation between a person and an organization. The idea was to explore the regularities in the SCRIBO corpus, parsed with FRMG. Two extraction methods have been tried, a semi-supervised one and a fully non-supervised one. The recurring syntactic dependency paths are then abstracted into patterns for a given relation, the motivation being to use the patterns to extract new occurrences of the relation in other corpora during information extraction task.

6.3.7. Word-sense disambiguation and integration

Marianna Apidianaki joined lately the Alpage team in September 2010. Since then, she has been working on the automatic acquisition of lexical semantic knowledge from text corpora for the development of resources in the SCRIBO project. More precisely, she has pursued her previous research on exploiting parallel corpora for unsupervised word sense induction and disambiguation for enriching the ontology used in the SCRIBO project.

Furthermore, she worked on the integration into existing lexical semantic resources of the results of monolingual semantic analysis methods (based on clustering techniques) carried out by the other members of ALPAGE involved in SCRIBO.

6.3.8. Validation methodologies

The above mentioned experiments return ranked lists of candidates (terms, word pairs, clusters, verb class, verb/noun pairs, ...) that need to be validated as some point before being used to enrich some reference lexical resource. More precisely, in the context of SCRIBO, the idea is that a candidate should be phrased as an enrichment of the seed ontology (for instance the addition of a new concept through an IS-A ontological relation with an existing concept). To ease the validation, a ticked should be attached to each candidate, describing the scope of the candidate, providing explaining features (such as the syntactic contexts used to regroup a pair a words) and illustrative sentences (from the SCRIBO corpus). This notion of tickets has been formalized and the acquisition algorithm are or will be modified to return such tickets. Already, the notion of tickets has guided the development of prototype Web validation interfaces (for terms, cluster labels, and noun/verb pairs).

6.4. Automatic construction of distributional thesauri for French

Participants: Enrique Henestroza Anguiano, Pascal Denis, Marie Candito.

This work involves the automatic construction of distributional thesauri for French, with an eye toward use in statistical parsing. The distributional hypothesis states that words occurring in the same contexts tend to have similar meanings, as posited by Harris (1954). Additionally, distributional similarity based on syntactic contexts naturally indicates shared selectional preference, which may be useful for statistical parsing. Finally, distributional lexical resources are appealing because they can be constructed automatically from raw text corpora using unsupervised approaches, avoiding the problem of limited lexical coverage found in hand-built resources.

6.4.1. Thesaurus creation and evaluation

Following primarily from the distributional similarity work of Lin (1998) and Curran (2004), we use a raw corpus of text to extract context relations consisting of a primary word, a relation, and a secondary word. We chose to use the L'Est Républicain corpus, a 125 million word journalistic corpus, freely available at CNRTL (http://www.cnrtl.fr/corpus/estrepublicain). Extracted context relation frequency counts are weighted using a function such as relative frequency or t-test, then similarities between pairs of primary words are calculated using a measure function such as Jaccard or cosine. Testing different combinations of weight and measure functions, we evaluated each resulting thesaurus using average cosine similarity against synsets from the WOLF and the French EuroWordNet, two existing wordnet resources for French. Our results indicate that the best approach for constructing distributional thesauri uses both linear bigram and syntactic dependency context relations, the t-test weight function, and the jaccard similarity function.

6.4.2. Thesaurus and software availability

A major motivation behind our work is to make freely available a wide-coverage distributional thesaurus for French, as well as software for the construction of distributional thesauri using different corpora, settings, or languages. This effort has been named FreDist, with an associated project webpage (http://alpage.inria.fr/~henestro/fredist.html). The initial release includes a distributional thesaurus for French, and software is currently being consolidated and packaged for inclusion in the next release. The project webpage also includes a technical report detailing this work.

6.5. Improving the lexical coverage of statistical parsers

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

Probabilistic parsers are trained on treebanks, namely syntactically annotated sentences, and this training allows to capture syntactic regularities. Yet, though lexical information is known to play a crucial role in determining the syntactic structure of a sentence, many lexical phenomena cannot be learned simply by training on a treebank of a few thousands of sentences (the French treebank we use contains about 12000 sentences). First because treebanks cover only a small part of the French vocabulary. Second, because lexical data is very sparse : a corpus contains a few very frequent words, and a lot of rare words. Compared to English, this is even truer for French, or more generally inflected languages : morphological marks for gender, number, tense etc... drastically augment the vocabulary size.

To cope with this inherent limitation of statistical parsing techniques, we have investigated the use of word clusters instead of words as input to the parser. Our work was inspired by [Koo et al. 08] , who have shown that word clusters obtained with unsupervised techniques could improve statistical dependency parsing, when used as features for classifiers determining the weights of dependency arcs. We tried in 2009 to use word clusters within the framework of generative statistical parsing [Candito and Crabbé, 09]. We continued this investigation en 2010: we tested the impact on parsing performance of two morphological clustering techniques on both lexicalized models [38] and PCFG-LA models [19]: the "desinflection" process proposed in 2009 (a lexicon-based morphological clustering), and "lemmatisation"[4], which is the more classic technique that groups inflected word forms into part-of-speech+lemma pairs. Though with this second technique the oracle obtained with gold part-of-speech and gold lemmas is higher, the results obtained in a realistic setting (predicted part-of-speech and lemmas) are comparable to the rougher desinflection method. We have also analyzed the improvement in performance for both techniques with respect to word frequency. We found that replacing word forms with clusters improves attachment performance for words that are originally either unknown or low-frequency, since these words are replaced by cluster symbols that tend to have higher frequencies. Furthermore, clustering also helps significantly for medium to high frequency words, suggesting that training on word clusters leads to better probability estimates for these words.
It shall be noted that this work serves as a basis of a grant proposal to the FP7 ICT-Call, named *parse4real*, jointly made by Dublin City University, Alpage, Uppsala University and others European leaders in parsing morphologically-rich languages on which augmenting the lexical coverage of statistical parsers if of crucial importance. Djamé Seddah is Alpage's leader for that proposal, to which Marie Candito, Éric de La Clergerie and Benoît Sagot also participate.

6.6. Dependency parsing

Participants: Marie Candito, Benoît Crabbé, Pascal Denis, Enrique Henestroza Anguiano.

Dependency trees are often preferred to syntagmatic trees for many NLP tasks, such as information extraction, question answering, lexical acquisition. We started in 2008, and continued in 2009, to work on the conversion of the syntagmatic trees of the French treebank into surface dependency trees. We have now a stabilized version of a dependency treebank : the French treebank converted to dependencies [17].

The constituent-to-dependency conversion procedure can also be applied to syntagmatic trees as output by a parser trained on the syntagmatic treebank. Hence, we have various ways to obtain a parser outputting dependency trees : (i) training a parser on syntagmatic trees, and converting the output of this parser into dependencies. And (ii) directly using existing algorithms to train a dependency parser on the treebank converted to dependencies. We have performed a comparison of the two approaches (i) and (ii). Approach (i) is tested with an architecture where a parser is trained on the French treebank (using Petrov's algorithm), and output trees from this parser are converted to dependencies. Approach (ii) is tested with two dependency parser training algorithms : MST [McDonald and Pereira, 06] and MaltParser [Nivre et al., 06]. First bare results [17] showed that directly training a dependency parser with the MST algorithm [McDonald and Pereira, 06] outperforms the architecture base on Petrov's algorithm. We then performed a more sophisticated benchmark

[4] Conducted by a data driven lemmatizer [67] which we adapted for French during a visit at Saarbrucken in 2009.

between the three architectures (Petrov, MSTparser, MaltParser), with the integration of morphological features (lemmas, gender, number...) and unsupervised word clustering features (cf. results "Improving the lexical coverage of statistical parsers"). We found [18] that with these additional information, the differences between parsing architectures are generally small, and there is no consistent trend favoring either constituency-based or dependency-based methods. The best performance is achieved using MSTParser, enhanced with predicted part-of-speech tags, lemmas, morphological features, and unsupervised clusters of word forms. MaltParser achieves slightly lower labeled accuracy, but is probably the best option if speed is crucial

6.7. New results on Mildly-Context Sensitive formalisms

Participants: Benoît Sagot, Djamé Seddah.

Understanding the properties of formal languages more sophisticated than Context-Free Grammars has been a research topic for several Alpage members for a long time. Indeed, although probablistic models based on simple CFG backbones or TAG-based systems are quite successful for parsing tasks, they fail to model correctly some linguistic constructions that are complex but not necessarily rare. So-called Mildly Context-Sensitive formalisms are a particular class of grammars that have reasonable expressive power (strictly higher than TAGs, for example, but strictly less than Range Concatenation Grammars that cover PTIME) although they have a reasonable parsing complexity (polynomial). Studying these formalisms is therefore a very important research track.

Following up on work started back in 2006 [99], Djamé Seddah and Benoît Sagot have introduced a new extension of the so-called Multi-Component TAGS, and more precisely an extension of the MCTAGs with Local Shared Deriva- tion introduced in [100] which can handle non local elliptic coordinations. Based on a model for control verbs that makes use of so-called ghost trees, one can show that this extension leads to an analysis of argument cluster coordinations that provides an adequate derivation graph. This is made possible by an original interpretation of the MCTAG derivation tree mixing the views of [76] and [111].

MCTAGs and some of their variants are equivalent to a well-studied MCS formalism, that of Linear Context-Free Rewriting Systems or LCFRSs [105]. Having NLP applications and language modeling in mind, studying the formal properties of LCFRSs is relevant. In collaboration with Giorgio Satta, from the University of Padova (Italy), Benoît Sagot has finalized and presented at ACL 2010 an optimal algorithm for reducing the rank of a sub-class of LCFRSs (namely those with fan-out 2) [35].

6.8. Temporal information processing

Participants: André Bittar, Pascal Denis, Philippe Muller.

An important task in temporal processing is to recover the chronology of the events that are described in a text. Most recent work has focused on learning temporal relations (e.g., precedence, inclusion) between given events in a text, while rarely ensuring that these separate pieces of information remain consistent. The target representations also differ, according to the distinctions they make or the inferences they allow.

In [22], we investigate the impact of using different temporal algebras for learning temporal relations between events. Specifically, we compare three interval-based algebras: Allen (1983) algebra, Bruce (1972) algebra, and the algebra derived from the TempEval-07 campaign. These algebras encode different granularities of relations and have different inferential properties. They in turn behave differently when used to enforce global consistency constraints on the building of a temporal representation. Through various experiments conducted on the English TimeBank/AQUAINT corpus, we show that although TempEval smaller relation set leads to the best classification accuracy performance, it is too vague to be used for enforcing consistency. By contrast, the other two relation sets are similarly harder to learn, but more useful when global consistency is important. Overall, the Bruce algebra is shown to give the best compromise between learnability and expressive power. In forthcoming work, we consider the use of yet another temporal algebra, namely point algebra, in the context of a global inference problem to address the problem of predicting temporal structure for texts.

This first work has been performed on English, for until recently there was no available TimeBank for French. One of the contributions of André Bittar's PhD dissertation, supervised by Laurence Danlos and co-supervised by Pascal Denis that was successfully defended this year, was to develop such a resource and to make it publicly available to the community. The current version of the French TimeBank consists of 109 journalistic texts (16,208 tokens) from 7 different sub-genres which have been annotated according to the ISO-TimeML standard. Bittar's dissertation also suggests a number of improvements to the ISO-TimeML schema in order to account for linguistic phenomena which apply across languages, as well as adaptations necessary for the processing of French texts. The French TimeBank is soon to be made available (via the INRIA Forge) to the wider scientific community and will provide a useful basis for studying the linguistic expression of temporal phenomena in French, as well as providing data for evaluating automatic temporal annotation systems.

6.9. Discourse processing

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

Discourse interpretation is often automated in two steps: a *segmentation* step where discourse units (DU) are extracted, and a *parsing* step where these DUs are related to derive a discourse structure. The relations are labelled with functions reflecting the underlying intention of the producer of the discourse.

Within that perspective, one can proceed bottom-up, and isolate the elementary discourse units (EDU) before building a structure made of groupings of such EDUs. Alternatively, a text can be segmented in a top-down fashion, separating topically coherent parts (this is called *text segmentation* or *text tiling*) before applying the more fine-grained approach mentioned before. Ideally the two methods complement each other, while this is rarely done in practice.

6.9.1. Discourse Unit Segmentation

A contribution to the bottom-up approach was made in [13], with an additional constraint on the EDUs. Previous research on discourse segmentation have relied on the assumption that elementary discourse units (EDUs) in a document always form a linear sequence. Unfortunately, this assumption turns out to be too strong, for some theories of discourse allow for nested discourse units. To address this problem, we developed a system using standard multi-class classification techniques making use of a regularized maximum entropy model, combined with a simple repairing heuristic that enforces global coherence. Our system was developed and evaluated on the first round of annotations provided by the French Annodis project (an ongoing effort to create a discourse bank for French) with an encouraging performance (an F-score of 73% for finding EDUs) on a small set of about 50 documents.

Another approach which is currently explored consists in examining whether discourse unit segmentation can be automatically obtained from the output of a deep syntactic analyzer, i.e. a parser which produces analyses closed to semantic dependency trees, like the parser FRMG developed by E. de La Clergerie in the formalism of Tree Adjoining Grammar [106]. This approach requires linguistic insights before being tested on real corpora.

6.9.2. Text Segmentation

From the other end of the scale, we have tried to contribute to a better definition of the requirements of the text segmentation (TS) task [12], by stressing the need for taking into account the types of texts that can be appropriately considered. Our hypothesis is that while TS is indeed relevant to analyze texts with a thematic organization, this task is ill-fitted to deal with other modes of text organization (temporal, rhetorical, etc.). By comparing the performance of a TS system on two corpora, with either a "strong" or a "weak" thematic organization, we show that TS is sensitive to text types.

6.9.3. Lexical Resource for Discourse Processing

The intentions of the producer of a discourse are often expressed by discourse connectives (or discourse markers): they explicitly signal that a relation is holding between two discourse units. We have focused on these linguistic cues, and have manually built a lexicon of discourse connectives for French, named LEXCONN

and described in [31]. The lexicon contains 330 connectives, collected with their syntactic category (conjunction, adverbial, preposition) and the discourse relation(s) they express. We are now improving LEXCONN by adding some information about the position(s) than can be occupied by the markers in a syntactic clause. Indeed, some discourse markers (especially adverbial ones) are ambiguous: they don't have a discourse function in all the position(s) they can occur.

This lexicon is used in various tasks: it provides automatically annotated data for discourse parsing, and permits to extract specific discourse structures for linguistic empirical studies.

6.9.4. Determining Equivalent Discourse Structures

Two tasks are crucial for the development of discourse parsing systems: the creation of gold-standard annotations, and the evaluation of annotations produced by discourse parsers. These two tasks require the development of metrics for accurately comparing distinct annotations of the same text(s); this involves in particular determining the conditions under which discourse structures are equivalent. While discourse theories provide some indication as to how to compare different discourse structures for the same text, they haven't studied this issue to its full extent. We try to investigate this question by studying discourse structures extracted using LEXCONN. In our study, we examine the possibility of deducing an annotation from another, assuming that an annotation can contain implicit information.

Given distinct discourse annotations of the same text, composed by relations between text segmenst, our goal is to calculate, using deduction rules about relations, the *discourse closure* of these annotations. In effect, this allows us to compare these annotations by considering all the implicit information (relations) they contain. Our overall goal is to build a discourse relations algebra, as it has been done by Allen (1893) for temporal relations. At least two types of rules seem necessary. Considering a discourse containing three adjacent segments α, β, and γ, either we know the relations $R_x(\alpha, \beta)$ and $R_y(\beta, \gamma)$, and try to deduce the discourse relation R_z holding between α and γ; or we know the relations $R_x(\alpha, \beta)$ and $R_z(\alpha, \gamma)$, and we try to deduce the discourse relation R_y holding between β and γ. These rules rely on theoretical relation's description and empirical data, gathered with our lexicon of connectives.

6.10. "Wrong" strong punctuation signs

Participants: Laurence Danlos, Benoît Sagot.

Some strong punctuation signs are "wrongly" used instead of weak punctuation signs, leading to graphic sentences which are not grammatical sentences, see the following discourse which includes five graphic sentences only made up of adverbial phrases.

> On avait donné dans le Nord un grand coup de pied dans la fourmilière, et les fourmis s'en allaient. Laborieusement. Sans panique. Sans espoir. Sans désespoir. Comme par devoir. [Saint Exupéry, Pilote de guerre]

In [20], we present a corpus study of this phenomenon — sometimes called "épexégèse" — and a tool in the early stages to automatically detect wrong strong punctuation signs. The goal of this tool is to automatically categorize these punctuation signs as weak, so that a parser can make a standard syntactic and semantic analysis.

6.11. Lexical incompleteness: typology and exploration of unknown words

Participants: Benoît Sagot, Rosa Stern, Gaëlle Recourcé.

In an attempt to cope with lexical incompleteness within the EDyLex project (8.2.2), a typology of unknown words (i.e. words, forms, tokens) has been proposed. This typology reflects both linguistic structures and operational issues regarding unknown words. It is indeed intended to guide automatic processing modules in the handling of out-of-lexicon words . Unknown words are defined relatively to the kind of unknown tokens which compose them. Those can be either productive sequences such as dates and recognizable for example by a local grammar, productive lexical creations (*red-hair*), lexicalized forms (*cupboard*), proper names, borrowings, dependant components (*priori*) or errors (spelling, misprint).

In order to verify and improve this typology, a corpus annotation of unknown words based on it has been carried out in two phases. First the corpus, consisting of xxx news items from Agence France Presse in French, English and Spanish, have been automatically annotated with the Alpage surface processing chain SxPipe, able to recognize a series of out-of-lexicon sequences and forms (dates, URLs, numerical values, proper names, some prefixed compositions...); this automatic annotation then marked remaining out-of-lexicon forms, based on the lexicons available in the three languages within the Alexina framework (Lefff, Enlex and Leffe). In the second phase the latter forms have been reviewed by human annotators who assigned to each of them the appropriate category from the typology. A high inter-annotators agreement showed the relevance of this typology and can therefore help the building of a complete and modular processing chain fo lexical resources enrichment.

6.12. Named entities recognition and resolution: a modular system and its resources

Participants: Benoît Sagot, Rosa Stern.

Within the Alpage surface processing chain SxPipe, a series of modules based on local grammars enable the analysis of sequences at the token- and form-level such as dates, URLs, addresses... A new version of NP, the named entity recognition tool, has been integrated to SxPipe in order to cope with named entities of different types (person, location, organization and company names). NP's local grammar consists of 130 rules, including contextual patterns and triggers, for detection and typing of named entities. In order for those rules to operate efficiently, a lexicon of proper nouns can be associated to it. The lexicon consists in a list of forms acceptable by the grammar as named entities variants and can be obtained by several means.

In this prospect, a database called Aleda has been built, based on two main external and freely available resources: Wikipedia and Geonames. Aleda intends to gather information and knowledge on entities and not only a list of entity names. This knowledge (precise type of the entity (for example capital, country, museum are subtypes of location), decomposition and variation of its name (title, last name, pseudonym, short form...), main sector of organizations and companies...) is reflected in its structure by dedicated typed fields. The base can then be converted into a lexicon, linking each entity to its variants with a unique identifier. This structure allows for the ulterior use of information contained in Aleda by the client application.

In order to make named entities recognition more complete and usable as an information extraction module, the denotational aspect of named entities must be handled, i.e. the entity to which a recognized name refer to must be identified. This process of *entity resolution* can be based on an entity base such as Aleda. After the recognition phase, the result of which being a set of entity mentions in text, the resolution phase intends to link each mention to the adequate entity in the reference base. This operation must deal with possible ambiguities among homonyms or names refering to entities absent from the base. Disambiguation can benefit from the knowledge gathered in Aleda, for example by comparing information associated to entities and named entities contexts in the text.

6.13. Developing language resources for Persian and Kurdish languages

Participant: Benoît Sagot.

In 2010, Alpage has taken part actively in the PerGram French-German project (co-funded by ANR and DFG, see 8.3.2). In particular, the development of an Alexina description of the morphology of Persian and the development of an associated Persian Alexina lexicon, named PerLex, has been pursued following a first preliminary version in end-2009 [50]. This work, in collaboration with Géraldine Walther (from LLF, Université Paris 7) and Pollet Samvelian, Pegah Fagiri and Ariel Gutman (MII, CNRS and Université Paris 3), has also lead to adapting the shallow processing chain SxPipe to Persian [37]. Ongoing work has started for the manual validation of PerLex and for its extension to the syntactic level (sub-categorization frames, entries for "complex predicates"...).

Benoît Sagot and Géraldine Walther have then pursued their efforts for resourcing Western Iranian languages. First, they formalized a methodology for developing a morphological lexicon for a language lacking any resource, and applied its first steps to Sorani Kurdish, hence creating a preliminary small-scale Alexina lexicon for this language [50]. Second, they performed experiments on later stages of this methodology, but in Kurmanji Kurdish [49], which is another Kurdish variant for which lexical resources do exist, although not formalized. Their work include preliminary experiments on developing a POS tagger using a lexicon but no annotated training corpus. These experiments will be pursued in 2011, and both Kurdish lexicons should reach a significant size.

6.14. Word ordering

Participants: Juliette Thuilier, Benoît Crabbé.

We study the problem of choice in the ordering of French words building upon two case studies. Both studies try to identify the factors that come into play when one has two choose among several possible orderings : the first inquiry is dedicated to the position of attributive adjectives wrt the noun. The second is dedicated to the relative order of postverbal dependants. Both questions have almost never been adressed for French.

In collaboration with Gwen Fox (Université Paris 3), the first investigation in this direction has been led towards identifying the importance of constraints that drive the placement of attributive adjectives wrt the noun in the noun phrase in French. This study brings an additional element to Bresnan's thesis, according to which the syntactic competence of human beings is indeed probabilistic. This year we enhanced our previous statistical models with a qualitative study trying to shed light (1) on semantic effects and (2) on word independant versus word specific constraints.

Another study on preferences in verbal complementation has started this year. We are planning to study the preferences between postverbal dependants (direct objects and indirect objects). This year we started by extracting data suitable for statistical modeling mainly from the French Treebank and started to annotate the treebank with missing information (of semantic nature). The expected outcome from this work is first to identify factors that are relevant in French and beyond in a crosslinguistic perspective we plan to compare the constraints observed in other languages such as german. This is intended to be realised in collaboration with psycholinguistic teams both in Paris 5 (LPNCog) and in Germany (Frias, Freiburg) and the Laboratoire de Linguistique Formelle (LLF) in Paris 7 also intends to set up experimental work in connection with our modeling results in this framework for French.

As can be seen from the outline above, this line of research brings us closer to cognitive sciences and more specifically to frameworks inspired by construction grammar. We hope in the very long run that these investigations will bring – among other – further insights on the design of probabilistic parsers. In NLP the framework that is closest to implementing construction grammar is Data Oriented Parsing.

6.15. Unsupervised acquisition of allophonic rules

Participants: Luc Boruta, Benoît Crabbé.

This is an exploratory work on modelling the acquisition of allophonic rules. It is made in collaboration with the Laboratoire de Sciences Cognitives de Paris, LSCP). It explores the acquisition of the phonological system made by children relying upon the assumption that word segmentation and phonological learning are dependant processes feeding each other. Hence this work splits into two subproblems, that of segmentation and that of phonological clustering.

State-of-the-art models of the acquisition of word segmentation have been evaluated using phonemically transcribed corpora. As such, they implicitly assume that children know how to undo phonetic variation when they learn to extract words from fluent speech. Moreover, whereas models of language acquisition should perform similarly across languages, evaluation is often limited to English samples. We first argue that online learning is a sound desideratum for any model of language acquisition and use this criterion to select candidate segmentation models. Then, using child-directed corpora of English, French and Japanese,

we evaluate the models' performance given inputs where phonetic variation has not been reduced. To do so, we propose a parametric benchmark where segmentation robustness can be measured across different levels of noise, simulating uniform errors in phoneme recognition or systematic allophonic variation. We show that statistical models do not resist noisy inputs and do not generalize to typologically different languages. From the perspective of early language acquisition, the results strengthen the hypothesis according to which phonological knowledge is acquired in large part before the construction of a lexicon.

It is recognized that infants learn phonemes through some kind of unsupervized clustering of the speech signal. Previous work has shown the feasibility of unsupervized clustering, using as input manually segmented phonetic parameters [103]. [104] showed that an HMM state-splitting algorithm run on conversational speech coded using standard MFCC coefcients automatically grows a network of HMM states which successfully encoded speech sounds with no loss of informaton compared to supervized HMM training. However the obtained states did not map one-to-one to phonemes: phoneme-size strings of states did not yield abstract phonemes, but rather, context dependant allophones. This problem is serious enough to impede subsequent unsupervized learning of words [16]. To adress this issue, we will recluster allophonic variants using higher order information in additon to acoustic distance in order to improve the metric used in the clustering algorithm. Two such informaton have been shown to be helpful: distributional information about adjacent segments [87] and (pseudo)-lexical informaton obtained through approximatng a lexicon using frequent n-grams [80].

7. Contracts and Grants with Industry

7.1. Contracts with Industry

Alpage has developed several collaborations with industrial partners. Apart from grants described in the next section, specific collaboration agreements have been set up with Verbatim Analysis (license agreement, see section 4.3), Kwaga (ARITT contract, see section 4.4), TNS-Sofres (see section 4.6) and possibly soon Lingua et Machina (see section 4.5).

8. Other Grants and Activities

8.1. Regional Initiatives

8.1.1. Action Scribo (2007 – 2009, extended until 2010)

Participants: Éric Villemonte de La Clergerie, Kata Gábor, Marianna Apidianaki, Tim van de Cruys, Yayoi Nakamura-Delloye, Rosa Stern, Benoît Sagot.

Scribo Homepage: http://www.scribo.ws/xwiki/bin/view/Main/WebHome

Scribo aims at algorithms and collaborative free software for the automatic extraction of knowledge from texts and images, and for the semi-automatic annotation of digital documents. Scribo has a total budget of 4.3M Euros and is funded by the French "Pôle de compétivité" Systematic from Mid 2008 til end 2010. It brings 9 participants together: AFP, CEA LIST, INRIA, LRDE (Epita), Mandriva, Nuxeo, Proxem, Tagmatica and XWiki.

8.2. National Initiatives

8.2.1. ANR project Sequoia (2009 – 2011)

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

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

8.2.2. ANR project EDyLex (2010 – 2012)

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

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

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

8.2.3. ANR project Rhapsodie (2008 – 2010)

Participants: Sylvain Kahane, Éric Villemonte de La Clergerie, Marie Candito, Benoît Crabbé, Benoît Sagot.

Rhapsodie is an ANR project headed by Anne Lacheret (University Paris X). The aim of the project is to study the matching of prosody and syntax on a 30 hours corpus of spoken French by providing prosodic and syntactic annotations. Alpage participates to the project at two different levels: the specification of the transciption and syntactic annotation framework and the use of parsers for preparing the manually validated syntactic corpus annotation.

8.2.4. ANR project PASSAGE (2007 – mid 2010)

Participant: Éric Villemonte de La Clergerie.

PASSAGE Homepage: http://atoll.inria.fr/passage
EASy homepage: http://www.elda.org/easy

PASSAGE is an action in ANR MDCA program (*Masse de Données Connaissance Ambiantes*) started in 2007 and extended till mid 2010. The participants are Alpage (coordinator), LIR (LIMSI, Orsay), "Langue & Dialogue" (LORIA, Nancy), LI2CM (CEA-LIST), plus several contractors (ELDA, TAGMATICA and several providers of parsing systems).

PASSAGE stands for "*Large Scale Production of Syntactic Annotations to move forward*" . Its main objectives are to parse a large corpus (100 to 200 million words) with several parsers (around 10 systems), combine the results provided by these parsers and use the resulting annotations to acquire new linguistic knowledge (semantic classes, subcategorization frames, disambiguation probabilities, ...). A small part of the corpus (around 400000 words) will be manually validated to be used as a reference treebank. Two evaluation campaigns based on the work done during the Technolangue action EASy will be conducted during PASSAGE to assess the performances of the parsing systems. The annotations and derived linguistic resources will be made available.

8.3. European Initiatives

8.3.1. Galician government research project Victoria (2008 – 2010)

Participants: Éric Villemonte de La Clergerie, Benoît Sagot.

As a follow-up of a long lasting collaboration with Galician universities, ALPAGE, Éric de La Clergerie and Benoît Sagot are strongly involved as associate researchers in the Galician government research project Victoria on the development of Spanish and Galician linguistic resources by adapting tools, methods and resources developed by ALPAGE.

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

Participant: Benoît Sagot.

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

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

Participants: Benoît Sagot, Marianna Apidianaki.

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

8.4. International Initiatives

8.4.1. ISO subcommittee TC37 SC4 on "Language Resources Management"

Participant: Éric Villemonte de La Clergerie.

The participation of ALPAGE to French Technolangue action Normalangue has resulted in a strong implication in ISO subcommittee TC37 SC4 on "Language Resources Management" (http://www.tc37sc4.org/). Éric de La Clergerie has participated to ISO events and has played a role of expert (in particular on morpho-syntactic annotations [MAF], feature structures [FSR & new FSD], and syntactic annotations [SynAF]).

9. Dissemination

9.1. Animation of the scientific community

- Éric de La Clergerie is an elected substitute member of INRIA's "Conseil scientifique".
- Alpage is involved in the French journal T.A.L. (AERES linguistic rank: A). Éric de La Clergerie is "Rédacteur en chef" and was the editor of the regular issue 52/1 (2011). Laurence Danlos and Philippe Muller are members of the editorial board. Benoît Sagot was "Secrétaire de rédaction" of the journal until September 2010. He has been invited to be the guest editor, with Nuria Bel, for a special issue on Language Resources.

- Alpage is deeply involved in a forthcoming special issue of the major journal in our field of research, Computational Linguistics. Djamé Seddah is one of the guest editor of this issue devoted to “Parsing of morphologically-rich languages” while Marie Candito is a reviewer for this issue.
- Alpage members were involved in many Program, Scientific or Reviewing Committees for other journals and conferences, such as TALN 2010 (Laurence Danlos, Benoît Sagot, Éric de La Clergerie, Sylvain Kahane, Benoît Crabbé, Djamé Seddah, Philippe Muller), the 2nd CMLF (Laurence Danlos), ACL 2010 (parsing area, Éric de La Clergerie), ECAI 2010 (Éric de La Clergerie), TAG+10 (Éric de La Clergerie), STAIRS 2010 (Éric de La Clergerie), Computational Linguistics (Éric de La Clergerie), LRE (Éric de La Clergerie), CoLing 2010 (Pascal Denis, Éric de La Clergerie), NAACL-HLT 2010 (syntax and parsing area, Éric de La Clergerie) and EMNLP 2010 (Pascal Denis)
- Djamé Seddah and Benoît Sagot are elected board member of the French NLP society (ATALA); Djamé Seddah is Program Chair of the “Journées ATALA” (one day long workshops in NLP, 4 or 5 per year); Benoît Sagot is Deputy Secretary since September 2010.
- Laurence Danlos is a member of the Permanent Committee of the TALN conference organized by ATALA
- Éric de La Clergerie has reviewed a proposal for the French program ANR CONTINT.
- Djamé Seddah is one of the founders of the statistical parsing of morphologically rich language initiative that started during IWPT’09. He was the program co-chair of the successful SPMRL 2010 NAACL-HLT Workshop (2nd most successful workshop of this conference in terms of attendees) and will be as well for its next utterance (that will take place during IWPT’11). Alpage is deeply involved in this initiative (with Marie Candito part of its core members and Benoît Sagot member of its review committee).
- Marianna Apidianaki has undertaken the organization of an ACL Workshop on “Syntax, Semantics and Structure in Statistical Translation” together with Marine Carpuat (National Research Council, Canada), Lucia Specia (University of Wolverhampton, UK) and Prof. Dekai Wu (Hong Kong University).
- Benoît Sagot has started the organization of WoLer 2011, an ESSLLI 2011 workshop on Lexical Resources (http://alpage.inria.fr/~sagot/woler2011/), to be held in Ljubljana, Slovenia in August 2011.
- Marianna Apidianaki and Benoît Sagot are organizing a one-day ATALA workshop on the “Extraction of lexical, semantic and syntactic, information from multilingual corpora” that will take place in May 2011.
- Laurence Danlos, Benoît Sagot and Éric de La Clergerie co-organized with Laure Aït-Ali the Journée Iliatech on "Apport des technologies de la langue pour l’accès à l’information". Laurence Danlos presented Alpage’s work in a talk entitled "Au-delà de l’analyse syntaxique, analyse sémantique et discursive". Éric de La Clergerie has presented demonstrations of Alpage tools, and Benoît Sagot presented his work on *vera*.
- Laurence Danlos attented to the Journée Club Ina SUP (Cap-Digital), Paris, on "Moteur de recherche ; de l’accès à la maîtrise des contenus"
- Alpage is an active member in the LabEx proposal on Experimental Linguistics, headed by Jacqueline Vaissière (Univ. Paris 3) and supported by the PRES Paris-Cité. Benoît Sagot is in charge of one of the 6 scientific “strands”, the strand on Language Resources.

9.2. Participation to workshops, conferences, and invitations

+ Laurence Danlos, Charlotte Roze and Philippe Muller co-presented a paper at the Conference on Multidisciplinary Approaches to Discourse (MAD 2010) in Moissac, France, on "LEXCONN: a French Lexicon of Discourse Connectives"

+ Several Alpage members attended LREC 2010 at Valetta, Malta. Laurence Danlos presented a paper on "Learning recursive segments for discourse parsing" on behalf of herself, Pascal Denis and Philippe Muller and a fourth co-author. Marie Candito presented a poster on "Statistical French dependency parsing: treebank conversion and first results" on behalf of herself, Benoît Crabbé and Pascal Denis. Benoît Sagot presented a paper on "The Le*fff*, a freely available, accurate and large-coverage lexicon for French". Laurence Danlos, Benoît Sagot and Rosa Stern co-presented a poster on "A Lexicon of French Quotation Verbs for Automatic Quotation Extraction". Benoît Sagot co-presented a poster on "A morphological lexicon for the Persian language". Other communications were presented during LREC workshops, and in particular a poster on "Resources for Named Entity Recognition and Resolution in News Wires" (Rosa Stern and Benoît Sagot), a poster on "Developing a large-scale lexicon for a less-resourced language: general methodology and preliminary experiments on Sorani Kurdish" (Benoît Sagot and a co-author), "Creating and maintaining language resources: the main guidelines of the Victoria project" (Benoît Sagot and co-authors).
+ Several Alpage members attended the NAACL workshop on statistical parsing of morphologically rich languages (SPMRL), co-organised by Djamé Seddah. Marie Candito presented a paper on "Parsing word clusters" on behalf of herself and Djamé Seddah. Djamé Seddah presented a paper on "Lemmatization and Statistical Lexicalized Parsing of Morphologically-Rich Languages" on behalf of himself, Gregorz ChrupaŁa, Ozem Cetinoglu, Josef van Genabith and Marie Candito.
+ Laurence Danlos attended the 2ème Congrès Mondial de Linguistique Française, New Orleans, United States, where she presented a paper on "Analyse discursive des incises de citation" on behalf of herself, Benoît Sagot and Rosa Stern.
+ Several Alpage members attended the TALN conference in Montréal, Canada. Pascal Denis, on behalf of himself and Benoît Sagot, presented a paper on "Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français". Benoît Sagot and his co-author presented a paper on "Développement de ressources pour le persan: lexique morphologique et chaîne de traitements de surface". Laurence Danlos and Benoît Sagot co-presentation a poster on "Ponctuations fortes abusives". Benoît Sagot and Rosa Stern co-presented a poster on "Traitement des inconnus : une approche systématique de l'incomplétude lexicale" with other co-authors. Rosa Stern and Benoît Sagot co-presented a poster on "Détection et résolution d'entités nommées dans des dépêches d'agence". Éric de La Clergerie presented on behalf of himself and Yayoi Nakamura-Delloye a poster on "Exploitation de résultats d'analyse syntaxique pour extraction semi-supervisée des chemins de relations". Juliette Thuilier presented on behalf of herself, Benoît Crabbé and one co-author a paper on "Approche quantitative en syntaxe : l'exemple de l'alternance de position de l'adjectif épithète en français". Philippe Muller presented two communications with co-authors.
+ Pascal Denis presented two papers at the 23rd International Conference on Computational Linguistics (COLING 2010), held in Beijing, China. The first, full paper "Comparison of different algebras for inducing the temporal structure of texts" was presented on his and Philippe Muller's behalf. The second, poster paper "Benchmarking of statistical dependency parsers for French" was presented on behalf of Marie Candito, himself, Enrique Enestroza and a fourth co-author.
+ Laurence Danlos and Benoît Sagot attented the 28th Conference on Lexis and Grammar, Belgrade, Serbia. Laurence Danlos and Benoît Sagot co-presented a paper on "Les verbes de citation dans le Lexique-Grammaire", and Benoît Sagot and his co-authors presented a paper on "Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish".
+ Éric de La Clergerie and Djamé Seddah have participated (with presentations) to TAG+10.
+ Djamé Seddah, Marie Candito have presented their respective and joint works at the NAACL-HLT SPRML'2010 workshop, LREC'2010.
+ Éric de La Clergerie has presented results and demonstrations relative to the SCRIBO action at the SCRIBO-CoLab Barcamp (November 2010).
+ Éric de La Clergerie has presented results on the PASSAGE action at the Grand Colloque STIC (January 2010).

+ Participation of Éric de La Clergerie to ISO TC37SC4 meetings (Hong Kong, January 2010; Berlin, October 2010), to the related FlareNet meeting (Barcelona, February 2010) and during LREC 2010 (May 2010).
+ Laurence Danlos made a 5-day visit at the Université de Montréal, Canada, where she gave an invited talk on "Présentation du formalisme D-STAG"
+ Éric de La Clergerie has presented a tutorial on MAF (Morpho-syntactic Annotation Framework) at LREC 2010 (day on “Standards for Language Resources. Overview and Use”).
+ Éric de La Clergerie has presented his work on “Comment améliorer une chaîne de traitement syntaxique ?” at Institut Gaspard Monge (University Marne la Vallée, January 2010) and on “Concevoir, améliorer et exploiter une méta-grammaire factorisée du français” at Geneva University (November 2010).
+ Laurence Danlos gave an invited talk at the Université de Genève on "La traduction (automatique) des connecteurs de discours"
+ Benoît Sagot made a 2-day visit at the University of Vigo, Spain, where he gave an invited talk
+ Philippe Muller was invited to give a class, titled "Sémantique du discours et interprétation temporelle", at the "Atelier Jeunes Chercheurs des Journées Sémantique et Modélisation" at Nancy in march 2010
+ Pascal Denis was invited to give an opening class (tutorial), titled "Résolution automatique des anaphores", at the "Atelier Jeunes Chercheurs des Journées Sémantique et Modélisation" at Nancy in march 2010
+ Djamé Seddah made a 3 days visit at Dublin City University and presented his work on lemmatization of lexicalized parsing models for MRLs. During his stay, he prepared the writing of the parse4real FP7 project proposal.

9.3. Teaching

Alpage is in charge of the prestigious cursus of Computational Linguistics of Paris 7, historically the first cursus in France in this domain. This cursus, which starts in License 3 and includes a Master 2 (research) and a professional Master 2, is directed by Laurence Danlos. Marie Candito is in charge of the License 3, and Laurence Danlos is in charge of both Master 2. All faculty members of Alpage are strongly involved in this cursus, but some Inria members also participate in teaching and supervizing internships. Unless otherwise specified, all teaching done by Alpage members belong to this cursus. Teaching by associate members in other universities are not indicated.

Laurence Danlos (INRIA partial delegation): Introduction to NLP (3rd year of License, 28h); Discourse, NLU and NLG (2nd year of Master, 28h).

Marie Candito: Information retrieval (2nd year of professional Master, 12h); Clustering and Classification (2nd year of professional Master, 12h); Probabilistic methods for Natural language processing (1st year of Master, 48h); Machine translation (1st year of Master, 48h); Probabilities and statistics for Natural language processing (3rd year of Licence, 24h);

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

Benoît Crabbé (INRIA delegation): Probabilistic methods for NLP (1st year of Master, 48h); Introduction to programming II (3rd year of Licence, 24h).

Pascal Denis: Computational Semantics (2nd year of Master, 24h).

Charlotte Roze: Introduction to Programming (3rd year of License, 24h); Algorithmics (3rd year of License, 24h).

Juliette Thuilier: Introduction to Linguistics (1st year of License in “Lettres modernes”, TD, 24h);

François-Régis Chaumartin: Modélisation (UML) et bases de données (SQL) (2rd year of professional Master, 24h).

Djamé Seddah, as an Assistant Professor in CS in the University Paris 4 Sorbonne, member of the UFR ISHA, mainly teaches “Generic Programming and groupware”, “Distributed Application and Object Programming”, “Syntaxic tools and text Processing for NLP”, “Machine Translation Seminars” in both years of the Master “Ingénierie de la Langue pour la Gestion Intelligente de l’Information”. Djamé Seddah is also the “Directeur des études” of a CS transversal module for the Sorbonne’s undergraduate students (ie “Certificat Informatique et Internet”).

André Bittar is an ATER at Université Paris-Est Marne-la-Vallée, where he taught “Introduction to Operating Systems” (1st year of DUT, 52h), “Unix/HTML” (1st year of License, 48h) and “Programming with Python” (1st year of Master, 12h) during the first semester of the university year 2009-2010.

9.4. PhD committees

- Laurence Danlos was a reviewer for Eric Charton’s dissertation on "Génération de phrases multilingues par apprentissage automatique de modèles de phrases", (Université d’Avignon, LIA, Computer Science Department)
- Laurence Danlos was President of Mario Barcala’s PhD Committee (Computer Science Department, University of Vigo, Spain); the title of his dissertation is "Création de corpus et fouille de textes en Galicien"
- Laurence Danlos was a member of Anne-Laure Jousse’s PhD Committee (Université de Montréal, Department of Linguistics, and Université Paris Diderot, UFR of Linguistics); the title of her dissertation is "Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales"
- Laurence Danlos was in the PhD defense committees as PhD supervisor for André Bittar ("Building a TimeBank for French: A Reference Corpus Annotated According to the ISO-TimeML Standard", UFR de Linguistique de l’Université Paris Diderot) and SinWon Yoo ("Une grammaire TAG du Coréen", UFR de Linguistique de l’Université Paris Diderot). Philippe Muller was member and Pascal Denis invited member of the committee for the former. Benoît Crabbé was member of the committee for the latter.
- Benoît Sagot was a member of Lionel Nicolas’s PhD Committee (Université de Nice, Computer Science Department); the title of his dissertation is "Efficient production of linguistic resources: the Victoria Project"
- Benoît Sagot was a member of Claire Mouton’s PhD Committee (Université Paris-Sud Orsay, LIMSI); the title of her dissertation is "Ressources et méthodes semi-supervisées pour l’analyse sémantique de texte en français"

9.5. Commissions

- Laurence Danlos was a member of the Comité de Sélection for a Full Professor position at INALCO (team Ertim) specialized on “Sémantique textuelle outillée multilingue” (CNU section 07), for an Assistant Professor position at Université de Montpellier 2 (team LIRMM-CNRS) specialized on "Traitement Automatique des Langues (syntaxe)" (CNU section 27) and for a Full Professor position at Université de Marne la Vallée (team Informatique Linguistique) specialized on "Informatique Linguistique" (CNU sections 07 et 27).
- Benoît Sagot was a member of the Comité de Sélection for an Assistant Professor position at University of Marseilles (team LIF), for an Assistant Professor position in Computer Science at University of Marne-la-Vallée (Institut Gaspard Monge) and for an Engineer position at University of Marne-la-Vallée (team Informatique Linguistique).
- Marie Candito was a member of the Comité de Sélection for an Assistant Professor position at University of Marseilles (team LIF).

- Laurence Danlos is a member of the Scientific Committee of the Linguistics UFR of University Paris Diderot
- Laurence Danlos is a member of the Scientific Committee of the LIF (Laboratoire d'Informatique Fondamentale de Aix-Marseille) and participated to this team's scientific days (17-18 Juin in Agay, France)
- Laurence Danlos is a member of the Conseil de l'Ecole Doctorale "Sciences du Langage" from University Paris Diderot

10. Bibliography

Major publications by the team in recent years

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

[2] P. BOULLIER, B. SAGOT. *Are very large grammars computationnaly tractable?*, in "Proceedings of IWPT'07", Prague, Czech Republic, 2007, (selected for publication as a book chapter).

[3] M. CANDITO, B. CRABBÉ, D. SEDDAH. *On statistical parsing of French with supervised and semi-supervised strategies*, in "EACL 2009 Workshop Grammatical inference for Computational Linguistics", Athens, Greece, 2009.

[4] L. DANLOS. *D-STAG : un formalisme d'analyse automatique de discours fondé sur les TAG synchrones*, in "Traitement Automatique des Langues", 2009, vol. 50, n^{o} 1.

[5] L. DANLOS, B. SAGOT. *Constructions pronominales dans Dicovalence et le lexique-grammaire – Intégration dans le Lefff*, in "Linguisticæ Investigationes", 2009, vol. 2, n^{o} 32.

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

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

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

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

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

Publications of the year

Articles in International Peer-Reviewed Journal

[11] C. MESSIANT, K. GÁBOR, T. POIBEAU. *Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français*, in "Traitement Automatique des Langues (T.A.L.)", 2010, vol. 51, n° 1, p. 65–96, http://hal.inria.fr/hal-00538752/en.

International Peer-Reviewed Conference/Proceedings

[12] C. ADAM, P. MULLER, C. FABRE. *Une évaluation de l'impact des types de textes sur la tâche de segmentation thématique*, in "Traitement Automatique des Langues Naturelles - TALN 2010", Canada Montréal, Association pour le Traitement Automatique des Langues, 2010, http://www.iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_60.pdf, http://hal.inria.fr/inria-00511605/en.

[13] S. AFANTENOS, P. DENIS, P. MULLER, L. DANLOS. *Learning Recursive Segments for Discourse Parsing*, in "Language Resources and Evaluation", Malte La Valette, N. CALZOLARI, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER, D. TAPIAS (editors), European Language Resources Association (ELRA), 2010, http://hal.inria.fr/hal-00468210/en.

[14] A. BITTAR. *Annotation of Events and Temporal Expressions in French Texts*, in "Computational Linguistics in the Netherlands 19", Pays-Bas Groningen, Feb 2010, http://hal.inria.fr/inria-00522315/en.

[15] H. BLANCAFORT SAN JOSÉ, G. RECOURCÉ, J. COUTO, B. SAGOT, R. STERN, D. TEYSSOU. *Traitement des inconnus : une approche systématique de l'incomplétude lexicale*, in "Traitement Automatique des Langues Naturelles : TALN 2010", Canada Montréal, 2010, http://hal.inria.fr/inria-00521237/en.

[16] L. BORUTA, S. PEPERKAMP, B. CRABBÉ, E. DUPOUX. *Testing the Robustness of OnlineWord Segmentation:Effects of Linguistic Diversity and Phonetic Variation*, in "Proceedings of CogSci 2011", 2011, submitted.

[17] M. CANDITO, B. CRABBÉ, P. DENIS. *Statistical French dependency parsing: treebank conversion and first results*, in "Seventh International Conference on Language Resources and Evaluation - LREC 2010", Malte La Valletta, European Language Resources Association (ELRA), May 2010, p. 1840-1847, http://hal.inria.fr/hal-00495196/en.

[18] M. CANDITO, J. NIVRE, P. DENIS, E. HENESTROZA ANGUIANO. *Benchmarking of Statistical Dependency Parsers for French*, in "23rd International Conference on Computational Linguistics - COLING 2010", Chine Beijing, Coling 2010 Organizing Committee, Aug 2010, p. 108-116, 9 pages, http://hal.inria.fr/hal-00514815/en.

[19] M. CANDITO, D. SEDDAH. *Parsing word clusters*, in "NAACL/HLT-2010 Workshop on Statistical Parsing of Morphologically Rich Languages - SPMRL 2010", États-Unis Los Angeles, Association for Computational Linguistics, Jun 2010, p. 76-84, http://hal.inria.fr/hal-00495177/en.

[20] L. DANLOS, B. SAGOT. *Ponctuations fortes abusives*, in "Traitement Automatique des Langues Naturelles : TALN 2010", Canada Montréal, 2010, http://hal.inria.fr/inria-00521235/en.

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

[22] P. DENIS, P. MULLER. *Comparison of different algebras for inducing the temporal structure of texts*, in "Proceedings of the 23rd International Conference on Computational Linguistics - Coling 2010", Chine Beijing, COLING 2010 ORGANIZING COMMITTEE (editor), 2010, p. 250–258, http://www.aclweb.org/anthology/C10-1029.pdf, http://hal.inria.fr/inria-00511586/en.

[23] P. DENIS, B. SAGOT. *Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morpho-syntaxique état-de-l'art du français*, in "Traitement automatique des langues naturelles", Canada Montréal, Association pour le Traitement Automatique des Langues, 2010, http://hal.inria.fr/inria-00514364/en.

[24] P. DENIS, B. SAGOT. *Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français*, in "Traitement Automatique des Langues Naturelles : TALN 2010", Canada Montréal, 2010, http://hal.inria.fr/inria-00521231/en.

[25] K. FORT, B. SAGOT. *Influence of Pre-annotation on POS-tagged Corpus Development*, in "The Fourth ACL Linguistic Annotation Workshop", Suède Uppsala, 2010, http://hal.inria.fr/hal-00484294/en.

[26] G. FOX, J. THUILIER. *Predicting the Position of Attributive Adjectives in the French NP*, in "Student session of the European Summer School for Logic, Language and Information", Danemark Copenhague, M. SLAVKOVIK (editor), Marija Slavkovik, Sep 2010, p. 173-183, http://marija.gforge.uni.lu/proceedings.pdf, http://hal.inria.fr/inria-00515393/en.

[27] P. MULLER, P. LANGLAIS. *Comparaison de ressources lexicales pour l'extraction de synonymes*, in "Traitement Automatique des Langues Naturelles - TALN 2010", Canada Montréal, Association pour le Traitement Automatique des Langues, 2010, papier court, http://www.iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_124.pdf, http://hal.inria.fr/inria-00511599/en.

[28] Y. NAKAMURA-DELLOYE. *Extraction des chemins entre deux entités nommées en vue de l'acquisition des patrons de relations*, in "21es Journées francophones d'Ingénierie des Connaissances - IC2010", France Nîmes, 2010, P120_Poster62, http://hal.inria.fr/hal-00511481/en.

[29] Y. NAKAMURA-DELLOYE, É. VILLEMONTE DE LA CLERGERIE. *Exploitation de résultats d'analyse syntaxique pour extraction semi-supervisée des chemins de relations*, in "17e Conférence sur le Traitement Automatique des Langues Naturelles - TALN 2010", Canada Montréal, 2010, taln2010_submission_164, http://hal.inria.fr/hal-00511541/en.

[30] L. NICOLAS, M. MOLINERO, B. SAGOT, N. FERNÁNDEZ FORMOSO, V. VIDAL CASTRO. *Creating and maintaining language resources: the main guidelines of the Victoria project*, in "Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management (LREC 2010 workshop)", Malte Valletta, 2010, http://hal.inria.fr/inria-00521241/en.

[31] C. ROZE, L. DANLOS, P. MULLER. *LEXCONN: a French Lexicon of Discourse Connectives*, in "Multidisciplinary Approaches to Discourse - MAD 2010", France Moissac, 2010, http://w3.workshop-mad2010.univ-tlse2.fr/MAD_files/papers/RozeDanlosMuller.pdf, http://hal.inria.fr/inria-00511615/en.

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

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

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

[35] B. SAGOT, G. SATTA. *Optimal rank reduction for Linear Context-Free Rewriting Systems with Fan-Out Two*, in "48th Annual Meeting of the Association for Computational Linguistics - ACL 2010", Suède Uppsala, Jul 2010, http://hal.inria.fr/inria-00515455/en.

[36] B. SAGOT, G. WALTHER. *A morphological lexicon for the Persian language*, in "7th international conference on Language Resources and Evaluation (LREC 2010)", Malte Valletta, 2010, http://hal.inria.fr/inria-00521243/en.

[37] B. SAGOT, G. WALTHER. *Développement de ressources pour le persan: lexique morphologique et chaîne de traitements de surface*, in "Traitement Automatique des Langues Naturelles : TALN 2010", Canada Montréal, 2010, http://hal.inria.fr/inria-00521233/en.

[38] D. SEDDAH, G. CHRUPALA, Ö. ÇETINOGLU, J. VAN GENABITH, M. CANDITO. *Lemmatization and Statistical Lexicalized Parsing of Morphologically-Rich Languages*, in "Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages - SPMRL 2010", États-Unis Los Angeles, CA, 2010, http://hal.inria.fr/inria-00525754/en.

[39] D. SEDDAH, B. SAGOT, L. DANLOS. *Control Verbs, Argument Cluster Coordination and MCTAG*, in "10th International Conference on Tree Adjoining Grammars and Related Formalisms (TAG+10)", États-Unis New Haven, 2010, http://hal.inria.fr/inria-00521230/en.

[40] D. SEDDAH. *Exploring the Spinal-Stig Model for Parsing French*, in "Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)", Malte Malta, 2010, http://hal.inria.fr/inria-00525753/en.

[41] R. STERN, B. SAGOT. *Détection et résolution d'entités nommées dans des dépêches d'agence*, in "Traitement Automatique des Langues Naturelles : TALN 2010", Canada Montréal, 2010, http://hal.inria.fr/inria-00521234/en.

[42] R. STERN, B. SAGOT. *Resources for Named Entity Recognition and Resolution in News Wires*, in "Entity 2010 Workshop at LREC 2010", Malte Valletta, 2010, http://hal.inria.fr/inria-00521240/en.

[43] J. THUILIER, G. FOX, B. CRABBÉ. *Approche quantitative en syntaxe : l'exemple de l'alternance de position de l'adjectif épithète en français*, in "Traitement Automatique des Langues Naturelles", Canada Montréal, 2010, http://hal.inria.fr/inria-00515411/en.

[44] J. THUILIER, G. FOX, B. CRABBÉ. *Fréquence, longueur et préférences lexicales dans le choix de la position de l'adjectif épithète en français*, in "2ème Congrès Mondial de Linguistique Française", États-Unis La Nouvelle-Orléans, F. NEVEU, V. M. TOKE, T. KLINGLER, J. DURAND, L. MONDADA, S. PRÉVOST (editors), 2010 [*DOI :* 10.1051/CMLF/2010161], http://hal.inria.fr/inria-00515415/en.

[45] J. THUILIER, G. FOX, B. CRABBÉ. *Fréquence, longueur et préférences lexicales dans le choix de la position de l'adjectif épithète en français*, in "Congrés Mondial de Linguistique Française - CMLF 2010", États-Unis New Orleans, 2010, p. 2197-2210 [*DOI :* 10.1051/CMLF/2010161], http://hal.inria.fr/inria-00525768/en.

[46] R. TSARFATY, D. SEDDAH, Y. GOLDBERG, S. KUBLER, 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, p. 1–12, http://hal.inria.fr/inria-00525751/en.

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

[48] É. VILLEMONTE DE LA CLERGERIE. *Convertir des dérivations TAG en dépendances*, in "17e Conférence sur le Traitement Automatique des Langues Naturelles - TALN 2010", Montreal Canada, ATALA, 2010, Projet SCRIBO, financement Pôle SYSTEM@TIC, http://hal.inria.fr/inria-00551973/PDF/dep.pdf, http://hal.inria.fr/inria-00551973/en/.

[49] G. WALTHER, B. SAGOT, K. FORT. *Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish*, in "International Conference on Lexis and Grammar", Serbie Belgrade, Sep 2010, http://hal.inria.fr/hal-00510999/en.

[50] G. WALTHER, B. SAGOT. *Developing a large-scale lexicon for a less-resourced language: general methodology and preliminary experiments on Sorani Kurdish*, in "SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 workshop)", Malte Valetta, 2010, http://hal.inria.fr/inria-00521238/en.

Workshops without Proceedings

[51] Y. NAKAMURA-DELLOYE. *Subordonnants japonais : réflexion sur les caractères substantifs des mots*, in "Colloque international Morphologie, syntaxe et sémantique des subordonnants", France Clermont-Ferrand, Mar 2010, http://hal.inria.fr/hal-00540541/en.

[52] Y. NAKAMURA-DELLOYE. *Étude contrastive français-japonais : comportements syntaxiques des interrogatifs et indéfinis*, in "Neuvième colloque de la société française des études japonaises", France Paris, Dec 2010, http://hal.inria.fr/hal-00540543/en.

[53] Y. NAKAMURA-DELLOYE. *Étude sur les connecteurs syntaxiques inter-propositionnels du japonais : définition et catégorisation*, in "XXIIIèmes Journées de Linguistique d'Asie Orientale", France Paris, Jul 2010, http://hal.inria.fr/hal-00540542/en.

Scientific Books (or Scientific Book chapters)

[54] *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, Association for Computational Linguistics, 2010, 113, http://hal.inria.fr/inria-00525752/en.

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

[56] L. DANLOS. *Extension de la notion de verbe support*, in "Les Tables, La grammaire par le menu, Volume d'hommage à Christian Leclère", T. NAKAMURA, É. LAPORTE, A. DISTER, C. FAIRON (editors), Presses Universitaires de Louvain, 2010, p. 81–90, http://hal.inria.fr/inria-00524746/en.

Scientific Popularization

[57] L. DANLOS. *Écriture automatique*, in "La Recherche. Les Cahiers de l'Inria", Jul 2010, n^{o} 443 Juillet-Août 2010, http://hal.inria.fr/inria-00511267/en.

References in notes

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[73] P. DENIS, B. SAGOT. *Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort*, in "Proceedings of PACLIC 2009", Hong Kong, China, 2009, http://atoll.inria.fr/~sagot/pub/paclic09tagging.pdf.

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

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

[76] L. KALLMEYER. *Tree-Local Multicomponent Tree-Adjoining Grammars with Shared Nodes*, in "Computational Linguistic", 2005, vol. 31, n^{o} 2, p. 187–226.

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

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

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

[80] A. MARTIN, S. PEPERKAMP, E. DUPOUX. *Learning Phonemes with a Pseudo-Lexicon*, in "Workshop on Computational Modelling of Sound Pattern Acquisition", Edmonton, Canada, 2009.

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

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

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

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

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

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

[87] S. PEPERKAMP, R. LE CALVEZ, J.-P. NADAL, E. DUPOUX. *The acquisition of allophonic rules: statistical learning with linguistic constraints*, in "Cognition", 2006, vol. 101, n^{o} 3, p. B31–B41.

[88] 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.

[89] S. PETROV, D. KLEIN. *Improved Inference for Unlexicalized Parsing*, in "Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference", Rochester, New York, Association for Computational Linguistics, April 2007, p. 404–411, http://www.aclweb.org/anthology/N/N07/N07-1051.

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

[91] 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.

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

[93] 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, to appear.

[94] 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.

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

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

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

[98] B. SAGOT. *Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish*, in "LNAI 5603, selected papers presented at the LTC 2007 conference", Springer, 2009.

[99] D. SEDDAH, B. SAGOT. *Modeling and Analysis of Elliptic Coordination by Dynamic Exploitation of Derivation Forests in LTAG parsing*, in "Proceedings of TAG+8", Sydney, Australia, July 2006, p. 147-152.

[100] D. SEDDAH. *The Use of MCTAG to Process Elliptic Coordination*, in "In Proceeding of the Ninth International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+9)", Tüebingen, Germany, June 2008.

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

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

[103] G. VALLABHA, J. MCCLELLAND, F. PONS, J. WERKER, S. AMANO. *Unsupervised learning of vowel categories from infant-directed speech*, in "Proceedings of the National Academy of Sciences", 2007, vol. 104, n^{o} 33, p. 13273–13278.

[104] B. VARADARAJAN, S. KHUDANPUR, E. DUPOUX. *Unsupervised Learning of Acoustic Sub-word Units*, in "Proceedings of ACL-08: HLT, Short Papers", Columbus, Ohio, Association for Computational Linguistics, June 2008, p. 165–168.

[105] K. VIJAY-SHANKER, D. J. WEIR, A. K. JOSHI. *Characterizing structural descriptions produced by various grammatical formalisms*, in "Proceedings of the 25th annual meeting on Association for Computational Linguistics", Stroudsburg, PA, USA, Association for Computational Linguistics, 1987, p. 104–111, http://dx.doi.org/10.3115/981175.981190.

[106] É. VILLEMONTE DE LA CLERGERIE, B. SAGOT, L. NICOLAS, M.-L. GUÉNOT. *FRMG: évolutions d'un analyseur syntaxique TAG du français*, in "Actes électroniques de la Journée ATALA sur "Quels analyseurs syntaxiques pour le français ?"", ATALA, October 2009.

[107] É. VILLEMONTE DE LA CLERGERIE. *DyALog: a Tabular Logic Programming based environment for NLP*, in "Proceedings of 2nd International Workshop on Constraint Solving and Language Processing (CSLP'05)", Barcelona, Spain, October 2005.

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

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

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

[111] D. J. WEIR. *Characterizing mildly context-sensitive grammar formalisms*, University of Pennsylvania, Philadelphia, PA, USA, 1988, Supervisor-Aravind K. Joshi.

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

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