ANR MDCA proposal

PASSAGE

Produire des annotations syntaxiques à grande échelle pour aller de l'avant
Large scale production of syntactic annotations to move forward

November 20, 2006

Contents

A Description courte du projet

A.1 Contexte et motivation du projet

The main motivations of the **PASSAGE** project are twofold :

1. to **improve the accuracy and robustness of existing French parsers** by using them on **large scale corpora** (several millions of words) and
2. to exploit the resulting syntactic annotations to **create richer and more extensive linguistic resources.**

The adopted methodology consists of a feedback loop between parsing and resource creation as follows:

1. parsing is used to create syntactic annotations
2. syntactic annotations are used to create or enrich linguistic resources such as lexicons, grammars or annotated corpora
3. the linguistic resources created or enriched on the basis of the syntactic annotations are then integrated into the existing parsers
4. the enriched parsers are used to create richer (e.g., syntactico-semantic) annotations
5. etc.

More generally, the **PASSAGE** project should help seeing the emergence of linguistic processing chains exploiting richer lexical informations, in particular semantic ones.

PASSAGE will build up on the results of the **EASy** French parsing evaluation campaign (EASy/Evalda action[1], Technolangue program). This campaign has shown that several parsing systems are now available for French but that robustness and accuracy can still be largely improved, especially for oral data.

Furthermore, although the initial plan was to combine the results produced by each participant to construct a treebank for French (a corpus annotated with syntactic information), the creation of this treebank is still to be achieved and the expected output, while very valuable, remains relatively small (around 40K sentences with a subset of around 4K sentences manually validated), compared to emerging international standards (10M to 100M words, i.e. 0.5M to 5M sentences).

PASSAGE aims at pursuing and extending the line of research initiated by the **EASy** campaign. In particular, **PASSAGE** aims at:

- running new evaluation campaigns to assess and improve the parsers for French on large scale corpora (billions of words)
- finalising a methodology for comparing and merging the output of several parsers
- using the merged output of the best parsers to construct a treebank for French

[1] `http://www.technolangue.net/article198.html`

- validating this automatically constructed treebank either manually or automatically
- using both the validated treebank and the non validated large scale corpora annotated with syntactic information to extract linguistic information
- integrating the thus acquired linguistic information into parsers
- developing methodologies for evaluating the quality of the acquired resources

The participation of around 10 parsing systems in a collective effort geared towards improving parsing robustness and acquiring linguistic knowledge from large scale corpora is a rather unique event. We believe that the combination of so many sources of information over a relatively long period of adaptation ensures good chances of success for the proposal.

A.2 Retombées scientifiques et techniques attendues

The expected results of the **PASSAGE** project include:

- the emergence of more robust, efficient and accurate linguistic processing chains for French, with a better evaluation of their level of performance
- the identification of methodologies and protocols to perform linguistic knowledge acquisition tasks. These methodologies should be adaptable for other languages than French, in particular to handle resources poor languages and overcome the famous bottleneck problem in NLP.
- a French dependency treebank for parsing technology improvement.
- the enrichment of French linguistic resources (grammars, syntactico-semantic lexica, a prototype PropBank)
- the consolidation of a strong parsing community in France, familiar with the systematic use of large scale evaluation procedures.

A.3 Retombées industrielles et economiques escomptées (le cas écheant)

Parsing is an important phase of linguistic processing that is not yet so widely deployed in industrial applications, because of its complexity and because of the requirement in terms of resources. **PASSAGE** could alter this situation thanks to:

- the emergence of more robust, efficient, and accurate linguistic processing chains for French. These systems should be natural candidates for industrial transfer and exploitation in more industrial applications.
- the availability of more linguistic resources for French (lexica, grammars)
- the assessment of syntactic annotations, and in particular syntactic dependencies, as an emerging source of data for information extraction applications (as is already the case for other languages, esp. English).
- the validation of a method for improving language processing technologies, through evaluation and combination of parsing systems.

A.1 Contexte et motivation du projet (Fr)

Les motivations principales de la proposition **PASSAGE** sont doubles:

- **améliorer la précision et la robustesse des analyseurs syntaxiques existants** pour le Français, en les utilisant sur de **gros corpus** (plusieurs million de mots) et
- exploiter les annotations syntaxiques résultantes pour créer des ressources linguistiques plus riches et plus extensives.

La méthodologie adoptée consiste en une boucle de rétroaction (*feedback*) entre analyse syntaxique et création de ressources, comme suit:

1. l'analyse syntaxique est utilisée pour créer des annotations syntaxiques
2. les annotations sont utilisées pour créer ou enrichir des ressources linguistiques comme des lexiques, grammaires ou corpus annotés
3. les ressources créées ou enrichies sur la base des annotations sont ensuite intégrées dans les systèmes d'analyse.
4. les analyseurs enrichis sont utilisés pour créer des ressources encore plus riches (par exemple syntactico-sémantiques)
5. etc.

Plus généralement, le projet **PASSAGE** devrait aussi aider à faire émerger des chaînes de traitement linguistique exploitant des informations lexicales plus riches, en particulier sémantiques.

PASSAGE s'appuie sur les résultats de la campagne d'évaluation des analyseurs syntaxiques menée dans le cadre de l'action **EASy/EVALDA**[2] (programme Technolangue). Cette campagne a montré que plusieurs systèmes d'analyse existent désormais pour le Français. Néanmoins, bien que les résultats furent meilleurs que prévus, cette campagne a confirmé que la robustesse et la précision peuvent encore être largement améliorées, en particulier pour les données orales.

De plus, bien que le plan initial de **EASy** était de combiner les résultats produits par chaque participant pour construire une treebank du Français (un corpus annoté syntaxiquement), cette phase reste à venir, et le résultat, malgré son intérêt certain, restera relativement limité (environ 40K phrases avec un sous-ensemble de 4K phrases manuellement validées), au regard des standards internationaux qui émergent (10M à 100M mots, i.e. 0.5M à 5M phrases).

PASSAGE vise à poursuivre et à étendre la ligne de recherche initiée par la campagne **EASy**. En particulier, **PASSAGE** cherche à:

- organiser des nouvelles campagnes d'évaluation pour évaluer et améliorer les systèmes d'analyse syntaxiques du Français sur de gros corpus (millions de mots)
- finaliser une méthodologie pour comparer et fusionner les résultats fournis par plusieurs analyseurs

[2] `http://www.technolangue.net/article198.html`

- utiliser les résultats fusionnés des meilleurs analyseurs pour construire une treebank du Français
- valider cette treebank soit manuellement soit automatiquement
- utiliser à la fois cette treebank et la partie non-validée du gros corpus annoté syntaxiquement pour extraire des informations linguistiques
- intégrer les ressources ainsi acquises dans les analyseurs
- développer les méthodologies pour évaluer la qualité des ressources ainsi acquises

La participation d'une dizaine systèmes d'analyse syntaxique dans un effort collectif tourné vers l'acquisition de ressources linguistiques est un occasion plutôt unique. Nous pensons que la combinaison d'autant de sources d'information sur une période d'adaptation relativement longue renforce les chances de succès de cette proposition.

A.2 Retombées scientifiques et techniques attendues (Fr)

Les retombées attendues du projet **PASSAGE** incluent:

- l'émergence de chaînes de traitement linguistique pour le Français qui soient plus robustes, efficaces, et précises, avec de plus une meilleur évaluation de leur niveau de performance.
- l'identification de méthodologies et de protocoles pour effectuer des tâches d'acquisition de connaissances linguistiques. Ces méthodologies devraient être adaptables pour d'autres langues que le Français, en particulier pour traiter des langues pauvrement dotées, aidant ainsi à surmonter le fameux problème du goulet d'étranglement en Traitement Automatique des Langues (TAL)
- une banque d'annotations syntaxiques (en dépendances) pour le Français, utiles pour améliorer le traitement syntaxique
- l'enrichissement de ressources linguistiques pour le Français (lexiques et grammaires)
- l'acquisition de connaissances linguistiques aidant au développement d'applications mieux adaptées aux utilisateurs.
- la consolidation d'une forte communauté française en analyse syntaxique, familière avec l'utilisation systématique de procédure d'évaluation à grande échelle.

A.3 Retombées industrielles et économiques escomptées (Fr)

L'analyse syntaxique est une phase importante de traitement linguistique qui n'est pas actuellement largement déployée dans le cadre d'applications industrielles, en partie à cause de sa complexité et des besoins en termes de ressources. **PASSAGE** pourrait altérer cette situation grâce à:

- l'émergence de chaîne de traitement linguistique pour le Français, plus robustes, efficaces et précises. Ces systèmes sont des candidats de choix pour des transferts industriels et leur exploitation dans des applications industrielles;
- l'accès à plus de ressources linguistiques pour le Français (lexiques, grammaires);
- l'évaluation des annotations syntaxiques, et en particulier sous forme de dépendances, comme une source émergent de données pour des applications d'extraction d'information (comme c'est déjà le cas pour d'autres langues, en particulier l'anglais);
- la validation d'une méthode pour améliorer les technologies de traitement du langage, au travers l'évaluation et la réunion de systèmes d'analyse syntaxique.

B Description scientifique et technique detaillée du projet

B.1 But du projet

The main motivations of the **PASSAGE** project are twofold :

1. to **improve the accuracy and robustness of existing French parsers** by using them on **large scale corpora** (several millions of words) and
2. to exploit the resulting syntactic annotations to **create richer and more extensive linguistic resources.**

The adopted methodology consists of a feedback loop between parsing and resource creation as follows:

1. parsing is used to create syntactic annotations
2. syntactic annotations are used to create or enrich linguistic resources such as lexicons, grammars or annotated corpora
3. the linguistic resources created or enriched on the basis of the syntactic annotations are then integrated into the existing parsers
4. the enriched parsers are used to create richer (e.g., syntactico-semantic) annotations
5. etc.

More generally, the **PASSAGE** project should help seeing the emergence of linguistic processing chains exploiting richer lexical informations, in particular semantic ones.

PASSAGE will build up on the results of the **EASy** French parsing evaluation campaign (EASy/Evalda action, Technolangue program). This campaign has shown that several parsing systems are now available for French but that robustness and accuracy can still be largely improved, especially for oral data.

Furthermore, although the initial plan was to combine the results produced by each participant to construct a treebank for French (a corpus annotated with syntactic information), the creation of this treebank is still to be achieved and the expected output, while very valuable, remains relatively small (around 40K sentences with a subset of around 4K sentences manually validated), compared to emerging international standards (10M to 100M words, i.e. 0.5M to 5M sentences).

PASSAGE aims at pursuing and extending the line of research initiated by the **EASy** campaign. In particular, **PASSAGE** aims at:

- running new evaluation campaigns to assess and improve the parsers for French on large scale corpora (billions of words)
- finalising a methodology for comparing and merging the output of several parsers
- using the merged output of the best parsers to construct a treebank for French
- validating this automatically constructed treebank either manually or automatically

- using both the validated treebank and the non validated large scale corpora annotated with syntactic information to extract linguistic information
- integrating the thus acquired linguistic information into parsers
- developing methodologies for evaluating the quality of the acquired resources

The participation of around 10 parsing systems in a collective effort geared towards improving parsing robustness and acquiring linguistic knowledge from large scale corpora is a rather unique event. We believe that the combination of so many sources of information over a relatively long period of adaptation ensures good chances of success for the proposal. The parsing systems will be provided by participants or contractants, including:

- *ATOLL / INRIA* – Éric de la Clergerie – **FRMG** (based on TAGs) and **SxLFG** (based on LFG)
- *Langue et Dialogue / LORIA* – Azim Roussanaly – **LLP2**
- *LIC2M / CEA-LIST*[3] – Gaël de Chalendar – LIMA
- TAGMATICA – Gil Francopoulo
- Équipe de Recherche en Syntaxe et Sémantique [ERSS][4] – Didier Bourigault – **Syntex**
- Laboratoire Langage & Parole [LPL][5] – Philipe Blache – 2 or 3 parsing systems, based on property grammars
- SYNAPSE[6] – Dominique Laurent
- Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier [LIRMM][7] – Jacques Chauché
- Xerox Research Center Europe[8] (XRCE) – Salah Ait-Mokhtar – **XIP**

This proposal clearly falls in the NLP part (Thematic 5) of the 2006 ANR MDCA call, in particular covering the points:

- production of linguistic resources
- comparative evaluation of linguistic systems
- (semi-) automatic construction of lexicon and grammars
- definition of parsing systems (improvement) and evaluation criteria
- some elements of standardization, essentially for syntactic annotations
- structuring of parsers output (syntactic annotations as dependencies)

But **PASSAGE** also addresses datawarehouse issues such as large corpora processing (using grids) and large syntactic annotation banks.

[3] http://www-list.cea.fr/
[4] http://www.univ-tlse2.fr/erss/
[5] http://cnrs.oxcs.fr/
[6] http://www.synapse-fr.com/
[7] http://www.lirmm.fr/xml/fr/lirmm.html
[8] http://www.xrce.xerox.com/

B.2 Contexte et état de l'art

At the international level, the last decade has seen the emergence of a very strong trend of researches on statistical methods in Natural Language Processing. This trend results from several reasons but one of them, in particular for English, is the availability of large annotated corpora, such as the Penn Tree bank (1M words extracted from the Wall Street journal, with syntactic annotations; 2nd release in 1995[9]), the British National Corpus (100M words covering various styles annotated with parts of speech[10]), or the Brown Corpus (1M words with morpho-syntactic annotations). Such annotated corpora were very valuable to extract stochastic grammars or to parametrize disambiguation algorithms. These successes have led to many similar proposals of corpus annotations. A long (but non exhaustive) list may be found on `http://www.ims.uni-stuttgart.de/projekte/TIGER/related/links.shtml`, but we can here mention:

- the French treebank[11] developed by Anne Abeillé at University Paris 7
- The Dutch Alpino treebank[12]: it is worth mentioning that this treebank contains a subset manually corrected and a much larger uncorrected one (5M sentences) built with a stochastic parser in the Dutch ALPINO project[13] (*Algorithms for Linguistic Processing*)
- The German NEGRA treebank[14]
- The German TIGER treebank[15]
- The Prague Dependency Bank[16]
- The UAM Spanish Treebank[17]
- The Italian Turin University Treebank[18]

However, the development of such treebanks is very costly from an human point of view and represents a long standing effort. The volume of data that can be manually annotated remains limited and is generally not sufficient to learn very rich information (sparse data phenomena). Furthermore, designing an annotated corpus involves choices that may block future experiments to acquire new kinds of linguistic knowledge. Last but not least, it is worth mentioning that even manually annotated corpora are not error prone.

We believe that a new option becomes possible. The French parsing evaluation campaign EASy has shown that parsing systems are now available for French, implementing shallow to deep parsing. Some of these systems were neither based on

[9] `http://www.cis.upenn.edu/~treebank/`
[10] `http://www.natcorp.ox.ac.uk/`
[11] `http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php`
[12] `http://www.let.rug.nl/vannoord/trees/`
[13] `http://www.let.rug.nl/vannoord/alp/`
[14] `http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/`
[15] `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/`
[16] `http://ufal.mff.cuni.cz/pdt/`
[17] `ttp://www.lllf.uam.es/~sandoval/UAMTreebank.html`
[18] `http://www.di.unito.it/~tutreeb/`

statistics, nor extracted from a treebank. While needing to be improved in robustness, coverage, and accuracy, these systems has nevertheless proved the feasibility to parse medium amount of data (40K sentences, 1M words). Some preliminary experiments made by some of the participants with deep parsers ([7]) indicate that processing several thousand sentences is not a problem, especially by relying on clusters of machines. These figures can even be increased for shallow parsers. In other words, it means that there now exists in France several parsing systems that could parse (and re-parse if needed) large corpora between 10 to 100M words (around 500K to 5M sentences).

While the quality of the analysis produced by these parsers remains to be assessed and improved (this is an important objective of this proposal), it should already be possible to learn valuable linguistic knowledge from the analysis of a large corpus, for instance about sub-categorization frames, restriction of selections, statistical distribution of syntactic phenomena, ..., as advocated by others (*Deriving Linguistic Resources from Treebanks*[19]; LREC'02 Workshop[20] on "*Linguistic Knowledge Acquisition and Representation : Bootstrapping Annotated Data*"). Furthermore, the knowledge thus acquired is meant to be integrated in the parsing systems to make them more accurate. Hopefully entering a virtuous circle, corpora may then be re-parsed to learn new knowledge.

The quality of the acquired linguistic knowledge will potentially be reinforced by combining (merging) the results produced by all the parsers involved in **PASSAGE**. To facilitate this merging, we will focus on dependency-based representations of the results, that also seem to be better adapted for the acquisition of linguistic knowledge. This combination process requires to assess the performance level of the different parsers involved in order to compute a confidence factor associated to the information provided by each parser. One way to assess this confidence factor is through quantitative black-box evaluation, a paradigm that has been used successfully for more than 20 years in speech processing community and is recognized as one of the factor that drove technology progresses in that field. **PASSAGE** will be the opportunity to build on the exploratory effort of EASY in order to promote the evaluation paradigm for parsing natural language.

Human validation by expert linguists is needed by the evaluation procedure and remains an important issue, first to assess the quality of the acquisition techniques but also because unsupervised acquisition does not seem reasonable since the improvement target has to be provided by humans, the only condition for breaching technological barriers. Furthermore, linguistic expertise and linguistic theories are necessary to guide the acquisition experiments. Part of the objectives of **PASSAGE** is to understand how this expertise may take place and be efficiently used through adequate validation interfaces (for instance on the model of the one developed for error mining[21]).

[19] `http://www.computing.dcu.ie/~away/Treebank/treebank.html`

[20] `http://www.lrec-conf.org/lrec2002/lrec/wksh/CfP-WP16.html`

[21] http://atoll.inria.fr/perl/results5/errorscgi.pl

The parsers may be improved by acquiring probabilistic information on natural language (for reducing ambiguities) but also by improving and enriching their underlying linguistic resources, lexica or grammars. Thus, as a very important side effect of **PASSAGE**, we should get richer and more extensive linguistic resources (or at least, get an improvement of existing ones).

Very recently, suggestions have been made to marry symbolic and statistics approaches [3]. Symbolic approaches provide ways to express linguistic knowledge and move to richer levels of descriptions (syntax, semantic) while statistics provide ways to capture the fact that languages are human artifacts partly characterized by their usage in a community. This proposal **PASSAGE** should help us going into this direction.

At the end of the project, the final set of syntactic annotations will also be made freely available to the community and, hopefully, boost new acquisition experiments. A subpart of this annotation set (around 500K words) should be manually validated and be also distributed.

B.2.1 ATOLL/INRIA

The INRIA project-team *ATOLL* (Software Tools for Linguistic Processing) is strongly involved in the exploration and development of parsing technologies for NLP, for various syntactic formalisms. Two parsing systems (FRMG/DYALOG and SXLFG/SYNTAX), exploiting a common pre-parsing processing chain SXPIPE and a syntactic lexicon Le*fff* [10], have been developed for French and tried during the EASy campaign [20, 19]. These systems have been since used in large scale parsing experiments (*Monde Diplomatique*, 300K to 600K sentences; botanical corpus, 100K sentences) [11, 7], partly relying on the use of clusters. The results of these experiments have already been used to mine errors [9, 8], in order to improve our linguistic resources. The Le*fff* lexicon developed by ATOLL is partly the result of experiments using linguistic knowledge acquisition techniques ([21, 10]), whose latest version has been applied on Slovak ([15]). ATOLL is also involved in the current standardization efforts at the level of ISO TC37 SC4, on morpho-syntactic annotations (MAF) [18] and on syntactic annotations (SynAF). With Le*fff*, FRMG and SXLFG, through several national actions (LexSynt, MOSAIQUE, GrammoFun proposal) and trough the current proposal **PASSAGE**, ATOLL tries to contribute to the development of wide-coverage linguistic resources for French.

B.2.2 LIR/LIMSI

The "Langues, Information et Représentations" group at LIMSI has a strong competence in the organization of evaluation campaigns, being the scientific organizer of the **GRACE** campaign for Part-of-Speech taggers and the scientific organizer of **EASY/EVALDA**, the parsing evaluation campaign for French.

B.2.3 Langue & Dialogue / LORIA

Langue et Dialogue / LORIA's research focuses on computational semantics and in particular, on the development of a computational infrastructure for the semantic processing of French. Within **PASSAGE**, *Langue et Dialogue / LORIA* will bring expertise on lexical information acquisition ; parsing ; semantic parsing and semantic annotation. Thus Azim Roussanaly was one of the participant in the EASy campaign (parser LLP2) and Claire Gardent has worked on extracting valency information from the LADL tables [17, 6, 12]. Further work of that group includes the development of a semantic processing chain for French including a grammar writing environment for tree based grammar [**?**], a Tree Adjoining Grammar for French integrating a semantic dimension [13] and the coupling with this grammar of two parsers [16] and a generator [5]. Finally, *Langue et Dialogue / LORIA* also has some experience in corpora annotation [4].

B.2.4 *LIC2M / CEA-LIST*

The LIC2M group from CEA-LIST laboratory is specialized in multilingual information search and extraction. As such, it has developed a multilingual linguistic analyzer. Its French version has particpated in the Easy campaign. The new evolution will also participate to the campaigns of the Passage project. The LIC2M is also involved in other projects where a very important computing architecture, like clusters or super computers, is necessary. This expertise will be used within **PASSAGE** in the building of the handler of multiple parsers. Finally, our researchs on the acquisition of semantic data from semantic analysis will be used in **PASSAGE** and developped thanks to our participation to this project.

B.2.5 ELDA

ELDA aims at collecting, commercializing and distributing Language Resources, as well as collecting and disseminating general information related to the field of Human Language Technologies, with the mission of providing a central clearing house to the players of the field. As a matter of fact, ELDA launched evaluation activities in the past few years, distributing the language resources appropriate to evaluation experiments in language engineering. ELDA is now involved in this field to a large extent, developing its skills to all aspects of evaluation: participation to the evaluation of systems, applications, involvement in European evaluation projects, such as CLEF'2026 To make this new activity more concrete, ELDA changed its former name "European Language Resources Distribution agency" into "Evaluation and Language Resources Distribution Agency". In the same way as ELDA made a great number of Language Resources available, and carries on by distributing them widely, it aims at building a permanent and long-lasting evaluation infrastructure. In **PASSAGE**, ELDA will act as a contractant.

B.2.6 Tagmatica

Tagmatica is a specialized SME in the standardization of Natural Language Processing (`www.tagmatica.com`). Gil Francopoulo (from Tagmatica) works in lexicon management, sentence parsing and search engines for 20 years. He is editor of the ISO standard dedicated to lexicons (LMF aka CD24613). He participates in the work in progress about ISO syntactic annotation framework (SynAF aka WD24615). In **PASSAGE**, Tagmatica will act as a contractant.

B.3 Organisation du projet – description des sous-projets

B.3.1 General principles

The project development and evaluation strategy is based on a feedback loop between parsing and resource creation. More specifically, the output of parsing is used to extract linguistic information which once validated is integrated into the parsers which are then re-run on large scale corpora thus providing the trigger for another acquisition-evaluation-improvment cycle. In this way, (i) parsing is used to acquire richer and larger linguistic resources and (ii) the acquired resources are used to improve parsing performance, coverage and precision.

Meetings will be held every 4 months to evaluate the evolution of the project and to favor exchange of information between the participants. The participants will have access to a maximum of information, resources and tools to conduct their parsing experiments and, in return, will return a maximum of information (syntactic annotations) to be shared by others (possibly in an anonymous way).

B.3.2 Tasks

The project will be based around the following tasks.

WP1 – Identification and preliminary preparation of corpora

Coordinator: *ATOLL / INRIA* (with a subcontract to ELDA)

Objectives: The aim of this work package is the creation of a freely available large scale corpora containing various text styles. The subtasks involved will include:

1. The selection of several corpus covering various kinds of styles (including oral transcription) totaling 10M words to 100M words. Corpora will be selected for their style but also for their possibility to be freely available (or at least easily available at reasonable cost). Natural candidates include:

    - the **EASy corpora** (around 1M words for 40K sentences, already covering various styles). This corpus includes a subset of around 4K sentences that have been manually validated.

- **Wikipedia Fr**, a freely available corpus of almost 500K entries covering various domains of knowledge, with various styles though biased toward descriptions.
- **Wikisources**, a collection of several thousand freely available French texts covering various thematics.
- **Wikilivres**, a collection of 1956 freely available educational French books.
- **Monde Diplomatique**, a low cost journalistic corpus already used by various teams and covering many thematics.
- **ABU : la Bibliothèque Universelle**, a collection of digitised French books such as Jule Vernes's novels available from the Web.
- Speech transcriptions available from the **Freebank**[22]
- **Europarl**[23], a corpus of parallel multilingue text from the European Parliament (around 28M words per language), for a long term preparation of linguistic transfer (outside the scope of **PASSAGE**)

2. Corpus cleaning: while the original corpora should remain available, some scripts should be made available to achieve a minimum of normalization on their form (removal of some meta-data, removal of HTML/XML mark-ups, possibly sentence segmentation, ...).

3. Corpus-annotation anchoring tools: for cases where the original corpora cannot be freely distributed, we will devise a method for distributing separatly the original material for one part and annotations and anchoring information for another part. More generally, it implies that annotations have to clearly reference portions of the original corpus (using standoff annotations and robust portable addressing schemes for linguistic units referred to in the annotations).

WP2 – Handling morpho-syntactic and syntactic annotations

Coordinator: *LIR / LIMSI* (with a subcontract to TAGMATICA)

Objectives: this work package will concentrate on defining a project internal annotation standard for syntactic annotations, on making the link between this standard and other existing standards and in setting up the basic computational infrastructure necessary to store, edit and search syntactically annotated corpora. More specifically, the aim will be:

1. To propose formats for the annotations produced at various stages, essentially for syntactic annotations but maybe also for morpho-syntactic annotations. These formats are to ease the distribution of annotations but also to

[22] `http://freebank.loria.fr/`

[23] `http://people.csail.mit.edu/koehn/publications/europarl/`

ease the comparison and fusion of the results provided by the parsers. They will be defined in relation with the ongoing standardization efforts promoted by ISO TC37 SC4, and relayed by the former Technolangue action **Normalangue/RNIL**, the European action **Lirics**, and the ANR MDCA proposal **Nortal** (if accepted). In particular, the most important emerging standards for **PASSAGE** are MAF "Morpho-syntactic Annotation Framework" and SynAF "Syntactic Annotation Framework".

It is not reasonable to quickly hope for a complete **SynAF** proposal. It seems more reasonable to look forward a relatively simple dependency-based model (possibly with a chunk level), extending the one proposed for the **EASy** campaign. The experience acquired on the **EASy** campaign should help the comparison and fusion of parsing results.

2. To develop the format conversion tools between the internal format used by the project to combine parsers' annotations and the standard defined by standardization institutions as they will exist when the project will end.

3. To identify technologies and tools to store, edit and search large repositories of syntactic annotations. Particularly, we will use the ongoing modifications of the LIC2M search engine that will be able to handle complex structured data.

Note: The selection and/or development of formats and tools will be guided by a preliminary investigation of existing treebank projects and annotation platforms (GATE, Linguistic Data Consortium [LDC], Alembic Workbench, MATE, TIGER, Prague Dependency Treebank, etc.).

WP3 – Large scale processing on clusters

Coordinator: *ATOLL / INRIA*

Objectives: This work package will aim at

1. developing tools and environments to allow for cluster based parsing and

2. identifying potential clusters such as GRID 5000 (resulting from the ACI Grid).

Note: This task may require some adaptation of the processing chains to be able to run on a cluster-based environment. However, there is no obligation for a participant to use a cluster-based environment if the performances on a single machine are good enough.

WP4 – Parser output comparisons and fusion

Coordinator: *LIR / LIMSI*

Objectives: To develop a merging protocol for the annotations provided by the various parsers. This procotol will use the performance assessment provided by

the two evaluation campaigns of the project (one a the beginning and one towards the end) to define a confidence factor to associate to each system annotation. When needed, automatic error correction scripts will be developped to improve the data produced by individual parsers for the most frequent and systematic errors. This task will build on the experience gained during the EASy campain.

WP5 – Acquisition tasks

Coordinators: *Langue et Dialogue / LORIA& ATOLL / INRIA*

Objectives: Exploration of various techniques to extract information from the syntactically annotated corpora resulting from the parsing process. The work will mainly focus on exploring the creation of a knowledge rich lexicon for French including :

- valency information (for verbs, nouns and adjectives)
- identification of verbal alternation classes ([1])
- disambiguation information
- weighted selectional restrictions

In particular, one aim will be to use the information derived from a large scale syntactically annotated corpus to create a lexicon that associate with each syntactic functor both valency and theta grid information. The idea is to first derive valency information, then use this information and its lexical distribution to create Beth Levin's type alternation classes and finally to use these classes to systematically assign a common thematic grid to all verbs of a given class. Corpus derived information will be compared and combined with information made available by already existing resources such as the syntactic lexicon Le*fff* [10], the Synlex lexicon derived from the LADL tables [17, 6, 12] and Patrick Saint-Dizier's manually constructed alternation classes [**?**].

Other kinds of information are susceptible to be acquired, such as

- semantic classes
- derivational morphology with transfer of syntactic information
- probabilities of some syntactic constructions
- extraction of stochastic grammars [22, 2]
- parametrization of meta-grammars ([14])
- automatons recognizing specific phenomena through grammar inference (as done by *LIC2M / CEA-LIST*)

Note:

- This part is obviously the most prospective and challenging one. This idea is to explore both at the theoretical and concrete level methodologies to acquire linguistic knowledge from syntactic annotations produced by parsers.
- Some of the experiences will be performed on the merged results while other will be performed on each parser results (for instance, to identify the probabilities of syntactic constructions, often related yo the underlying grammar).

WP6 – Integration and validation tasks

Coordinator: *Langue et Dialogue / LORIA*

Objectives:

1. To provide ways to validate the NLP lexicon provided by the acquisition task (WP5)

2. To integrate some or all of this NLP lexicon into at least one parsing system (but the information will be available to all participants for integration within their systems).

This work package will focus on assessing the correctness of the syntactic and semantic information contained by the lexicon acquired in WP5. Specifically, we intend (i) to integrate this lexicon into a parsing system, (ii) to use the resulting parsing system to build a pilot corpora annotated with semantic information and (iii) to evaluate the results of the parsing system against a manually created gold standard.

The parsing system used will build on the SEMTAG system developed in Nancy [16, 13] which consists of a lexicon, a Tree Adjoining Grammar integrating syntax and semantics and Eric de la Clergerie's DYALOG parser. The lexicon will be extended with the syntactic (valency) and semantic (thematic grid) information acquired from corpora, the grammar coverage will be extended to deal with constructions not yet covered and corpus extracted probabilistic information will be used to disambiguate the parser results.

The resulting parsing system will then be used to semi-automatically create a **PropBank**[24] like corpora (that is, of a corpora annotated with semantic functor/arguments dependencies). That is, annotators will be asked to choose from amongst the parser output, the parse yielding the most appropriate thematic grid for each basic clause. The resulting annotated corpus will then be compared against a manually created gold standard using standard precision and recall measures.

Note: The aim here is not to construct and make available a Propbank style corpora but rather to conduct some pilot experiments on the usefulness of the acquired information for constructing such a corpora. Specifically, the construction of a Propbank style corpora will permit a first assessment of the quality of the valency and thematic grid information contained in the lexicon by adressing the question: does the constructed parsing system yield correct thematic grids for most cases?

WP7 – Administrative and Scientific Management of Evaluation campaign 1

Coordinator: *LIR / LIMSI*

Objectives: The partner will deploy an open evaluation campaign for parsers of French. The first evaluation campaign will take place at the beginning of the project. Building on the results of **EASy** it will provide a performance assessment

[24] `http://www.cis.upenn.edu/~mpalmer/project_pages/ACE.htm`

of the available parsing technology. In particular it will gauge the progress of the technology made since the end of EASy. The performance information (confidence factor) associated to each parser will be used to drive the combination process (weighted voting procedure) and specific error analysis of the parsers output will be used to improve parsing quality of the corpus.

WP8 – Manually validated reference subcorpus

Coordinator: *ATOLL / INRIA* (with a subcontract to ELDA)

Objectives: Identification of a sub-corpus (500K words) and stabilization and validation of its syntactic annotations. Validation will be performed by human annotators. Human validation will be checked with inter-annotator agreement measures performed on randomly selected excerpts of corpus (amounting to 10% of the corpus). First, we will need to determine what is more efficient between correcting annotated material and hand-annotating anew the original material (for instance for speech transcriptions, if the error rate is too high, it is much cheaper to re-do the transcription from scratch than to try to correct the existing one).

Methodology and specific software for hand annotation/correction will be investigated in order to optimize the annotation task (in particular, consistency checking tools). The tools developped, if any, will be made available to the community at the end of the project along with the reference corpus which should be composed of copyright free material.

The dependency bank so produced will be used in the second evaluation campaign and be an invaluable resource to assess in the future the quality of parsers and the robustness of various acquisition tasks with respects to parsing errors (comparing the use of manually versus automatically annotated material).

WP9 – Administrative and Scientific Management of Evaluation campaign 2

Coordinator: *LIR / LIMSI*

Objectives: The partner will deploy an open evaluation campaign for parsers of French. The second evaluation campaign will take place at the end of the project. It will provide a performance assessment of the progress made during the project. The performance information (confidence factor) associated to each parser will be used to drive the combination process (weighted voting procedure) for the final release of the treebank. An evaluation package will be made freely available (free open source) with the treebank, to enable parsing technology developer to assess the performance of their system against the annotations of the treebank.

WP10 – Preparation of final results for distribution

Coordinator: *ATOLL / INRIA* (with a subcontract to ELDA and the CNRS "Centre National de Ressources Textuelles et Lexicales" (CNRTL) of Nancy)

Objectives based on designed format, ensure possibility to distribute results (individual, merged) at destination of both the scientific community and the indus-

try. The “Centre de compétence” will be in charge of distributing the free material and managing contacts with the research community while ELDA will provide a commercial interface to the SMEs and industry.

B.3.3 Membres permanents impliqués

ATOLL / INRIA

Nom & Prénom	Statut	implication
De la Clergerie Éric	CR1	30%
Pierre Boullier	DR	15%

LIR / LIMSI

Nom & Prénom	Statut	implication
Paroubek Patrick	IR1	60%
Vilnat Anne	MdC	10%
Robba Isabelle	MdC	20%

Langue et Dialogue / LORIA

Nom & Prénom	Statut	implication
Claire Gardent	CR1 CNRS	20%
Azim Roussanaly	MC Nancy 2	20%

LIC2M / CEA-LIST

Nom & Prénom	Statut	implication
de Chalendar Gaël	Chercheur	20%
Ferret Olivier	Chercheur	10%

B.4 Principaux “délivrables”

- **[1]** A WEB site for the project, including a Wiki area, up-to-date documentation of the formats, and access to tools and resources.
- **[2]** A central repository where to store and access corpora, annotations (raw, converted, and possibly validated) and results of acquisition experiments. This repository is to keep a memory of what is done and to provide an easy access to each participant to a maximum of information.
- **[14** A non-validated dependency bank (between 10 to 100 Mwords) that will be extracted from the final version of **[2]** after the second evaluation campaign.
- **[8]** A manually validated dependency bank (around 500K words)

- **[15]** A toolkit to manage dependency banks (storing, browsing, querying) and to perform manual annotation/correction/validation of syntactic annotations
- **[4]** A documentation on the dependency banks (formats, organization, toolkit)
- **[9,20]** A collection of partial lexical resources with some documentation. The pieces of information have vocation to be distributed to be integrated by interested people in lexical resources. This collection could include a lexicon containing both valency and thematic grid information.
- **[19]** A prototype parsing system SEMPARSER integrating the acquired semantic information
- **[10,18]** technical reports to describe the acquisition methodologies, the experiments and an evaluation of the results.
- **[5]** Results of the open evaluation campaign 1 published in scientific conference or journals
- **[16]** Results of the open evaluation campaign 2 published in scientific conference or journals
- **[17]** An evaluation package for free stand-alone self evaluation against the treebank annotations.
- Semestrial short and long reports (SSR **[3,7,12]** & SLR **[6,11,13]**)

	Libellé	Type	Partenaire(s) pilote(s)	Date
1	Web Site	web site	ATOLL	T02
2	Initial Repository	online database	ATOLL	T06
3	SSR	report	ATOLL	T06
4	Intermediary Documentation	report	ATOLL	T12
5	Report evaluation camp. 1	report	LIR	T12
6	SLR	report	ATOLL	T12
7	SSR	report	ATOLL	T18
8	Reference subcorpus	database	LIR	T24
9	Acquired Lexical Resources (intermediate version)	database	LED & ATOLL	T24
10	Report on Acquisition	report	LED & ATOLL & LIC2M	T24
11	SLR	report	ATOLL	T24
12	SSR	report	ATOLL	T30
13	Final Report	report	ATOLL	T36
14	ROVER Corpus	online database	LIR	T36
15	Treebank toolkit mgmt.	software	ATOLL	T36
16	Report evaluation camp. 2	report	LIR	T36
17	Evaluation package	software	LIR	T36
18	Report on Propbank experiment	report	LED	T36
19	Semantic parser	software	LED	T36
20	Acquired Lexical Resources	database	LED & ATOLL	T36

B.5 Résulats escomptés – perspectives

The expected results of the **PASSAGE** project include:

- the emergence of more robust, efficient and accurate linguistic processing chains for French, with a better evaluation of their level of performance through the two scheduled evaluation campaigns.
- the identification of methodologies and protocols to perform linguistic knowledge acquisition tasks. These methodologies should be adaptable for other languages than French, in particular to handle resources poor languages

and overcome the famous bottleneck problem in NLP. The quality of these methodologies will be assessed through partial human evaluation.
- a French dependency treebank for parsing technology improvement, including a large part with merged uncorrected annotations (but an estimation of the error rate) and a smaller manually validated part
- the enrichment of French linguistic resources (grammars, a syntactico-semantic lexicon, a prototype PropBank).
- the consolidation of a strong parsing community in France, familiar with the systematic use of large scale evaluation procedures.

Even if incomplete or partially incorrect, all the linguistic resources resulting from **PASSAGE** (dependency bank, lexicon, prototype PropBank) should be very valuable for the French NLP community.

At a more prospective level, the emergence of several efficient and evaluated parsing systems for French, able to parse large corpora, should boost their use in industrial applications, especially the information extraction ones. Furthermore, we believe that dependency-based representations of parser's output, as advocated in **PASSAGE**, should be a good basis for such applications and for moving forward more semantic representations.

The acquisition techniques explored in **PASSAGE** have vocation to be re-exploited and extended for other languages, in particular at the European level. Furthermore, a strong expertise in both parsing technologies and acquisition techniques could open the way for linguistic knowledge acquisition techniques through transfer, where using multilingual aligned corpora (such as **Europarl**) and good parsers on one side may help designing or improving parsers on the other side. Here again, by marrying symbolic and statistical techniques, we could maybe move forward efficient statistical transfer-based translation techniques.

B.6 Propriété intellectuelle

We are looking forward to producing freely available annotations and derived linguistic resources, at least for academic purpose. Resources will be distributed through the "Centre de Competence CNRS" (ATILF, Nancy) and, through ELDA/ELRA.

References

[1] English verb classes and alternations: A preliminary investigation. University of Chicago Press, Chicago, 1993.
Beth Levin.

[2] Analyse syntaxique probabiliste pour grammaires de dépendances extraites automatiquement. Hdr, Université Paris 7, 2004.
Alexis Nasr.

[3] Fast and scalable HPSG parsing. *Traitement Automatique des Langues (T.A.L.)*, 2005. to appear,
Takashi Ninomiya, Yoshimasa Tsuruoka, Yusuke Miyao, Kenjiro Taura, and Jun´ichi Tsujii.

[4] Création d'un corpus annoté pour le traitement des descriptions définies. *Traitement Automatique des Langues*, 46(1), 2005.
C. Gardent and H. Manuélian.

[5] Generating and selecting grammatical paraphrases. *ENLG, Aberdeen*, 2005.
C. Gardent and E. Kow.

[6] Extracting subcategorisation information from Maurice Gross' grammar lexicon. *Archives of Control Sciences*, 15(LI):253–264, 2005.
C. Gardent, B. Guillaume, G. Perrier, and I. Falk.

[7] Analyse syntaxique profonde à grande échelle: SxLFG. *Traitement Automatique des Langues (T.A.L.)*, 2005. to appear,
Pierre Boullier and Benoît Sagot.

[8] Error mining in parsing results. In *Proc. of COLING-ACL 2006*, July 2006. to appear,
Benoît Sagot and Éric Villemonte de la Clergerie.

[9] Trouver le coupable : Fouille d'erreurs sur des sorties d'analyseurs syntaxiques. In *Proc. of TALN'06*, pages 287–296, 2006. Prix du meilleur papier,
Benoît Sagot and Éric Villemonte de la Clergerie.
[**URL:**`ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/TALN06.pdf`]

[10] The Lefff 2 syntactic lexicon for french: architecture, acquisition, use. In *Proc. of LREC'06*, 2006.
Benoît Sagot, Lionel Clément, Éric Villemonte de la Clergerie, and Pierre Boullier.
[**URL:**`http://atoll.inria.fr/~sagot/pub/LREC06b.pdf`]

[11] Efficient parsing of large corpora with a deep LFG parser. In *Proc. of LREC'06*, 2006.
Benoît Sagot and Pierre Boullier.
[**URL:**`http://atoll.inria.fr/~sagot/pub/LREC06a.pdf`]

[12] Extraction d'information de sous-catégorisation à partir des tables du LADL. In *Actes de La 13ème édition de la conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006)*, 2006.
C. Gardent, B. Guillaume, G. Perrier, and I. Falk.

[13] Intégration d'une dimension sémantique dans les grammaires d'arbres adjoints. In *Actes de La 13ème édition de la conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006)*, 2006.
C. Gardent.

[14] Comment obtenir plus des méta-grammaires. In *Proceedings of TALN'05*, Dourdan, France, June 2005. ATALA.
François Thomasset and Éric Villemonte de la Clergerie.
[**URL:**`ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/mg05.pdf`]

[15] Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658 (© Springer-Verlag), Proceedings of TSD'05*, pages 156–163, Karlovy Vary, Czech Republic, September 2005.
Benoît Sagot.
[**URL:**`http://atoll.inria.fr/~sagot/pub/TSD05.pdf`]

[16] Large scale semantic construction for tree adjoining grammar. In *Proceedings of Logical Aspects in Computational Linguistics*, Bordeaux, France, 2005.
C. Gardent and Y. Parmentier.

[17] Maurice Gross' grammar lexicon and natural language processing. In *Proceedings of the 2nd Language and Technology Conference*, Poznan, Poland, 2005.
C. Gardent, B. Guillaume, G. Perrier, and I. Falk.

[18] MAF: a morphosyntactic annotation framework. In *proc. of the 2nd Language & Technology Conference (LT'05)*, pages 90–94, Poznan, Poland, April 2005.
Lionel Clément and Éric Villemonte de la Clergerie.
[**URL:**`ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/LTC05.pdf`]

[19] « Simple comme EASy :-) ». In *Proceedings of TALN'05 EASy Workshop*, pages 57–60, Dourdan, France, June 2005. ATALA.
Pierre Boullier, Lionel Clément, Benoît Sagot, and Éric Villemonte de La Clergerie.

[**URL:**http://atoll.inria.fr/~sagot/pub/TALN05easyworkshop.pdf]

[20] Chaînes de traitement syntaxique. In *Proceedings of TALN'05*, pages 103–112, Dourdan, France, June 2005. ATALA.
Pierre Boullier, Lionel Clément, Benoît Sagot, and Éric Villemonte de la Clergerie.
[**URL:**http://atoll.inria.fr/~sagot/pub/TALN05easy+sxpipe.pdf]

[21] Morphology based automatic acquisition of large-coverage lexica. In *Actes de LREC'04*, pages 1841–1844, Lisbonne, Portugal, May 2004.
Lionel Clément, Benoît Sagot, and Bernard Lang.

[22] A uniform method of grammar extraction and its applications. In *the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, Hong Kong, October 2000.
Fei Xia, Martha Palmer, and Aravind Joshi.

C Annexe: curriculum vitae des membres du projet

C.1 Éric de la Clergerie (*ATOLL / INRIA*)

Éric de la Clergerie defended his PhD in 1993 and is now Research Officer (CR) at INRIA. Since 2002, he is the scientific leader of project-team ATOLL (Sotware Tools for Natural Language Processing). His researches focus on the development of parsing techniques for several grammatical formalisms. With ATOLL, he also got involved in the development of wide coverage linguistic resources for French (lexicon, grammar), exploited during the parsing evaluation campaign **EASy/EVALDA**. He has participated to several national actions (Technolangue/Normalangue-RNIL & Evalda-EASy; ACI MD Biotim; INRIA ARCs RLT, GENI & MOSAIQUE; ILF Lexsynt). In particular, he was coordinator for RLT and is co-coordinator for LexSynt. He is also involved in the international standardization efforts within ISO TC37 SC4 where he is the editor of the MAF proposal (Morpho-syntactic Annotation Framework, CD 24611), has actively contributed to the FSR proposal (Feature Structure Representation), and has acted as head of the French delegation for several international ISO meetings.

Éric de la Clergerie is also a member of the editorial committee of the French review T.A.L. (Traitement Automatique des Langues) and was guest editor of this review for a thematic issue on "*Evolutions in Parsing*".

References

[23] Trouver le coupable : Fouille d'erreurs sur des sorties d'analyseurs syntaxiques. In *Proc. of TALN'06*, pages 287–296, 2006. Prix du meilleur papier, Benoît Sagot and Éric Villemonte de la Clergerie.
[**URL:**`ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/TALN06.pdf`]

[24] Comment obtenir plus des méta-grammaires. In *Proceedings of TALN'05*, Dourdan, France, June 2005. ATALA.
François Thomasset and Éric Villemonte de la Clergerie.
[**URL:**`ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/mg05.pdf`]

[25] Chaînes de traitement syntaxique. In *Proceedings of TALN'05*, pages 103–112, Dourdan, France, June 2005. ATALA.
Pierre Boullier, Lionel Clément, Benoît Sagot, and Éric Villemonte de la Clergerie.
[**URL:**`http://atoll.inria.fr/~sagot/pub/TALN05easy+sxpipe.pdf`]

C.1 Pierre Boullier (*ATOLL / INRIA*)

Pierre Boullier is a Research Director (DR) at INRIA and a member of project-team ATOLL. He is a world-wide recognized specialist in parsing and has developed SYNTAX, a very efficient parser compiler for CFGs that has been extended to handle Lexical Functional Grammars (LFGs), resulting in the SXLFG parser compiler. An LFG parser for French built with a preliminary version of SXLFG was exploited during the parsing evaluation campaign EASy.

References

[26] Analyse syntaxique profonde à grande échelle: SxLFG. *Traitement Automatique des Langues (T.A.L.)*, 2005. to appear,
Pierre Boullier and Benoît Sagot.

[27] Efficient and robust LFG parsing: SxLfg. In *Proceedings of IWPT'05*, pages 1–10, Vancouver, Canada, October 2005.
Pierre Boullier and Benoît Sagot.
[**URL:**`http://atoll.inria.fr/~sagot/pub/IWPT05.pdf`]

[28] Chaînes de traitement syntaxique. In *Proceedings of TALN'05*, pages 103–112, Dourdan, France, June 2005. ATALA.
Pierre Boullier, Lionel Clément, Benoît Sagot, and Éric Villemonte de la Clergerie.
[**URL:**`http://atoll.inria.fr/~sagot/pub/TALN05easy+sxpipe.pdf`]

C.1 Patrick Paroubek (*LIR / LIMSI*)

- CNRS September 2002 to present, Research Engineer in the Language, Information and Representation group of the Human-Machine Communication Department, where my activities have regained their previous focus on basic Natural Language Processing (POS tagging, parsing), particularly in the context of evaluation, but now address also machine understanding (human-machine dialog) and language emergence in communities of artificial autonomous entities .
- Limsi - CNRS July 1997 to September 2002 Research Engineer in the Spoken Language Processing group of the Human-Machine Communication department where I have been involved in the DISC DISC European project (best practice in Spoken Language Dialog System Engineering), the ELSE preparatory action (crafting of a generic blue print for evaluating Natural Language Processing Systems using a semi-automatic quantitative black-box approach), the CLASS project and in the following national projects: GRACE, MULTITAG (production of a tagged corpus of 1 million words

from the GRACE results, in coordination with the CLIFF project), SILFIDE, TECHNOLANGUE scientific committee.

- INaLF - CNRS - November 1993 to June 1997 Research engineer at Institut National de la Langue Francaise (INaLF) a unit of Centre National de la Recherche Scientifique (CNRS). My activities there included some system administration but dealt mainly with dictionary computerization, graphic interfaces design and implementation, evaluation of Part Of Speech taggers (coordinator of the french GRACE action), as well as building, management, automatic analysis and distribution (WWW access) of large text corpora. Contribued to the French Corpus task of the European project LE2-4017-LE-PAROLE, after having participated to the preparatory project PAROLE (MLAP-63-386).
- Rouen University, Lecturer at Laboratoire d'Informatique de Rouen (LIR)- October 1992 to September 1993.
- University of Pennsylvania, Invited researcher - July 1992 to September 1992.
- Computers Communications and Visions (c2v), Research Engineer - February 1992 to May 1992.
- BULL SA FRANCE, Software Engineer - October 1991 to December 1991.
- Rouen University, Lecturer at Laboratoire d'Informatique de Rouen (LIR)- February 1991 to September 1991.
- University of Pennsylvania, Post-doctoral research. February 1990 to December 1990.
- University Malaya, Kuala Lumpur, Malaysia, Lecturer - November 1986 to December 1987.

C.2 Anne Vilnat (*LIR / LIMSI*)

Anne Vilnat, is Maître de Conférences (HDR) at University Paris Sud since January 1988 and was in charge of the research topic "Analysis Processes, Generation and Dialogue" of the LIR group of LIMSI-CNRS, up to June 2005. She has already co-directed 6 PhD. thesis and has got her "Habilitation à diriger les recherches" in December 2005. Her memoire is entitled "Dialogue et analyse de phrases". She got her PhD. in Computer Science from University Paris 6 in 1984 and since has made 73 scientific publications. She participated in the organizationof the EASY evaluation campaign, for which she designed the reference formalism. Her most recent publications include:

- Brigitte Grau, Olivier Ferret, Martine Hurault-Plantet, Laura Monceaux, Isabelle Robba, Anne Vilnat, Christian Jacquemin,
  Coping with Alternate Formulations of Questions and Answers, chapitre dans Advances in Open-Domain Question-Answering, Strzalkowski & Harabagiu (eds), Series Text, Speech and Language Technology, Vol.32, Springer, Juillet 2006.

- Patrick Paroubek, Isabelle Robba, Anne Vilnat et Christelle Ayache,
  Data, Annotations and Measures in EASY, the Evaluation Campaign for Parsers of French, LREC 2006, Gênes, Italie, Mai 2006.
- Christelle Ayache, Brigitte Grau et Anne Vilnat,
  EQueR : the French Evaluation Campaign of Questions Answering Systems LREC 2006, Gênes, Italie, Mai 2006.
- Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Anne Vilnat et Laura Monceaux
  FRASQUES: A Question-Answering System in the EQueR Evaluation Campaign LREC 2006, Gênes, Italie, Mai 2006.
- Anne-Laure Ligozat, Brigitte Grau, Isabelle Robba et Anne Vilnat
  Evaluation and Improvement of Cross-Lingual Question Answering Strategies, WS MLQA de EACL 2006, Trento, Italie, Mai 2006.
- Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Anne Vilnat,
  Term Translation Validation by Retrieving Bi-terms, LNCS, (Version étendue du WS Clef 05) 2006.
- Anne-Laure Ligozat, Brigitte Grau, Isabelle Robba et Anne Vilnat
  L'extraction des réponses dans un système de question-réponse, TALN, Leuven, Avril 2006.
- Anne-Laure Ligozat, Brigitte Grau, Isabelle Robba et Anne Vilnat
  Question-Réponse multilingue : évaluation et amélioration des stratégies de changement de langue, CORIA, Lyon, Mars 2006.

C.3 Isabelle Robba (*LIR / LIMSI*)

Isabelle Robba is Maître de Conférences at IUT d'Orsay where she teaches basic Computer Science. She got her PhD. in Computer Science from University Paris Sud in 1992. The title is "L'étude des mécanismes de raisonnement par analogie dans l'analyse de phrase: le système MIRA". She was involved in the EASY evaluation campaign as an organizer and worked more particulary on the evaluation protocole and reference annotations.

Her most recent publications include:

- Anne Vilnat, Patrick Paroubek, Laura Monceaux, Isabelle Robba, Véronique Gendner, Gabriel Illouz, Michèle Jardino,
  The Ongoing Evaluation Campaign of Syntactic Parsing of French: EASY, actes de la Fourth International Conference on Language Resources and Evaluation (LREC), Lisboa, May 2004, vol. 6, pp 2023-2026.
- V. Gendner, G. Illouz, M. Jardino, L. Monceaux, P. Paroubek, I. Robba, A. Vilnat
  PEAS, the first instanciation of a comparative framework for evaluating parsers of French, actes de la 10th Conference of The European Chapter of the Assciation for Computational Linguistics, Budapest, Hungary, April 12-17 2003, Companion Volume, pp 95-98.

- Véronique Gendner, Gabriel Illouze, Michèle Jardino, Laura Monceaux, Patrick Paroubek, Isabelle Robba, Anne Vilnat
  A Protocol for Evaluating Analyzers of Syntax (PEAS) in Proceedings of the Third International Conference on Language Ressources and Evaluation (LREC), Las Palmas de Gran Canaria, Spain, May 27th-June 2nd 2002, vol 2, pp 590-597

C.4 Claire Gardent (*Langue et Dialogue / LORIA*)

Claire Gardent graduated in linguistics at the University of Toulouse in 1986, obtained an MSc in Artificial Intelligence from the University of Essex in 1987 and defended a PhD in Cognitive Science at the University of Edinburgh in 1991. From 1991 to 2000, she worked as a researcher for the Universities of Utrecht and Amsterdam (The Netherlands), Clermont-Ferrand (France) and Sarrebruecken (Germany). Since 2000, she is a Chargée de Recherche de 1ère classe at the CNRS.

Claire Gardent's research focuses on the computational treatment of natural language meaning i.e., on computational semantics. She has worked on the role of inference in Natural Language Processing (NLP) and is currently focusing on developing a computational infrastructure for the semantic treatment of French.

Claire Gardent has published a book on analysis and generation (with Karine Baschung) and about 50 articles in (mainly international) journals and conference proceedings. She has been nominated Chair of the European Chapter for the Association of Computational Linguistics, editor in chief of the french journal "Traitement Automatique des Langues" and member of the editorial board of the journals "Computational Linguistics", "Journal of Semantics". Each year she is on the programme committee for half a dozen international conferences or workshops. She also acted as scientific chair for various international conferences, workshops and summer schools (e;g., ESSLLI, the European Summer School for Logic, Language and Information).

References

[29] Création d'un corpus annoté pour le traitement des descriptions définies. *Traitement Automatique des Langues*, 46(1), 2005.
C. Gardent and H. Manuélian.

[30] Extracting subcategorisation information from Maurice Gross' grammar lexicon. *Archives of Control Sciences*, 15(LI):253–264, 2005.
C. Gardent, B. Guillaume, G. Perrier, and I. Falk.

[31] Large scale semantic construction for tree adjoining grammar. In *Proceedings of Logical Aspects in Computational Linguistics*, Bordeaux, France, 2005.
C. Gardent and Y. Parmentier.

C.1 Azim Roussanaly (*Langue et Dialogue / LORIA*)

Since 1988, Azim Roussanaly is Assistant Professor (Maître de conférences) at University Nancy2 and a member of the "Langue et Dialogue" project-team at LORIA . He conducts his researches in the area of the Natural Language Processing (NLP) and, for the last years, his works are mainly focused on parsing techniques especially for the Lexicalized Tree Adjoining Grammars (LTAG) formalism in the context of a wide coverage French grammar. He has developed LLP2, a LTAG-based parser for French which was exploited during the parsing evaluation campaign EASy.

References

[32] Représentation et gestion de grammaires TAG. *Traitement Automatique des Langues (T.A.L.)*, 2004.
B Crabbé, B Gaiffe, and A. Roussanaly.

[33] Premier bilan de la participation du LORIA à la campagne d'évaluation EASy. In *Proceedings of TALN'05 EASy Workshop*, Dourdan, France, June 2005. ATALA.
A. Roussanaly, B. Crabbé, and J. Perrin.

[34] A new metagrammar compiler. In *TAG+6 (International Workshop on Tree Adjoining Grammars and Related Frameworks)*, Venice, Italy, May 2002.
B Crabbé, B Gaiffe, and A. Roussanaly.

C.1 Gaël de Chalendar (*LIC2M / CEA-LIST*)

Gaël de Chalendar is a researcher at CEA in Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue (LIC2M) CEA-LIST, Centre de Fontenay-aux-Roses. Gael got his PhD. in computer science in 2001 from University of Paris Sud. The title is: "Généralisation de Graphes Conceptuels à l'aide d'Heuristiques et Apprentissage de Relations Sémantiques entre Concepts". From 2001 to 2002, Gael was Attaché Temporaire d'Enseignement et de Recherche at IUT d'Orsay where he tought basic computer science and natural language processing.

Publications:

- Romaric Besançon et Gaël de Chalendar (2005)
  L'analyseur syntaxique de LIMA dans la campagne d'évaluation EASY, Actes de la 12ème conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2005), 6-10 juin 2005, Dourdan, France.
- Romaric Besançon, Gaël de Chalendar, Olivier Ferret, Christian Fluhr, Olivier Mesnard et Hubert Naets (2004)
  Concept-Based Searching and Merging for Multilingual Information Retrieval: First Experiments at CLEF 2003, 4th Workshop of the Cross- Language Evaluation Forum, Springer Verlag.

- Gaël de Chalendar, Tiphaine Dalmas, F. Elkateb-Gara, Olivier Ferret, Brigitte Grau, Martine Hurault-Plantet, Gabriel Illouz, Laura Monceaux, Isabelle Robba, Anne Vilnat (2002)
  The Question Answering System QALC at LIMSI, Experiments in Using Web and WordNet. TREC 2002.
- Gaël de Chalendar, Brigitte Grau and Olivier Ferret (2000)
  A cost-bounded algorithm to control events generalization, in Conceptual Structures : Logic, Linguistic, and Computational Issues. 8th International Conference on Conceptual Structures ICCS, Springer, Bernhard Ganter and Guy W. Mineau editors, Lectures Notes in Artificial Intelligence, vol. 1867, pages 555-568
- Gaël de Chalendar and Brigitte Grau (2000)
  SVETLAN' or how to Classify Words using their Context, Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2000, Springer, Rose Dieng and Olivier Corby eds., Lectures notes in artificial intelligence vol. 1937, pages 203-216

C.2 Olivier Ferret (*LIC2M / CEA-LIST*)

Olivier Ferret is a researcher at CEA in the Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue (LIC2M) CEA-LIST, Centre de Fontenay-aux-Roses. He got his PhD. in computer science in 1998 from University Paris Sud. The title is: "ANTHAPSI : un système d'analyse thématique et d'apprentissage de connaissances pragmatiques fondé sur l'amorçage". From 2000 to 2001, he did a post-doc at Direction des Technologies de l'Information of CEA, in the context of the PRISME project about information filtering. In 1999 he was research engineer in AEGIS company for the project Eurêka PVS 98 about multi-agent based flexible information systems.

Publications:

- Olivier Ferret (2004)
  Discovering word senses from a network of lexical cooccurrences, COLING 2004, Genève.
- Sana Châar, Olivier Ferret et Christian Fluhr (2004)
  Filtrage pour la construction de résumés multi-documents guidée par un profil, revue TAL. Olivier Ferret (2002) Using collocations for topic segmentation and link detection, COLING 2002, Tapei.
- Olivier Ferret et Brigitte Grau (2002)
  A Bootstrapping Approach For Robust Topic Analysis, Journal of Natural language Enginering (NLE), Special issue on robust methods of corpus analysis.
- O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, C. Jacquemin, L. Monceaux, I. Robba, A. Vilnat (2002)
  How NLP Can Improve Question Answering, Knowledge Organization, Vol. 29, N°3-4, p. 135-155

C.3 Gil Francoploulo

Gil Francopoulo got a PhD thesis at University Paris-6 in 1988 on "induction of grammar rules" at CNRS-LIMSI laboratory. He works in lexicon management, sentence parsing and search engines for 20 years: 4 years as Tagmatica Director (www.tagmatica.com), 12 years for LexiQuest (formely ERLI and GSI-ERLI) and 4 years for various banks and software companies. He was co-author of the Genelex lexicon model in 1991 and managed the five languages lexicons within LexiQuest. He participated in several National and International projects (Eureka/Genelex, Eagles, eContent/LIRICS, Technolangue/Normalangue-RNIL, Evalda-Easy, ILF Lexsynt). Currently, he is editor of the ISO standard for the lexicons dedicated to Natural Language Processing (aka Lexical Markup Framewrok, LMF). He is convenior of the ISO group for the management of morpho-syntactic data categories within ISO-12620 revision. He participates in the work in progress about SynAF (ISO WD24615). He is liaison officer between ISO-TC37/SC4 and W3C.

Recent Publications :

- Francopoulo G., George M. *Lexical Markup Framework*. ISO-CD24613 rev-9. 2006 ISO Geneva.
- Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. *Lexical Markup Framework* (LMF). LREC-2006
- Francopoulo G., Declerck T., Monachini M., Romary L. *The relevance of standards for research infrastructures*. LREC-2006
- Romary L., Salmon-Alt S., Francopoulo G. *Standards going concrete: from LMF to Morphalou*. COLING-2004 workshop Enhancing and using electronic dictionaries.