Large scale production of syntactic annotations for French

Éric Villemonte de la Clergerie, Christelle Ayache, Gaël de Chalendar, Gil Francopoulo, Claire Gardent, Patrick Paroubek

Large scale production of syntactic annotations for French

Éric Villemonte de la Clergerie
INRIA Paris-Rocquencourt
Eric.De_La_Clergerie@inria.fr

Christelle Ayache
ELDA
ayache@elda.org

Gaël de Chalendar
CEA-LIST
Gael.de-Chalendar@cea.fr

Gil Francopoulo
TAGMATICA
gil.francopoulo@wanadoo.fr

Claire Gardent
CRNS/LORIA
Claire.Gardent@loria.fr

Patrick Paroubek
CNRS/LIMSI
Patrick.Paroubek@limsi.fr

Abstract

We present the motivations and objectives of French **Passage** project that ambitions the *large scale production of syntactic annotations* by repeatedly combining the outputs of 10 French parsing systems.

1 Introduction

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 (NLP). 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[1]), the British National Corpus (100M words covering various styles annotated with parts of speech[2]), 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`.

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.

With the **Passage** project, we believe that a new option becomes possible. Funded by the French ANR program on Data Warehouses and Knowledge, **Passage** is a 3-year project (2007–2009), coordinated by INRIA project-team Alpage. Its main objective is the large scale production of syntactic annotations to move forward (*Produire des annotations syntaxiques à grande échelle*). It builds up on the results of the **EASy** French parsing evaluation campaign, funded by the French Technolangue program, which has shown that French parsing systems are now available, ranging from shallow to deep parsing. Some of these systems were neither based on 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 (1M words). Preliminary experiments made by some of the participants with deep parsers (Boullier and Sagot, 2005) indicate that processing more than 10 Mwords is not a problem, especially by relying on clusters of machines. These figures can even be increased for shallow parsers. In other words, there now exists several French parsing systems that could parse (and re-parse if needed) large corpora between 10 to 100M words.

Passage aims at pursuing and extending the line of research initiated by the **EASy** campaign. Its main objective is to use 10 of the parsing systems that have participated to EASy. They will be used to parse and possibly re-parse a French corpus of more

[1] `http://www.cis.upenn.edu/~treebank/`
[2] `http://www.natcorp.ox.ac.uk/`

than 100 Mwords. More precisely, as illustrated by Figure 1 the proposed 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. (going back to step 1)

These objectives should be helped by running two new evaluation campaigns to precisely assess the quality of the available French parsers and by using the information to combine, through a *ROVER* (*Recognizer output voting error reduction*), the annotations produced by parsing a very large corpus. Furthermore, a subcorpus of around 500K words shall be manually validated to get an even better idea of the quality of the rover annotation set. Methodologies shall also be deployed for evaluating the quality of the acquired resources.

We believe the **Passage** project should help seeing the emergence of linguistic processing chains exploiting richer lexical information, in particular semantic ones. 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.

The remaining of the paper presents the various components and tasks of **Passage**.

2 Selecting and cleaning corpora

While not looking for a perfect distribution of existing French styles, the corpora used for **Passage** will provide a relatively large diversity of styles (including oral transcriptions) totalling over 100 Mwords. Corpora are selected for their style but also for their possibility to be freely available (or, at least, available at reasonable cost). The current selection is not yet fully closed but should include:

- the **EASy corpora** (1M words). This corpus used for the EASy campaign already cover various styles and includes a subset of around 4K sentences (76K words) that have been manually validated.
- **Wikipedia Fr**, a freely available corpus of almost 500K entries (estimated to 86M words) covering many domains of knowledge and collectively written by many authors, with various styles though biased toward descriptions.
- **Wikisources**, a collection of several thousand freely available French texts covering various thematics (estimated 84M words).
- **Wikilivres**, a collection of 1956 freely available educational French books (estimated 800K words).
- **Monde Diplomatique**, a low cost journalistic corpus already used by various teams and covering many thematics (18M words).
- **FRANTEXT**, a collection of 500 digitised French books such as Jule Vernes's novels ($\sim$ 20M words).
- **Europarl**[3], a corpus of parallel multilingual texts extracted from the proceedings of the European Parliament (28M words for French)
- **The JRC-Acquis Multilingual Parallel Corpus**[4] which is is the total body of European Union laws (39M words for French).
- **Ester**, a corpus of oral transcriptions (1M words)

The size estimation of this corpus selection is already over 270M words. While the primary objective of **Passage** is to parse the corpora, it should be obvious that they also provide longer terms perspectives, such as the acquisition of knowledge from Wikipedia or the transfer of linguistic knowledge from the multilingual aligned corpora (Europarl, JRC-Acquis, and, to some extent, wikipedia).

Given the diversity of parsing systems to be used, we have to assume the simpler possible textual format for the corpora, hence involving to clean them by removing HTML/XML markup elements, wiki syntax, and meta-data. However, the cleaning process should keep traces of some of these removed pieces of information in companion files, using standoff pointers to refer to segments of the cleaned versions. The companion files will be available to

[3] http://www.statmt.org/europarl/

[4] http://langtech.jrc.it/JRC-Acquis.html

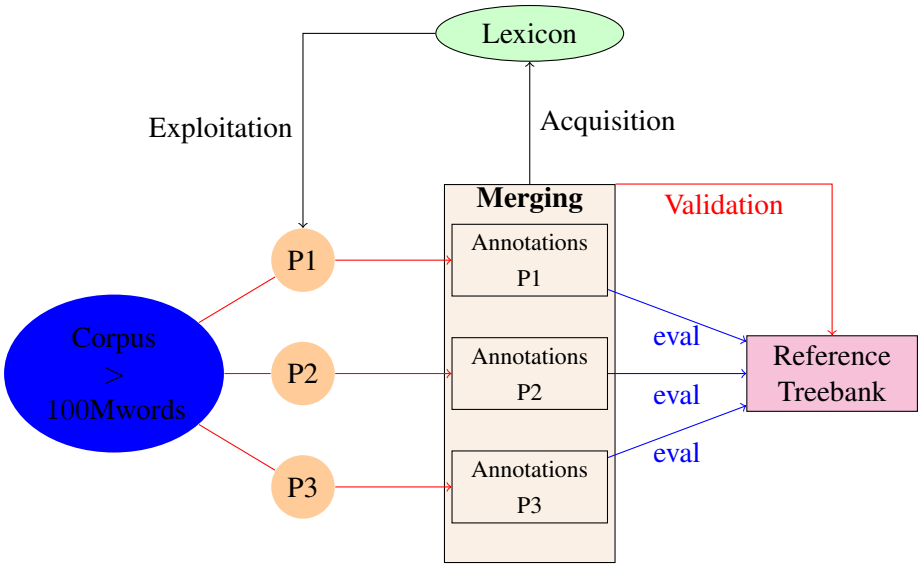

Figure 1: A bootstrap model for **Passage**

the parsing systems that can exploit them.

3 The parsing systems

The participation of 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 **Passage**. The parsing systems are provided by participants or contractants, including:

- FRMG, an hybrid TIG/TAG parser derived from a metagrammar, developed at INRIA (Villemonte de La Clergerie, 2005b; Thomasset and Villemonte de la Clergerie, 2005; Boullier et al., 2005)
- SXLFG, a very efficient LFG-based parser, developed at INRIA (Boullier and Sagot, 2005; Sagot and Boullier, 2006; Boullier et al., 2005);
- LLP2, a TAG parser also derived from a metagrammar, developed at LORIA (Roussanaly et al., 2005);
- LIMA, developed at *LIC2M / CEA-LIST*[5] (Besancon and Chalendar, 2005);
- TAGParser, an hybrid statistical/symbolic parser developed by Gil Francopoulo (TAGMATICA[6]) (Francopoulo, 2005);
- SYNTEX, a rule based parser, developed by Didier Bourigault at ERSS[7];
- Two parsers based on Property Grammars, developed at [LPL][8] and using constraint satisfaction (Blache, 2005). The first one is symbolic and deterministic while the second one is statistical and trained thanks to the results of the parsers during the Easy campaign (Vanrullen et al., 2006);
- CORDIAL, a commercial parser developed by SYNAPSE[9];
- SYGMART, a parser developed at LIRMM[10];
- XIP, a cascade rule-based parser developed at Xerox Research Center Europe[11], (Aït-Mokhtar et al., 2002)

It may be noted that these parsing systems are based on very different paradigms and produce different kinds of output. While keeping their specificities, the parsers will be compared using a common syntactic annotation format and this experience by itself should provide useful information about

[5] http://www-list.cea.fr/

[6] www.tagmatica.com
[7] http://www.univ-tlse2.fr/erss/
[8] http://cnrs.oxcs.fr/
[9] http://www.synapse-fr.com/
[10] http://www.lirmm.fr/xml/fr/lirmm.html
[11] http://www.xrce.xerox.com/

the expected requirements of a syntactic annotation standard.

4 Extending the EASy annotation format

For the **EASy** project, a mixed constituency/dependency XML annotation format was designed providing information about:

- a segmentation of corpora into **sentences**;
- a segmentation of sentences into **forms**;
- non-recursive **chunks** embedding forms and typed with a type of Table 1(a);
- labeled **dependencies** that are anchored by either forms or chunks. All dependencies have a binary arity but for COORD dependencies that are usually ternary (and sometimes binary where there is no leftwards coordinated anchor). The 14 kinds of dependencies are listed in Table 1(b).

Type	Explanation
GN	Nominal Chunk
NV	Verbal Kernel
GA	Adjectival Chunk
GR	Adverbial Chunk
GP	Prepositional Chunk
PV	Prepositional non-tensed Verbal Kernel

(a) Chunks

Type	Anchors	Explanation
SUJ-V	suject,verb	Subject-verb dep.
AUX-V	auxiliary, verb	Aux-verb dep.
COD-V	object, verb	direct objects
CPL-V	complement, verb	other verb arguments/complements
MOD-V	modifier, verb	verb modifiers (such as adverbs)
COMP	complementizer, verb	subordinate sentences
ATB-SO	attribute, verb	verb attribute
MOD-N	modifier, noun	noun modifier
MOD-A	modifier, adjective	adjective modifier
MOD-R	modifier, adverb	adverb modifier
MOD-P	modifier, preposition	prep. modifier
COORD	coord., left, right	coordination
APPOS	first, second	apposition
JUXT	first, second	juxtaposition

(b) Dependencies

Table 1: EASy format

For **Passage**, the **EASy** format should be enriched, in particular to be closer to the emerging ISO TC37 SC4 standards (Ide et al., 2003). Forms should be built upon tokens referring spans of the original documents through standoff pointers, following the *Morphosyntactic Annotation Framework* [MAF] proposal (Clément and Villemonte de La Clergerie, 2005). Besides chunks, constituency should be completed by allowing nested recursive groups as proposed in the *Syntactic Annotation Framework* [SynAF], following the TIGER model. Structured content represented by feature structures relying on a common tagset such as MULTEXT (Ide and Romary, 2001) could be attached to forms, groups, and possibly dependencies.

A rather complete annotation guide has been developed for **EASy** and will be extended for **Passage**, taking into account the extensions of the format.

5 Evaluating the parsing systems

Two evaluation campaigns will be run during the project. The first one will take place before the end of 2007 and mostly reuse the annotated data used by the **EASy** campaign, with the addition of 400 freshly annotated sentences. The campaign 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).

The second campaign will take place at the end of **Passage** (in 2009) on the manually annotated reference corpus (Section 7) and use the enriched syntactic annotation format. This campaign should show the evolutions of the parsers during the project, and, in particular, their capacity to integrate the linguistic knowledge that will be acquired (Section 8).

For this second campaign, we will try to avoid two of the main problems of **EASy**, namely the explicit segmentations into forms and sentences. The form segmentation will be parser-dependent, but the use of span-referred tokens, completed by dynamic alignment techniques, should allow us to align the forms. The notion of sentence will also be more dynamic and derive from the dependencies and groups, a sentence being a connected set of dependencies from a main governor.

6 Combining Annotations

Each parsing system has its weaknesses and strengths. To get more accurate annotations, we therefore plan to combine (merge), using a rover, the results produced by all the parsers involved in **Passage** as advocated by others (Henderson and Brill, 1999; Brunet-Manquat, 2004; Sagae and Lavie, 2006).

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 (Brunet-Manquat, 2004). This combination process requires to assess the performance level of the different parsers involved in order to compute a confidence factor associated to the annotations provided by each parser, possibly at the level of syntactic phenomena. Combination should be based on majority voting, pondered by the confidence factor, and taking into account, topological and integrity constraints over dependencies.

The quality of the rover will also be assessed through the manual validation of a reference corpus (Section 7) and through the use of feedback indicators. For example, if it is observed that the confidence factor selects the annotations of a system very often in contradiction with the majority of the other systems, an alarm should be raised, possibly leading to a re-evaluation of the confidence factor. More generally, when needed, automatic error correction scripts will be developed to improve the data produced by individual parsers for their most frequent and systematic errors.

7 Validating a reference subcorpus

A sub-corpus of 500K words will be selected and stabilized through human correction of the rover annotations. Human validation will be checked using inter-annotator agreement measures performed on randomly selected excerpts of corpus (amounting to 10% of the corpus). Annotating 500K words is an burdensome task but we assume that starting from the rover annotations should greatly help. The annotation process should itself provide feedback information to evaluate and improve the rover annotations (by running the rover again). Methodology and specific software such as EASYREF (Section 10) for hand annotation/correction will be investigated in order to speedup the annotation task with, in particular, consistency checking tools.

The reference annotations 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).

8 Acquiring linguistic knowledge

While the quality of the analysis produced by these parsers remains to be assessed and improved during **Passage**, it should already be possible to learn valuable linguistic knowledge from the analysis of a large corpus as advocated by others (*Deriving Linguistic Resources from Treebanks*[12]; LREC'02 Workshop[13] on "*Linguistic Knowledge Acquisition and Representation : Bootstrapping Annotated Data*").

Various techniques will be explored to extract information from the syntactically annotated corpora resulting from the parsing process, with the ambition to prepare the creation of a knowledge rich lexicon for French. The idea is to first derive valency information, then use this information and its lexical distribution to create Beth Levin's type alternation classes (Levin, 1993) 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* (Sagot et al., 2006), the Synlex lexicon derived from the LADL tables (Gardent et al., 2005b; Gardent et al., 2005a; Gardent et al., 2006) and Patrick Saint-Dizier's manually constructed alternation classes (Saint-Dizier, 1999).

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

- weighted selectional restrictions for disambiguation (van Noord, 2007)

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

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

- semantic classes, using Harris distributional hypothesis over syntactic contexts
- derivational morphology with transfer of syntactic information
- probabilities of syntactic constructions
- extraction of stochastic grammars (Xia et al., 2000; Nasr, 2004)

Human validation by expert linguists remains an important issue, first to assess the quality of the acquisition techniques but also because fully 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 theories are necessary to guide the acquisition experiments. Part of the objectives of **Passage** is to understand how this expertise may efficiently take place through adequate validation interfaces, as tried for error mining (Sagot and Villemonte de La Clergerie, 2006).

9 Integrating knowledge

The knowledge thus acquired is meant to be integrated in some of the parsing systems to make them more accurate. Hopefully entering a virtuous circle, corpora may then be re-parsed to learn new knowledge.

The parsers may be improved by acquiring probabilistic information for disambiguation 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).

Recently, suggestions have been made to marry symbolic and statistics approaches (Ninomiya et al., 2005). 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. **Passage** should help us going into this direction by:

1. providing ways to validate the lexical information that has been acquired

2. integrating some or all of this NLP lexicon into at least one parsing system.

In particular, a prototype parsing system used will build on the SEMTAG system (Gardent and Parmentier, 2005; Gardent, 2006) which consists of a lexicon, a Tree Adjoining Grammar integrating syntax and semantics and DYALOG parser (Villemonte de La Clergerie, 2005a). The lexicon will be extended with the syntactic (valency) and semantic (thematic grid) information acquired from corpora, the grammar will be extended to deal with constructions not yet covered and corpus extracted probabilistic information used to disambiguate the parser results.

The resulting parsing system will then be used to semi-automatically create a prototype **PropBank**[14] 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.

The aim 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.

10 Deploying an infrastructure

Many teams are involved in **Passage** with several tasks to be conducted, which requires an excellent coordination, for instance to ensure interoperability. A shared and solid infrastructure is needed to access, process, compare, view, discuss, validate, exploit and distribute syntactic annotations. Existing annotation platforms such as AGTK, Mate, Atlas, ... are investigated but we have also started the development of EASYREF, a lightweight WEB-based collaborative environment to handle syntactic annotations. We indeed believe that a server-based infrastructure will be needed with a server powerful enough to serve large annotated corpora. Scalability issues will obviously have to be solved, for instance

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

by building good indexes and using efficient XML technologies such as XML databases.

Parsing several hundred million words remains a difficult tasks for most parsers. An obvious solution seems to distribute the load on clusters of machines (Ninomiya et al., 2005). While already tried by some of the **Passage** participants, this idea will be further explored and possibly tried on large clusters (such as the French GRID 5000[15]), implying the use or development of specific tools and environments.

11 Conclusion

It is much too early to judge the results of **Passage** but we believe that this project proposes a pertinent methodology to bootstrap the creation of large annotated corpora. It relies on the rather unique long-term cooperation of 10 French parsing systems and the expertise of the **EASy** evaluation campaign. The project should prove that is now possible to make parsing systems cooperate through an interchange syntactic annotation format and to use the resulting annotations to acquire new linguistic knowledge, hence entering a virtuous circle. Even if human interaction remains important, annotated corpora should then become larger and more dynamic, evolving with the improvement of the parsing systems. It would be an encouragement for other parsing systems (in France and worldwide) to join the process, in order to progressively and simultaneously develop better systems and linguistic resources for French. To achieve this last ambition, the resources developed within **Passage**, such as annotated corpora and lexical resources, will be made freely available, if possible through adequate collaborative interfaces.

Acknowledgments

This research is supported by the French National Research Agency (ANR) in the context of the Passage project (ANR-06-MDCA-013).

References

S. Aït-Mokhtar, J.-P. Chanod, and C. Roux. 2002. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(3):121–144.

R. Besancon and G. De Chalendar. 2005. L'analyseur syntaxique de LIMA dans la campagne d'évaluation EASy. In *Proceedings of TALN'05*, Dourdan, France. ATALA.

Philippe Blache. 2005. Property grammars : A fully constraint-based theory. In & J.Villadsen In H. Christiansen, P. Skadhauge, editor, *Proc. of Constraint Satisfaction and Language Processing (CSLP)*. Springer-Verlag.

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

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

Francis Brunet-Manquat. 2004. Syntactic parser combinaison for improved dependency analysis. In *Proceedings of ROMAND-2004, Workshop COLING*, pages 24–31, Geneva, Switzerland.

Lionel Clément and Éric Villemonte de La Clergerie. 2005. MAF: a morphosyntactic annotation framework. In *proc. of the 2nd Language & Technology Conference (LT'05)*, pages 90–94, Poznan, Poland.

Gil Francopoulo. 2005. TagParser et technolangue-easy. In *Proc. Atelier EASy, 12th Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2005)*, Dourdan, France.

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

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

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

C. Gardent, B. Guillaume, G. Perrier, and I. Falk. 2006. 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)*.

C. Gardent. 2006. 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)*.

[15] https://www.grid5000.fr/mediawiki/index.php/Grid5000:Home

J. Henderson and E. Brill. 1999. Exploiting diversity in natural language processing: combining parsers. In *Fourth Conference on Empirical Methods in NLP*, College Park MD.

Nancy Ide and Laurent Romary. 2001. A common framework for syntactic annotation. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 306–313, Morristown, NJ, USA. Association for Computational Linguistics.

N. Ide, L. Romary, and Éric Villemonte de La Clergerie. 2003. International standard for a linguistic annotation framework. In *Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology*. Journal version submitted to the special issue of JNLE on Software Architecture for Language Engineering.

Timo Järvinen. 1994. Annotating 200 millions words: the bank of english project. In *Proceedings 15th COLING*, pages 565-568, Kyoto.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

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

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

Patrick Paroubek, Isabelle Robba, Anne Vilnat, and Christelle Ayache. 2006. Data, annotations and measures in EASY - the evaluation campaign for parsers of french. In ELRA, editor, *proc. of LREC'06*, pages 315–320, Genoa, Italy.

Patrick Paroubek, Anne Vilnat, Isabelle Robba, and Christelle Ayache. 2007. Les résultats de la campagne easy d'évaluation des analyseurs syntaxiques du français. In *Actes de la 14ème conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN)*, volume 2, pages 243–252, Toulouse, France.

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

K. Sagae and A. Lavie. 2006. Parser combination by reparsing. In *Proceedings of the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics - short papers (HLT-NAACL'06)*, New York, NY.

Benoît Sagot and Pierre Boullier. 2006. Efficient parsing of large corpora with a deep LFG parser. In *Proc. of LREC'06*.

Benoît Sagot and Éric Villemonte de La Clergerie. 2006. 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*, pages 329–336, Sydney, Australia. Association for Computational Linguistics.

Benoît Sagot, Lionel Clément, Éric Villemonte de la Clergerie, and Pierre Boullier. 2006. The Lefff 2 syntactic lexicon for french: architecture, acquisition, use. In *Proc. of LREC'06*.

Patrick Saint-Dizier. 1999. A Generative Modelling of Sense variations. *Informatica*, 10:10–30.

François Thomasset and Éric Villemonte de la Clergerie. 2005. Comment obtenir plus des méta-grammaires. In *Proceedings of TALN'05*, Dourdan, France. ATALA.

Gertjan van Noord, Ineke Schuurman, and Vincent Vandeghinste. 2006. Syntactic annotation of large corpora in STEVIN. In *proc. of LREC'06*, Genoa, Italy.

Gertjan van Noord. 2007. Using self-trained bilexical preferences to improve disambiguation accuracy. In *Proceedings of the Tenth International Conference on Parsing Technologies. IWPT 2007*, pages 1–10, Prague.

Tristan Vanrullen, Philippe Blache, and Jean-Marie Balfourier. 2006. Constraint-based parsing as an efficient solution: Results from the parsing evaluation campaign easy. In *proc. of LREC'06*, Genoa, Italy.

Éric Villemonte de La Clergerie. 2005a. 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.

Éric Villemonte de La Clergerie. 2005b. From metagrammars to factorized TAG/TIG parsers. In *Proceedings of IWPT'05 (poster)*, pages 190–191, Vancouver, Canada.

Fei Xia, Martha Palmer, and Aravind Joshi. 2000. 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.