# PASSAGE: from French Parser Evaluation to Large Sized Treebank

**Éric Villemonte de la Clergerie**[1],
**Olivier Hamon**[2,3], **Djamel Mostefa**[2], **Christelle Ayache**[2],
**Patrick Paroubek**[4] **and Anne Vilnat**[4]

[1] INRIA-Rocquencourt
Domaine de Voluceau Rocquencourt, B.P. 105, 78153 Le Chesnay, France
eric.de_la_clergerie@inria.fr

[2] Evaluation and Language Resources Distribution Agency (ELDA)
55-57 rue Brillat-Savarin, 75013 Paris, France
{hamon, mostefa, ayache}@elda.org

[3] Laboratoire d'Informatique de Paris-Nord (UMR 7030)
Université Paris 13 & CNRS 99 av. J.-B. Clément, 93430 Villetaneuse, France

[4]LIMSI-CNRS
BP 133 91403 Orsay cedex, FRANCE
{pap,anne}@limsi.fr

**Abstract**

In this paper we present the PASSAGE project which aims at building automatically a French Treebank of large size by combining the output of several parsers, using the EASY annotation scheme. We present also the results of the of the first evaluation campaign of the project and the preliminary results we have obtained with our ROVER procedure for combining parsers automatically.

## 1. Introduction

At the international level, the last decade has seen the emergence of a very strong research trend on statistical methods in Natural Language Processing (NLP). Among its origins we find, in particular for English, the availability of large annotated corpora such as the Penn Tree bank[1] (1M words extracted from the Wall Street journal, with syntactic annotations; 2nd release in 1995), or the British National Corpus[2] (100M words covering various styles annotated with parts of speech). Such annotated corpora are very valuable to extract stochastic grammars or to parametrize disambiguation algorithms. However, the development of large Treebanks is very costly from a human point of view and represents a long standing effort. The three year project PASSAGE[3] aims at building the first large sized Treebank for French and making it available to the community.

The paper is divided in two parts: firstly, we will describe the PASSAGE project in general, and then we will focus on the first evaluation campaign within the project and the results of systems. Finally, we describe our first ROVER experiments, where we combine parser outputs.

## 2. The PASSAGE Project

Funded by the French ANR program on Data Warehouses and Knowledge, PASSAGE is a three years project (2007-2009), coordinated by INRIA project-team Alpage. Its main objective is the large scale production of syntactic annotations to move forward (the acronym stands for *"Produire des Annotations Syntaxiques à Grande Échelle* in French). It builds up on the results of the EASY French parsing evaluation campaign (Paroubek et al., 2006), funded by the French Technolangue program, which has shown that French parsing systems are now available, ranging from shallow to deep parsers, and that it is possible to run an evaluation campaign from end to end using a common syntactic formalism with a very positive impact for the field, inspired from (Carroll et al., 2002). PASSAGE aims at pursuing and extending the line of research initiated by the EASY campaign. Its main objective is to use 10 of the participating parsing systems to EASY to jointly parse a French corpus of more than 100 million words. The proposed methodology (illustrated by Figure 1) consists of a feedback loop

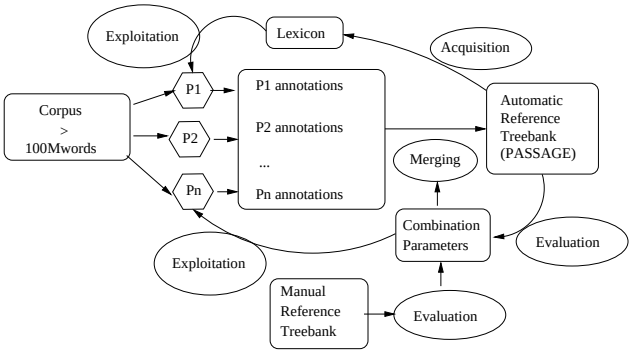


Figure 1: A bootstrap model for PASSAGE

[1]http://www.cis.upenn.edu
[2]phttp://www.natcorp.ox.ac.uk
[3]ANR-06-MDCA-013, http://atoll.inria.fr/passage

between parsing and resource creation as follows:

1. Parsing is used to create syntactic annotations for a large corpus;

2. Parsers are evaluated against a small reference Treebank manually built to produce parser combination parameters;

3. Syntactic annotations produced by the parsers are used to create or enrich linguistic resources such as lexicons, grammars extracting information from the large sized Treebank automatically produced by combining the parsers' output according to the parameters from the previous step (ROVER);

4. The linguistic resources created or enriched on the basis of the syntactic annotations are then integrated into the existing parsers;

5. The enriched parsers are used to create better and richer annotations, going back to step 1.

In addition to building the first large sized French Treebank, and validating automatic parser combination for the purpose, we believe that PASSAGE will help seeing the emergence of linguistic processing chains exploiting richer lexical information (de la Clergerie, 2005a), in particular semantic ones bootstrapped from the large Treebank by acquisition. 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.
As in the EASY project, the corpora used for PASSAGE will provide a relatively large diversity of styles of texts totalling over 100 millions words. The complete corpus has been defined to reflect the different styles of texts (not only newspapers, but also literary books, specialized texts, web or email texts, oral transcriptions, etc.). For each style we chose preferably corpora which were freely available. Two evaluation campaigns were planned to take place during the project. The first one ended in December 2007. As reference data, it reused the annotated data of EASY campaign, with the addition of 400 freshly manually annotated sentences (roughly 10% of the initial size of the EASY reference corpus). This campaign was also the occasion to test the automatic combination of parser outputs to produce a reference corpus of 45,000 sentences from a new corpus and to test with it the new evaluation protocole without imposed segmentation boundaries for words and sentences.
The second campaign will take place at the end of Passage (in 2009) on a manually annotated reference corpus including both the corpus of the first campaign and about 500 sentences extracted from the corpus under construction. This campaign will use the enriched syntactic annotation format as well as the new evaluation protocol without imposed segmentation for word forms and sentences.

## 3. The First Evaluation Campaign

In this section we describe the development of the first evaluation campaign: corpora, protocol, schedule, the parsing systems and the setting up of an evaluation service.

### 3.1. Corpora

The different corpora used for the first campaign are presented in Table 3.1.. The diversity of the resources allows the comparison of corpora from specific domain and with particular styles.

### 3.2. Evaluation Protocol

In addition to the "classical" EASY track, there is an exploratory track for testing the new PASSAGE evaluation protocol avoiding the explicit segmentation into forms and sentences which EASY revealed to be problematic. The segmentation is parser-dependent, but the use of span-referred tokens, completed by dynamic alignment techniques, is applied to align the forms returned by the participants with the ones of the reference corpus, which is built automatically by combining the output of all the parsers according to a majority vote.

The evaluation period has two phases. The first one consists in adapting the parsers to the test conditions. The participants use an evaluation server deployed at ELDA, which give them instant feedback on the performance of their parser for each run they upload. It uses the 4,000 sentences of the EASY corpus that were already annotated for reference data. Each output uploaded by a participant is evaluated automatically and the results are returned almost instantly. A maximum of ten submissions and their results are kept on the server, but the overall amount of submissions is not limited. Before the end of that development phase, participants are invited to select a primary submission from which the evaluation results will be computed in the second phase.

After one month of development, the server is closed and performance of the primary submission is computed automatically against the new complementary EASY corpus (classical EASY track results), this time using the 400 sentences newly annotated. The participants, as well as the organisation, are right away informed of their results.

Performance is also computed offline against the PASSAGE corpus obtained by combining the output of all parsers (exploratory PASSAGE track results). This is the occasion to compare results obtained with the two tracks, providing both a feedback on the confidence we can put in the automatic Treebank construction process and a verification that the participant did not overtrained their parsers during the initial development phase.

The annotation formalism used in PASSAGE is the EASY formalism. It has six types of constituents: nominal, adjectival, prepositional, adverbial, verbal and prepositional-verbal (i.e. an infinitive verb introduced by a preposition) and 14 types of relations: subject/verb, auxiliary/verb, direct object, complement/verb, noun modifier, verb modifier, adjective modifier, adverb modifier, preposition modifier, complement, subject attribute, coordination, apposition and collocation. More details can be found in (Vilnat et al., 2004).

The evaluation metrics used are precision, recall and f-measure, with 15 various relaxation constraints (Paroubek et al., 2006).

| Resource | #words | Description |
|---|---|---|
| WIKIPEDIA | 200K | A freely available corpus covering many domains of knowledge and collectively written by many authors, with various styles though biased toward descriptions. |
| WIKINEWS | 18.2K | A collection of short journalistic news |
| WIKILIVRES | 170K | A collection of 1956 freely available educational French books from WIKIBOOKS. |
| EUROPARL | 200K | A corpus of parallel multilingual texts extracted from the Proceedings of the European Parliament French part. |
| JRC-ACQUIS | 120K | Part of the total body of European Union laws, existing in several languages of the European Commission. |
| ESTER | 100K | A corpus of oral transcriptions from the ESTER project (Galliano et al., 2006) |
| LE MONDE | 100K | A journalistic corpus, with worldwide news. |
| EASY: | 1M | The corpus used for the EASY campaign already covers various genres and includes a subset of around 4K sentences (76K words) that have been manually validated and 400 new sentences freshly manually validated. |
| - LE MONDE | 86K | |
| - Parlementiary | 82K | |
| - Literary | 230K | French novels |
| - Oral from DELIC | 9K | oral transcripted |
| - Oral from Ester | 12K | |
| - Medical | 50K | medical texts from different domains |
| - Questions | 52K | Questions from different sources |
| - Web | 17K | web pages |
| - Mails | 150K | |

Table 1: Descriptions of corpora used in the PASSAGE project

### 3.3. The Participants' Parsing Systems

The participation of 11 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 contractors, including:

- FRMG, an hybrid TIG/TAG parser derived from a metagrammar, developed at INRIA[4] (Boullier and Sagot, 2005), (Thomasset and de la Clergerie, 2005), (de la Clergerie, 2005b);
- SXLFG, a LFG-based parser, developed at INRIA (Boullier and Sagot, 2005), (Boullier et al., 2005),
- LLP2 a TAG parser also derived from a metagrammar, developed at LORIA[5] (Roussanaly et al., 2005);
- LIMA, dependency based parser developed at LIC2M / CEA-LIST[6] (Besançon and de Chalendar, 2005);
- TAGParser, an extended chunker developed at TAGMATICA[7] (Francopoulo, 2005);
- Two parsers based on Property Grammars, developed at LPL[8]and using constraint satisfaction (Blache, 2005). The first parser 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 rule based parser developed by SYNAPSE[9];
- SYGMART, developed at LIRMM[10];
- XIP, a cascade rule-based parser developed at Xerox Research Center Europe[11] (Ait-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 are compared using a common syntactic annotation format and this experience by itself should continue to bring useful information about the expected requirements of a syntactic annotation standard, as EASY began to do.

### 3.4. Constituents results of the classical EASY Track

Ten systems participated to the constituents annotation task. Figure 2 shows the results they obtained. For most of the systems, F-measure is up to 90% and only three systems are between 80% and 90%. The trend is quite the same for Recall and Precision. For all the constituents, P05 is the best system. Around 96.5% of the constituent it returns are correct and it found 95.5% of the constituents present in the reference data. Figure 3 presents the scores for each constituents and each genre specific subcorpus.

[4]http://www.inria.fr/rocquencourt
[5]http://www.loria.fr/
[6]http://www-list.cea.fr/
[7]http://www.tagmatica.com/
[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/

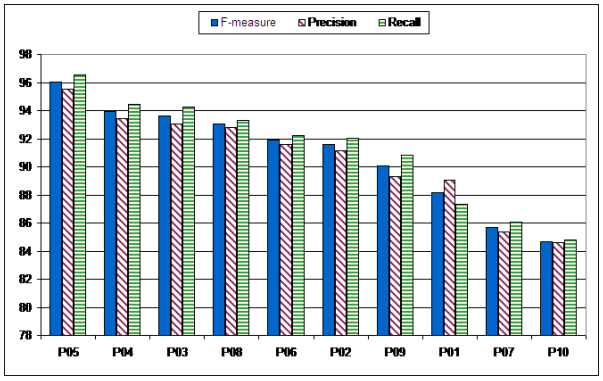


Figure 2: Overall constituents results

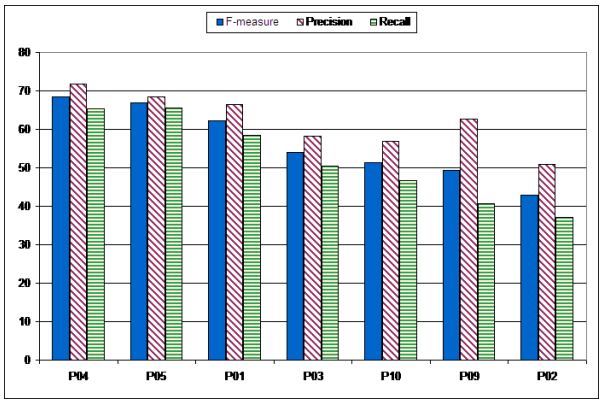


Figure 4: Overall functional relations results

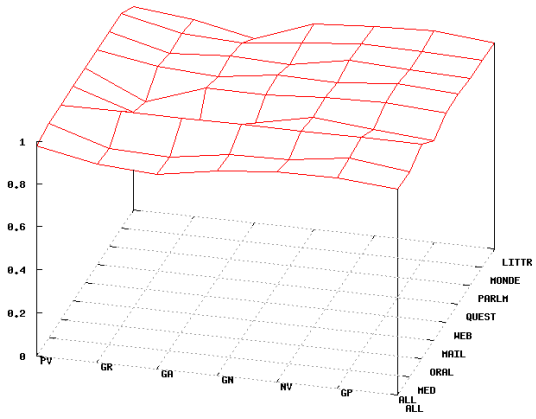


Figure 3: Constituents results for P05

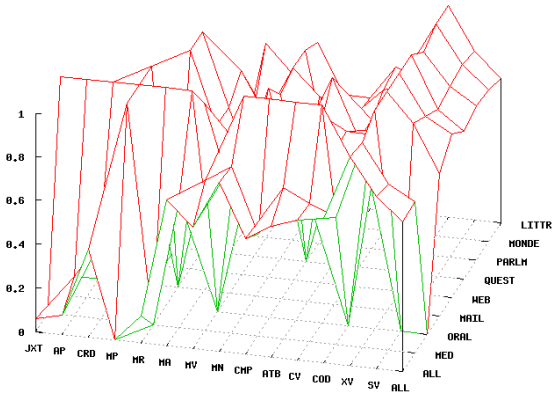


Figure 5: Functional relations results for P04

Even if scores are higher than those of the other systems, it represents the general trend:

- scores are lower for the adjectival (GA) and adverbial (GR) constituents;
- scores are higher for verbal, prepositional and prepositional-verbal constituents.

Note that the performances of two systems (P02 and P08) fall down for prepositional-verbal constituents nevertheless.
The systems obtain in general good results with the *mail*, *parliament* and *questions* corpora, while *le monde*, *litteraire* or *medical* corpora are associated with slightly lower results. The scores for the *mail* corpus are the lowest. Note that the scores of each system do not vary much across corpora, e.g. the P05 system obtains from 97.6% of F-measure for the *questions* corpus to 92.9% for the *mail* corpus.

### 3.5. Functional relations results of the classical EASY Track

Figure 4 shows the results of the seven systems that participated to the functional relations annotation task.
Performance is lower than for constituents and differences between systems are increased, an evidence that the task is more difficult. No systems gets a performance above 70% in F-measure, three are above 60% and two above 50%.

The last two systems are above 40%. System P04 has the highest results, whose details by relation and genre specific subcorpus are shown in Figure 5.
Performance change a lot with the type of relation, but a general trend can be identified:

- systems succeed for the auxiliary/verb relation (around 96% of F-measure for P04), and in a lesser degree for the noun modifier relation (around 77% for P04) and the subject/verb relation (around 78%);
- for some relations the scores are really low: adverb modifier (12% of F-measure for P04), prepositional modifier (0%), apposition (9%), and collocation (5%)
- scores for other relations lie between 40% and 70% with systems P04, P05 and P01, but are less than 50% for the other systems.

Results are similar whatever the corpus, the P04 system obtains scores from 74.7% of F-measure (for the *web* corpus) to 64.9% (for the *mail* corpus).

### 3.6. Stability of Systems on Corpora

In order to observe the stability of the systems' performance over the the different genre specific subcorpora, we computed the variance of the F-measures with a weighting depending on the population. The weights were drawn from information presented in Table 2, which shows the number

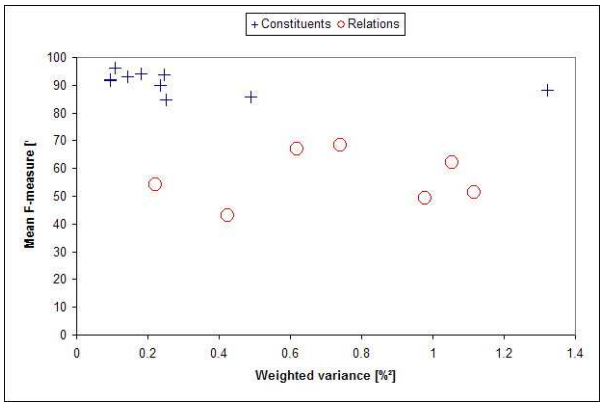


Figure 6: Weighted variances by corpus for each system

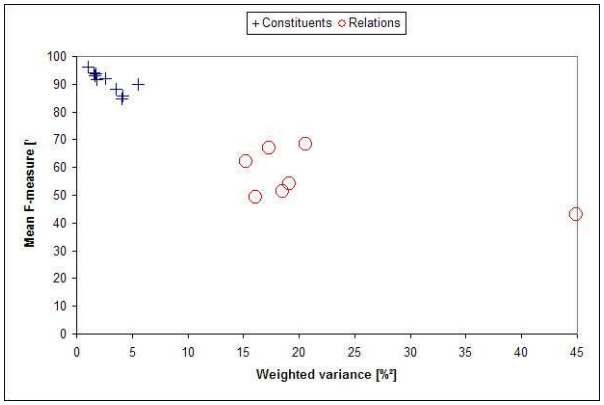


Figure 7: Weighted variances by type of constituent/relation for each system

of utterances, constituents and relations by genre specific subcorpus in the reference data. Then, we compared the weighted variances to the mean F-measures.

| Corpus | #Utterances | #Constituents | #Relations |
|---|---|---|---|
| *lemonde* | 52 | 681 | 746 |
| *litteraire* | 163 | 1680 | 1971 |
| *mail* | 21 | 194 | 240 |
| *medical* | 47 | 600 | 613 |
| *oral_delic* | 1 | 3 | 2 |
| *parlement* | 79 | 1093 | 1226 |
| *questions* | 28 | 252 | 257 |
| *web* | 14 | 177 | 183 |
| Total | 405 | 4680 | 5246 |

Table 2: Number of utterances by corpus.

We did not consider the oral_delic corpus in our study because it has only one sentence, a consequence of time pressure. The analysis of corpus stability is done as follow. First, weighted variances by corpus are computed for each system for both constituents and relations types, according to the formula $V = \frac{\sum_{i=1}^{N}(F_i - F_m)^2 * W_i}{N}$, where $W_i = \frac{C_i}{C_{total}}$, $N$ is the total number of corpora, $F_i$ is the system F-measure for the corpus $i$ for constituents (resp. relations), $F_m$ is the mean F-measure for the constituents( resp. relations), $C_i$ is the number of constituents (resp. relations) in the reference data for corpus $i$ and $C_{total}$ is the number of constituents (resp. relations) in the reference data. Both kinds of variance scores are shown by system in Figure 6.

Weighted variances by corpus are low, meaning systems' F-measures are stable and there is no real dispersion through the corpora. If we exclude one system (P01, which has a higher variance on constituents than other systems), the weighted variances by corpus are lower when observed on the constituents than on the relations. Moreover, systems react in the same way when weighted variances on corpora are those for constituents. Both the web subcorpus (in a positive way) and the mail subcorpus (in a negative way) are associated with extreme scores, causing dispersion of scores. When we are dealing with relations, systems are slightly more spreadout but the two previous subcorpora behave in the same way. Note that one system (P09) had a very low performance on the questions corpus.

Similarly, we then compute the weighted variances by type of constituent or relation. The variances are then weighted by the types of constituents, instead of the overall number of constituents previously used. Results are shown in Figure 7.

Weighted variances by type of constituent or relation are much higher than by type of corpus. But here again systems tend to aggregate, except for one system (P02 for the variance on relations) and react in the same way. Dispersion of F-measure scores on constituents is quite high, while dispersion on relations is very high (it means a standard deviance up to 4). This representation enables to diagnose in more details the origin of the main problem that the systems encountered:

- for constituents, dispersion of systems generally comes from adverbial (inducing lower scores) and prepositional (including higher scores) constituents,
- for relations, dispersion of systems generally comes from collocation, complement/verb on one side (inducing lower scores), and subject/verb, auxiliary/verb, noun modifier relations on another side (inducing higher scores).

On the previous graph, we can notice that relations, like collocation (in a negative way) or auxiliary/verb (in a positive way), are outlayers that modify drastically the mean F-measure.

## 4. ROVER

### 4.1. Background and motivation

The idea to combine the output of systems participating to an evalauation campaign in order to obtain a combination with better performance than the best one is not new. What now is known as the ROVER (Reduced Output Voting Error Reduction) algorithm was invented to our knowledge by J. Fiscus (Fiscus, 1997) in a DARPA/NIST evaluation campaign about speech recognition. He found out that by aligning the output of the participating speech transcription systems with a dynamic programming algorithm (Allison et

al., 1990) and by selecting the hypothesis which was proposed by the majority of the systems, he obtained better performances than with the best system. Since, the idea gained support, first in the speech processing community (Lööf et al., 2007), where people now work on refined versions of the algorithm, using the performance of the different speech recognizers as confidence weights in the hypothesis lattice obtained by combining the different ouptuts and by applying language models to guide the final stage of best hypothesis selection (Schwenk and Gauvain, 2000). In general better results are obtained with retaining only the output of the two or three best performing systems, in which case the relative improvement can go up to 20% with respect to the best performance (Schwenk and Gauvain, 2000). For text processing, examples of use of ROVER procedure are more rare, one such instance is for POS tagging, where the algorithm was applied to provide POS tags with confidence annotation to yield a validated language resource from data produced in an evalation campaign (Paroubek, 2000). Machine translation evaluation is another area where ROVER algorithms are used (Matusov et al., 2006).

In our case, we will use the text itself to realign the annotations provided by the various parser before computing their combination, as we did for our first experiments with the EASY evaluation campaign data (Paroubek et al., 2008). Note that if the different parser do not necessarily use the same word and sentence segmentation, we will need to first realign all the data with a common word and sentence segmentation (which can be computed here again by majority vote) before computing the ROVER data itself.

But our motivation for applying a ROVER procedure, is not only concerned with the obtention of a better performance, but also with the obtention of a confidence measure for the annotation. If all systems agree on a particular annotation to associate to a given word, this annotation is very likely to be true. This confidence measure is an information essential to have when building automatically a large treebank as we want to do.

At this stage many options are open for the way we want to apply the ROVER algorithm, since we have both constituents and relations in our annotations. We could:

- select first the relations, then the consitutents needed by these,
- select first the constituents, then the relations they carry,
- use different comparison functions for the equality of the text spans corresponding to constituents or relations source or target, with various degrees of constraint relaxation on their limits (Patrick Paroubek, 2006), and thus modifying the number of votes for each relation or constituent,
- merge all the annotations together, then perform a majority vote,
- perform an incremental merging of the various annotation, incorporating, each one a time; using of course different presentation sequences,
- use various weightings for the annotation of each systems,
- use various thesholds in the annotation selection process, e.g. a global threshold or different thresholds according to genre specific subcorpora or the annotation themselves.

In passage, ROVER experiments are only beginning and we have yet to determine which is the best strategy before applying it to word and sentence free segmentation data. In the next section, we report one of the latest experiment we did on the "EASY classic" PASSAGE track corpus.

### 4.2. First ROVER Experiment

The experiment we report here was done taking the data returned by 6 participants (all the data we had available at the time of the experiment). The "EASY classic" PASSAGE track uses a fixed word and sentence segmentation we could apply our ROVER algorithm straighaway on the parsers output. For the results presented here, we

- first selected the constituents, then the relations they bore,
- merged all the annotation together, then performed a majority vote,
- have used a detailed confidence measure for the annotation, at the level of both specific genre subcorpora (litterature subcorpus) and particular annotation (e.g. subject/verbe relation)
- used a weighting scheme that combines both the rank of the parser when comparing its performance to the one of the other parsers involved in the ROVER and the precision measure obtained by that parser for a particular subcorpus and a particular annotation. Here is the formula we used to compute the confidence to put in the annotation of a parser: $c_{s,a} = (N - (r - 1)) * p_{s,a}$, where $cs, a$ is confidence value given to annotation $a$ in the subcorpus $s$, $N$ is the total number of parsers of the ROVER, $r$ is the rank of the parser when comparing its precision measure to the one of the other parsers of the ROVE, and $p_{s,a}$ is the value of the precision measure obtained by the parser on subcorpus $s$ for annotation $a$.
- for given annotation in a particular subcorpus to be selected by the ROVER procedure, it needed to have its confidence value averaged over the number voters to be over the maximum value for this sucorpus and this annotation over all systems.

ROVER experiments are only beginning, we have yet to find a parameter combination that extend the area where the ROVER is better performing and to generalize its application to unsegmented data, but preliminary results are nonetheless encouraging.

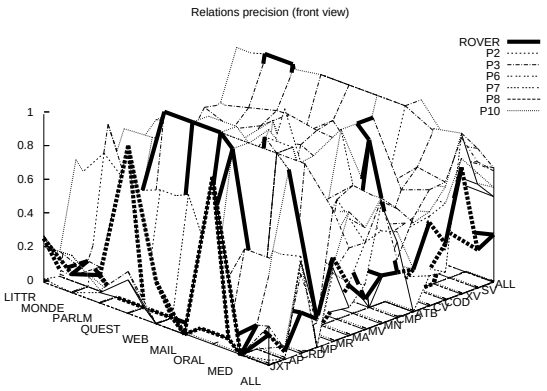


Figure 8: Plot of the precision performance of all the parsers together and the ROVER

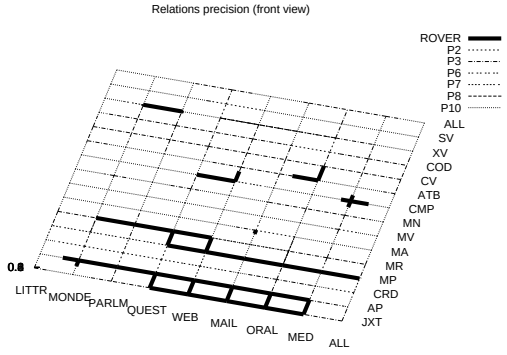


Figure 9: Plot the areas (bold line) of the precison plot in Figure 8 where the ROVER has a better performance thant the best one.

## 5. WEB Evaluation Server

Such server is useful for parsers development, and automatic systems in general. With the help of such evaluation servers, participants can perform a large number of evaluations during the development of their systems, without having to run scripts, check results into log files, etc. The interface allows participants to submit a file then observe the results obtained almost instantly. Moreover, after the development phase the results can be computer with the same tools also very rapidly.

The Participants could probably not carry out so much development to improve their parser without the help of the evaluation server, in particular to perform non-regression tests and diagnostic failures. For instance, the best system for the constituent evaluation, P05, had around 92.5% of F-measure at the beginning of the development phase, and after one month of development its F-measure increased to 96% (after more than 50 submissions).

In the same way, the WEB evaluation server is obviously useful for organizers. It permits to have a look on the number of development submissions and to follow progress of the participants. But most of all, it permits to carry out the test of the systems automatically, without any manual and time-consuming intervention such as: collecting participants' data, organizing data, starting evaluation scripts, waiting for the results, checking the correctness, sending the results to the participants, etc.

The other kind of interest of such an infrastructure is to have a perennial evaluation server that can be used by participants even after the end of the campaign (which they requested in PASSAGE). We will continu developping the WEB evaluation server and improve its interface for the second evaluation campaign.

## 6. 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 experience of the EASY evaluation campaign. The project should show that it is now possible to make parsing systems cooperate through an interchange syntactic annotation format and use the resulting annotations to acquire new linguistic knowledge, hence entering a virtuous circle.

## 7. References

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

L. Allison, C. S. Wallace, and C. N. Yee. 1990. When is a string like a string? In *Proceedings of International Symposium on Artificial Intelligence in Mathematics (AIM)*, Ft. Lauderdale, Florida, January.

R. Besançon and G. de Chalendar. 2005. L'analyseur syntaxique de lima dans la campagne d'évaluation easy. Dourdan, France, June. TALN'05.

P. Blache. 2005. Property grammars: A fully constraint-based theory. In H. Christiansen, editor, *Constraint Solving and Language Processing*, volume 3438. LNAI, Springer.

P. Boullier and B. Sagot. 2005. Analyse syntaxique profonde à grande échelle: Sxlfg. *Traitement Automatique des Langues*, 46(2):65–89.

P. Boullier, L. Clément, B. Sagot, and E. Villemonte de la Clergerie. 2005. Simple comme easy. pages 57–60, Dourdan, France, June. TALN'05.

J. Carroll, D. Lin, D. Prescher, and H. Uszkoreit. 2002. Proceedings of the workshop beyond parseval - toward improved evaluation measures for parsing systems. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain.

E. Villemonte de la Clergerie. 2005a. Dyalog: a tabular logic programming based environment for nlp. Barcelona, Spain, October. CSLP'05.

E. Villemonte de la Clergerie. 2005b. From metagrammars to factorized tag/tig parsers. Vancouver, Canada, October. IWPT'05.

Jonathan G Fiscus. 1997. A post-processing system to yield reduced word error rates: recognizer output voting error reduction (rover). In *In proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–357, Santa Barbara, CA.

G. Francopoulo. 2005. Tagparser et technolangue-easy. Dourdan, France, June. TALN'05.

S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri. 2006. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In ELRA, editor, *In proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May. ELRA.

J. Lööf, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, , and H. Ney. 2007. The rwth 2007 tc-star evaluation system for european english and spanish. In *In proceedings of the Interspeech Conference*, pages 2145–2148.

Evgeny Matusov, N. Ueffing, and Herman Ney. 2006. Automatic sentence segmentation and punctuation prediction for spoken language translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 158–165, Trento, Italy.

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, *In proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, pages 315–320, Genoa, Italy, May. ELRA.

Patrick Paroubek, Isabelle Robba, Anne Vilnat, and Christelle Ayache. 2008. Easy, evaluation of parsers of french: what are the results? In *Proceedings of the 6$^{th}$ International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morroco.

Patrick Paroubek. 2000. Language resources as by-product of evaluation: the multitag example. In *In proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000)*, volume 1, pages 151–154.

Anne Vilnat Christelle Ayache Patrick Paroubek, Isabelle Robba. 2006. Data, annotations and measures in easy - the evaluation campaign for parsers of french. In ELRA, editor, *In proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, pages 315–320, Genoa, Italy, May. ELRA.

A. Roussanaly, B. Crabbé, and J. Perrin. 2005. L'analyseur syntaxique de lima dans la campagne d'évaluation easy. Dourdan, France, June. TALN'05.

Holger Schwenk and Jean-Luc Gauvain. 2000. Improved rover using language model information. In *In proceedings of the ISCA ITRW Workshop on Automatic Speech Recognition: Challenges for the new Millenium*, pages 47–52, Paris, September.

F. Thomasset and E. Villemonte de la Clergerie. 2005. Comment obtenir plus des meta-grammaires. Dourdan, France, June. TALN'05.

T. Vanrullen, P. Blache, and J.-M. Balfourier. 2006. Constraint-based parsing as an efficient solution: Results from the parsing evaluation campaign easy. In ELRA, editor, *In proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May. ELRA.

A. Vilnat, P. Paroubek, L. Monceaux, I. Robba, V. Gendner, G. Illouz, and M. Jardino. 2004. The ongoing evaluation campaign of syntactic parsing of french: Easy. In *Proceedings of the 4$^{th}$ International Conference on Language Resources and Evaluation (LREC)*, pages 2023–2026, Lisbonne, Portugal.