EASY, Evaluation of Parsers of French: what are the results?

P. Paroubek*, I. Robba*, A. Vilnat*, C. Ayache**

(*)LIMSI-CNRS
Bât 508 Université Paris-Sud,
91403 ORSAY Cedex France
{firstname.name}@limsi.fr

(**) ELDA
55-57, rue Brillat-Savarin,
75013 Paris
ayache@elda.org

Abstract

This paper presents EASY, which has been the first campaign evaluating syntactic parsers on all the common syntactic phenomena and a large set of dependency relations. During this campaign, an annotation scheme has been elaborated with the different actors: participants and corpus providers, then a corpus made of several syntactic materials has been built and annotated. Both corpus and annotation scheme are here briefly presented, moreover, evaluation measures are explained and detailed results are given. To conclude, a first experiment aiming to combine the outputs of the different systems is shown.

1. Introduction

In NLP, research on syntactic parsing appeared very early. Parsing is indeed a necessary step in many complex tasks such as translation, dialog systems or question-answering systems. Since 1992, syntactically annotated corpora have been built for English, e.g. the SUSANNE corpus (Sampson, 1995) or the much bigger Penn Treebank (Marcus et al., 1993). More recently one can observe the development of such corpora in many other languages (as of today there are 30 languages listed in the treebank page of Wikipedia). Despite the existence of these corpora, there has been no large scale evaluation campaign for syntactic parsing. Indeed, no campaign has yet evaluated all the common syntactic phenomena nor has it taken into account constituent or dependency relation evaluation and none has regrouped more than a few parsing systems. The objective of EASY (Syntactic Parser Evaluation) (Vilnat et al., 2004) was to organize a campaign with such characteristics for parsers of French.

EASY is one of the 8 evaluation campaigns of the EVALDA platform, which itself is part of the TECHNOLANGUE [1] project (Chaudiron and Mariani, 2006). The aim of the EVALDA platform is to constitute a framework for the NLP system evaluation covering all domains of text or speech processing..

EASY gathered 5 corpus providers to collect and annotate a corpus of various genres in the EASY annotation format (Vilnat et al., 2004) and 12 participants who were interested in the evaluation of their parsers. At the beginning of EASY, there were probably as many output formats as participants. Moreover, not only were the formats different but so were the studied syntactic phenomena. Some of the participants had to specifically produce a version of their parser that could respect the demands of EASY. A first step in EASY, was then to come to an agreement concerning what to annotate and how to evaluate it. Corpus providers and participants had first to determine together which form of output a syntactic parser should produce. Thus, EASY allowed the setting up of a common base supplied to the participants for comparing the performances of their parsers. Although this base is relatively limited in linguistic terms as compared to the complexity of a fine grained full parse that a linguist could produce, it can easily be enriched and reused for other evaluation campaigns. In this paper, we give a brief description of the corpus, and the annotation scheme, then we present the evaluation measures and detail the results obtained during this first campaign. We conclude by listing the benefits drawn from the campaign and give a short presentation of how EASY results will be used in the new PASSAGE project whose objective is to build automatically a large sized French treebank of several hundred million words annotated with the EASY annotation scheme by combining the output of several parsers.

2. The corpus

Contrary to what is normally done in evaluation campaigns, our objective was to evaluate parsers on different syntactic materials. Hence, instead of being made of only one kind of text (generally newspaper articles), our corpus includes texts belonging to several different linguistic types. First of all, we gathered newspaper articles from *Le Monde*, which is the kind of text often used for evaluation as well as during development of parsers. Then, in order to take into account more sophisticated syntactic constructions, we collected a set of literary texts, extracted from ATILF databases. For more technical material we included medical texts, whose vocabulary is very specific. Collaborating with EQUER (another campaign of TECHNOLANGUE which provides an evaluation framework for Question/Answering systems for the French language), we included a sub-corpus composed of questions, because their very particular structure often creates problems for syntactic parsers. We also added handmade transcriptions of parliamentary debates, whose hybrid form is somewhere between oral and written and is therefore interesting as well. To study some texts with imperfect structures, we added Web pages and e-mails. Finally, we included oral transcriptions, coming from two different sources, on the one hand the TECHNOLANGUE ESTER

[1] TECHNOLANGUE is financed in an interdepartmental framework. Its objective is to set up an infrastructure for producing and circulating language resources and evaluation technologies for natural language (oral or written), http://www.technolangue.net.

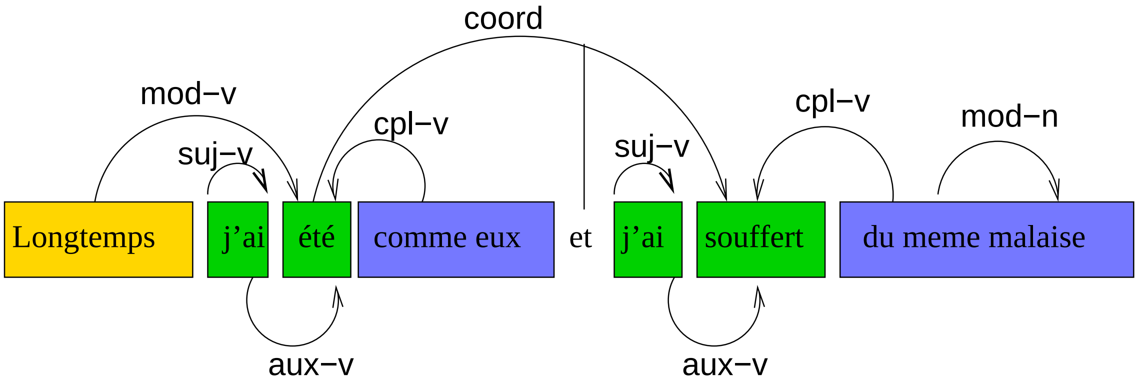

Figure 1: Annotation of a sentence extracted from the literary corpus

(campaign evaluating the performances of broadcast news transcription systems [2]), on the other hand transcriptions of oral interviews provided by the DELIC laboratory. At the end, the corpus was composed of about 40,000 sentences, composed of 770,000 words.

3. Annotation of the reference

The list of the phenomena we wanted to annotate has been completed by all the actors of the campaign (participants who develop a parser and corpus providers). The syntactic formalism has to make possible all kinds of syntactic annotation (shallow or deep parsing, complete or partial analysis), without giving any advantage to any particular approach. The EASY annotation formalism allows the annotation of minimal continuous and non recursive constituents, as well as relations encoding syntactic functions. Those relations have sources and targets which may be either forms or constituents (grouping several forms).

For the EASY campaign, 6 kinds of constituents have been considered:

- GN for Noun Phrase (*Groupe Nominal* in French), as *le petit chat* [3],
- GP for Prepositional Phrase (*Groupe Prépositionnel* in French), as *de la maison* or *comme eux*[4],
- NV for Verb Kernel (*Noyau Verbal* in French), including clitics as *j'ai*, or *souffert* [5]
- PV for Verb Kernel introduced by a Preposition (*Groupe Verbal Prépositionnel* in French), as *de venir* [6],
- GA for Adjectival Phrase (*Groupe Adjectival* in French), used for postponed adjectives in French, which are not included in GN,
- GR for Adverb Phrase (*Groupe Adverbial* in French) as *longtemps* [7]

The dependencies establish all the syntactic links between the minimal constituents described above. Participants, corpus providers and organizers agreed on a list of 14 kinds of dependencies:

- SUJ_V (subject),
- AUX_V (auxiliary),
- COD_V (direct object), CPL_V (verb complement) and MOD_V (verb modifier) for the different verb complements,
- COMP (complementor),
- ATB_SO (attribute of the subject or of the object),
- MOD_N, MOD_A, MOD_R, MOD_P (modifier respectively of the noun, the adjective, the adverb or the proposition),
- COORD (coordination),
- APP (apposition),
- JUXT (juxtaposition).

To find details on this annotation process, see (Paroubek et al., 2006).

The figure 1 gives an example of a literary sentence annotation.

We carried out an estimation of the annotation error rate concerning the relations. For each sub-corpus of the reference, we asked an expert to examine about one tenth of its sentences. An annotated sentence was considered as erroneous if it contained at least one relation wrongly annotated. For sub-corpora having an error rate bigger than 6 %, we made corrections both of the observed errors and the most frequent errors.

4. Evaluation measures

In EASY, precision/recall performance measurement can be obtained independently for constituents, for relations or both, in order to be able to evaluate any parser, whatever the kind of annotation it produces. Furthermore, performance measurements are computed for each kind of constituent, each kind of relation, and for each genre of sub-corpus, as well as globally.

[2] Nevertheless, those transcriptions could not be taken into account in the evaluation because of sentence segmentation problems.

[3] the small cat

[4] respectively: from the house and as they

[5] respectively: I have and suffered

[6] from coming

[7] for a long time

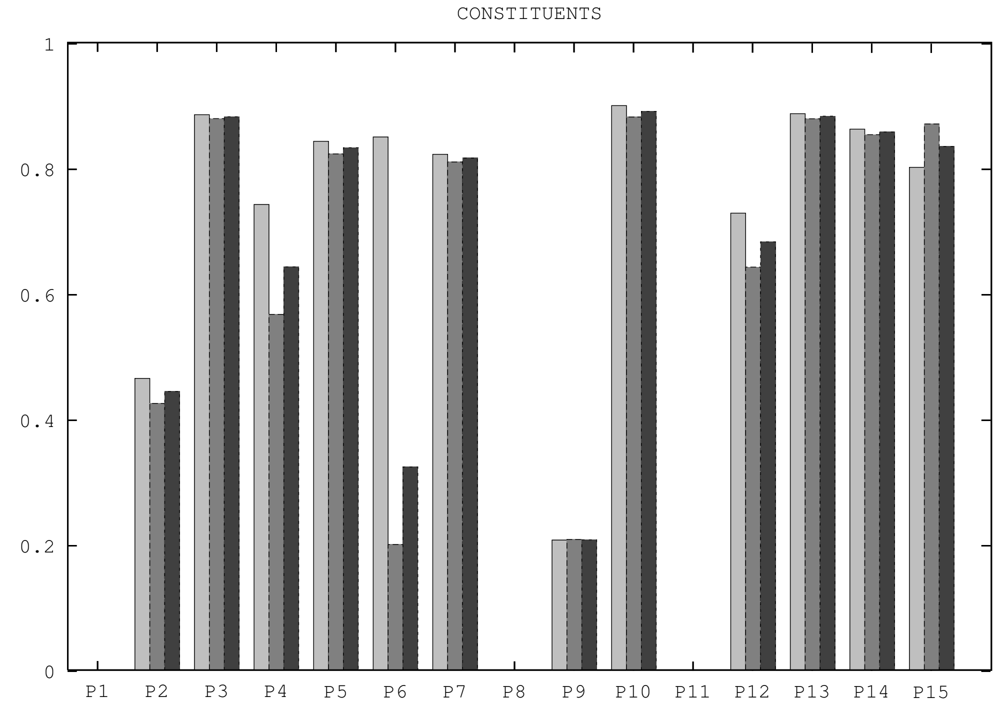

Figure 2: Results of the 15 parsers for constituents in precision/recall/f-measure (in this order), globally for all sub-corpora and all annotations together.

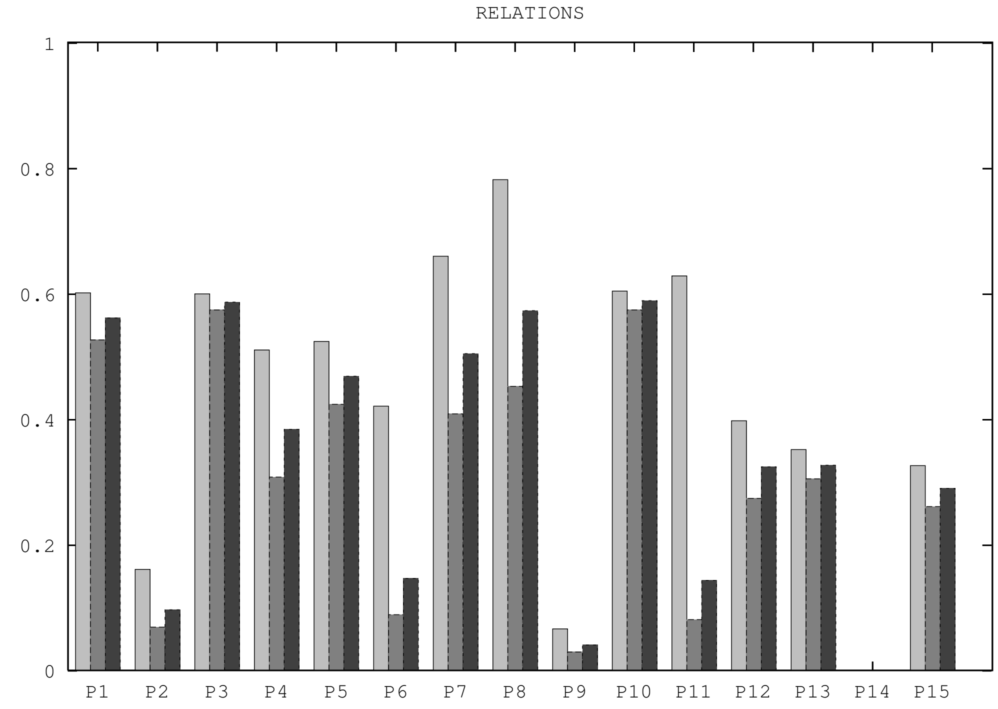

Figure 3: Results of the 15 parsers for relations in precision/recall/f-measure (in this order), globally for all sub-corpora and all annotations together.

Two constituents are considered equal if they have the same type (e.g. nominal group GN) and have equal text spans. To compare the text spans, we test different equality functions, between the manually annotated corpus (R, for reference), and the result proposed by the parser (H, for hypothesis):

- EQUALITY: $H = R$, the less permissive
- UNITARY FUZZINESS $|H \backslash R| \leq 1$
- INCLUSION: $H \subset R$
- BARYCENTER: $\frac{2*|R \cap H|}{|R|+|H|} > 0.25$

- INTERSECTION: $R \cap H \neq \emptyset$, the most lenient

These functions have also been used to evaluate the relations. Two relations are considered equal if they have the same type (e.g. subject-verb SUJ-V) and if their respective source and target have equal text spans (with the meaning just described above). To take into account the fact that some parsers do not build constituents, but only give relations, we consider that the target or the source of a relation may be a constituent or a word inside this constituent. So we try three different comparisons, to consider the address of the source or target of each relation:

- HYP: considering the encompassing constituent in the hypothesis, if any,
- HYP-REF: considering the encompassing constituent in the hypothesis, if any, else considering the encompassing constituent in the reference, if any,
- REF: considering the encompassing constituent in the reference, if any.

After testing these different comparison methods, we observe that the differences do not affect the ranking of the parsers. So the following results will consider the BARYCENTER to compare the text spans, and the HYP-REF:measure to compare the target and source addresses in the hypothesis and the reference.
Thus we obtain a fine grained picture of the performance of a parser with which we can correlate the influence that the kind of annotation and the text genre have on the performance. We end up with a measurement whose granularity is variable, from global measurements where all annotations and all genres are considered on the same footing to measurements specific to a particular annotation and a given text genre, e.g. measure of the performance of a parser for the subject-verb relation in the literature subcorpus.

5. Performance results of EASY campaign

We have collected 15 sets of results from 12 teams (some participants have submitted two runs). The figures 2 and 3 illustrate the results of the different parsers by giving for each, the global performance computed for all sub-corpora and all annotations together, for precision, recall and f-measure. The graph in figure 2 gives the results for the constituents, and contains only 12 results because 3 parsers did not provide any constituent result but only relation information. Similarly, in the graph of figure 3 (giving the results for the relations) we see that one parser did not output any relation.
As was expected, we observe a greater variability of results and inferior performances for the annotation of relations than for constituents. For constituents, most parsers obtained good results, with quite similar results in precision and recall (thus in f-measure too). Considering the relations, the contrasts are more important. The parsers obtained various results, and the it is clear that some have made the choice of the precision, while others obtained a better recall.

To precise the results study, we observe the results of the parsers on the different sub-corpora, for the different relations. We draw the results of the parser obtaining the best precision (P8), the best recall (P3), and the best f-measure (P10). The resulting graphs are drawn on the figures 4, 5, 6, giving the results of the best parser respectively for precision, recall and f-measure. For the three figures, the different types of sub-corpus are on the left: from literary (LITTR) at the background to medical (MED) corpus, and the complete corpus (ALL) at the foreground. The different dependencies are on the right: from JUXT (juxtaposition) at the foreground to SV (subject) and all of them (ALL), at the background. re In figure 4, we see a “valley” corresponding to the fact that this parser gives no result for the oral sub-corpus. For the other genres, the results are quite similar, and rather good for the simplest relations (i.e. subject or verb complements or modifiers which are on the background of the graphs). The results do not depend on the corpus genres, even if they are lower for the other relations. In figure 5, we see that the parser gives results for all corpus genres, even if they are rather bad for oral corpus,

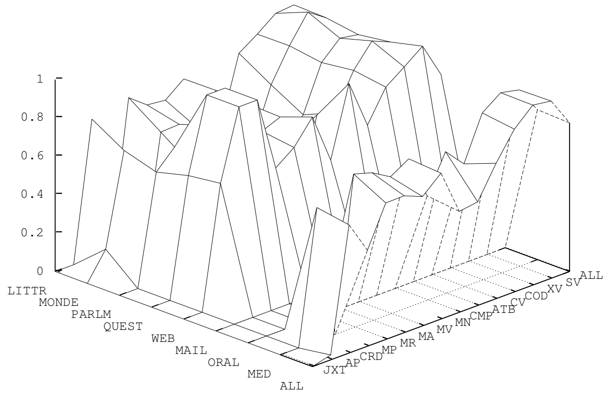

Figure 4: Results for relations of the parser obtaining the best **precision** measure

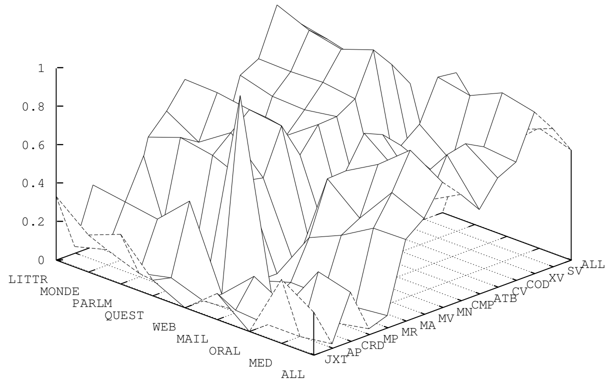

Figure 5: Results for relations of the parser obtaining the best **recall** measure

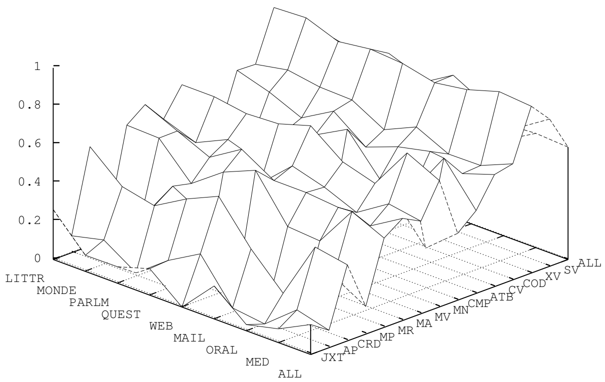

Figure 6: Results for relations of the parser obtaining the best **f-measure**

and for "difficult" relations (the ones on the foreground of the graphs, which are coordination, apposition or juxtaposition) in all sub-corpora. The last figure 6 presents a flatter profile: the results are more similar for all the relations and all the corpus, even if the best ones are lower than those of P8, in figure 4.

Of course, this is only a picture of the field taken at a given time (taken with an imperfect camera) which needs to be put into perspective, as the results of any quantitative evaluation.

But despite these imperfections, the results of EASY are very interesting for the following reasons:

- we see that for relation annotations, the best systems have an average f-measure near 0.60, which is considered as a significant threshold in the MUC evaluations on natural language understanding (Hirschman, 1998), thus a campaign which measures results in a comparable domain. .

- although the variability of results for relation annotation is high (the detailed performance profile according to text genre and annotation is more chaotic than for constituents), some parsers manage to preserve the same level of performance across text genres.

- the results show that there is still an important part of work to do for analyzing syntactic phenomena which are rarely or never handled by the actual parsers because they are judged too complex, like for instance the apposition or juxtaposition relation, or when coordination are combined together or mixed up with ellipses, as we saw it before on the three figures (4, 5, 6) where the results are better in the background, than in foreground.

- the best performances are obtained by different parsers, which exhibit different performance profiles, so there is *a priori* a relatively important margin for performance increase which could be obtained by combining the annotations of different parsers: this is the object of the following paragraph. .

6. ROVER

The EASY evaluation campaign was also the occasion to test the idea of performance improvement by combining the output of the parsers using what now begins to be known in some circles as a ROVER (Reduced Output Voting Error Reduction) algorithm. The acronym and the first experiment of the kind are due to J. Fiscus (Fiscus, 1997) in a DARPA/NIST evaluation campaign on 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 outputs 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 evaluation campaign (Paroubek, 2000).

Since we are processing text the problem seems to be simpler than for speech, because we can use the words of the text to be annotated for realigning the different annotations, provided the parsers respect the text they annotate. In the EASY evaluation campaign, our work is made easier by the fact that all parsers had to use the same word and sentence segmentation, providing *de facto* aligned data. But there are a great variety of ways we could combine the outputs of the parsers, to name a few, we can:

- select first the relations, then the constituents need 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 constrain relaxation on their limits as mentioned in section4., 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 annotations, incorporating each one a time, of course using different presentation sequences,

- use various weightings for the annotation of each system,

- use various thresholds in the annotation selection process, a global threshold, different thresholds according to the sub-corpus or the annotation itself.

Following the advice we found in the literature (see below), we performed our first experiments with only the three best performing systems. The combination procedure we used consisted in merging the annotations of all the systems together and selecting the ones for which a majority of systems voted. First by selecting the relations, then the constituents needed by this relations as source or targets, with the provision that constituents could not intersect nor be nested, as is required by the EASY annotation guidelines.

But this crude algorithm did not work well, we then weighted the vote of each system by its average performance at the evaluation, still not much success.

What we found to work best was by weighting the annotation of a system proportionally to the rank the system obtained at the evaluation, in a way that the annotation of the best system could be changed only if the majority of the other systems voted against it. For this experiment, we used also a strict equality on the text spans corresponding to constituents or the target or source of a relation. With this algorithm, Figure 7 illustrated the relative gain of performance in precision against the best performance shown in Figure 4.

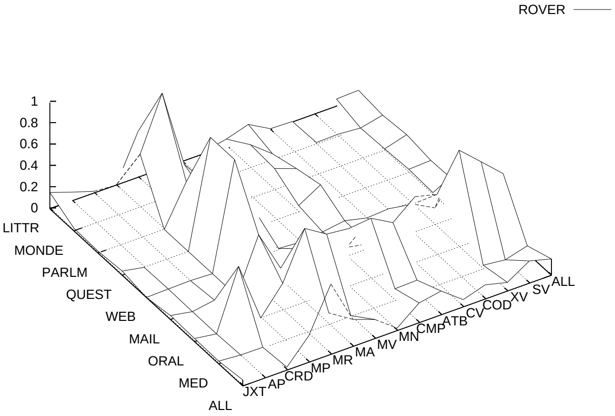

Figure 7: Relative gain of performance in precision against the best **precision** result

While, Figure 8 displays the performance surface of the ROVER and the three systems together.

We see from this graph that our ROVER procedure improves the performance, but not for all points of the surface, there are places where the ROVER results are slightly behind the performance of the best system, the most important gains being observed for sub-corpus or annotations where most of the systems had problems. This encouraging results are only preliminary, since we have to run complete ROVER experiments with all the systems together and try the variations given previously in the list. In particular, selecting first the constituent according to a majority vote then the relations having these consituents as source or target should improve results, because in general the average performance of all the systems is much better for constituents. Further ROVER experiment will be the basis of the algorithm used to combine parsers annotations in the project PASSAGE[8].

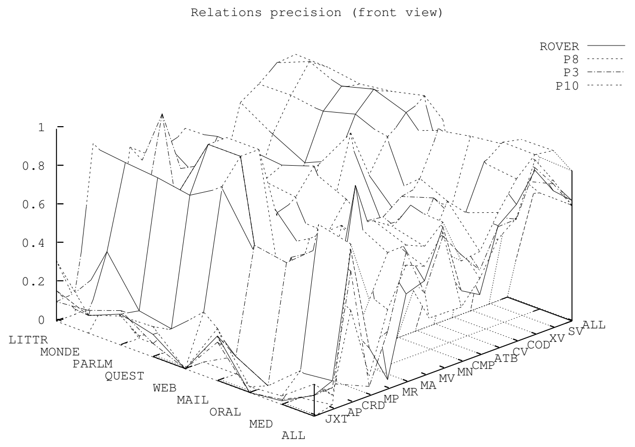

Figure 8: Comparative precision results of the ROVER and the three best systems

7. Conclusion

EASY has been the first campaign deploying the evaluation paradigm in real size for syntactic parsers of French with a black-box evaluation scheme using objective quantitative measures. Concerning the obtained performances on syntactic relations, we observe that, 3 different systems obtained respectively the best results in recall, precision and f-measure, which leads us to think that these systems have complementary characteristics which could be further exploit. We also notice from EASY results that some systems are robust across genre variation, a characteristic which many parsers lack nowadays, those being often designed to process newspaper articles of rather high quality.

EASY was also the occasion to create a working group on parsing evaluation gathering a majority of actors of the domain, and finding an extension in the PASSAGE project (2007-2009), which regroups a kernel of EASY participants and organizers. PASSAGE has the aim to produce a large sized French treebank (of several hundred million words) by combining automatically the output of different parsers according to parameters obtained as the results of 2 evaluation campaigns based on improved version of EASY, one at the beginning of the project the other at the end. The annotation scheme and the evaluation procedure is enhanced version of the ones used in EASY. The details on PASSAGE may be found in the paper on this project in the same conference proceedings.

8. References

L. Allison, C. S. Wallace, and C. N. Yee. 1990. When is a string like a string? In *Proceedings of International Sym-*

[8] PASSAGE is a project supported by ANR (ANR-06-MDCA-013), the national research funding agency, in the subprogram Data Mass and Ambient knowledge (*Masses de données et Connaissances Ambiantes*).

posium on Artificial Intelligence in Mathematics (AIM), Ft. Lauderdale, Florida, January.

S. Chaudiron and J. Mariani. 2006. Techno-langue: The french national initiative for human language technologies (hlt). In *Proceedings of the 5th International Conference on Language Ressources and Evaluation (LREC)*, pages 767–772, Genoa, Italy, may.

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.

L. Hirschman. 1998. Language understanding evaluations: lessons learned from muc and atis. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain.

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.

M. Marcus, B. Santorini, and M. Marcinkiewciz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19:313–330.

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. 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.

G. Sampson. 1995. *English for the Computer: The Susanne Corpus and Analytic Scheme*. Oxford University Press, USA.

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.

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 4th International Conference on Language Resources and Evaluation (LREC)*, pages 2023–2026, Lisboa, Portugal.