

The Ongoing Evaluation Campaign of Syntactic Parsing of French: EASY

Anne Vilnat*, Patrick Paroubek*, Laura Monceaux*[†],
Isabelle Robba*, Véronique Gendner*, Gabriel Illouz*, Michèle Jardino*,

*LIMSI-CNRS

bp 133 91403 Orsay Cedex, France
{firstname.lastname}@limsi.fr

[†]LINA

2, rue de la Houssinière bp 92208 44322 Nantes Cedex 3 France
Laura.Monceaux@lina.univ-nantes.fr

Abstract

This paper presents EASY (Evaluation of Analyzers of SYntax), an ongoing evaluation campaign of syntactic parsing of French, a subproject of EVALDA in the French TECHNOLOGUE program. After presenting the elaboration of the annotation formalism, we describe the corpus building steps, the annotation tools, the evaluation measures and finally, plans to produce a validated large linguistic resource, syntactically annotated.

1. Introduction

In this paper, we present EASY (Evaluation of Analyzers of SYntax) an open evaluation campaign for parsers of French. EASY is a part of TECHNOLOGUE¹, the EVALDA language engineering multi-evaluation campaign project, which is a program launched in spring 2002 jointly by the three French Ministries of Industry, Culture and Research. In this paper, we present the choices that were made in the design of the formalism: remaining as much as possible independent from any particular parsing theory, the presence of both continuous non-recursive constituents and functional relations, the possibility to have either words or constituents as arguments in relations and the absence of clausal mark-up. We give a description of the corpus, and we present the tools built for annotation. Then we provide the evaluation measures and plans to produce a validated large linguistic resource, syntactically annotated.

2. Annotation formalism

The formalism that we have adopted for annotation has to respect two strong constraints. On the one hand it has to allow encoding most of the syntactic phenomena of French, and not only the most simple or frequent ones. On the other hand, it has to remain as independent as possible from any particular parsing theory, in order to allow the participation of any kind of parser: deep or shallow, rule-based or not, relying on supervised or unsupervised training algorithm.

As it is the case in other syntactic evaluation formalisms, we have in EASY two types of information: constituents and functional relations. We choose to adopt small, neither recursive nor discontinuous constituents. The relations between these minimal constituents are annotated thanks to relations, which link these constituents inside complex syntagms. Thus we are able to evaluate from chunkers (which only annotate simple constituents) to deep parsers (which are able to recognize complex syntagms).

There are 6 types of constituents: nominal, adjectival, prepositional, adverbial, verbal and prepositional-verbal, the last being used for infinitive verb introduced by a preposition. These constituents are illustrated in figure 3. Let us examine the first utterance: “*Il arrive en retard, avec, dans sa poche, un discours qu’il est obligé de garder*”². To give some examples, we annotate there a nominal constituent (*un discours*), a verbal constituent, which includes clitics, (*Il arrive*), a prepositional constituent (*dans sa poche*³), an adjectival constituent (*obligé*⁴) and a prepositional-verbal constituent (*de garder*⁵). It is worth noticing that the annotation of a prepositional constituent is only a shortcut: it is equivalent to the annotation of a nominal constituent and a relation between the preposition which introduces it and this noun phrase. This case is encountered with the discontinuous prepositional phrase *avec,...,un discours*⁶: at this step, as we cannot use the shortcut of a prepositional phrase because of the discontinuity, we only annotate the noun phrase *un discours*; the relation with the preposition *avec* will be annotated at the following step. In the third utterance: “*Et même, je serai d’avis qu’on usât pour les calmer, les endormir d’appareils profondément bousculatoires*”⁷, we have adverbial constituents *même* or *profondément*⁸.

EASY uses 14 types of functional relations. Among them, we find the traditional functions such as subject, auxiliary verb, verb object, verb complement, noun/adjective/adverb modifiers etc. These relations may link forms as well as constituents. To come back on our example, we annotate a subject between *il* and *arrive*,

²A free translation could be: “He arrives late, with, in his pocket, a discourse, that he is compelled to keep”.

³in his pocket

⁴compelled, but the english translation is not an adjective!

⁵to keep

⁶with,..., a discourse

⁷A free translation could be: “And moreover, my opinion will be that one uses to make them quiet, to make them sleep, deeply rocking apparatus”

⁸moreover or deeply

¹<http://www.recherche.gouv.fr/technologue>

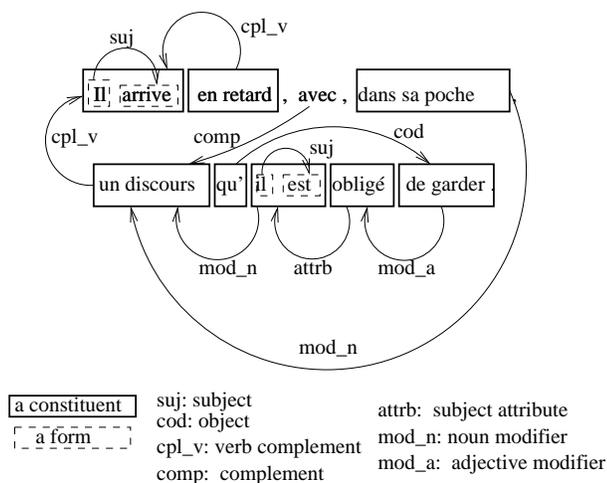


Figure 1: Annotated relations

that means between the two forms included inside a verbal phrase. The relative *qu'* is annotated as the object of *garder*⁹. The constituents *en retard* and *un discours* are linked to *il arrive* as verb complement. The constituent *dans sa poche* modifies the noun *un discours*, *de garder* modifies the adjective *obligé*. The link between the relative clause *qu'il est obligé de garder* and the noun *un discours* that it modifies, is annotated between the verb phrase of the relative clause *il est* and this noun phrase *un discours*. This annotation is always used when we have to link a secondary clause to a constituent, for example to link a temporal subordinate clause to the verb of the principal clause by a *verb complement* relation. We also annotate at this step a *complement* relation between the preposition *avec* and the noun phrase *un discours*. All these annotations are illustrated in figure 1. EASY distinguishes also apposition, coordination and juxtaposition that are less frequently encountered in annotation schemes, since probably few parsers are able to make such subtle distinctions; but these phenomena are frequently encountered in French corpora. Each kind of constituents may be linked by a relation of coordination. This relation concerns three elements : the coordination and the two coordinated constituents (two noun phrases, or verb phrases...). When two clauses are coordinated, the relation links their verbs. When the coordination links two utterances, the relation only implies the coordination and the second coordinate, as it is the case in utterance 3, where the *et* and the verb phrase *je serai* are related by a coordination where the first coordinated term is lacking. Apposition links noun phrases whose referents are identical, as in “Le président Jacques Chirac”, where the second noun phrase *Jacques Chirac* is in apposition to the first one *le président*. Juxtaposition is used to link forms or constituents which are neither coordinated, neither subordinated, nor in apposition. Utterance 3 shows an example of juxtaposition between *les calmer* and *les endormir*¹⁰. More examples may be found in the PEAS annotation manual¹¹.

⁹to keep

¹⁰to make them quiet, to make them sleep

¹¹<http://www.limsi.fr/Recherche/CORVAL/easy>

3. Corpus

Given the ubiquitous character that basic functionalities for language analysis have nowadays with the spreading of information technology throughout all aspects of society, we think it is primordial to have an heterogeneous corpus whose genre distribution would be as close as possible to the one that parsers are likely to encounter. Thus, our corpus contains: newspaper articles (as utterance 2 in figure 2, novel excerpts (as utterance 3 in figure 2, a set of interrogative forms translated from a question-answering evaluation corpus (as utterance 8 in figure 2), audio transcriptions (as utterance 9 in figure 2), and in a smaller amount medical texts (because the difficulties which arise from them are mainly due to the specialized vocabulary they contain).

As it is the case in most evaluation campaigns, the participants know what kinds of texts the corpus contains, but they will only receive it when they will have to analyze it.

4. Tools for annotation

Since syntactic annotation is a rather difficult and fastidious task, we wanted to propose simple and easy-to-use tools. Moreover, in such a large evaluation campaign, annotation is made by several persons using different operating systems, so we had to elaborate tools with minimal requirements for installation and porting. This two constraints lead us to choose generic HTML editors, which are widespread and whose usage is intuitive. Annotation is done in two steps: first, constituents are marked by highlighting the corresponding text portion using constituent specific colors, selected from a predefined palette (we see in figure 2 a snapshot illustrating the constituent annotation); second, the annotator, using these constituents, fills in a set of predefined tables corresponding to the 14 relations, with the relation arguments addresses that were automatically generated during the first step. The figures 3 and 4 illustrate the second step, with the argument addresses automatically generated and the tables the annotator has to fill in. The relations corresponding to the first utterance, graphically represented in figure 1, are annotated in figure 4.

With such tools whose handling is quite intuitive and easy, the only problem left concerns exhaustivity checking of the annotation.

Finally, the annotator's work is translated in an XML format into which all participants in the evaluation will also have to map their parser output. This format designed in collaboration with all the participants (using the results of PEAS (Gendner et al., 2003) as a starting point), is already by itself a valuable result of EASY for the parsing community, since up to day, there doesn't exist any standard for annotating syntax of French.

5. Evaluation

As a first approach, we plan in EASY to use classical parsing evaluation measures. Crossing-brackets measure will be computed for constituent boundaries, and precision-recall measures will be employed for functional relations. Parseval propositions (Black et al., 1991) as well as the criticisms that were made about them (Gaizauskas et al., 1998) and (Lin, 1998) will be considered. At the time of writing, we do not expect to have to deploy new measures in

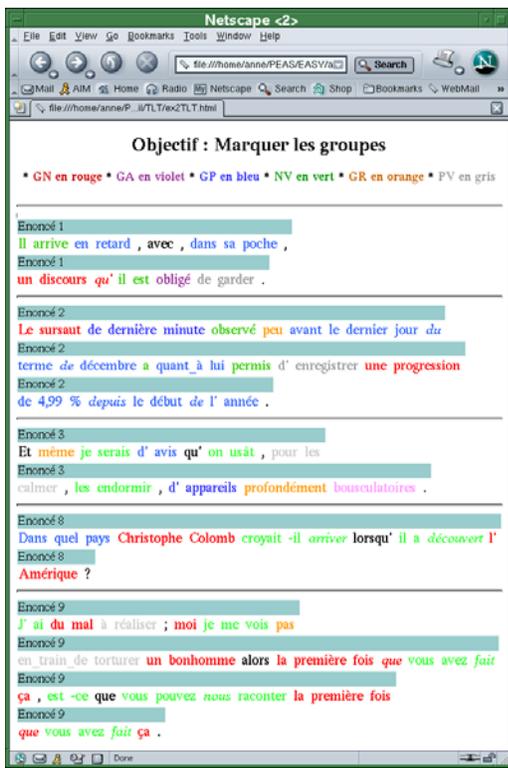


Figure 2: First step: constituent annotation

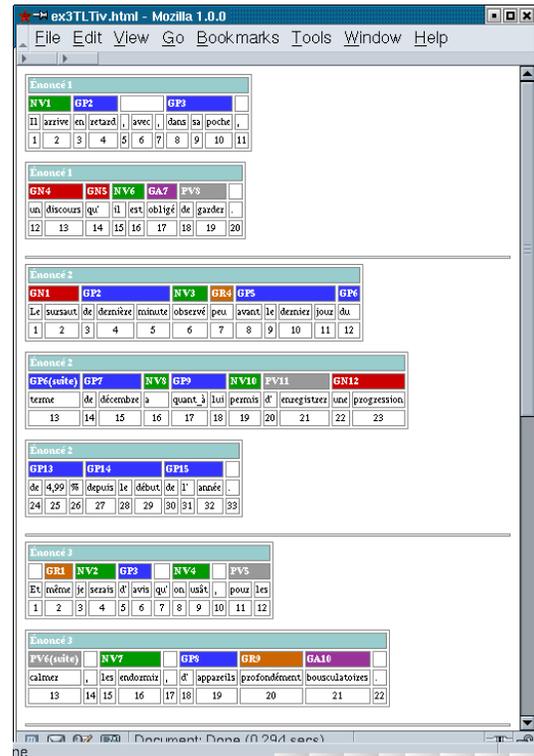


Figure 3: Result of the first step

addition to precision and recall, but since we will enter the phase of the project where we choose the evaluation measures in collaboration with the participants, there still exists the possibility for changes regarding this topic. Whatever the measures we will use at the end, we will rely on variety of text genre and on richness of syntactic annotation to produce parametrized multi-dimensional results. To take into account, small variations of annotations particular to some system, we intend to apply constraint relaxation techniques on constituent boundaries and functional relation arguments to allow for segmentation variations across participants, in addition to the computation of comparative evaluation measures on word segmentation. Some very preliminary tests were run on the precision/recall evaluation measures applied to constituents, during PEAS(Gendner et al., 2003), the pre-project of EASY, on a small corpus of approximately 5,000 words (300 sentences) annotated with 2,000 constituents. The figure 5 shows the strict¹² application of the precision measures on the constituents identified by the two parsers, computed on a per constituent, per genre basis. The results show clearly that one system is better than the other, but we have not yet investigated the pertinence of the difference observed, neither have we applied any constraint relaxation on the constituent boundaries to see whether the extra leeway would confirm or infirm this result.

Furthermore, it is far too early to draw any hasty conclusion from a result based only on constituent comparison, since most of the information produced by the parsers is

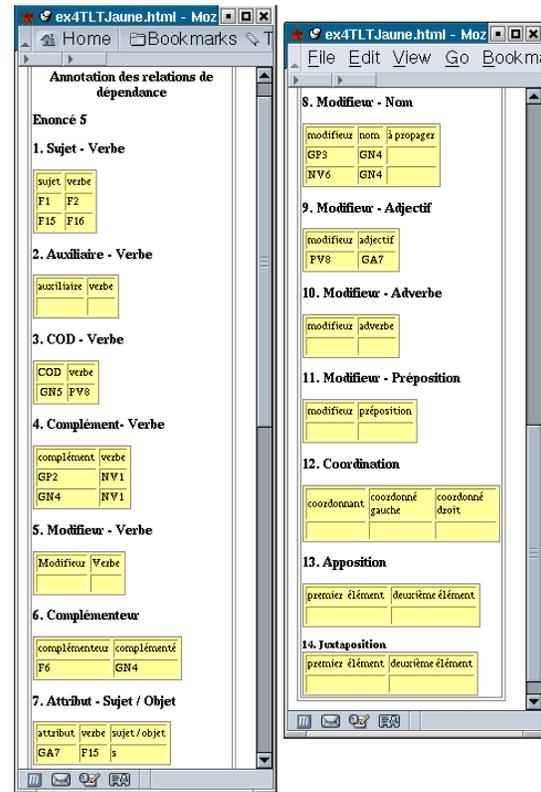


Figure 4: Second step: relation annotation

¹²strict equality on constituents boundaries without any constraint relaxation.

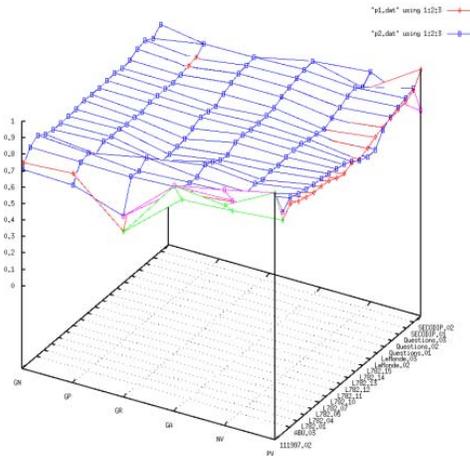


Figure 5: Precision measure applied to constituents by 2 parsers.

held in the relations.

6. Linguistic resource constitution

The linguistic resource that EASY will yield, will be two fold. As in any corpus based comparative evaluation campaign, there will be the evaluation results and the reference corpus built by the corpus providers, but also the corpus annotated by all the participants (its size will be much greater than the one of the reference corpus in order to prevent the participants from making any last minute adjustment to their parser). From this bigger corpus multiply-annotated, we plan to extract, with a fusion algorithm similar to the one described in (Monceaux and Vilnat, 2003), a syntactic corpus with confidence annotations. They will be used to guide human annotators who will validate the parts where confidence measures are too low (hopefully, less that 10% of the whole text will have to be proofread). Our goal is to obtain a syntactically annotated corpus for French, which will be rich enough to constitute an interesting common resource, and moreover which will be potentially extended by following campaigns. Furthermore, there is very few existing resources of this kind for French, to compare with other countries which are developing such resources for their own languages, following the precursor PennTreeBank (Marcus et al., 1993) and its descendant the PropBank (Kingsbury and Palmer, 2002). We can cite, without being exhaustive, the following projects : DDT (Danish Dependency Treebank) (Kromann et al., 2003), Tiger (Brants et al., 2002) for German, and all the treebanks presented at the TLT Workshop (Nivre and Hinrichs, 2003).

7. Conclusion

Although, we cannot give a formal proof that our annotation scheme is generic across grammar formalism, we consider that the number of participants in EASY is by itself a proof that this formalism is sufficient to express the

major part of the information produced by most of the existing parsers of French. Indeed, fourteen parsers are about to be evaluated, while five corpus providers are annotating the corpus, using the formalism and the tools we built. The annotated corpora that EASY will yield, represent a valuable linguistic resource which could be used in the future for training new parsers or making new evaluations. And as it was said in section 3, the proposition of a common syntactic formalism, seems very promising to us, because it constitutes a first step towards a standardized format for representing syntactic parsing of French, contributing to efforts like (Ide and Romary, 2003).

8. References

- Black, E., S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, F. Jelinek R. Ingria, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski, 1991. A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*. DARPA, Pacific Grove, California: Morgan Kaufmann.
- Brants, S., S. Dipper, S. Hansen, W. Lezius, and G. Smith, 2002. The tiger treebank. In *Workshop on Treebanks and Linguistic Theories*. Sozopol.
- Gaizauskas, R., M. Hepple, and C. Huyck, 1998. A scheme for comparative evaluation of diverse parsing systems. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, volume 1. Granada, Spain: ELRA.
- Gendner, V., G. Ilouz, M. Jardino, L. Monceaux, P. Paroubek, I. Robba, and A. Vilnat, 2003. Peas, the first instantiation of a comparative framework for evaluating parsers of french. In *Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics*, volume Companion. Budapest, Hungary.
- Ide, N. and L. Romary, 2003. *Encoding Syntactic Annotation*, chapter 16. Language and Speech series. Dordrecht: Kluwer, pages 281–296.
- Kingsbury, P. and M. Palmer, 2002. From treebank to propbank. In *LREC2002*. Las Palmas, Canary Islands, Spain.
- Kromann, M., L. Mikkelsen, and S. Lyng, 2003. *Danish Dependency Treebank Annotation Manual*. Dpt of Computer Linguistics, Copenhagen Business School, <http://www.id.cbs.dk/mtk/treebank>.
- Lin, D., 1998. Dependency based method for evaluating broad-coverage parsers. *Natural Language Engineering*, 4 (2):97–114.
- Marcus, M., B. Santorini, and M.A. Marcinkiewicz, 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19.
- Monceaux, L. and A. Vilnat, 2003. Multi-analyse, vers une analyse syntaxique plus fiable. In *TALN2003*. Batz-surmer: TALN.
- Nivre, J. and E. Hinrichs (eds.), 2003. *2nd Workshop on Treebanks and Linguistic Theories*. Vaxjo, Sweden: Vaxjo University Press.