PEAS, the first instantiation of a comparative framework for evaluating parsers of French

V. Gendner, G. Illouz, M. Jardino, L. Monceaux, P. Paroubek, I. Robba, A. Vilnat
LIMSI – CNRS, BP 133, 91403 Orsay – France
`{gendner,gabrieli,jardino,monceaux,pap,isabelle,anne}@limsi.fr`

Abstract

This paper presents PEAS, the first comparative evaluation framework for parsers of French whose annotation formalism allows the annotation of both constituents and functional relations. A test corpus containing an assortment of different text types has been built and part of it has been manually annotated. Precision/Recall and crossing brackets metrics will be adapted to our formalism and applied to the parses produced by one parser from academia and another one from industry in order to validate the framework.

1 Introduction

In natural language understanding, many complex applications use a syntactic parser as a basic functionality. Today, in particular for the French language, the developers face the great diversity of the offer in the domain. Therefore, the need for a complete comparative evaluation framework – including a pivot annotation formalism, a reference treebank, evaluation metrics and the associated software – is increasing.

It is worth noting that most of the recently developed parsers use a robust approach. Consequently, they do not always produce a complete parse of the sentence, but they are able to produce a result, whatever the size, the particularities and the grammaticality of the input. For this reason, it is essential to be able to compare in a fair way the parses they produce against those produced by other parsers whatever their characteristics. One possible solution is to offer a common reference annotation formalism along with a fully parsed reference corpus and a set of robust metrics, allowing for both complete and selective evaluation over an assortment of different text types and syntactic phenomena.

The aim of our research is to build such evaluation framework, which to date is missing for French. Figure 1 presents the different modules of our evaluation protocol as it stands today.
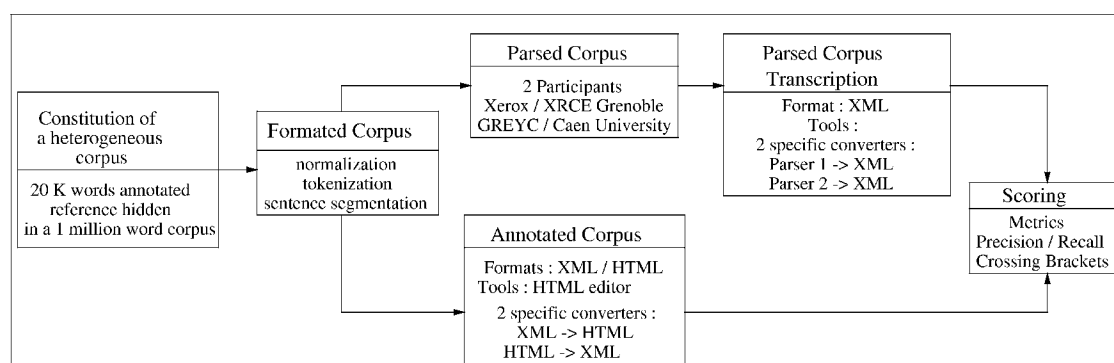

Figure 1: Evaluation protocol modules

2 Annotation formalism

The definition of the annotation formalism is the core element of the evaluation process. Indeed, the formalism must have a coverage of syntactical phenomena as broad as possible in order to allow any parser to participate, whatever the grammatical formalism it uses.

We have decided immediately upon a two-steps annotation: first the chunks annotation is carried out, second functional information is annotated through relations between words, words and chunks or between chunks. The constituents or chunks are continuous and non-embedded. They are as small as possible to allow any segmentation chosen by a parser to be converted into our formalism. For the same reason, the information that is not expressed in the constituents is expressed through a large number of functional relations: twelve in all. Such formalism is closer to a dependency-based formalism than to a constituent based formalism (Sleator and Temperley, 1991). It neither prevents the "deep" parsers to be evaluated, nor disadvantages them, but the transcription of their parses could be more complex. The six types of chunks and twelve functional relations are given in table 1. They were mainly inspired by Abeillé et al. (2000), and have been adapted while annotating corpus excerpts.

Chunks	Functional relations
NV – verbal	subject-verb
GN – nominal	auxiliary-verb
GR – adverbial	argument-verb
GA – adjectival	modifier-verb
GP – prepositional introducing a nominal phrase	modifier-noun
	modifier-adjective
PV – prepositional introducing a verbal phrase	modifier-adverb
	attribute-subject/object
	Coordination
	Apposition
	Complementer

Table 1: Annotated chunks and relations

No clausal or sentential segmentation is identified, because as in a dependency-based formalism, the complex structure of the sentence is obtained through the whole chain of relations. The following sentence[1] that contains three noun phrases (NP) gives an example: *<NP1> la porte de la chambre fermée à clef à l'intérieur </NP1><NP2> les volets de l'unique fenêtre fermés, eux aussi, à l'intérieur </NP2> et <NP3> par-dessus les volets, les barreaux intacts </NP3>, [...]*. In our formalism, the noun phrases are described through the following chunks and relations:

> *<GN1> la porte </GN1>*
> *<GN2> les volets </GN2>*
> *<GN3> les barreaux </GN3>*
> *coordination (",", GN1, GN2)*
> *coordination (et, GN2, GN3)*

And the noun phrase NP1 is expressed through:

> *<GN1> la porte </GN1>*
> *<GP1> de la chambre </GP1>*
> *<GA1> fermée </GA1>*
> *<GP2> à clef </GP2>*
> *<GP3> à l'intérieur </GP3>*
> *modifier-noun (GP1, porte)*
> *modifier-noun (GA1, porte)*
> *modifier-adjective (GP2, fermée)*

Moreover, since our chunks are not embedded, all the modifiers placed before a noun are included in the same nominal group as the noun itself. And here again, the relations are used to express the links between the particular terms, as in the annotated example of *mon très riche et très proche ami*[2]:

> *<GN> mon très riche et très proche ami </GN>*
> *modifier-adjective (très, riche)*
> *modifier-adjective (très, proche)*
> *coordination (et, riche, proche)*
> *modifier-noun (et, ami)*

The formalism gives the possibility to annotate ambiguities at dependency level (by duplicating the relation tables). Note that we are

---

[1] This original sentence is extracted from (Leroux, 1907), and may be translated as: *the shutters of the single window also closed from inside, and over the shutters, the bars intact*.

[2] Translation: *my very rich and very close friend*.

still studying how our evaluation will handle this phenomenon.

3 Corpus and tools for annotation

The corpus retained for annotation is a set of texts whose nature is as diverse as possible. Indeed the corpus contains excerpts from: newspapers, novels, Web pages, automatic audio transcriptions, and a set of questions translated from the question-answering track of TREC. The whole corpus contains 1 million words; each text has been segmented in sentences and tokenized in words. Each participant to the evaluation protocol has received the texts both in pre-segmented and original format.

The part of the corpus that has been annotated contains about 20,000 words. The annotation tools, that we have developed, use an HTML editor. For chunk marking, the annotator selects chunks and colors them (each type of chunk corresponding to a particular color). For the twelve functional relations, the annotator has a set of twelve tables to fill in for each sentence; giving for each relation the address of its parameters. Of course, all of them are not to be filled in. All the information thus annotated is then translated into an XML format. Annotation of the example of §2 is translated in:

```
<E id="0">
<constituants>
<Groupe type="GN" id="G0">
 <F id="F0"> la </F>
 <F id="F1"> porte </F>
</Groupe>
<Groupe type="GP" id="G1">
 <F id="F2"> de </F>
 <F id="F3"> la </F>
 <F id="F4"> chambre </F>
</Groupe>
<Groupe type="GA" id="G2">
 <F id="F5"> fermée </F>
</Groupe>
<Groupe type="GP" id="G3">
 <F id="F6"> à </F>
 <F id="F7"> l' </F>
<F id="F8"> intérieur </F>
</Groupe>
<F id="F9"> , </F>
<Groupe type="GN" id="G4">
 <F id="F10"> les </F>
 <F id="F11"> volets </F>
</Groupe>
<Groupe type="GP" id="G5">
 <F id="F12"> de </F>
 <F id="F13"> l' </F>
```

```
 <F id="F14"> unique </F>
 <F id="F15"> fenêtre </F>
</Groupe> ...
</constituants>
<relations>
<rel xmlns:xlink="extended" type="MOD-N" id="R0">
<modifieur xmlns:xlink="locator" href="G1">
<nom xmlns:xlink="locator" href="F1">
</rel>
<rel xmlns:xlink="extended" type="MOD-N" id="R1">
<modifieur xmlns:xlink="locator" href="G2">
<nom xmlns:xlink="locator" href="F1">
</rel>
<rel xmlns:xlink="extended" type="MOD-A" id="R2">
<modifieur xmlns:xlink="locator" href="G3">
<adjectif xmlns:xlink="locator" href="F5">
</rel> ...
<rel xmlns:xlink="extended" type="COORD" id="R9">
<coordonnant xmlns:xlink="locator" href="F9">
<coord-g xmlns:xlink="locator" href="F1">
<coord-d xmlns:xlink="locator" href="F11">
</rel> ...
</relations>
</E>
```

For the French language, Abeillé et al. (2000) is the only other attempt at building a treebank. In this case, the corpus is homogeneous in text genre, since it contains only newspaper articles extracted from *Le Monde* although it covers various domains from politics to sports. The approach is however ambitious and interesting: the corpus contains 1M words, 17 000 different lemma; it is annotated both with morpho-syntax and grammatical functions.

4 Evaluation metrics

The first proposals for parser evaluation were made in Parseval (Black et al., 1991). Carroll et al. (1998) gave a survey and proposed a new evaluation scheme. Since, two orientations have emerged. The first, inspired by Parseval, is based on phrase boundaries and uses recall plus crossing-bracket measures. Although it has been criticized (Gaizauskas 1998, Lin 1998), it is still in use nowadays. The second one is based on dependency relations, (on which recall and precision can also be computed) and seems to be more and more in favor (see the workshop *Beyond Parseval* 2002).

Since our annotation formalism has both constituents and functional relations, there is no reason to dismiss either approaches. Nevertheless, we have to outline that the transcription of the parses will be more systematic for the relations than for the

constituents. Indeed, in our formalism, relations can associate words, chunks or words and chunks, but it is always possible to match any relation argument with the reference parse, because we always know to which chunk a word belongs. On the other hand, for the segmentation, the chunk boundaries may vary a lot from one parse to another. So we have to foresee either an important set of matching rules, or flexible evaluation methods.

5 Prospective

Based on this preliminary research, a larger project for syntactic parser evaluation, named EASY/EVALDA has been accepted by TECHNOLANGUE, a joint program of the three French Ministries of Industry, Culture and Research. A rather large francophone community has declared its interest for the project, fourteen participants (belonging to universities or to private institutions) are ready to evaluate their parser, while five corpus providers are interested in annotating large size corpora both in syntax and in functional relations. This community will contribute to enrich every aspect of our proposal: annotation formalism, tools and metrics.

Moreover, the participation of a sufficient number of parsers will allow the production of a good quality validated linguistic resource. Indeed, we will produce the automatic fusion of all annotated data of the parsers, and then manually correct the divergent parses.[3]

Last of all, the XML format into which we translate the parses is an open exchange format. It is an important asset for portability and reuse of parsing technolgy. E.g. for question answering application, where a parser is often needed to parse both the questions and the huge set of candidate answers, the use of XML makes easier the selection of the paser for the task at hand.

6 Conclusion

At the time of writing, we have developed all the different phases of our evaluation process except for the evaluation metrics tools. The two candidate parsers have parsed the corpus, and we are now translating their outputs within our

formalism. Here, the difficulty is neither to loose information nor to miss incorrect parses. The application of our metrics and the results examination will constitute a first validation of our framework.

References

A. Abeillé, L. Clément and A. Kinyon. 2000. *Building a treebank for French*. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC), (1):87-94, Athens, Greece, May. ELRA

Beyond Parseval – Towards Improved Evaluation Measures for Parsing Systems. Workshop of the 3rd International Conference LREC. Las Palmas, Spain. John Carroll editor.

E. Black et al., *A procedure for quantitatively comparing the syntactic coverage of English grammars*. In DARPA, editor Proceedings of the Fourth Darpa Speech and Natural Language Workshop, pages 306-311, Pacific Grove, California, February, Morgan Kaufmann.

R. Gaizauskas, M. Hepple and H. Huyck. 1998. *A scheme for comparative evaluation of diverse parsing systems*. In Proceedings of the 1st International Conference LREC, (1):143-149, Granada, Spain, May. ELRA

V. Gendner, G. Illouz, M. Jardino, L. Monceaux, P. Paroubek, I. Robba, A. Vilnat. 2002. *A Protocol for Evaluating Analyzers of Syntax (PEAS)*. In Proceedings of the 3rd LREC, May 2002, Las Palmas, Spain.

G. Leroux. 1907. *Le mystère de la chambre jaune*. L'Illustration, Paris.

D. Lin. 1998. *Dependency based method for evaluating broad-coverage parsers*. Natural Language Engineering 4 (2):97-114.

L. Monceaux. 2002. *Adaptation du niveau d'ana-lyse des interventions dans un dialogue. Application à un système de question-réponse*. PhD thesis Paris 11, December 2002.

D. Sleator and D. Temperley. 1991. *Parsing English with a Link Grammar*. Research report CMU-CS-91-196, Carnegie Mellon U., School of Computer Science, 91 p.

J. Carroll, T. Briscoe and A. Sanfilippo. 1998. *Parser Evaluation: a Survey and a New Proposal*. In Proceedings of the 1st International Conference LREC (1):447-454, Granada, Spain, May. ELRA.

---

[3] Monceaux (2002) proposes a rover algorithm, which merges in one parse the outputs of several parsers.