Data, Annotations and Measures in EASY the Evaluation Campaign for Parsers of French

Patrick Paroubek*, Isabelle Robba*, Anne Vilnat*, Christelle Ayache†

* LIMSI-CNRS
Bât 508 Université Paris XI
91403 Orsay Cedex France
{pap,isabelle.robba,anne.vilnat,}@limsi.fr
† ELDA
55-57 rue Brillat Savarin
75013 Paris, France
ayache@elda.fr

Abstract

This paper presents the protocol of EASY the evaluation campaign for syntactic parsers of French in the EVALDA project of the TECHNOLANGUE program. We describe the participants, the corpus and its genre partitioning, the annotation scheme, which allows for the annotation of both constituents and relations, the evaluation methodology and, as an illustration, the results obtained by one participant on half of the corpus.

1. Introduction

EASY is one of the 8 evaluation campaigns about language technology of EVALDA, a project of the TECHNOLANGUE national program[1].The aim of the EASY campaign (Vilnat et al., 2004) (Paroubek et al., 2005) is to design and test an evaluation methodology for comparing parsers of French and to produce a treebank by combining automatically all the data annotated by the participants. The corpus consists of texts taken from various domains (literature, medicine, technical, etc.) and with different genres (newspapers, questions, websites, oral transcriptions, etc.). EASY is a complete protocol of evaluation including corpora constitution, manual corpora annotation, evaluation and production of the treebank. In this paper, we describe the corpus and its genre partitioning, the annotation scheme which allows for the annotation of both constituents and relations, the evaluation methodology and, as an illustration, the results of one of the 16 systems participating in the campaign, on half of the corpus, since at the time of writing, all the results were not yet computed (all results will be presented at the conference).

2. State of the art

In the early days, parsing evaluation was done by experts who built their opinion from the observation of parses. In many cases, they were using a grid (Blache and Morin, 2003) of parsing features to guide their analysis. Concerning the parsing of French, it seems that the first attempt at comparative evaluation dates back to (Abeillé, 1991). In an attempt at reducing the objectivity introduced by the particular views that experts might entertain about particular approaches and to improve the reuse of linguistic knowledge, people started to employ specific test suites, of which TSNLP is a good example (Oepen et al., 1996). But test suites do not reflect the distribution of the phenomena encountered in real corpora since they hold in general a limited number of examples without statistical information. Further, they can only be reused for non-regression tests, because once they have been utilized, it is relatively simple to adapt one's parser to the specific language items present in the test suite. Finally, they often require a mapping of syntactic annotations, since there is a good chance that the test suite will encode the syntactic information in a formalism different from the one used by the parser and in general such mapping induces an information loss or is complex to perform. With the advances in computer technology and markup standards, a new solution emerged to get rid of these drawbacks: treebanks. The first and certainly the most famous is the Penn Treebank (Marcus et al., 1993), which was followed by many other developments for different languages, including French (Brant et al., 2002) (Abeillé et al., 2000). Since 2002, (Palmer et al., 2005) propose to add semantic role labels to the Penn Treebank. Then in 2004, (Miltsakaki et al., 2004) proposed a large-scale discourse annotation project: the Penn Discourse Treebank, which aims at identifying discourse connectives and their arguments. Although treebanks can solve the problem of language coverage and representation of the linguistic phenomena distribution, if they are large enough and their genre is representative of the material parsed; they do not provide a solution for finding easily an appropriate pivot formalism in case the ones used by the parser under test and the treebank are different. To be faithful, an evaluation must preserve both the information present in the reference data and the one output by the parsers. Devising a universal syntactic formalism that enables the description of all linguistic phenomena generally encountered is precisely one of the research objective of parsing. Many proposals have been made, some use annotation mappings (Gaizauskas et al., 1998), other compare information amounts like (Musillo and Sima'an, 2002) (which unfortunately requires the building of one parallel corpus per for-

[1]TECHNOLANGUE (december 2002 - april 2006) is supported by the 3 French ministries of Culture, Industry and Research.

malism), others propose to use automatic grammar learning procedures (Xia and Palmer, 2000) or computations based on the "edit" distance (Roark, 2002). If we go back a little farther in time, (Black et al., 1991) focused on evaluation measures and used the constituent boundaries to compare parsers, by measuring the percentage of crossing brackets [2] and recall[3]. If you add precision to two previous measures you get the GEIG[4] protocol, (Srinivas et al., 1996), also called PARSEVAL measures (Carroll et al., 2002). But in practice these measures have been applied only on unlabeled constituents, because no common ground could be found between all the different categories of constituents used by the different parsers that were tested in the few campaigns where these measures have been used. To answer this problem, (Lin, 98) proposed to consider dependencies instead of constituents for evaluation (Briscoe et al., 2002), (Carroll et al., 1998). Going even further, (Carroll et al., 2003) proposes to annotate tagged grammatical relations between lemmatized lexical heads, in order to work on both the logic and grammatical relations present in the sentence, instead of working on the topology of the parse trees. The EASY annotation scheme was inspired by (Carroll et al., 2003). As we will explain in section 4., it has an initial level of constituents and grammatical relations, but without any explicit notion of head (Gendner et al., 2003), (Vilnat et al., 2004).

3. Corpus building

Our corpus is constituted of different kinds of texts to assess the ability of parsers in processing different kinds of material. First, we collected some archives of the French newspaper *Le Monde*, to obtain journalistic style texts, very often used for evaluation purposes. To bring some diversity and to take into account perhaps more elaborate sentences, we collected texts of French literature. To add technical style papers, we included medical texts. As a link with EQUER (Question-Answering campaign in TECHNOLANGUE), we included a corpus of questions. Transcribed debates of the French Senate constitute another genre, which lies between real oral transcription and written texts. To take into account texts with a more and more relaxed syntax, we include extracts of web pages, emails and oral transcriptions, some coming from ESTER (another TECHNOLANGUE evaluation campaign on automatic speech transcription). Table 1 give excerpts of all genres present in the EASY corpus. Five corpus providers participated in the EASY evaluation campaign: ATILF (Analyse et Traitement Informatique de la Langue Française), DELIC (DEscription Linguistique Informatisée sur Corpus), LLF (Laboratoire de Linguistique Française), STIM-AP/HP (Assistance Publique / Hôpitaux de Paris), and ELDA (Evaluations and Language resources Distribution Agency), which co-organized the campaign with LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur). Their tasks in the campaign were to collect a large corpus of various genres and to annotate a part of it. The ratio between the size of the annotated part and the total size of the corpus had to be sufficient to discourage the participants from using any kind of processing other than automatic ones, since the participant do not know on which part of the corpus they will be evaluated.
Table 2 gives the size of the various genre specific subcorpora and table 3 indicates for each the amount of data that has been annotated.

Genre	Sentence example in French (*free English translation in italics*)
Newspaper	Le gouvernement intérimaire a décidé d'asphyxier économiquement le ≪Taylorland≫, en imposant un embargo total sur les marchandises à destination des zones sous contrôle du FNLP. *The temporary government decided to smother economically, "Taylorland" by imposing a total embargo on the goods to zones under control of the FNLP.*
Literature	Longtemps j'ai été comme eux, et j'ai souffert du même malaise. *For a long time I felt like them, and I suffered from the same unease.*
Medical	La sensibilité de l'échotomographie pour la définition des calculs vésiculaires de plus de 2mm de diamètre est de 98% environ. *The sensibility of the echotomography for the definition of vesicular calculi of more than 2mm in diameter is approximately 98 %.*
Parliament	- Monsieur le Président, mes chers collègues, je tiens simplement à faire un rappel au Règlement. *Mister President, my dear colleagues, I would only like to raise a point of order.*
E-mail	Alors moi je dis chapeau bas pour tes explications mon Jean. *Me, I take off my hat to you for your explanations my dear Jean*
Oral	euh l'intervention c'est quoi *hum the operation what is it*

Table 1: Number of sentences and word forms per genre in the EASY corpus.

4. 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

[2] the number of constituent boundaries output by the parser that cross a constituent boundary of the reference

[3] the number of constituent boundaries output by the parser that do exist in the reference data

[4] Grammar Evaluation Interest Group.

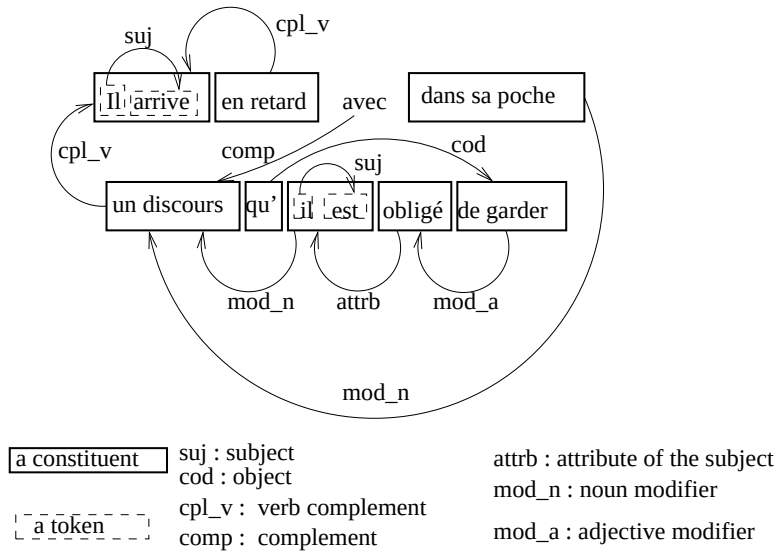

Figure 1: Annotated relations

Genre	Provider	Sentences	Words
WEB	ELDA	836	16786
LE MONDE	LLF	2950	86273
PARLIAMENT	ELDA	2818	81310
LITERATURE	ATILF	8062	229894
EMAIL	ELDA	7976	149328
MEDICAL	STIM	2270	48858
ORAL_DELIC	DELIC	522	8106
ORAL_ESTER	ELDA	11298	97053
QUESTIONS	ELDA	3528	51546
total		40260	769154

Table 2: Number of sentences and word forms per genre in the EASY corpus.

Genre	Provider	Sentences	Words
WEB	ELDA	77	2104
LE MONDE	LLF	380	10081
PARLIAMENT	ELDA	276	7551
LITTERATURE	ATILF	892	24358
EMAIL	ELDA	852	9243
MEDICAL	STIM	554	11799
ORAL_DELIC	DELIC	505	8117
ORAL_ESTER	ELDA	502	5365
QUESTIONS	ELDA	203	4116
total		4241	82734

Table 3: Number of sentences and word forms annotated per genre in the EASY corpus.

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 syntactic links between these minimal constituents are annotated by means of relations, which associate these constituents to form complex syntagmas. Thus we are able to evaluate from chunkers (which only annotate simple constituents) to deep parsers (which are able to recognize complex syntagmas). The details on the annotation process may be found in the annotation guide[5]. We will only illustrate them on an example. 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 1. Let us examine the sentence : "Il arrive en retard, avec, dans sa poche, un discours qu'il est obligé de garder"[6]. To give some examples, we annotate there a nominal constituent (*un discours*[7]), a verbal constituent, which includes clitics, (*Il arrive*[8]), a prepositional constituent (*dans sa poche*[9]), an adjectival constituent (*obligé*[10]) and a prepositional-verbal constituent (*de garder*[11]). 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 of a relation between this nominal constituent and the preposition it is introduced by. Note that in the example, we have a discontinuous prepositional phrase *avec,...,un discours*[12]. Since during constituent annotation, we cannot use the shortcut of a prepositional phrase here, we only annotate the noun phrase *un discours*, and the relation with the preposition *avec* will be annotated by means of a relation (see below).

[5] www.limsi.fr/Recherche/CORVAL/easy

[6] A free translation could be: "He arrives late, with, in his pocket, a discourse, that he is compelled to keep".

[7] a discourse

[8] he arrives

[9] in his pocket

[10] compelled, but the english translation is not an adjective!

[11] to keep

[12] with,..., a discourse

EASY uses also 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 indifferently forms or constituents or a mix of both. To come back on our example, we annotate a subject between *il* and *arrive*[13], that means between the two forms included inside a verbal phrase. The relative *qu*[14] is annotated as the object of *garder*[15]. The constituents *en retard* and *avec,..., 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 *il est* and the noun *un discours*. This solution is always adopted when we have to link a secondary clause to a constituent, such as a to link by a *verb complement* a temporal subordinate clause to the verb of the principal clause, for instance. 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 may be rather frequent in some French corpora.

5. Parsing Evaluation

In EASY we have collected 16 runs from 13 different teams (9 research laboratories and 4 private companies). The participant are: CEA-LIST, ERSS, FRANCE TELECOM R&D, GREYC, INRIA-ATOLL, LATL, LIRMM, LORIA, LPL, PERTIMM, SYNAPSE DEVELOPMENT, TAGMATICA, XEROX RESEARCH CENTER EUROPE. In the following we present the results of one participant over half of the corpus as example, since at the time of writing all the results were not yet computed. To assess the performances, we use precision and recall measures with various constraint relaxations on constituent boundaries for the independent evaluation of both constituent and grammatical relation annotations. A parser may produce constituent annotations, relation annotations or both and still be evaluated under the same conditions. Different results are computed for both constituents and relations: over the whole corpus, over each genre specific subcorpus, and separately for each type of constituent or relation. The figure 2 displays different values of f-measures[16] for one of the participants according to different subcorpora and relations. The table 6 gives the number of occurrences of each relation in the evaluation corpus taken as example.

As was expected, we observe an important decrease of performance for ORAL_DELIC data (see table 4 and second row of results from the left in figure 2).

[13] between the pronoun *he* and the verb *arrives*

[14] that

[15] to keep

[16] $F = \frac{1}{\frac{\alpha}{P}+\frac{(1-\alpha)}{R}} = \frac{2\times P\times R}{P+R}$ with $\alpha = 0.5$ (Manning and Schütze, 2002)

Corpus	Sentences	Words	Av. f-measure
LITERATURE	892	24358	$0.64(\pm 0.004)$
ORAL_DELIC	505	8117	$0.26(\pm 0.007)$
PARLIAMENT	276	7551	$0.63(\pm 0.003)$
QUESTIONS	203	4116	$0.61(\pm 0.002)$
total	1875	44142	$0.53(\pm 0.004)$

Table 4: Number of sentences and word forms along with the average f-measure obtained by the participant we chose as sample over all relations for each subcorpus.

Function	Formula
EQUAL	$H = R$
FUZZY	$\lvert H \backslash R \rvert \leq 1$
INCLUDE	$H \subset R$
INTERSECTION	$R \cap H \neq \emptyset$
BARYCENTER	$\frac{2*\lvert R\cap H\rvert}{\lvert R\rvert+\lvert H\rvert} > 0.25$

Table 5: With H the hypothesis range and R the reference range, the table gives formulas for the different range equality functions.

During results computation, all information pertaining to relation sources or targets, as well as information about the constituents is mapped onto a unique representation made of word form address ranges. When comparing the ranges present in the hypothesis with the one in the reference, we use various equality functions (see table 5), which allow some latitude in the specification of the beginning and end address of the hypothesis range.

In addition, we allow three different modes of comparison between the reference and hypothesis constituents[17] (before mapping the constituent onto a word range):

1. HYP, in which only the hypothesis constituents are used for the hypothesis data,
2. DEFHYP, in which the corresponding reference constituent is used when hypothesis data mentions a form not included in any hypothesis constituent,
3. REF, in which systematically the reference constituent are used instead of the hypothesis constituent.

All the different manners of combining the 3 previous constituent evaluation modes with the various constraint relaxations possible on the word address ranges produce 15 different ways of computing the precision and recall measures for relations (see figure 3).

The variation due to the more or less strict ways of computing the equality of a relation between reference and hypothesis data is negligible for what concerns the global performance measure [18], which lead us to think that our protocol does not introduce any bias due to constituent boundaries differences in the relation performance evaluation. Of

[17] also for the evaluation of relation sources and targets.

[18] in the graph the most signifi cant variation is for mode number 2, i.e. HYP-EQUAL, the strictest way which considers only the hypothesis constituents and uses strict equality on constituents boundaries.

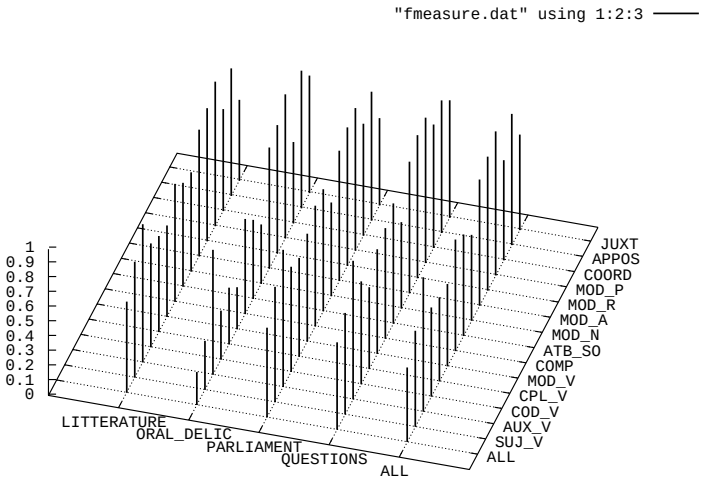

Figure 2: Different values of f-measures computed in the HYP-EQUAL evaluation mode for a participant according to PARLIAMENT, LITERATURE, ORAL_DELIC and QUESTIONS subcorpus and all syntactic relations.

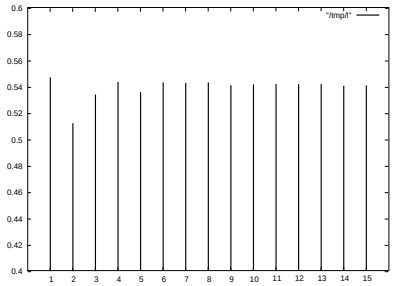

Figure 3: Different values of the average f-measure for a participant over all subcorpora for all relations according to the 15 different modes of computing equality of relation between reference and hypothesis data.

course, this partial result needs to be validated on the whole corpus and for the other participants.

relation	#	relation	#
SUJ-V	4156	MOD-N	5777
AUX-V	743	MOD-A	478
COD-V	2858	MOD-R	168
CPL-V	3294	MOD-P	14
MOD-V	1747	COORD	1358
COMP	697	APPOS	238
ATB-SO	754	JUXT	1186
total 23468			

Table 6: Distribution of the different relations in the evaluation corpus used as example.

6. Conclusion

EASY has proved the feasibility of deploying the evaluation paradigm in an evaluation campaign for parsing of French on a large corpus of various genres. The 13 different teams were able to map the output of their 16 parsers onto the EASY annotation scheme, with which parsers may produce constituent annotations, relation annotations or both and still be evaluated under the same conditions. Detailled results were computed: over the whole corpus, over each genre specific subcorpus, and separately for each type of constituent or relation. By putting more or less constraints on the evaluation of constituent boundaries, we preserved the relation performance evaluation from being biased by the constituent annotations.

7. References

A. Abeillé, L. Clément, and A. Kinyon. 2000. Building a treebank for french. In *Proceedings of the 2nd International Conference on Language Ressources and Evaluation (LREC)*, pages 1251–1254, Athen, Greece.

A. Abeillé. 1991. Analyseurs syntaxiques du français. *Bulletin Semestriel de l'Association pour le Traitement Automatique des Langues*, 32:107–120.

P. Blache and J.Y. Morin. 2003. Une grille d'évaluation pour les analyseurs syntaxiques. In *Acte de l'atelier sur l'Evaluation des Analyseurs Syntaxiques dans les actes de la 10e conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN)*, Batz-sur-Mer, juin.

E. Black, S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harison, , D. Hindle, R. Ingria, F. Jelineck, J. Klavan, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and Strzalkozskijl. 1991. A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pages 306–311, Pacific Grove, California. Morgan Kaufman.

S. Brant, S. Dipper, S. Hansen, W. Lezius, and G. Simth. 2002. The tiger treebank. In *Proceedings of the 1st Workshop on Treebank and Linguistics Thories (TLT)*, Sozopol, Bulgaria.

E. Briscoe, J. Carroll, J. Grayham, and A. Copestake. 2002. Relational evaluation schemes. In *Proceedings of the Workshop Beyond Parseval - Toward improved evaluation measures for parsing systems at the 3rd International Conference on Lnaguage Resources and Evaluation (LREC)*, Las Palmas.

J. Carroll, T. Briscoe, and A Sanfilipo. 1998. Parser evaluation: a survey and a new proposal. In *Proceedings of the 1st International Conference on Linguistic Resources and Evaluation (LREC)*, pages 447–454, Granada, Spain.

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.

J. Carroll, G. Minnen, and E. Briscoe, 2003. *Parser evaluation using a grammtical relation annotation scheme*, pages 299–316. Treebanks: Building and Using Parsed Corpora. Kluwer, Dordrecht.

R. Gaizauskas, M. Hepple, and C. Huyck. 1998. Modifying existing annotated corpora for general comparative evaluation of parsing. In *Proceedings fo the Workshop on Evaluation of Parsing Systems in the Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain.

V. Gendner, G. Illouz, M. Jardino, L. Monceaux, P. Paroubek, I. Robba, and A. Vilnat. 2003. Peas the first instanciation of a comparative framework for evaluating parsers of french. In *Proceedings of the 10th Conference of the European Chapter fo the Association for Computational Linguistics*, pages 95–98, Budapest, Hungarie. Companion Volume.

D. Lin. 98. Dependency-based evaluation of minipar. In *Proceedings of the Workshop on Evaluation of Parsing Systems*, Granada, Spain.

C. D. Manning and H. Schütze. 2002. *Foundation of Statistical Natural Language Processing*. Massachusetts institute of Technology Press, 5 edition.

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

E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2004. Annotating discourse connectives and their arguments. In *Proceedings of the Frontiers in Corpus Annotation NAACL/HLT Conference Workshop*, Boston.

G. Musillo and K. Sima'an. 2002. Toward comparing parsers from different linguistic frameworks - an information theoretic approach. In *Proceedings of the Workshop Beyon Parseval - Toward improved evaluation measures for parsing systems at the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain.

S. Oepen, K. Netter, and J. Klein. 1996. Test suites for natural language processing. In *CSLI Lecture Notes*. Center for the Study of Language and Information.

M. Palmer, D. Gildea, and P.Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1).

P. Paroubek, L.-G. Pouillot, I. Robba, and A. Vilnat. 2005. Easy : Campagne d'évaluation des analyseurs syntaxiques. In *Proceedings of the 12e conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 3–12, Dourdan, France.

B. Roark. 2002. Evaluating parser accuracy using edit distance. In *Proceedings of the Workshop Beyon Parseval - Toward improved evaluation measures for parsing systems at the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain.

B. Srinivas, C. Doran, B.A. Hockey, and K. Joshi. 1996. An approach to robust partial parsing and evaluation metrics. In *Proceedings of the Workshop on Robust Parsing*, Pragues. ESSLI.

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.

F. Xia and M. Palmer. 2000. Evaluating the coverage of ltags on annotated corpora. In *Proceedings of the Workshop on Using Evaluation within HLT Programs: Results and Trends in Proceedings of the 2nd International Conference on Language Ressources and Evaluation (LREC)*, Athen.