# A Protocol for Evaluating Analyzers of Syntax (PEAS)

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

*LIMSI-CNRS
Batiment 508 Universite Paris XI, BP 133 - 91403 ORSAY Cedex - France
{gendner, jardino, monceaux, pap, robba, vilnat}@limsi.fr

†Columbia University
2960 Broadway New York (NY)
gabrieli@cs.columbia.edu

**Abstract**

Providing a comparative framework for parsers is a task that has already been tried in the past, e.g. (Abeillé, 1991), (Atwell and Sutcliffe, 1997), (Black et al., 1991), and studied in the literature (Black, 1993), (Black, 1994), (Carroll et al., 1998), (Gaizauskas et al., 1998), (WEPS-98, ), (Mengel and Lezius, 2000), but mainly for English. In this paper, we present PEAS: a Protocol for Evaluating Analyzers of Syntax (in French: Protocole d'Evaluation pour les Analyseurs Syntaxiques), based on an ongoing experiment at LIMSI which aims at developing and testing a generic quantitative black-box evaluation protocol for parsers of French. Two fully operational parsers will be used to test the evaluation protocol; they are: the parser (Giguet and Vergne, 1997) developed at GREYC (Caen University) and the latest version of the parser developed at Rank Xerox Research Center in Grenoble (Ait-Mokhtar and Chanod, 1997)

## 1. Introduction

After a presentation of the problem through a literature review, we will present how PEAS builds upon previous efforts, and describe in details the current stage of development of the PEAS project[1]. In particular we will review:

1. the syntactic information taken into account by the evaluation process
2. the way it is encoded,
3. the evaluation metric and the protocol,
4. the data and the systems under test,
5. the points left aside and the reasons for doing so,
6. and the current results.

The most commonly mentioned evaluation scheme computes precision and recall metrics on syntactic information and the number of crossing brackets on phrase boundaries. However, (Lin, 1995) has outlined the shortcomings of this scheme and has proposed a new method using dependency relations instead of phrase boundaries. In a similar way (Gaizauskas et al., 1998) have argued that the current dominant paradigm which combines the Penn Treebank and the Parseval scoring scheme (here again recall, precision and number of crossing brackets) is not well suited for the evaluation of parsers. Hence they propose to use flatter annotations in the reference corpus and the new evaluation metrics recall and conformance (proportion of the gold standard constituents that are not crossed by any constituent in the response).

But the key issue of any parser metrics lies with the reference annotation. To guide the specification of these, one can use either:

1. the target application requirements,
2. the syntactic phenomena studied in the literature or
3. the sentences of a corpus chosen as representative of the target task.

Since we aim at remaining as much as possible task independent, we chose the second option, and decided to start from the results of TSNLP [2] (Lehmann et al., 1996a) (Lehmann et al., 1996b), and its successor DiET (Klein et al., 1998). We will use the syntactic phenomena (maximum coverage with minimum description length) to guide the design of the reference annotation while remaining as close as possible to the information produced by the two test parsers used in PEAS. For the evaluation, the parsers output will be mapped onto the reference formalism with the information that will have been provided in advance by the parser developers. In order to get an insight into the robustness of the systems under test and to use material more likely to be encountered in an information search (an application task at the center of today NLP preoccupations), we decided to use as corpus a set of large sized excerpts from various media and genres: automatic transcripts of audio sources, Q&A TREC questions, Le Monde newspaper, Internet newsgroups, literature, texts from free online corpora, excerpts from ELRA corpora, and some randomly selected WEB pages.

Our evaluation metrics will be applied to sentences (segmentation provided in the reference, although a participant may choose to ignore the sentence segmentation provided) annotated with the following syntactic information: first level constituents bracketing, constituent attributes (syntactic function tags) and attachments. Note that all the constituents of level higher than two will be automatically transformed into attachments both in the reference

[1] http://www.limsi.fr/Recherche/CORVAL/PEAS

[2] http://cl-www.dfki.uni-sb.de/tsnlp/

and the system parses. In addition to sentence boundaries, PEAS will start by using the five syntactic tags described in section 5 which were inspired by the ones ones mentioned in SPARKLE (Carroll et al., 1996) and the 12 tags proposed by A. Abeillé in her Treebank (Abeillé et al., 2000) (Abeillé et al., 2001).

Parser evaluation schemes propositions are almost as numerous as propositions for syntactic annotation (Mengel and Lezius, 2000), (Ide and Romary, 2001), (Lenci et al., 2000). The problem here is to remain as neutral as possible with respect to the formalisms used by the parsers. Finally, we will present our conclusion on the feasibility of deploying our evaluation protocol in a larger context.

## 2. Robust Syntactic Parsing

Over the past years, syntactic parsing knew an important evolution. Whereas, the first syntactic parsers were developed in order to recognize "linguistic phenomena" present in various test suites, today, the necessary priority of syntactic parsers is the robustness: to be able to return a syntactic parse for all sorts of input; even when they are ungrammatical or contain extralinguistic phenomena like markup tags, hesitations etc.

These robust parsers are generally deterministic, incremental but do not necessarily return the same results (minimal constituent segmentation or more complex constituents and sometimes dependency relations) and the strategies they use vary also greatly. We can distinguish two families of shallow parsers: the symbolic / linguistic parsers, based on grammatical formalism and the probabilistic / statistical parsers, based on corpus learning. The linguistic parsers themselves can be futher divided into three categories:

- constituent-based parsers (Abney, 1996) (Wehrli, 1992) which return the constituent segmentation of the sentence (recognition of noun phrase, prepositional phrase...)
- dependence-based parsers (Link Grammar (Sleator and Temperley, 1991)) which return the word (or phrase) dependences (recognition of Subject relation, ...)
- and constituent-and-dependence-based parsers which return the constituent segmentation and the links between these constituents.

To test our evaluation protocol, we will use the linguistic constituent-and-dependence-based parsers of institutions which collaborate on PEAS: the parser developped at GREYC (Caen University) in the team of J. Vergne and the latest version of the parser developped at Rank Xerox Research Center in Grenoble, in the team of J.P Chanod.

### 2.1. Incremental Deep Parsing of Xerox

The parser developed at Xerox is a robust incremental deep parsing. The system aims at producing a set of dependency relations, by applying dependences rules defined by a grammarian. These dependency relations correspond to linguistic relations between several words or groups of words. Before performing these dependency analysis, the system has three optional modules : tokenization and morphological analysis, POS disambiguation and chunking which are activated depending on the type of input. Indeed one of the system advantages is the possibility to take as input different kinds of linguistic objects such a ASCII text, a sequence of consitutent structure...

The role of the chunking rules is to group sequences of categories into structures (chunks) in order to facilitate the dependency analysis. This chunking module is determinist, i.e it returns one constituent segmentation. This dependency analysis has a bipartite input: the initial constituent tree and the incremental set of dependencies. Indeed the dependency rules are applied in sequence (they are applied by level) and rely on the evolving background knowledge stored in the syntactic tree and in the dependency set. This dependency analysis module is not determinist, for example for the PP (prepositional phrase) attachment.

### 2.2. Syntactic parsing of unrestricted French of GREYC

The parser developed at GREYC is a determinist robust system for syntactic parsing of unrestricted French. In this system, syntactic parsing means identifying constituents, called non-recursive phrases (nr-phrases) and linking them together. This system is based on the work of Tesniere (59) but in the parser developed at GREYC, the notion of dependency corresponds to relations between nr-phrases and not between word forms. The system architecture combines two techniques:

- Part-of-Speech Tagging and Chunking techniques at word-level that build a constituent structure (each constituent is an nr-phrase),
- Linking rules at nr-phrase-level that link nr-phrases to build a functional structure.

These two structures are build simultaneously by two interactive processes. The analysis is carried out deterministically from left to right.

## 3. Previous evaluation campaigns

### 3.1. Test Suite Evaluation of Parsers of French

(Abeillé, 1991) compared six French-language parsers, in a limited experiment. Since it was impossible to deal with all the aspects of syntactic analyses, the focus was put on linguistic formalization and on the difficulties encountered when trying to build grammars covering a large amount of phenomena. The corpus used to compare the parsers was made of 30 sentences, each one dedicated to one or more syntactic phenomena, such as past participle agreement, prepositional phrases attachment, relative clauses, and so on. The results were only given in terms of the number of analyses produced, going from 0 to 16, too many analyses being as bad as no result at all. But when an analysis was built, there was no judgement given on its pertinence. Even if most of the parsers were based on unification grammars (lexical functional grammar, tree adjoining grammar or categorial unification grammar), the underlying syntactic theories gave rise to different choices (for example, an infinitive could be be a verb phrase or a

sentence with an empty subject). This experiment, which to our knowledge was the first published on parsers of French, illustrates the different problems encountered when trying to peform a real evaluation: even with a limited corpus, the authors had to modify some sentences to be able to compare the performances of the parsers on the phenomena they wanted to examine. For instance, some were not able to deal with numeral determiners, while others transformed particular expressions into compound nouns, etc. With these modifications, only about 16 sentences among the 30 initially given, were analyzed. In our opinion, the lessons to draw from this experiment are first that one should not try to test too many different linguistic phenomena per test sentence (later TSNLP (Lehmann et al., 1996a) was an attempt at solving this particular point by providing a diagnostic copus, holding at most 1 phenomena per sentence), and second, that an evaluation metric relying only upon the number of analyses returned is not informative enough to provide a sound comparison basis.

### 3.2. Test Suite Evaluation of Parsers of English

A rather similar experiment has been made by (Atwell and Sutcliffe, 1997) on English-language parsers. Eight parsers were evaluated on a small corpus of sentences from software manuals (IPSM corpus). The goal was to compare speed, efficiency, coverage and accuracy. After the tests, the authors realized that restricting the evaluation to such objectively quantifiable surface features was too limiting. It was impossible to rely on what was counted as a correct parse: different amounts of customization were necessary, the numbers of parses yielded were very different, and most significantly, the outputs were very different in both format and contents. Then, the authors argued that a fair comparison can be made by mapping each parser output onto the standard proposed by EAGLES (Leech et al., 1995) for grammatical annotation, organized in 8 hierachical layers:

1. Bracketing of segments
2. Labelling of segments
3. Showing dependency relations
4. Indicating functional labels
5. Marking subclassification of syntactic segments
6. Deep or "logical" information
7. Information about the syntactic unit rank
8. Special syntactic characteristics of spoken languages

With this generic framework, the authors were able to give a weighted score to the parsing schemes in terms of how many of the EAGLES levels were covered, giving more importance to higher layers. They concluded that it is not sensible to apply a single accuracy metric across all domains, but that users should first decide what they want from a parser, in terms of generic functionality as exemplified by EAGLES standard, and then gauge the parsers according to their specific criteria.

### 3.3. Parseval

(Black et al., 1991) describes the comparison of 10 parsers using only constituent boundaries, in particular the tags associated to the constituents were not taken into account. Initially planned to be made against hand-parsed sentences from Penn TreeBank, the evaluation reported was done against a reference build from a majority vote out of the output parses. For each system, 2 measures were computed: 1) the number of crossing parenthesis and 2) the recall (number of parses both in the output and the reference over the number of parses in the reference). The measures were performed on versions of the output and of the reference that had been first submitted to normalization (auxiliaries deletion, deletion of "not", deletion of punction marks etc.). The test held 14 sentences and the average results obtained were respectively of 4% crossing-brackets and 94% recall. In the litterature, this evaluation scheme is usually referred to as "Parseval" (Harrison et al., 1991).

In a follow-up paper (Black, 1994), the author takes a different track and advocates, instead of evaluating on common data, to perform a distinct evaluation for each systems againts his own goals and notations. He opposes parsing evaluation, with its measures that he qualifies of "subjective" (system formalism dependent), to language modeling evaluation, with its "objective" measures: error rate and perplexity. Here again the measures described for parser evaluation are the number of crossing brackets and recall. The author mention the problem of ambiguous reference parses, but list only properties of the proposed protocol for standalone evaluation without relating any specific evaluation or giving any solution for he extra cost that such evaluation scheme incur because of the multiple reference data sets required (1 per system). In (Black, 1993), the author draws a picture of the state of the art in parsing, and present the results of a separate evaluation of 5 systems which he extracted from the litterature. Depending on the systems, the measures used were either the percentage of sentences which had no crossing brackets or the percentage of sentences which matched exactly the reference (the results vary from 69% to 29%), obtained with a number of test sentences varying between 100 to 10,000. In comparison, the author quotes the results of the UPenn evaluation from 1992: with a best score of 41% and an average score of 22%.

### 3.4. Dependency Evaluation

(Lin, 1995) has criticized the Parseval evaluation methods based on phrase boundaries and proposed a whole new method based on the dependency relations. Moreover, her scheme can also be applied to constituent based parsers, since she has worked out a procedure to transform the constituents of a sentence into a dependency tree.

In her system, the key is a dependency tree in which each word of the sentence is a modifier of exactly one other word. And only the head of the sentence is not a modifier. The dependency relations are described using the following features: the modifier, its category, the position of the head with respect to the modifier, the relation between the modifier and its head. Since this last element is optional, it is not taken into account in the evaluation process. The answer is

also a dependency tree, in which if a head is unknown, its position is noted with a "?". Given the dependency tree, the evaluation is rather simple: the error count is the number of words that are assigned a different head in the key and in the answer.

### 3.5. Evaluation with Grammatical Relations

In LREC 1998, (Carroll et al., 1998) presented an evaluation scheme based on the use of grammatical relations. He argued that an independant language had to be build to represent the information given by the parsers, especially a language including all grammatical relations. He based this language on a LFG F-structure in AVM (Attribute-Value Matrix) notation and developed a three layered approach like in EAGLES. This new evaluation scheme, used in SPARKLE, was partially applied to 500 english sentences from SUZANNE. Recall and precision were computed from argument relations only, modification relations were not evaluated. Although they had few test results, the authors thought that their grammatical relation scheme present several improvments over a conventional constituency based scheme.

### 3.6. Evaluation with Flatter Keys

In 1998, (Gaizauskas et al., 1998) proposed a new scoring scheme for the evaluation of parsing systems. They argued that the dominant paradigm, combining the use of the Penn Treebank reference corpus and the Parseval scoring system, (computing three metrics: recall, precision and number of crossing brackets) was not well-adapted to the task of evaluation of parsing systems.

They proposed to use a more flatter corpus, only encoding the constituents for which "there is a broad agreement across a range of grammatical theories". This corpus would only be dedicated to evaluation whereas the Penn Treebank had also other other objectives. Moreover such a corpus could be derived from existing annotated corpus.The annotated contituents are: sentence, clause, noun phrase, verb phrase, prepositional phrase, adverb phrase and adjective phrase.

Then, the proposed metrics are:

1. the recall, which is the proportion of key constituents that are also present in the response.

2. the conformance, which is the proportion of key constituents that are "not crossed by any constituent of the response", it is indeed a modified calculus of the classical number of crossing brackets, which have the disadvantage to penalize several time the same error, especially when the parsing system assigns complex structures.

It is worth noting that since the proposed corpus encodes few constituents, there is no more reason to compute the precision which would penalize parsing systems assigning a more complex structure than the minimal one recorded in the corpus.

## 4. The treebanks

### 4.1. The Penn Treebank

The motivation in constructing the Penn Treebank was providing a research tool for natural language processing, speech recognition, and also theoretical linguistics. The Penn Treebank is a large annotated corpus, it contains over 4.5 million words of American English. It is entirely annotated with Part-Of-Speech tagging, and about two-third of it is annotated for skeletal syntactic structures (bracketing). The annotation process, both for part-of-speech and for syntactic annotation was done in two stages: an automated one followed by a manual correction stage. For the syntactic annotation stage in which we are particularly interested, a deterministic parser was used to assign an initial bracketing. This parser provides only one analysis, it never attaches a constituent if it cannot be sure of this attachment, and it has a rather good grammatical coverage. The syntactic tagset is precisely described in (Marcus et al., 1993). During the correction stage, the annotators use a mouse-based package to link together the unattached structures. After a first utilization phase, the users asked for a richer annotation, and an increased consistency in the corpus. So in 1994, (Marcus et al., 1994) proposed a new annotation scheme that enabled to annotate predicate-argument structures.

### 4.2. A.Abeillé (Le Monde)

The work presented by (Abeillé et al., 2000) (Abeillé et al., 2001) is the first attempt to build a treebank for French language. Their goal is to build a corpus completely annotated for morphosyntax and syntax (with a control of quality), which will be useful as well for computational linguists (for taggers and/or parsers training, parsers evaluation) than for traditional linguists or psycholinguists (to observe some rare constructions, or to measure the difference between what is possible and what happens). The corpus is made of articles extracted from "Le Monde" newspaper, between 89 and 93. It contains 1 million words, 17000 different lemma, 33000 sentences. It covers different domains, from economics to sports. At the tagging level, the authors chose to use 14 lexical tags and 12 types of phrases. It is worth noting that there is no Verbal Phrase, which is not useful for French. The annotation has been processed in two steps : a morpho-syntactic annotation (fixing word and sentence boundaries, tagging, stemming), and a syntatcic annotation chunking, and assigning grammatical functions to these chunks). At each step, there is first an automatic pass, and then a systematic human validation/correction, based on very detailed annotation guides. At the time of writing of this article, the first step is achieved. Concerning the second step, the corpus has been parsed by a shallow parser and the chunks have been validated or corrected by human annotators. The determination of the grammatical functions is not yet operational.

### 4.3. PEAS Corpus

In order to get an insight into the robustness of the systems under test and to use material more likely to be encountered in an information search (an application task at the center of today NLP preocupations), we decided in

PEAS to use as corpus a series or large sized excerpts from various media and genres:

- automatic transcripts of audio sources (AirFrance corpus)
- Q&A TREC track questions,
- Le Monde newspaper,
- Internet newsgroups (archives of LN-FR group),
- literature, texts from free online corpora (ABU[3]),
- and some randomly selected WEB pages.

For a first try, the total size of the corpus will be approximatively of 1 million words, out of which we will annotate 20,000 forms for computing the evaluation measures. Indeed, in these different corpus, parsers will have to deal with the extralinguistic phenomena such as hesitations, the markup tags, ungrammatical sentences and phenomena like titles, lists etc.

## 5. PEAS annotations

PEAS annotation scheme is meant for French (for an interlingua annotation scheme see (Lenci et al., 2000)). Our purpose is to determine basic elements that can express information produced by any linguistic theory. It is based on non recursive chunking and relations between words, between words and chunks or between chunks. The chunks are non recursive in order to simplify comparison metrics. They are as small as possible so that any segmentation chosen for a system can be converted into one or a combination of our basic chunks. We put the emphasis on expliciting all functional information: for example, the verb-object relation is annotated with a specific relation and does not have to be deduced from the embedding of a noun phrase in a verb phrase (Ide and Romary, 2001). The model presented in (Ide and Romary, 2001) involves four levels of information and is based on a tree-structure (called structural skeleton) which gives to the model a constituency based orientation. In order to remain as simple as possible during our preliminary test of our protocol, we chose an opposite solution and based our annotation scheme on dependency models (see also (Lin, 1995) (Lin, 1996), (Lin, 1998), (Carroll et al., 1998), (Lenci et al., 2000). In PEAS, we distinguish 5 types of chunks:

1. verbal <NV>,
2. prepositional <GP>,
3. nominal <GN>,
4. adjectival <GA>,
5. adverbial <GR>.

And 9 relations express functional information:

1. subject-verb,
2. auxiliary-verb,
3. argument-verb,
4. attribute-subject,
5. attribute-object,
6. modifier-verb,
7. modifier-noun,
8. modifier-adjective,
9. modifier-adverb.

The type (<GN> or <GP>) of the verb argument indicates whether it is direct or indirect. A third boolean argument of the argument-verb relation marks the agent in passive constructions.

Coordination is represented with a relation which specifies the coordinating element (conjunction, coma ...) and the two coordinated elements. Another relation expresses apposition.

As in a dependency based model, the complex structure of the sentence can be restituted using a chain of relations. For example in (1) three nominal syntagms (SN1, SN2 and SN3) are coordinated.

(1) [...] <SN1> <GN1> la porte </GN1> de la chambre fermée à clef à l' intérieur </SN1>, <SN2> <GN2> les volets </GN2> de l' unique fenêtre fermés , eux aussi , à l' intérieur , et , par-dessus les volets , <SN3> <GN3> les barreaux </GN3> intacts </SN3>, [...] [4]

With PEAS format, we note :
COORD(",", GN1, GN2) and
COORD("et", GN2, GN3).

The chain of relations that link all modifiers to the head noun allows the restitution of the limit of the syntagm, as illustrated in first part of figure 1 for the first nominal syntagm, and also given here in PEAS format:
MOD_N(GP1,"porte")
MOD_ADJ(GP2,"fermée")
MOD_ADJ(GP3,"fermée")
MOD_N(GA1,"porte")

For this reason, no clausal or sentential segmentation is identified.

Another relation marks complementizer and the two related <NV>. For now, other introducers (such as preposition) are not explicitly marked. This may be added later on if necessary to express information provided by some system.

In order to stick to a non recursive chunking, we decided to keep all modifiers placed before a noun in the same <GN> as the noun itself. The modification relations are

[3] http://abu.cnam.fr/

[4] As our evaluation concerns French language parsers,the examples are given in French. We will give here a literal translation of this sentence extracted from (Leroux, 1907) : the door of the room locked from inside, the shutters of the single window also closed from inside, and over the shutters, the bars intact,[...]

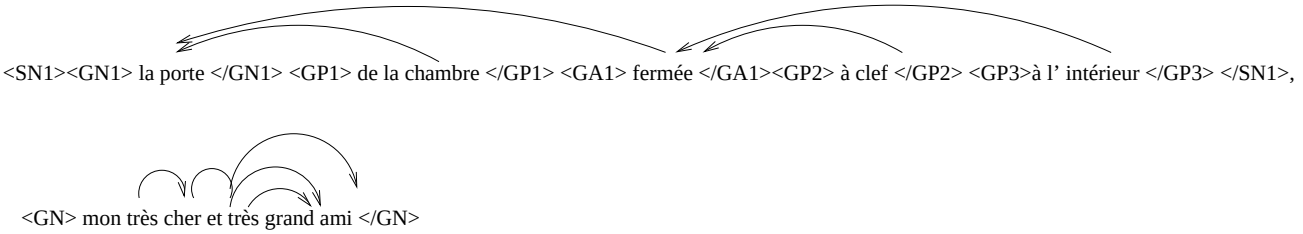


Figure 1: Chain of relations restituting complex syntagms.

then expressed between words, as illustrated below and in second part of figure 1.

(2) <GN> mon très cher et très grand ami </GN>[5]

MOD_ADJ("très","cher")
MOD_ADJ("très","grand")
MOD_N("et","ami")
COORD("et", "cher", "grand")

In a first step, it has been decided not to mark and not to evaluate determiners. Since they form a closed class, they can later be identified automatically. Also, contrary to (Lenci et al., 2000), only surface syntax phenomena are taken into account: subject control and raising object/subject relation are not annotated: in (4), the only subject-verb relation annotated is SUBJ_V("Jean "," promis ")

(3) Jean a promis à Marie de partir [6]

The first part of the manually tagged reference which is about to be produced and the first test of two systems are meant to test this scheme against real data and existing systems and their theoretical choices. If necessary, attribute such as those described in (Lenci et al., 2000) will be introduced to express more detailed information such as verb mode, diathesis, nominal quantification or completeness/missing arguments in subordinate clause.

Note that as for the annotation models described in (Ide and Romary, 2001) and (Mengel and Lezius, 2000), we use XML as an encoding format.

## 6. Conclusion

Since a decade, several scoring schemes have been elaborated and discussed. The first of them has been the origin of new proposals which are more fair and more open to the variability of parsing systems outputs. For instance the proposals like those of Lin or Gaizauskas will enable our protocole to evaluate a parser like the one recently developed at Xerox, which produces two outputs: the constituents and the dependency relations. As to the metrics that we will be using, they will be based on precision and recall, measured on specific classes of linguistic phenomena, after a preprocessing phase of annotation generalization/normalization which will bring the parses of the systems to the same level of complexity as the one of the reference annotations, transforming constituants into dependencies when necessary. Their exact nature will be determined once we have run the first tests with the annotated reference corpus (whose annotation is now in progress) and will be described in subsequent articles. We believe that deploying the evaluation paradigm using a comparative and quantitative black-box evaluation methodology (Mariani and Paroubek, 1999) is a key asset for both the development of parsing technology and the production of validated and high quality language resources (Paroubek, 2000).

[5] literal translation : my dearest and closest friend

[6] translation : John promised Mary to leave.

## 7. 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 (LREC2000)*, volume 1, pages 87–94, Athens, Greece, May. ELRA.

A. Abeillé, L. Clément, A. Kinyon, and F. Toussenel. 2001. Un corpus français arboré : quelques interrogations. In *Actes de la conférence sur le Traitement Automatique de la Langue Naturelle (TALN 2001)*, pages 33–42, Tours, juillet.

A. Abeillé. 1991. Analyseurs syntaxiques du français. *Bulletin Semestriel de l'ATALA (Association pour le Traitement Automatique des LAngues)*, 32(2):107–120. ISSN 0039-8217.

S. Abney. 1996. Partial parsing via finite-state cascades. *J. of Natural Language Engineering*, 2(4):337–344.

S. Ait-Mokhtar and J. Chanod. 1997. Incremental finite state parsing. In *Proceedings of ANLP-97*, Washington, March.

E. Atwell and R. Sutcliffe. 1997. Industrial parsing of software manuals: Empirical qualitative comparison of parsers and parsing schemes. In *Proceedings of the Speech and Language Technology (SALT) Club Workshop on Evaluation in Speech and Language Technology*, Halifax Hall, University of Sheffield, Sheffield, UK, June.

E. Black, S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, 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 DARPA, editor, *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, pages 306–311, Pacific Grove, California, February. Morgan Kaufmann.

E. Black. 1993. Parsing english by computer: The state of the art. In *Proceedings of the International Symposium on Spoken Dialog (ISDD-93)*, pages 77–81, Tokyo, November. Waseda University.

E. Black. 1994. A new approach to evaluating braod-coverage parser/grammars of english. In *Proceeding of the International Conference on New Methods in Language Processing (NemLap)*, pages 59–65, UMIST, Manchester UK, September. Centre for Computational Linguistics.

J. Carroll, Ted Briscoe, Nicoletta Calzolari, Stefano Federici, Simonetta Montemagni, Vito Pirrelli, Greg Grefenstette, Antonio Sanfilippo, Glenn Carroll, and Mats Rooth. 1996. Sparkle work package 1 - specification of phrasal parsing - prefinal report. Technical report, SPARKLE project, May. http://www.ilc.pi.cnr.it/sparkle/wp1-prefinal/wp1-prefinal.html.

J. Carroll, T. Briscoe, and A. Sanfilippo. 1998. Parser evaluation: a survey and a new proposal. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC98)*, volume 1, pages 447–454, Granada, Spain, May. ELRA.

R. Gaizauskas, 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 (LREC98)*, volume 1, pages 143–149, Granada, Spain, May. ELRA.

E. Giguet and J. Vergne. 1997. Syntactic analysis of unrestricted french. In *Proceedings of the International Conference on Recent Advances in Natural Languages Processing (RANLP'97)*, pages 276–281, Tzigov Chark, Bulgaria, September.

P. Harrison, S. Abney, E. Black, D. Flickinger, C. Gdaniec, R. Grishman, D. Hindle, B. Ingria, M. Marcus, and T. Strzalkowski B. Santorini. 1991. Evaluating syntax performance of parser/grammars of english. In G. Leech and R. Garside, editors, *Proceedings of the Workshop on Evaluating Natural Language Processing Systems*. ACL.

N. Ide and L. Romary. 2001. A common framework for syntactic annotation. In *Proceeding of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 298–305, Toulouse, France, July. ACL.

J. Klein, S. Lehmann, K. Netter, and T. Wegst. 1998. Diet in the context of mt evaluation. In *Proceedings of KONVENS-98*, Bonn, October.

G.N. Leech, R. Barnett, and P. Kahrel. 1995. Eagles final report and guidelines for the syntactic annotation of corpora. Technical Report EAGLES Document EAG-TCWG-SASG/1.5, Istituto di Linguistica Computazionale, Pisa. http://www.ilc.pi.cnr.it/EAGLES/home.html.

S. Lehmann, D. Estival, and S. Oepen. 1996a. Tsnlp des jeux de phrases-test pour l'evaluation d'applications dans le domaine du taln. In *Actes de la conférence sur le Traitement Automatique de la Langue Naturelle (TALN 1996)*, Marseille, May.

S. Lehmann, S. Oepen, S. Regnier-Prost, K. Netter, V. Lux, J. Klein, K. Falkedal, F. Fouvry, D. Estival, E. Dauphin, H. Compagnion, J. Baur, L. Balkan, and D. Arnold. 1996b. Tsnlp — test suites for natural language processing. In *Proceedings of COLING'96*, Cophenhagen, July.

A. Lenci, S. Montemagni, V. Pirrelli, and C. Soria. 2000. Where opposites meet. a syntactic meta-scheme for corpus annotation and parsing evaluation. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000)*, volume 2, pages 625–632, Athens, Greece, May. ELRA.

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

D. Lin. 1995. A dependency-based method for evaluating broad-coverage parsers. In *Proceedings of IJCAI-95*, pages 1420–1425, Montréal, Canada.

D. Lin, 1996. *Industrial Parsing of Software Manuals*, chapter Dependency-based parser evaluation: a study with software ma nual corpus, pages 25–46. Rodopi, Amsterdam.

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

M. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330.

M. Marcus, G. Kim, M.A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The penn treebank: annotating predicate argument structure. In *Proceedings of the ARPA'94 conference.*

J. Mariani and P. Paroubek. 1999. Human language technologies evaluation in the european framework. In *Proceedings of the DARPA Broadcast News Workshop*, pages 237–242, Whashington, February. Morgan Kaufman.

A. Mengel and W. Lezius. 2000. An xml-based representation format for syntactically annotated corpora. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000)*, volume 1, pages 121–126, Athens, Greece, May. ELRA.

P. Paroubek. 2000. Language resources as by-product of evaluation: the multitag example. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000)*, volume 1, pages 151–154, Athens, Greece, May. ELRA.

D. Sleator and D. Temperley. 1991. Parsing english with a link grammar. Technical report Carnegie Mellon University, School of Computer Science CMU-CS-91-196.

E. Wehrli. 1992. The ips system. In *Proceedings of Coling-92*, pages 870–874, Nantes, July.

WEPS-98. 1998. Workshop on the evaluation of parsing systems / lrec'98. Granada, Spain, May. Editors: John Carroll and Roberto Basili and Nicoletta Calzolari and Robert Gaizauskas and Gegory Grefenstette.