Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC'98), Granada, 1998 pp. 143-149.

A Scheme for Comparative Evaluation of Diverse Parsing Systems

R. Gaizauskas, M. Hepple & C. Huyck

{R.Gaizauskas,M.Hepple,C.Huyck}@dcs.shef.ac.uk

Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello Street
Sheffield S1 4DP UK

Abstract

We argue that the current dominant paradigm in parser evaluation work, which combines use of the Penn Treebank reference corpus and of the Parseval scoring metrics, is not well-suited to the task of general comparative evaluation of diverse parsing systems. We propose an alternative approach which has two key components. Firstly, we propose parsed corpora for testing that are much flatter than those currently used, whose "gold standard" parses encode only those grammatical constituents upon which there is broad agreement across a range of grammatical theories. Secondly, we propose modified evaluation metrics that require parser outputs to be 'faithful to', rather than mimic, the broadly agreed structure encoded in the flatter gold standard analyses.

1. Introduction

Interest in the evaluation of language technology has grown immensely in the past few years. This interest varies depending on the perspective one has on the technology: users and suppliers want to know how accurate, usable and reliable the technology is with respect to the tasks they or their customers wish to perform; researchers want to understand the strengths and limitations of various techniques; funders want to assess the success or failure of research projects.

Clearly what gets evaluated and how one goes about it depends very much on which of these perspectives one adopts. In this paper we shall be taking the perspective of researchers proposing what we believe to be a generally useful way to (at least partially) evaluate parsing technology. In the terminology of Crouch *et al.* (1995), parsing is a *component language technology* that performs a *user transparent* task. That is, parsers appear as parameterisable components in a range of systems fulfilling some user-oriented, language-related task, but parser outputs are in themselves of no direct interest to a user. Appropriate metrics, therefore, are not user satisfaction or enhanced user performance at some task, since the contribution of a component like a parser to such outcomes is too difficult to measure, or at least too difficult to measure in a way that feeds back useful information to parser developers. Rather, parser output needs to be directly assessed in a way that enables researchers to better understand and develop the technology. Of course, this assessment needs to be globally tempered by the demonstration that parsing technology does in fact contribute to the construction of enhanced language processing application systems.

Given a reasonably broad definition of parsing, there is every evidence that it does. To take just a small set of examples, most of the entrants in the DARPA MUC evaluations (DARPA, 1995) and many of the current EC research amd technology development projects (LangEng, 1998), including FACILE, ECRAN, and SPARKLE do at least some form of phrasal recognition or partial parsing; many would do more if it were more reliable. The energy going in to this area of language engineering is a strong argument in favour of developing suitable approaches to evaluating it. Where there are no yardsticks, there is tendency for research efforts to be unfocused and repetitive. If objective measures can be agreed, winning techniques will come to the fore and better technology will emerge more efficiently. As developments in language technology now stand poised to spread rapidly beyond a core set of languages where interest has historically lain, such arguments in favour of developing effective evaluation measures for parsing technology become even stronger.

In the following, we argue that the current dominant paradigm in parser evaluation work, which combines use of the Penn Treebank reference corpus with use of the Parseval scoring scheme, is not well-suited to the task of general comparative evaluation of diverse parsing systems, and propose an alternative approach which has two key components. Firstly, we propose parsed corpora for testing that are much flatter than those currently used, whose "gold standard" parses encode only those grammatical constituents upon which there is broad agreement across a range of grammatical theories. Secondly, we propose modified evaluation metrics that require parser outputs to be 'faithful to', rather than mimic, the broadly agreed structure encoded in the flatter gold standard analyses.

2. Approaches to Parser Evaluation

There are a variety of parser evaluation approaches that have been proposed,[1] which can be subclassified in a number of ways. A key division is between approaches which employ a 'grammatical' notion of coverage, i.e. whether or not specified grammatical constructions are handled, and

[1] See Carroll *et al.* (1996) for a survey.

those which employ a corpus-oriented notion of coverage, i.e. where the performance on each sentence in some real corpus of text is averaged. The former perspective is exemplified by, for example, (Flickinger *et al.*, 1987) and (Lehmann *et al.*, 1996), who provide 'test suites', i.e. manually constructed collections of test examples, which include sentences that exemplify each of some specified list of grammatical constructions, as well as related ungrammatical strings (used to discover overgeneration). Such test suites have principally been used as a tool in large-scale grammar development, facilitating fault diagnosis and progress evaluation. The second class of approaches, employing corpus oriented notions of coverage, can be subclassified in terms of whether or not they require an annotated (i.e. parsed) reference corpus. Approaches using an unannotated corpus (e.g. a calculation of the proportion of sentences that receive at least one parse) have the advantage of simplicity, but provide a poor basis for evaluation, given that there is no discrimination of correct and incorrect analysis (where the latter might, in the worst case, involve overgeneration by the grammar). Annotated corpus approaches can make this distinction, but face a problem in terms of the expense of generating annotated resources, due to the manual effort required (even where there is just manual checking/correction of automatically generated initial parses).

Returning to the initial division, between grammatical and corpus-oriented notions of coverage, we note that there need be no simple correspondence between the relative performance of two systems with respect to a given test suite and their relative performance with respect to a corpus. For a parsing system to provide a basis for real-world NL applications, it must be able to perform well on real sentences, in which different linguistic constructions appear in 'natural' distribution. However useful test suites may be to the task of grammar development, we consider (annotated) corpus oriented approaches to provide the best basis for comparative evaluation of parsing systems in relation to their likely utility for NLP applications.

3. Evaluation Using Parse-Labelled Corpora: Limitations of the Dominant Paradigm

Although there are a number of metrics that have been used in work employing parse annotated corpora,[2] one approach has come to dominate in results reported at recent computational linguistics conferences — the Parseval scheme (Black *et al.*, 1991). This scheme compares a candidate parse (the response) with its reference parse from the annotated corpus, and delivers scores under three different metrics: *precision*, *recall* and *crossing brackets*. Precision is the proportion (or percentage) of constituents in the response that appear also in the key parse, i.e. what proportion of the parse found (the response) is correct. Recall is the proportion of constituents in the key that appear also in the response parse, i.e. what proportion of the 'correct' parse the parser has found. In the basic scheme, precision and recall are computed purely in terms of unlabelled brackettings identifying constituent units; a stricter scheme compares labelled brackettings. The third metric counts the number of bracket crossings, i.e. the number of response constituents that violate the boundaries of a constituent in the key (which occurs where the two bracketted sequences overlap but neither is properly contained in the other). Of course, the scheme can only be used to evaluate a parser with respect to a given annotated reference corpus, and the one that has been most widely used is the Penn Treebank (PTB) (Marcus *et al.*, 1993). This conjunction of Parseval scheme plus PTB reference corpus has effectively become the 'dominant paradigm' of parser evaluation in recent years. The existence of such a dominant paradigm has some perhaps predictable consequences in terms of encouraging work directed towards producing parsers and grammars that are very much tuned to the particular grammar implicit in this reference corpus.

Given its dominant status, it is important to consider whether the Parseval/PTB paradigm really does provide a suitable basis for comparison of parsing systems. For any given example, the way to optimise precision and recall scores is to produce a parse that is as close as possible, and preferably identical, to the key. This seems reasonable as a basis for evaluating the output of a system that is tuned to the PTB, and hence also for comparing two such systems. However, contemporary grammatical theory and practice is highly diverse, and any system, however effective, whose form of analysis is in general at variance with that of the PTB can only lose out when evaluated in this way.

This fact is recognised in (Black *et al.*, 1991), where the Parseval scheme is proposed. There, use of the evaluation metrics is preceded by a 'pre-processing' stage, which applies to both response and key, that serves to eliminate aspects of structure that are considered likely to be contentious. This process begins with the deletion of all instances of certain 'problematic' items, including auxiliaries, "not", pre-infinitival "to", possessive endings and null categories. Subsequent deletion of all unary and nullary structure serves to eliminate much of the structure that was associated with those problematic items. For certain constructions, where contention is not easily overcome by deletion of some lexical item (e.g. the attachment position of extraposed clauses), transformation rules are provided that convert structures to a canonical format. One criticism of this approach to parser evaluation is that it relies on a tree massaging process that is clearly language specific in its statement.

Even after pre-processing, there can still be substantial differences between parser output and the PTB standard, both in terms of more detailed structure assigned by the parser (damaging precision), and of structure in PTB analyses that not every parser will assign (damaging recall). In the case where such differences are systematic, they can be addressed by restructuring the parser's output to be more in line with the PTB, as in (Grishman *et al.*, 1992), who use a tree transducer that implements a set of tree rewriting rules. Clearly, the need for such post-processing of parser output to some extent undermines the utility of an evaluation approach for general comparative evaluation of diverse parsing systems.

There are two obvious routes that might be followed in

[2] Again see Carroll *et al.* (1996) for a survey.

seeking a better approach, namely, use of different metrics and use of different reference corpora. In the following sections, we make suggestions in both of these directions that lead toward an approach which does, we believe, provide some reasonable basis for comparing quite dissimilar parsing systems.

4. Flatter Keys: Articulating the Common Ground

4.1. General characteristics the scheme

The alternative to using an evaluation resource whose representations embody a particular grammatical viewpoint and processing them to eliminate and/or canonicalise contentious structure, as in the Parseval/PTB approach, is to begin with a gold standard that only encodes those grammatical constituents upon which there is broad agreement across a range of grammatical theories. We suggest that the 'flatter keys' of such a gold standard would have the following characteristics: (i) no intermediate projections, (ii) no null elements or associated structure, (iii) no one-word constituents and (iv) no unary structure in general.[3] Clearly, there is much in common between the gold standard here advocated and the output of Parseval's pre-processing stage. One difference is that, unlike the output of Parseval pre-processing (which deletes certain words), all words of the original sentence are present in the flatter keys of our approach — a characteristic that has obvious benefits when the need arises for direct inspection of representations. More significantly, we expect the keys of our approach to be somewhat flatter than those output by Parseval pre-processing, by a broader omission of contentious structure. Some cases of contentious structure are addressed in Parseval pre-processing by being transformed to a canonical form. For example, the extraposition sentence structure (i) below would be rewritten as (ii). From our perspective, the existence of the two incompatible analyses by definition demonstrates lack of consensus, and so we would expect the disputed structure to be omitted from our keys, e.g. as in (iii).[4]

i. (it (is (necessary (for us to leave))))

ii. (it (is necessary) (for us to leave))

iii. (it is necessary (for us to leave))

[3] Unary structure is essentially trivial where there is *unlabelled* comparison of key and response, as is appropriate for comparison of diverse parsing systems. Where unlabelled representations are used, brackets are best seen as indicating the *grouping* of elements into constituents, rather than as corresponding one-to-one with constituent nodes.

[4] An alternative is that we might use keys that encode *both* possibilities for such disputed structures, and to parameterise the bracket comparison process to either ignore disputed brackets in keys, or to treat the key as including one or other alternative. This idea, however, is somewhat out of the spirit of the general proposal, and we shall not pursue it here.

4.2. Components of an annotation scheme

Recall that the gold standard we advocate is one that encodes for each sentence only the constituents upon which there is broad agreement across a range of grammatical theories. A set of grammatical categories for which there is general agreement has been identified by the EAGLES Syntactic Annotation Group (Leech *et al.*, 1996), and includes sentence, clause, noun phrase, verb phrase, prepositional phrase, adverb phrase and adjective phrase. However, limiting analyses to contain only constituents of these consensual categories does not in itself serve to exclude all contentious constituents. Some observations relevant to this point will be made in the following remarks about each of the categories.[5]

sentence/clause: The EAGLES Syntactic Annotation Group recognises a distinction between *sentence*, the maximal independent segments into which a text is subdivided (typically beginning with a capital letter and ending with a terminal punctuation for written text), and *clause*, for units such as relative and adverbial clauses, which are embedded within some superordinate sentence. Both of these units will be recognised in the scheme.

noun phrase: Even a relatively simple NP such as "the man from Madrid" might be variously be analysed as:

```
(NP (NP DET N) (PP P NP))
```

or with intermediate projections as in:

```
(NP DET (N′ N (PP P NP)))
```

or in the comparatively flat form:

```
(NP DET N (PP P NP))
```

Note that for the first of these options, the embedded NP grouping is not a matter of consensus, even though it bears a category from the 'consensual' list. For such a case, we would include only the outermost NP node in the gold standard (giving a result identical to the third option).

verb phrase: The correct analysis of verb phrases is likewise contentious. Cases with multiple auxiliaries, for example, are variously analysed as having a structure that is flat, involves VP recursion, or clusters verbal elements into a 'verb group'. The VP recursion option again illustrates that use of only the 'consensual' categories does not exclude all contentious nodes. Again, we propose that only the outermost VP node should be included in the gold standard so that we would have e.g.:

```
(S He (VP will go home))
```

rather than either of:

```
(S He (VP will (VP go home)))
(S He (VP (VC will go) home))
```

prepositional phrase: This unit, consisting of a preposition and its complement, is recognised in the scheme and would be marked wherever found.

adverbial phrase: This unit, i.e. multiword phrases whose head is an adverb, would be recognised in the scheme.

[5] A further issue is whether the categories should be recorded in key files, given that the comparison method is not sensitive to labelling. We think that they should be, as this will be helpful for human inspection of files, and may also help configurable evaluation — see the remarks made in the next section.

adjectival phrase: The marking of adjectival phrases would be restricted to certain contexts, e.g. they would be marked when appearing in post-copula position, but would not be marked in prenominal position.

It should be emphasised that we are not advocating these flat structures as the 'correct' linguistic analyses, but instead merely suggesting that they include the appropriate aspects of grammatical structure to be included in a gold standard for use in comparative evaluation of diverse parsers. Even so, we recognise that even these 'consensual' units will not be present in the output of all parsing systems — most obviously of dependency-based systems[6] — and there may in such cases be need of some post processing of parser output (in the manner of Grishman *et al.*, 1992), although hopefully less than would be needed with evaluation approaches whose gold standard was more richly structured.

4.3. Benefits for resource creation

In the next section, we shall argue that the gold standard representations we have advocated, when used with suitable metrics, provide an appropriate basis for comparative evaluation of diverse parsing systems. Here we shall note a further advantage of our proposal, that is of relevance to the general pursuit of the evaluation enterprise for parsing systems.

One of the principal obstacles to the broader use of evaluation in all areas of NLP — not just parsing — is the cost of creating evaluation resources, since this creation can in general only be partially automated, if at all. The English language is currently in a somewhat privileged position in relation to parser evaluation, given the existence of parse-labelled corpora such as the PTB and (parts of) the SUSANNE Corpus (Sampson, 1995). Comparable resources are largely non-existent for other languages, even those for which there has been some significant investment in computational linguistic research (as there has been, for example, for some European Community languages).

A clear advantage of our proposal is that, given the relatively simple character of its keys, it should be substantially easier to create gold standards of this kind, as compared to creating representations comparable to those of, say, the PTB.[7] We hope that this consequence of our approach, of allowing evaluation resources to be generated both more rapidly and for less cost, improves the likelihood of a broader use of evaluation to inform research in the area of grammar and parsing for more languages than just English.

For languages for which a substantial investment of effort into the creation of parse evaluation resources has already been made (i.e. English), we would hope to avoid the need for any substantial further investment in producing the gold standard resources we advocate. This topic is addressed in Gaizauskas *et al.* (1998), where we suggest that for English resources of the form we advocate should be derivable in a largely automatic fashion from resources that already exist.

5. Evaluation Using Flatter Keys

Given gold standards of the kind that have been outlined, how should they be used in evaluation? Since we wish to be able to evaluate parsers having very different output representations, reflecting differing grammatical theories, we are faced with two options. We may either provide a potentially complex mapping of the parser's output structures into the gold standard format, or we can determine metrics which can be used for direct comparison of the output representation to the gold standard. The first option would require a different mapping to be specified for each parser, which is potentially a substantial obstacle to the broad use of the scheme. We have therefore chosen to pursue the second option.

In line with (Black *et al.*, 1991), comparison of parser output and gold standard is based on *unlabelled* brackettings. Thus, the parser's role as evaluated in this way is one of correctly identifying constituent *groupings* within a sentence; the category a parser may assign to such groupings is seen as a grammar-specific characteristic. The metrics we advocate for use with the flatter gold standards are recall and a modified crossing brackets measure; we reject use of the precision metric. The metric of unlabelled recall can be used under its standard definition, providing a measure of the extent to which the parser has found the consensual constituent units.[8]

A precision metric, however, which measures the percentage of proposed structure which is in the gold standard, does not make sense for use with the gold standard we have outlined. Parsers will of course assign more structure than the minimal, consensual structure recorded in the gold standard, and they should not be punished for doing so. To put this point differently, a precision metric makes sense where the gold standard specifies precisely the analysis that the parser should produce, but not where the gold standard specifies only minimal structural requirements that the parser's output should satisfy.[9]

Whilst a parser should not be punished for assigning more elaborate structure than is recorded in the key, it clearly should be punished for assigning structure that is incompatible with the key — a situation identified by crossing brackets measures. The standard crossing brackets measure, which counts the number of response constituents that

[6] See Lin (1995) for a dependency-oriented approach to parser evaluation.

[7] Of course, we recognise that the PTB is not designed to be used purely for parser evaluation. Its rich representation of linguistic structure provides a basis for investigating a range of phenomena, whereas the gold standard representations we advocate are geared specifically toward parser evaluation.

[8] Provided our gold standard files recorded one of the consensual categories for each bracket pair (c.f. footnote 5), and given a mapping from parser categories into the consensual categories, scores for *labelled* recall could be generated, but such a move seems against the general spirit of our approach in embracing the diversity of practice that is to be found in contemporary research.

[9] The limitations of the precision metric have already been noted by Grishman *et al.* (1992), who observe that most automatic parsing systems generate more structure than is present in the PTB, giving rise to apparently poor precision scores.

cross a key constituent (i.e. where the two bracketted sequences overlap but neither is properly contained in the other) has the undesirable characteristic that a parser which assigns elaborate structure will tend to be multiply penalised for violating a given gold standard constituent whereas a parser that assigns fairly flat structure will not. Instead, we suggest, the metric should address only whether or not each constituent of the key is 'violated' by the response. For consistency with the use of recall (and precision) where a high score is 'good', we suggest the *conformance* metric, which is simply the proportion of the gold standard constituents that are not 'crossed' by any constituent in the response. Together, recall and conformance fulfill the complementary roles (filled by recall and precision in the standard scheme) of rewarding the correct discovery of gold standard structure, whilst penalising structure that the gold standard forbids.[10]

In addition to supporting the evaluation of diverse parsing systems, this scheme has the further advantage that it readily supports 'configurable evaluation'. Since parser outputs are required only to capture the core structural information that keys record (and not to precisely mimic the key), we can therefore tailor our keys to include only those aspects of structure that we want to evaluate. For example, keys might contain only brackettings for specific categories or to specific levels of embedding. Thus, configurability is achieved without alteration of the underlying evaluation algorithm, only manipulation of keys is required (which might often involve simple 'filtering' of existing, more complete, keys).

We conclude this section by considering a simple example that illustrates the evaluation metrics we propose. Consider the sentence *The monthly sales have been setting records every month since March.* This sentence appears in the Penn Tree Bank in the file `wsj_0016.mrg` and receives the following analysis (functional tag extensions have been removed):

```
PTB:
( (S
    (NP (DT The) (JJ monthly)
              (NNS sales))
    (VP (VBP have)
      (VP (VBN been)
        (VP (VBG setting)
          (NP (NNS records) )
          (NP
            (NP (DT every) (NN month))
            (PP (IN since)
              (NP (NNP March) ))))))
    (. .) ))
```

Under the flat annotation that we have advocated above, this sentence would receive the minimal bracketting *((The monthly sales) (have been setting records (every month (since March))).)* Including phrasal and lexical tags for readability, this would be written:

```
Flat:
((S
   (NP (DT The) (JJ monthly)
         (NNS sales))
   (VP (VBP have) (VBN been)
         (VBG setting)
     (NNS records)
     (NP (DT every) (NN month)
       (PP (IN since)
         (NNP March))))
   (. .)))
```

Suppose now that we wish to evaluate a parser that is based on a rather different grammatical theory, one that differs from PTB-type analyses in

1. clustering verbal auxiliaries together with the main verb into a verb cluster (VC), and
2. employing NBAR and VBAR intermediate projection levels

Call this approach ALT. Note that we are neither advocating this style of analysis, nor are we suggesting that anyone else has. ALT simply serves the role of representing a plausible alternative approach to grammatical analysis which an evaluation scheme ought to be capable of sensibly addressing. ALT might propose the following as the correct analysis for the example.

```
ALT-good:
((S
   (NP (DT The)
       (NBAR (JJ monthly) (NNS sales)))
   (VP (VBAR
         (VBAR
           (VC (VBP have)
               (VC (VBN been)
                   (VBG setting)))
           (NP (NNS records)))
         (NP (DT every) (NN month)
           (PP (IN since)
               (NP (NNP March))))))
   (. .)))
```

Further, for the sake of illustration, we may imagine that a given ALT parser which we wish to evaluate in fact

[10] However, recall and conformance are not independent metrics in the same way that recall and precision are. While recall and precision may vary upwards or downwards independently, a decrease in conformance is associated with a reduced upper limit in the attainable recall, i.e. since the presence of a 'crossing bracket' in a parse rules out the presence also of the 'crossed' key constituent in that parse. This interaction holds for all crossing bracket measures. A separate observation is that the 'inhibitory' effect of conformance is absent in the case of one-word constituents, since they can never cross another constituent. Hence, a parser's output could mark each word as a one-word constituent, perhaps with a view to gaining some points of recall, and never risk being penalised for incontinent structure assignment by reduction in conformance. This observation provides a simple practical motivation for characteristic (iii) of the general annotation scheme, as stated in section 4.1, i.e. the exclusion of any one-word constituents from gold standard keys.

Key	Response	KeyCons	RespCons	Matched	KCV	Recall	Precision	Conformance
Flat	PTB	5	10	5	0	100	50	100
Flat	ALT-good	5	11	5	0	100	45	100
Flat	ALT-bad	5	11	3	1	60	27	80
PTB	Flat	10	5	5	0	50	100	100
PTB	ALT-good	10	11	7	2	70	64	80
PTB	ALT-bad	10	11	6	3	60	55	70
ALT-good	ALT-bad	11	11	9	1	82	82	91
ALT-good	PTB	11	10	7	3	64	70	73

Table 1: Comparison of Parse Evaluation Metrics for a Simple Example

generates the following 'incorrect' analysis in ALT terms – it misses the first NP constituent *The monthly sales* and wrongly attaches the PP *since March*.

```
ALT-bad:
((S
   (DT The)
   (NBAR (JJ monthly) (NNS sales))
   (VP (VBAR
         (VBAR
           (VBAR
             (VC (VBP have)
                 (VC (VBN been)
                     (VBG setting)))
             (NP (NNS records)))
           (NP (DT every) (NN month)))
         (PP (IN since)
             (NP (NNP March)))))
   (. .)))
```

Figures for the recall, precision and conformance metrics for this example, alternately considering the Flat, the PTB and the ALT-Good analyses to be the key and the PTB/Flat, ALT-Good and/or ALT-Bad analyses to be the response are presented in Table 1[11]. The table shows, for each key and response:

1. the number of constituents in the key (KeyCons);
2. the number of constituents in the response (RespCons);
3. the number of constituents matched (Matched);
4. the number of key constituents violated by the response (KCV);
5. the recall $(Matched/KeyCons) * 100$;
6. the precision $(Matched/RespCons) * 100$; and
7. the conformance $((KeyCons - KCV)/KeyCons) * 100$.

Examining the table, a number of observations can be made. First, proponents of PTB-style analyses and ALT-style analyses are going to score rather poorly if they use each others' annotated corpora as gold standards; further, the scores generated provide little help in disentangling what is the result of irreducibly different theoretical differences from what may be wrong in any particular analysis suggested by a parser implementing their theory (note the relatively small difference in results between the ALT-good vs PTB and the ALT-bad vs PTB figures). Second, note that both the PTB and ALT-good analyses score full marks as responses against the Flat key: this assures their proponents that their 'correct' analysis meets the minimal, consensual gold-standard. However, the ALT-bad analysis is marked down appropriately: recall and conformance both suffer and serve to alert the ALT grammar developer that something is wrong, exactly as the ALT-good key does. Finally, observe that precision figures are really only of use when both key and response are generated according to the same grammatical theory: comparing across theories, or even comparing to a minimal, consensually agreed structure leads to precision figures which are little help, since they penalise structure which may be viewed as entirely appropriate within a given grammatical framework.

6. Conclusion

We have proposed a scheme for comparative evaluation of diverse parsing systems whose principal distinguishing features are: (i) the use of parsed reference corpora whose constituency annotations are just those upon which there is broad agreement across a range of grammatical theories, and hence are much flatter than those currently used, and (ii) the use of recall and conformance as the chief evaluation metrics, with the aim of requiring parser outputs to conform to, rather than precisely mimic, the broadly agreed structure encoded in the flatter gold standard analyses.

The comparatively simple character of the gold standard representations we have proposed should make possible much cheaper generation of parse evaluation resources for languages for which none currently exist, as compared to the use of other annotation schemes. Even so, we would hope to avoid much of even this reduced effort in the case where there already exists parse-labelled resources for a language, annotated according to some other scheme. In Gaizauskas *et al.* (1998) we describe an approach to deriving gold standards of the kind we advocate from existing parse annotated resources in a comparatively inexpensive fashion.

[11] The figures in this table were generated automatically by a program which given a key file and a response file of bracketted sentences computes precisely the column headings in the table.

Software which calculates the metrics advocated here and which derives flattened keys of the sort proposed here from the Penn Treebank has been implemented and is available from the authors.

References

Black, E., Abney, S., Flickinger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B. & Strzalkowski, T. (1991). A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In Proceedings of the DARPA Speech and Natural Language Workshop (pp. 306–311).

Carroll, J., Briscoe, T., Calzolari, N., Federici, S., Montemagni, S., Pirrelli, V., Grefenstette, G., Sanfilippo, A., Carroll, G., and Rooth, M. (1996). Sparkle Work Package 1: Specification of Phrasal Parsing. Available at URL: http://www.ilc.pi.cnr.it/sparkle/wp1-prefinal/wp1-prefinal.html

Crouch, R., Gaizauskas, R. and Netter, K. (1995). Report of the Study Group on Assessment and Evaluation. Available at URL: http://xxx.lanl.gov/ps/9601003.

DARPA (1995). Defense Advanced Research Projects Agency. Proceedings of the Sixth Message Understanding Conference (MUC-6), Morgan Kaufmann.

Flickinger, D., Nerbonne, J., Sag, I.A. and Wassow, T. (1987). Toward Evaluation of NLP systems. Technical report. Hewlett-Packard Laboratories.

Gaizauskas, R., Hepple, M. and Huyck, C. (1998). Modifying Existing Annotated Corpora for General Comparative Evaluation of Parsing. In Proceedings of the Workshop on Evaluation of Parsing Systems held in conjunction with The First International Conference on Language Resources and Evaluation.

Grishman, R., Macleod, C. and Sterling, J. (1992). Evaluating Parsing Strategies Using Standardized Parse Files. In Proceedings of the 3rd ACL Conference on Applied Natural Language Processing.

LangEng (1998). European Commission Language Engineering RTD Projects. URL: http://www2.echo.lu/langeng/en/rtd.html

Leech, G., Barnett, R., and Kahrel, P. (1996). Provisional Recommendations and Guidelines for the Syntactic Annotation of Corpora. EAGLES Document EAG-TCWG-SASG/1.8, March 1996. Available by internet: contact eagles@ilc.pi.cnr.it for information.

Lehmann, S., Oepen, S., Regnier-Prost, S., Netter, K., Lux, V., Klein, J., Falkedal, K., Fouvry, F., Estival, D., Dauphin, E., Compagnion, H., Baur, J., Balkan, L. and Arnold, D. (1996). TSNLP — Test Suites for Natural Language Processing. In Proceedings of COLING'96, Copenhagen.

Lin, D. (1995). A Dependency-based Method for Evaluating Broad Coverage Parsers. In Proceedings of IJCAI–95 (pp. 1420-1425), Montreal, Canada.

Marcus, M.P., Santorini, B. and Marcinkiewicz, M.A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, 19(2), pp 313-330.

Sampson, G. (1995). English for the Computer: the SUSANNE corpus and analytic scheme. Clarendon.