A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars

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, T. Strzalkowski

IBM Research Division, Thomas J. Watson Research Center
Yorktown Heights, NY 10598

The problem of quantitatively comparing the performance of different broad-coverage grammars of English has to date resisted solution. Prima facie, known English grammars appear to disagree strongly with each other as to the elements of even the simplest sentences. For instance, the grammars of Steve Abney (Bellcore), Ezra Black (IBM), Dan Flickinger (Hewlett Packard), Claudia Gdaniec (Logos), Ralph Grishman and Tomek Strzalkowski (NYU), Phil Harrison (Boeing), Don Hindle (AT&T), Bob Ingria (BBN), and Mitch Marcus (U. of Pennsylvania) recognize in common only the following constituents, when each grammarian provides the single parse which he/she would ideally want his/her grammar to specify for three sample Brown Corpus sentences:
The famed Yankee Clipper, now retired, has been assisting (as (a batting coach)).
One of those capital-gains ventures, in fact, has saddled him (with Gore Court).
He said this constituted a (very serious) misuse (of the (Criminal court) processes).

Specific differences among grammars which contribute to this apparent disparateness of analysis include the treatment of punctuation as independent tokens or, on the other hand, as parasites on the words to which they attach in writing; the recursive attachment of auxiliary elements to the right of Verb Phrase nodes, versus their incorporation there en bloc; the grouping of pre-infinitival "to" either with the main verb alone or with the entire Verb Phrase that it introduces; and the employment or non-employment of "null nodes" as a device in the grammar; as well as other differences. Despite the seeming intractability of this problem, it appears to us that a solution to it is now at hand. We propose an evaluation procedure with these characteristics: it judges a parse based only on the constituent boundaries it stipulates (and not the names it assigns to these constituents); it compares the parse to a "hand-parse" of the same sentence from the University of Pennsylvania Treebank; and it yields two principal measures for each parse submitted.

The procedure has three steps. For each parse to be evaluated: (1) erase from the fully-parsed sentence all instances of: auxiliaries, "not", pre-infinitival "to", null categories, possessive endings ('s and '), and all word-external punctuation (e.g. " . , ; –); (2) recursively erase all parenthesis pairs enclosing either a single constituent or word, or nothing at all; (3) compute goodness scores (Crossing Parentheses, and Recall) for the input parse, by comparing it to a similarly-reduced version of the Penn Treebank parse of the same sentence.

For example, for the Brown Corpus sentence:
Miss Xydis was best when she did not need to be too probing. consider the candidate parse:
(S(NP-s(PNP(PNP Miss) (PNP Xydis))) (VP(VPAST was) (ADJP(ADJ best))) (S(COMP(WHADVP(WHADV when))) (NP-s (PRO she)) (VP ((VPAST did) (NEG not) (V need)) (VP((X to) (V be)) (ADJP(ADV too) (ADJ probing)))))) (?(FIN .))
After step-one erasures, this becomes:
(S(NP-s(PNP(PNP Miss) (PNP Xydis))) (VP(VPAST was) (ADJP(ADJ best))) (S(COMP(WHADVP(WHADV when))) (NP-s (PRO she)) (VP((VPAST) (NEG)

(V need)) (VP((X) (V be)) (ADJP(ADV too) (ADJ probing)))))) (?(FIN))
And after step-two erasures:
(S(NP-s Miss Xydis) (VP was best) (S when she (VP need (V be (ADJP too probing)))))
The University of Pennsylvania Treebank output for this sentence, after steps one and two have been applied to it, is:
(S(NP Miss Xydis) (VP was best (SBAR when (S she (VP need (VP be (ADJP too probing)))))))

Step three consists of comparing the candidate parse to the treebank parse and deriving two scores: (1) The Crossing Parentheses score is the number of times the treebank has a parenthesization such as, say, (A (B C)) and the parse being evaluated has a parenthesization for the same input of ((A B) C)), i.e. there are parentheses which "cross". (2) The Recall score is the number of parenthesis pairs in the intersection of the candidate and treebank parses (T intersection C) divided by the number of parenthesis pairs in the treebank parse T, viz. (T intersection C) / T. This score provides an additional measure of the degree of fit between the standard and the candidate parses; in theory a Recall of 1 certifies a candidate parse as including all constituent boundaries that are essential to the analysis of the input sentence. We applied this metric to 14 sentences selected from the Brown Corpus and analyzed by each of the grammarians named above in the manner that each wished his/her grammar to do. Instead of using the UPenn Treebank as a standard, we used the automatically computed "majority parse" of each sentence obtained from the set of candidate parses themselves. The average Crossing Parentheses rate over all our grammars was .4%, with a corresponding Recall score of 94%. We have agreed on three additional categories of systematic alteration to our input parses which we believe will significantly improve the correlation between our "ideal parses", i.e. our individual goals, and our standard. Even at the current level of fit, we feel comfortable allowing one of our number, the UPenn parse, to serve as the standard parse, since, crucially, it is produced by hand. Our intention is to apply the current metric to more Brown Corpus data "ideally parsed" by us, and then to employ it to measure the performance of our grammars, run automatically, on a benchmark set of sentences.

APPENDIX: EVALUATION PROCEDURE FOR COMPUTER ENGLISH GRAMMARS

0. *Input format*
   A parse for evaluation should consist initially of:
   - *(a) the input word string, tokenized as follows:*
     - (1) *Any tokens containing punctuation marks are enclosed by vertical bars, e.g.* |D'Albert| |4,000|
     - (2) *Contracted forms in which the abbreviated verb is used in the sentence under analysis as a main verb, as opposed to an auxiliary, are to be split:* you've -> you |'ve| (In "You've a good reason for that." but not in "You've been here often.") John's -> John |'s| (In "John's (i.e. is) a good friend" or "John's (i.e. has) a good friend" but not "John's (i.e. is) leaving" and not "John's (i.e. has) been here"
     - (3) *Hyphenated words, numbers and miscellaneous digital expressions are left as is (i.e. not split), i.e.* |co-signers| (and not "co |-| signers"); |2,000| (and not "2 |,| 0 0 0"); |all-woman|; |fifty-three|; |free-for-all|; 56th; |3/4|; |212-488-9027|;
   - *(b) the parse of the input word string with respect to the grammar under evaluation*
     - (1) *Each grammatical constituent of the input is grouped using a pair of parentheses, e.g.*

"(((I)) ((see) ((Ed))))"

(2) Constituent labels may, optionally, immediately follow left parentheses and/or immediately precede right parentheses, e.g.
(S (N' (N Sue))
(V' (V sees)
(N' (N Tom)))) =
= (((Sue))
((sees)
((Tom)))) etc.

1. Erasures of Input Elements
The first of the two steps necessary to prepare initial parsed input for evaluation consists of erasing the following types of word (token) strings from the parse:

(a) Auxiliaries
Examples are:
"would go there"
-> "go there",
"has been laughing"
-> "laughing",
"does sing it correctly"
-> "sing it correctly",
but not: "is a cup",
"is blue", "has a dollar",
"does the laundry"

(b) "Not"
E.g.
"is not in here"
-> "is in here",
"Not precisely asleep,
John sort of dozed"
->"precisely asleep,
John sort of dozed"

(c) Pre-infinitival "to"
E.g.
"she opted to retire"
-> "she opted retire",
"how to construe it"
-> "how construe it"

(d) Null categories
Example 1:
("getting more pro letters than con"):
(NXc (Qr more)
(NX (A pro)
(Npl letters))
(Than than)
(NX (A con) (Npl)))
NOTA BENE
->
(NXc (Qr more)
(NX (A pro)
(Npl letters))
(Than than)
(NX (A con) ( )));
NOTA BENE

Example 2:
("The lawyer with whom I studied law"):
(NP (DET The)
(N lawyer)
(S-REL (PP (P with)
(NP whom))
(NP I)
(VP (V studied)
(NP (N law))
(PP 0))))
NOTA BENE
->
(NP (DET The)
(N lawyer)
(S-REL (PP (P with)
(NP whom))
(NP I)
(VP (V studied)
(NP (N law))
(PP))))
NOTA BENE

(e) Possessive endings ('s, ')
E.g. "|Lori's| mother"
(i.e. the mother of Lori)
-> "Lori mother"

(f) Word-external punctuation
(quotes, commas, periods, dashes, etc.)
E.g.
The "blue book" was there
-> The blue book was there
Your first , second and third ideas -> Your first second and third ideas
This is it. -> This is it
All--or almost all--of them
-> All or almost all of them
But leave as is: |3,456| |8.29| |3/17/90| |11:30| |p.m.| |1)| |Ph.D.| |U.N.| |ne'er-do-well|

2. *Erasures of Constituent Delimiters, i.e. Parentheses*
The second of the two steps necessary to prepare initial parsed input for evaluation consists of erasing parenthesis pairs, proceeding recursively, from the most to the least deeply embedded portion of the parenthesization, whenever they enclose either a single constituent or word, or nothing at all.

Example:
"Miss Xydis was best when she did not need to be too probing."

1. *Original parse*

```
(S (NP-s (PNP (PNP  Miss )
               (PNP  Xydis )))
   (VP (VPAST  was )
       (ADJP (ADJ  best )))
   (S (COMP (WHADVP
       (WHADV  when )))
       (NP-s (PRO  she ))
       (VP ((VPAST  did )
            (NEG  not )
            (V  need ))
            (VP ((X  to )
                 (V  be ))
                 (ADJP
                 (ADV  too )
                 (ADJ
                 probing )
                 )))))
          (? (FIN  . ))
```

2. *Parse with all erasures performed except those of constituent delimiters (parentheses):*

```
(S (NP-s (PNP (PNP  Miss )
               (PNP  Xydis )))
   (VP (VPAST  was )
       (ADJP (ADJ  best )))
   (S (COMP (WHADVP
       (WHADV  when )))
       (NP-s (PRO  she ))
       (VP ((VPAST      )
            (NEG        )
            (V  need ))
            (VP ((X     )
                 (V  be ))
                 (ADJP
                 (ADV  too )
                 (ADJ
                 probing )
                 )))))
          (? (FIN     ))
```

3. *Parse with all constituent delimiters erased which are superfluous by the above definition:*

```
(S (NP-s Miss
         Xydis )
   (VP was
       best )
   (S when
      she
      (VP
       need
       (VP
        be
        (ADJP too
              probing)))))
```

NOTE: Any single-word adverbs which are left behind, as it were, by the erasure of auxiliary elements, are attached to the highest node of the immediately following verb constituent.

Example:

```
(will probably have)
   (seen Milton) ->
(     probably     )
   (seen Milton) ->
(probably seen Milton)
```

3. *Redefinition of Selected Constituents*
The third step in the process of preparing initial parsed input for evaluation is necessary only if the parse submitted treats any of three particular constructions in a manner different from the canonical analysis currently accepted by the group. This step consists of redrawing constituent boundaries in conformity with the adopted standard. The three constructions involved are extraposition, modification of noun phrases, and sequences of prepositions which occur constituent-initially and/or

particles which occur constituent-finally.

(a) Extraposition
The treatment accepted at present attaches the extraposed clause to the topmost node of the host (sentential) clause.

Example:
If initial analysis is:
(It (is (necessary (for us to leave))))
Then change to standard as follows:
(It (is necessary) (for us to leave))

NOTE: The following is not an example of extraposition, and therefore not to be modified, although it seems to differ only minimally from a genuine extraposition sentence such as: "It seemed like a good idea to begin early":
(It (seemed (like ((a good meeting) (to begin early)))))

(b) Modification of Noun Phrases
The treatment accepted at present attaches the modified "core" noun phrase and all of its modifiers from a single (noun phrase) node:
Example:
If initial analysis is:
((((the tree (that (we saw)))
(with (orange leaves)))
(that (was (very old))))
Then change to standard as follows:
((the tree) (that (we saw))
(with (orange leaves))
(that (was (very old))))

(c) Sequences of Constituent-Initial Prepositions and/or Constituent-Final Particles
For sequences of prepositions occurring at the start of a prepositional phrase, the currently accepted practice is to attach each individually to the prepositional-phrase node. For sequences of particles which come at the end of a verb phrase or other constituent with a verbal head, the adopted standard is, likewise, to attach each individually to the top node of the constituent:
Example:
If initial analysis is:
(We (were (out (of (oatmeal cookies)))))
Then change to standard as follows:
(We (were (out of (oatmeal cookies))))

4. Computation of Evaluation Statistics
(a) Number of Constituents Incompatible With Standard Parse
For the sentence under analysis, compare the constituents as delimited by the standard parse with those delimited by the parse for evaluation. The first statistic computed for each sentence is the number of constituents in the parse being evaluated which "cross", i.e. are neither subsstrings nor superstrings of, the constituents of the standard parse.

Example:
Standard parse:
((The prospect) (of (cutting back spending)))
Parse for evaluation:
(The (prospect (of ((cutting back) spending))))
The (non-unary) constituents of the parse for evaluation are:

1. *The prospect of cutting back spending*
2. *prospect of cutting back spending*
3. *of cutting back spending*
4. *cutting back spending*
5. *cutting back*

While both constituents 2 *and* 5 *differ from the standard, only* 2 *qualifies as a "crossing" violation, as* 5 *is merely a substring of a constituent of the standard parse. So the "Constituents Incompatible With Standard" score for this sentence is* 1.

(*b*) *"Recall" and "Precision" of Parse Being Evaluated*

As a preliminary to computing Recall:

Number of Standard-Parse Constituents in Candidate

Total Number of Standard-Parse Constituents

and Precision:

Number of Candidate-Parse Constituents in Standard

Total Number of Candidate-Parse Constituents

the total number of constituents in the standard parse, and in the candidate parse, are simply counted. Notice that "Number of Standard-Parse Constituents in Candidate" and "Number of Candidate-Parse Constituents in Standard" are merely different names for the same object--the intersection of the set of standard-parse constituents with the set of candidate-parse constituents. So the final count preliminary to the computation of Recall and Precision is the number of elements in that intersection. To return to the first example of the last subsection:

Standard parse:
((The prospect) (of (cutting back spending)))
Parse for evaluation:
(The (prospect (of ((cutting back) spending))))

there are 4 *standard-parse constituents, if the convention is adopted of excluding unary constituents; and* 5 *candidate-parse constituents, under the same convention. Three of these are common to both sets, i.e. the intersection here is* 3.

Computing Recall and Precision is accomplished for this parse as follows:

Recall = 3 / 4
Precision = 3 / 5 .

(*c*) *Combining Statistics Gathered*

In order to evaluate a set of parses, first simply compute a distribution over "Incompatible Constituents" scores for the parses in the set, e.g.

Incompatible Constituents:

0	1	2
Frequency:		
3	1	1

(Total = 5)

Next, average the Recall and Precision scores for the various parses in the set, e.g.

Average Recall = (3/4 + 7/8 + 2/4 + 5/8 + 3/4) / 5 = .700

Average Precision = (3/5 + 7/10 + 2/5 + 5/10 + 3/5) / 5 = .560