|  |  | Gold Tags |  |  |  |  |  | Parser Tags |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | all |  |  | <=40 |  |  | all |  |  | <=40 |  |  |  |
|  | Corpus | Prec | Rec | F | Prec | Rec | F | Prec | Rec | F | Prec | Rec | F | Tags |
| 1 | Rl/Dev | 83.7 | 83.7 | 83.7 | 86.3 | 86.4 | 86.3 | 83.8 | 83.1 | 83.4 | 86.2 | 85.8 | 86.0 | 96.9 |
| 2 | Rd/Dev | 84.9 | 84.5 | 84.7 | 86.6 | 86.7 | 86.7 | 84.5 | 83.7 | 84.1 | 86.5 | 86.2 | 86.3 | 96.9 |
| 3 | Rd/Tst | 85.8 | 85.2 | 85.5 | 87.9 | 87.3 | 87.6 | 84.8 | 83.9 | 84.3 | 86.7 | 85.8 | 86.2 | 97.1 |
| 4 | RdNPs/Dev | 87.1 | 86.3 | 86.7 | 88.9 | 88.5 | 88.7 | 86.3 | 85.1 | 85.7 | 88.4 | 87.6 | 88.0 | 96.9 |
| 5 | RdNPsVPs/Dev | 87.2 | 87.0 | 87.1 | 89.5 | 89.4 | **89.5** | 86.3 | 85.7 | 86.0 | 88.6 | 88.2 | **88.4** | 97.0 |
| 6 | PTB/23 | 90.3 | 89.8 | 90.1 | 90.9 | 90.4 | 90.6 | 90.0 | 89.5 | 89.8 | 90.6 | 90.1 | 90.3 | 96.9 |

Table 4: Parsing results with Berkeley Parser. The corpus versions used are Release (Rl), Reduced (Rd), Reduced+NPs (RdNPs), and Reduced+NPs+VPs (RdNPsVPs). Results are shown for the parser forced to use the gold POS tags from the corpus, and with the parser supplying its own tags. For the latter case, the tagging accuracy is shown in the last column.

| Corpus | Section | Avg len | % <=40 |
|---|---|---|---|
| PPCMBE | Dev | 24.1 | 85.5 |
|  | Test | 21.2 | 89.9 |
| PTB | Dev | 23.6 | 92.9 |
|  | Test | 23.5 | 91.3 |

Table 3: Average sentence length and percentage of sentences of length <=40 in the PPCMBE and PTB.

we will report the parsing results for sentences of length <= 40 and all sentences, as with the PTB.

## 5 Parsing Experiments

The PPCMBE is a phrase-structure corpus, and so we parse with the Berkeley parser (Petrov et al., 2008) and score using the standard evalb program (Sekine and Collins, 2008). We used the Train and Val sections for training, with the parser using the Val section for fine-tuning parameters (Petrov et al., 2006). Since the Berkeley parser is capable of doing its own POS tagging, we ran it using the gold tags or supplying its own tags. Table 4 shows the results for both modes.[8]

Consider first the results for the Dev section with the parser using the gold tags. The score for all sentences increases from 83.7 for the Release corpus (row 1) to 84.7 for the Reduced corpus (row 2), reflecting the POS tag simplifications in the Reduced corpus. The score goes up by a further 2.0 to 86.7 (row 2 to 4) for the Reduced+NPs corpus and up again by 0.4 to 87.1 (row 5) for the Reduced+NPs+VPs corpus, showing the effects of the extra NP and VP brackets. We evaluated the Test section on the Reduced corpus (row 3), with a result 0.8 higher than the Dev (85.5 in row 3 compared to 84.7 in row 2). The score for sentences of length <= 40 (a larger percentage of the PPCMBE than the PTB) is 2.4 higher than the score for all sentences, with both the gold and parser tags (row 5).

The results with the parser choosing its own POS tags naturally go down, with the Test section suffering more. In general, the PPCMBE is affected by the lack of gold tags more than the PTB.

In sum, the parser results show that the PPCMBE can be parsed at a level approaching that of the PTB. We are not proposing that the current version be replaced by the Reduced+NPs+VPs version, on the grounds that the latter gets the highest score. Our goal was to determine whether the parsing results fell in the same general range as for the PTB by roughly compensating for the difference in annotation style. The results in Table 4 show that this is the case.

As a final note, the PPCMBE consists of unedited data spanning more than 200 years, while the PTB is edited newswire, and so to some extent there would almost certainly be some difference in score.

## 6 Parser Analysis

We are currently developing techniques to better understand the types of errors is making, which have already led to interesting results. The parser is creating some odd structures that violate basic well-formedness conditions of clauses. Tree (7a) in Figure 6 is a tree from from the "Reduced" corpus, in which the verb "formed" projects to IP,

[8] We modified the evalb parameter file to exclude punctuation in PPCMBE, just as for PTB. The results are based on a single run for each corpus/section. We expect some variance to occur, and in future work will average results over several runs of the training/Dev cycle, following Petrov et al. (2006).