# The Penn Parsed Corpus of Modern British English: First Parsing Results and Analysis

**Seth Kulick**
Linguistic Data Consortium
University of Pennsylvania
skulick@ldc.upenn.edu

**Anthony Kroch** and **Beatrice Santorini**
Dept. of Linguistics
University of Pennsylvania
{kroch,beatrice}@ling.upenn.edu

## Abstract

This paper presents the first results on parsing the Penn Parsed Corpus of Modern British English (PPCMBE), a million-word historical treebank with an annotation style similar to that of the Penn Treebank (PTB). We describe key features of the PPCMBE annotation style that differ from the PTB, and present some experiments with tree transformations to better compare the results to the PTB. First steps in parser analysis focus on problematic structures created by the parser.

## 1 Introduction

We present the first parsing results for the Penn Parsed Corpus of Modern British English (PPCMBE) (Kroch et al., 2010), showing that it can be parsed at a few points lower in F-score than the Penn Treebank (PTB) (Marcus et al., 1999). We discuss some of the differences in annotation style and source material that make a direct comparison problematic. Some first steps at analysis of the parsing results indicate aspects of the annotation style that are difficult for the parser, and also show that the parser is creating structures that are not present in the training material.

The PPCMBE is a million-word treebank created for researching changes in English syntax. It covers the years 1700-1914 and is the most modern in the series of treebanks created for historical research.[1] Due to the historical nature of the PPCMBE, it shares some of the characteristics of treebanks based on modern unedited text (Bies et al., 2012), such as spelling variation.

The size of the PPCMBE is roughly the same as the WSJ section of the PTB, and its annotation style is similar to that of the PTB, but with differences, particularly with regard to coordination and NP structure. However, except for Lin et al. (2012), we have found no discussion of this corpus in the literature.[2] There is also much additional material annotated in this style, increasing the importance of analyzing parser performance on this annotation style.[3]

## 2 Corpus description

The PPCMBE[4] consists of 101 files, but we leave aside 7 files that consist of legal material with very different properties than the rest of the corpus. The remaining 94 files contain 1,018,736 tokens (words).

### 2.1 Part-of-speech tags

The PPCMBE uses a part-of-speech (POS) tag set containing 248 POS tags, in contrast to the 45 tags used by the PTB. The more complex tag set is mainly due to the desire to tag orthographic variants consistently throughout the series of historical corpora. For example "gentlemen" and its orthographic variant "gen'l'men" are tagged with the complex tag `ADJ+NS` (adjective and plural noun) on the grounds that in earlier time periods, the lexical item is spelled and tagged as two orthographic words ("gentle"/ADJ and "men"/NS).

While only 81 of the 248 tags are "simple" (i.e., not associated with lexical merging or splitting),

[1] The other treebanks in the series cover Early Modern English (Kroch et al., 2004) (1.8 million words), Middle English (Kroch and Taylor, 2000) (1.2 million words), and Early English Correspondence (Taylor et al., 2006) (2.2 million words).

[2] Lin et al. (2012) report some results on POS tagging using their own mapping to different tags, but no parsing results.

[3] Aside from the corpora listed in fn. 1, there are also historical corpora of Old English (Taylor et al., 2003), Icelandic (Wallenberg et al., 2011), French (Martineau and others, 2009), and Portuguese (Galves and Faria, 2010), totaling 4.5 million words.

[4] We are working with a pre-release copy of the next revision of the official version. Some annotation errors in the currently available version have been corrected, but the differences are relatively minor.



*Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 662–667,
Baltimore, Maryland, USA, June 23-25 2014. 

| Type | # Tags | # Tokens | % coverage |
|---|---|---|---|
| Simple | 81 | 1,005,243 | 98.7% |
| Complex | 167 | 13,493 | 1.3% |
| Total | 248 | 1,018,736 | 100.0% |

Table 1: Distribution of POS tags. Complex tags indicate lexical merging or splitting.

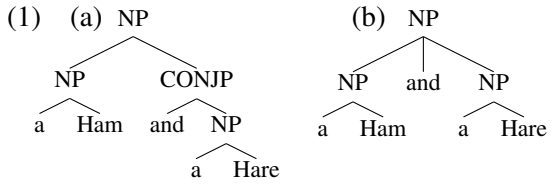


Figure 1: Coordination in the PPCMBE (1a) and the PTB (1b).

they cover the vast majority of the words in the corpus, as summarized in Table 1. Of these 81 tags, some are more specialized than in the PTB, accounting for the increased number of tags compared to the PTB. For instance, for historical consistency, words like "one" and "else" each have their own tag.

### 2.2 Syntactic annotation

As mentioned above, the syntactic annotation guidelines do not differ radically from those of the PTB. There are some important differences, however, which we highlight in the following three subsections.

#### 2.2.1 Coordination

A coordinating conjunction and conjunct form a CONJP, as shown in (1a) in Figure 1. (1b) shows the corresponding annotation in the PTB.

In a conjoined NP, if part of a first conjunct potentially scopes over two or more conjuncts (shared pre-modifiers), the first conjunct has no phrasal node in the PPCMBE, and the label of the subsequent conjuncts becomes NX instead of NP, as shown in (2a) in Figure 2. The corresponding PTB annotation is flat, as in (2b).[5]

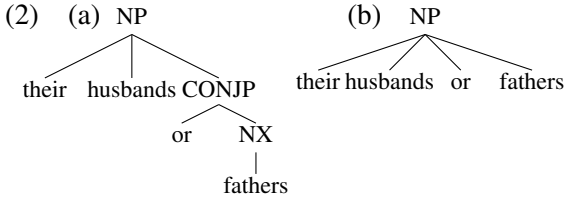


Figure 2: (2a) is an example of coordination with a shared pre-modifier in the PPCMBE, and (2b) shows the corresponding annotation in the PTB.

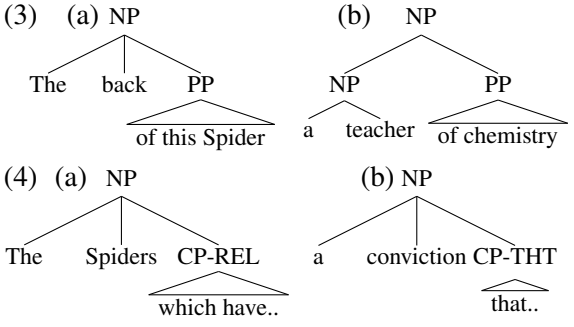


Figure 3: (3a) shows that a PP is sister to the noun in the PPCMBE, in contrast to the adjunction structure in the PTB (3b). (4ab) show that clausal complements and modifiers of a noun are distinguished by function tags, rather than structurally as in the PTB, which would adjoin the CP in (a), but not in (b).

#### 2.2.2 Noun Phrase structure

Neither the PPCMBE nor the PTB distinguish between PP complements and modifiers of nouns. However, the PPCMBE annotates both types of dependents as sisters of the noun, while the PTB adjoins both types. For instance in (3a) in Figure 3, the modifier PP is a sister to the noun in the PPCMBE, while in (3b), the complement PP is adjoined in the PTB.

Clausal complements and modifiers are also both treated as sisters to the noun in the PPCMBE. In this case, though, the complement/modifier distinction is encoded by a function tag. For example, in (4a) and (4b), the status of the CPs as modifier and complement is indicated by their function tags: `REL` for relative clause and `THT` "that" complement. In the PTB, the distinction would be encoded structurally; the relative clause would be adjoined, whereas the "that" complement would not.

#### 2.2.3 Clausal structure

The major difference in the clausal structure as compared to the PTB is the absence of a VP level[6], yielding flatter trees than in the PTB. An example clause is shown in (5) in Figure 4.

[5] Similar coordination structures exist for categories other than NP, although NP is by far the most common.

[6] This is due to the changing headedness of VP in the overall series of English historical corpora.

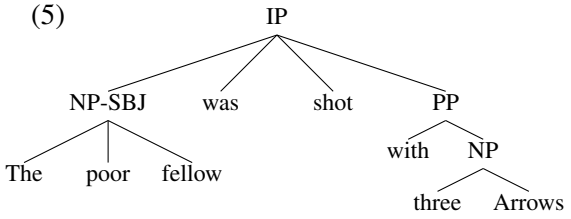


Figure 4: An example of clausal structure, without VP.

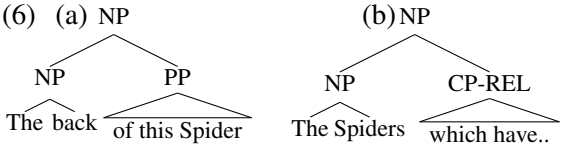


Figure 5: (6a) shows how (3a) is transformed in the "reduced +NPs" version to include a level of NP recursion, and (6b) shows the same for (4a).

## 3 Corpus transformations

We refer to the pre-release version of the corpus described in Section 2 as the "Release" version, and experiment with three other corpus versions.

### 3.1 Reduced

As mentioned earlier, the PPCMBE's relatively large POS tag set aims to maximize annotation consistency across the entire time period covered by the historical corpora, beginning with Middle English. Since we are concerned here with parsing just the PPCMBE, we simplified the tag set.

The complex tags are simplified in a fully deterministic way, based on the trees and the tags. For example, the POS tag for "gentleman", originally `ADJ+N` is changed to `N`. The `P` tag is split, so that it is either left as P, if a preposition, or changed to CONJS, if a subordinating conjunction. The reduced tag set contains 76 tags. We call the version of the corpus with the reduced tag set the "Reduced" version.

### 3.2 Reduced+NPs

As discussed in Section 2.2.2, noun modifiers are sisters to the noun, instead of being adjoined, as in the PTB. As a result, there are fewer NP brackets in the PPCMBE than there would be if the PTB-style were followed. To evaluate the effect of the difference in annotation guidelines on the parsing score, we added PTB-style NP brackets to the reduced corpus described in Section 3.1. For example, (3a) in Figure 3 is transformed into (6a) in Figure 5, and likewise (4a) is transformed into (6b). However, (4b) remains as it is, because the following CP in that case is a complement, as indicated by the `THT` function tag. This is a significant transformation of the corpus, adding 43,884 NPs to the already-existing 291,422.

| Section | # Files | Token count | % |
|---|---|---|---|
| Train | 81 | 890,150 | 87.4% |
| Val | 4 | 38,670 | 3.8% |
| Dev | 4 | 39,527 | 3.9% |
| Test | 5 | 50,389 | 4.9% |
| Total | 94 | 1,018,736 | 100.0% |

Table 2: Token count and data split for PPCMBE

### 3.3 Reduced+NPs+VPs

We carry out a similar transformation to add VP nodes to the IPs in the Reduced+NPs version, making them more like the clausal structures in the PTB. This added 169,877 VP nodes to the corpus (there are 131,671 IP nodes, some of which contain more than one auxiliary verb).

It is worth emphasizing that the brackets added in Sections 3.2 and 3.3 add no information, since they are added automatically. They are added only to roughly compensate for the difference in annotation styles between the PPCMBE and the PTB.

## 4 Data split

We split the data into four sections, as shown in Table 2. The validation section consists of the four files beginning with "a" or "v" (spanning the years 1711-1860), the development section consists of the four files beginning with "l" (1753-1866), the test section consists of the five files beginning with "f" (1749-1900), and the training section consists of the remaining 81 files (1712-1913). The data split sizes used here for the PPCMBE closely approximate that used for the PTB, as described in Petrov et al. (2006).[7] For this first work, we used a split that was roughly the same as far as timespans across the four sections. In future work, we will do a more proper cross-validation evaluation.

Table 3 shows the average sentence length and percentage of sentences of length $<= 40$ in the PPCMBE and PTB. The PPCMBE sentences are a bit longer on average, and fewer are of length $<= 40$. However, the match is close enough that

[7] Sections 2-21 for Training Section 1 for Val, 22 for Dev and 23 for Test.

| | | 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).

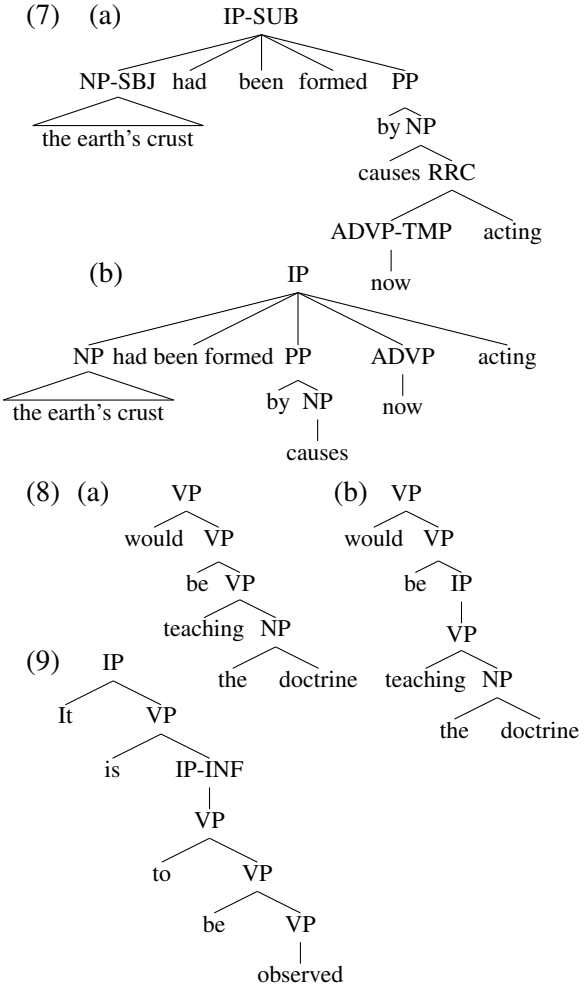


Figure 6: Examples of issues with parser output

with two auxiliary verbs ("had" and "been"). In the corresponding parser output (7b), the parser misses the reduced relative RRC, turning "acting" into the rightmost verb in the IP. The parser is creating an IP with two main verbs - an ungrammatical structure that is not attested in the gold.

It might be thought that the parser is having trouble with the flat-IP annotation style, but the parser posits incorrect structures that are not attested in the gold even in the Reduced+NPs+VPs version of the corpus. Tree (8a) shows a fragment of a gold tree from the corpus, with the VPs appropriately inserted. The parser output (8b) has an extra IP above "teaching". The POS tags for "be" (BE) and "teaching" (VAG) do not appear in this configuration at all in the training material. In general, the parser seems to be getting confused as to when such an IP should appear. We hypothesized that this is due to confusion with infinitival clauses, which can have an unary-branching IP over a VP, as in the gold tree (9). We retrained the parser, directing it to retain the INF function tag that appears in infinitival clauses as in (9). Overall, the evalb score went down slightly, but it did fix cases such as (8b). We do not yet know why the overall score went down, but what's surprising is one would have thought that IP-INF is recoverable from the absence of a tensed verb.

Preliminary analysis shows that the CONJP structures are also difficult for the parser. Since these are structures that are different than the PTB[9], we were particularly interested in them. Cases where the CONJP is missing an overt coordinating cord (such as "and"), are particularly difficult, not surprisingly. These can appear as intermediate conjuncts in a string of conjuncts, with the structure `(CONJP word)`. The shared premodifier structure described in (2a) is also difficult for the parser.

## 7 Conclusion

We have presented the first results on parsing the PPCMBE and discussed some significant annotation style differences from the PTB. Adjusting for two major differences that are a matter of annotation convention, we showed that the PPCMBE can be parsed at approximately the same level of accuracy as the PTB. The first steps in an investigation of the parser differences show that the parser is generating structures that violate basic well-formedness conditions of the annotation.

For future work, we will carry out a more serious analysis of the parser output, trying to more properly account for the differences in bracketing structure between the PPCMBE and PTB. There is also a great deal of data annotated in the style of the PPCMBE, as indicated in footnotes 1 and 3, and we are interested in how the parser performs on these, especially comparing the results on the modern English corpora to the older historical ones, which will have greater issues of orthographic and tokenization complications.

## Acknowledgments

This work was supported by National Science Foundation Grant # BCS-114749. We would like to thank Ann Bies, Justin Mott, and Mark Liberman for helpful discussions.

[9]The CONJP nonterminal in the PTB serves a different purpose than in the PPCMBE and is much more limited.

## References


Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank. LDC2012T13. Linguistic Data Consortium.

Charlotte Galves and Pabol Faria. 2010. Tycho Brahe Parsed Corpus of Historical Portuguese. http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html.

Anthony Kroch and Ann Taylor. 2000. Penn-Helsinki Parsed Corpus of Middle English, second edition. http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/index.html.

Anthony Kroch, Beatrice Santorini, and Ariel Diertani. 2004. Penn-Helsinki Parsed Corpus of Early Modern English. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html.

Anthony Kroch, Beatrice Santorini, and Ariel Diertani. 2010. Penn Parsed Corpus of Modern British English. http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174, Jeju Island, Korea, July. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3. LDC99T42, Linguistic Data Consortium, Philadelphia.

France Martineau et al. 2009. Modéliser le changement: les voies du français, a Parsed Corpus of Historical French.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of COLING-ACL*, pages 433–440, Sydney, Australia, July. Association for Computational Linguistics.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2008. The Berkeley Parser. https://code.google.com/p/berkeleyparser/.

Satoshi Sekine and Michael Collins. 2008. Evalb. http://nlp.cs.nyu.edu/evalb/.

Ann Taylor, Anthony Warner, Susan Pintzuk, and Frank Beths. 2003. The York-Toronto-Helsinki Parsed Corpus of Old English Prose. Distributed through the Oxford Text Archive. http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm.

Ann Taylor, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. 2006. Parsed Corpus of Early English Correspondence. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive. http://www-users.york.ac.uk/~lang22/PCEEC-manual/index.htm.

Joel Wallenberg, Anton Karl Ingason, Einar Freyr Sigursson, and Eirkur Rgnvaldsson. 2011. Icelandic Parsed Historical Corpus (IcePaHC) version 0.4. http://www.linguist.is/icelandic_treebank.