The effect of alternative tree representations on tree bank grammars

Mark Johnson*
Cognitive and Linguistic Sciences, Box 1978
Brown University
Providence, RI 02912, USA
Mark_Johnson@Brown.edu

Abstract

The performance of PCFGs estimated from tree banks is shown to be sensitive to the particular way in which linguistic constructions are represented as trees in the tree bank. This paper presents a theoretical analysis of the effect of different tree representations for PP attachment on PCFG models, and introduces a new methodology for empirically examining such effects using tree transformations. It shows that one transformation, which copies the label of a parent node onto the labels of its children, can improve the performance of a PCFG model in terms of labelled precision and recall on held out data from 73% (precision) and 69% (recall) to 80% and 79% respectively. It also points out that if only maximum likelihood parses are of interest then many productions can be ignored, since they are subsumed by combinations of other productions in the grammar. In the Penn II tree bank grammar, almost 9% of productions are subsumed in this way.

1 Introduction

Parsers which are capable of analysing unrestricted text are of considerable scientific interest, and have technological applications in areas such as machine translation and information retrieval as well. One way to produce such a parser is to extract a grammar from one of the larger tree bank corpora currently available.

The relative frequency estimator described below provides a simple way to estimate from a tree bank corpus a Probabalistic Context Free Grammar (PCFG) that generates Part Of Speech (POS) tags. Such a PCFG induced from a sufficiently large corpus typically generates all possible POS tag strings. A parsing system can be obtained by using a parser to find the maximum likelihood parse tree for an input string. Such parsing systems often perform as well as other broad coverage parsing systems for predicting tree structure from POS tags (Charniak, 1996). In addition, many more sophisticated parsing models are elaborations of such PCFG models, so understanding the properties of PCFGs is likely to be useful (Charniak, 1997; Collins, 1997).

It is well-known that natural language exhibits dependencies that Context Free Grammars (CFGs), and hence PCFGs, cannot describe (Shieber, 1985). But as explained below, the independence assumptions implicit in PCFGs introduce biases in the statistical model induced from a tree bank even in constructions which are adequately described by a CFG. The direction and size of these biases depend on factors such as the following:

- the precise tree structures used in the tree bank, and
- whether the set of well-formed trees according to the linguistic model used to assign trees to strings can be described with a CFG.

This paper explains how such biases can arise, and presents a series of experiments in which the trees of a tree bank corpus are systematically transformed to other tree structures to obtain a grammar used for parsing, and the inverse tree

* I would like to thank Chris Manning, whose observation that PCFG parsers do not accurately reproduce PP attachment preferences in their training data stimulated this work, as well as Eugene Charniak, Stuart Geman and our students at Brown.

Mark Johnson (1998) **The effect of alternative tree representations on tree bank grammars.** In D.M.W. Powers (ed.) *NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning,* ACL, pp 39-48.

transform is applied to the structures produced using this grammar before evaluation. One of the transformations described here improves the average labelled precision and recall on held out data from 73% (precision) and 69% (recall) to 80% and 79% respectively.

2 Probabalistic Context Free Grammars

A PCFG is a CFG in which each production $A \to \alpha$ in the grammar's set of productions R is associated with an emission probability $\mathrm{P}(A \to \alpha)$ that satisfies a normalization constraint

$$\sum_{\alpha : A \to \alpha \in R} \mathrm{P}(A \to \alpha) \;=\; 1$$

and a consistency or tightness constraint not discussed here.

A PCFG defines a probability distribution over the (finite) parse trees generated by the grammar, where the probability of a tree τ is given by

$$\mathrm{P}(\tau) \;=\; \prod_{A \to \alpha \in R} \mathrm{P}(A \to \alpha)^{\mathrm{C}_\tau(A \to \alpha)}$$

where $\mathrm{C}_\tau(A \to \alpha)$ is the 'count' of the local tree consisting of a parent node labelled A with a sequence of immediate children nodes labelled α in τ, or equivalently, the number of times the production $A \to \alpha$ is used in the derivation τ.

The PCFG which assigns maximum likelihood to a tree bank corpus $\tilde{\tau}$ is given by the relative frequency estimator.

$$\widehat{\mathrm{P}}_{\tilde{\tau}}(A \to \alpha) \;=\; \frac{\mathrm{C}_{\tilde{\tau}}(A \to \alpha)}{\sum_{\alpha' : A \to \alpha' \in R} \mathrm{C}_{\tilde{\tau}}(A \to \alpha')}$$

Here $\mathrm{C}_{\tilde{\tau}}(A \to \alpha)$ refers to the 'count' of the local tree in the tree bank, or equivalently, the number of times the production $A \to \alpha$ would be used in derivations of exactly the trees in $\tilde{\tau}$.

It is practical to induce PCFGs from tree bank corpora and find maximum likelihood parses for such PCFGs using relatively modest computing equipment. All the experiments reported here used the Penn II Wall Street Journal (WSJ) corpus, modified as described by Charniak (Charniak, 1996), i.e., empty nodes were deleted, and all other components of nodes labels except syntactic category were removed.

Grammar induction or training used the 39,832 trees in the F2–21 sections of the Penn II WSJ corpus, and testing was performed on the 1,576 sentences of length 40 or less of the F22 section of the corpus. Parsing was performed using an exhaustive CKY parser that returned a maximum likelihood parse. Ties between equally likely parses were broken randomly; on the tree bank grammar this leads to fluctuations in labelled precision and recall with a standard deviation of approximately 0.07%.

3 Different tree structure representations of adjunction

There is considerable variation in the tree structures used in the linguistic literature to represent various linguistic constructions. In this paper we focus on variations in the representation of adjunction constructions, particularly PP adjunction, but similiar variation occurs in other constructions as well.

Early analyses in transformational grammar typically adopted a 'flat' representation of adjunction structures in which adjuncts are represented as siblings of the phrasal head, as shown in Figure 1. This representation does not systematically distinguish between adjuncts and arguments, as both are attached as children of a single maximal projection.

The Penn II tree bank represents PP adjunction to VP in this manner, presumably because it permits the annotators to avoid having to determine whether the PP in question is an adjunct or an argument.

Because this representation attaches all of the adjuncts modifying the same phrase to the same node, distinct CFG productions are required for each possible number of adjuncts. Thus the set of all possible trees following this representation scheme can only be generated by a CFG if one imposes an upper bound on the number of PPs that can be adjoined to any one single phrase, but according to standard linguistic wisdom there is no natural bound on the number of PPs that may be adjoined to a single phrase.

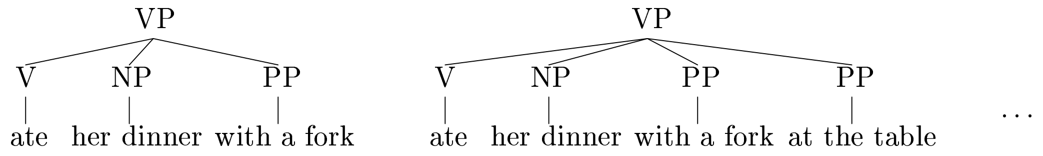


Figure 1: 'Flat' attachment representations of adjunction, where adjuncts are attached as siblings of a lexical head (in this case, the verb *ate*). The Penn II tree bank represents VP adjunction in this manner.

Later transformational analyses adopted the more complex 'Chomsky adjunction' representation of adjunction structures for theory-internal reasons (e.g., it was a corollary of Emmonds' "Structure Preserving Hypothesis"). This representation provides an additional level of recursive phrasal structure for each adjunct, as depicted in Figure 2.

Modern transformational grammar, following Chomsky's X' theory of phrase structure, represents adjunction with similiar recursive structures; the major difference being that the non-maximal phrasal nodes are given a new, distinct category label.

Because the Chomsky adjunction structure and the X' theory based on it use a single rule to recursively adjoin an arbitrary number of adjuncts, the set of all tree structures required by this representation scheme can be generated by a CFG.

The Penn II tree bank uses a mixed kind of representation for NP adjunction, involving two levels of phrasal structure irrespective of the number of adjuncts, as shown in Figure 3. This representation permits adjuncts to be systematically distinguished from arguments, although this does not seem to have been done systematically in the Penn II corpus.[1] Just as with the 'flat' representation, the set of all possible trees required by this mixed representation cannot be generated by a CFG unless the number of PPs adjoined to a single phrase is bounded.

[1]The tree annotation conventions used in the Penn II corpus are described in detail in (Bies et al., 1995). The mixed representation arises from the fact that "postmodifiers are Chomsky-adjoined to the phrase they modify" with the proviso that "consecutive unrelated adjuncts are non-recursively attached to the NP the modify". However, because constructions such as appositives, emphatic reflexives and phrasal titles are associated with their own level of NP structure, it is possible for NPs with more than two levels of structure to appear.

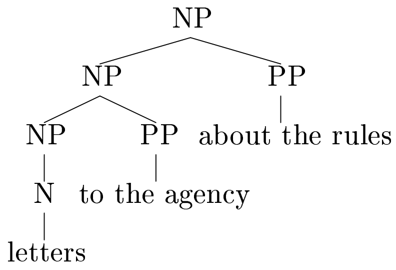


Figure 4: A tree structure generated by any PCFG that generates the trees in Figure 3, yet it does not fit the general representational scheme for adjunction structures used in the Penn II tree bank.

Perhaps more seriously for PCFG modelling of such tree structures, a PCFG which can generate a nontrivial subset of such 'two level' NP tree structures will also generate tree structures which are not instances of this representational scheme. For example, the NP production needed to produce the leftmost tree in Figure 3 can apply recursively, generating an alternative tree structure for the yield of the rightmost tree of Figure 3, as shown in Figure 4. It is not clear what interpretation to give tree structures such as these, as they do not fit the chosen representational scheme for adjunction structures.

4 PCFG models of PP adjunction

This section presents a theoretical investigation into the effect of different tree representations on the performance of PCFG models of PP adjunction. The analysis of four different models is presented here.

Clearly actual tree bank data is far more com-

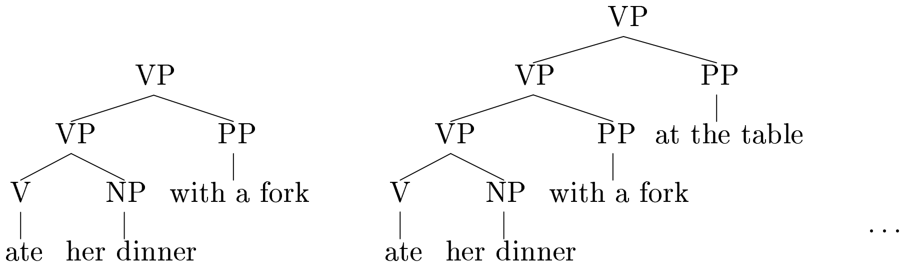


Figure 2: ‘Chomsky adjunction’ representations of adjunction, where each adjunct is attached as the unique sibling of a phrasal node (in this case, VP). Chomsky’s X' theory, used by modern transformational grammar, analyses adjunction in a structurally similiar way, except that the non-maximal (in these examples, non-root) phrasal nodes are given a new category label (in this case V$'$).

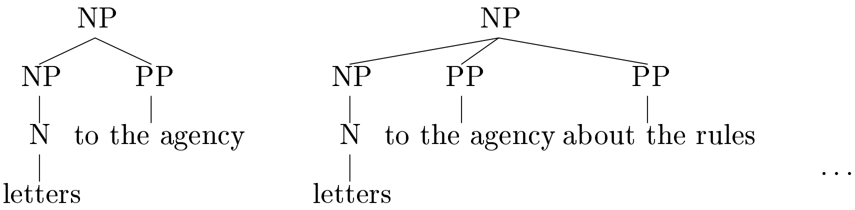


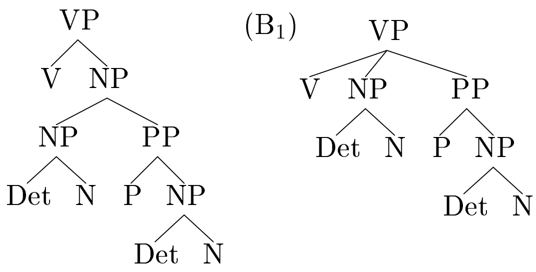
Figure 3: The representation of NP adjunction used in Penn II tree bank, where adjuncts are attached as siblings of a single NP node.

plicated than the simple models investigated in this section, and the next section investigates the effects of different tree representations empirically by applying tree transformations to the Penn II tree bank representations. However, the theoretical models discussed in this section show clearly that the choice of tree representation can in principle affect the generalizations made by a PCFG model.

4.1 The Penn II tree bank representations

Suppose we train a PCFG on a corpus $\tilde{\tau}_1$ consisting only of two different tree structures: the NP attachment structure labelled (A_1) and the VP attachment tree labelled (B_1).

(A_1) (B_1)



In the Penn II tree bank, structure (A_1) occurs 7,033 times in the F2-21 subcorpora and 279 times in the F22 subcorpus, and structure (B_1) occurs 7,717 times in the F2-21 subcorpora and 299 times in the F22 subcorpus. Thus $f \approx 0.48$ in both the F2-21 subcorpora and the F22 corpus.

Returning to the theoretical analysis, the relative frequency counts C_1 and the non-unit production probability estimates $\widehat{P}_1$ for the PCFG induced from this two-tree corpus are as follows:

R	$C_1(R)$	$\widehat{\mathrm{P}}_1(R)$
VP $\to$ V NP	f	f
VP $\to$ V NP PP	$1-f$	$1-f$
NP $\to$ Det N	2	$2/(2+f)$
NP $\to$ NP PP	f	$f/(2+f)$

Of course, in a real tree bank the counts of all these productions would also include their occurences in other constructions, so the theoretical analysis presented here is a crude idealization.

Thus the estimated likelihoods using $\widehat{\mathrm{P}}_1$ of the tree structures (A_1) and (B_1) are:

$$\begin{aligned}\widehat{\mathrm{P}}_1(\mathrm{A}_1) &= \frac{4f^2}{(2+f)^3}\\ \widehat{\mathrm{P}}_1(\mathrm{B}_1) &= \frac{4\,(1-f)}{(2+f)^2}.\end{aligned}$$

Clearly $\widehat{\mathrm{P}}_1(\mathrm{A}_1) < f$ and $\widehat{\mathrm{P}}_1(\mathrm{B}_1) < (1-f)$ except at $f = 0$ and $f = 1$, so in general the estimated frequencies using $\widehat{\mathrm{P}}_1$ differ from the frequencies of A_1 and B_1 in the training corpus. This is not too surprising, as the PCFG $\widehat{\mathrm{P}}_1$ assigns non-zero probability to trees not in the training corpus. For example, $\widehat{\mathrm{P}}_1$ assigns non-zero probability to the tree in Figure 4. We discuss the ramifications of this in section 6.

In any case, in the parsing applications mentioned earlier the absolute magnitude of the probability of a tree is not of direct interest; rather we are concerned with its probability relative to the probabilities of other, alternative tree structures. Thus it is arguably more reasonable to ignore the "spurious" tree structures generated by $\widehat{\mathrm{P}}_1$ but not present in the training corpus, and compare the estimated relative frequencies of (A_1) and (B_1) under $\widehat{\mathrm{P}}_1$ to their frequencies in the training data.

Ideally the estimated relative frequency $\hat{f}_1$ of (A_1)

$$\begin{aligned}\hat{f}_1 &= \widehat{\mathrm{P}}_1(\tau = \mathrm{A}_1 : \tau \in \{\mathrm{A}_1, \mathrm{B}_1\})\\ &= \frac{\widehat{\mathrm{P}}_1(\mathrm{A}_1)}{\widehat{\mathrm{P}}_1(\mathrm{A}_1) + \widehat{\mathrm{P}}_1(\mathrm{B}_1)}\\ &= \frac{f^2}{2-f}\end{aligned}$$

will be close to its actual frequency f in the training corpus. The relationship between f and $\hat{f}_1$ is plotted in Figure 5. The value of $\hat{f}_1$ can diverge substantially from f. For example, at $f = 0.48$ (the estimate obtained from the Penn II corpus presented above) $\hat{f}_1 = 0.15$.

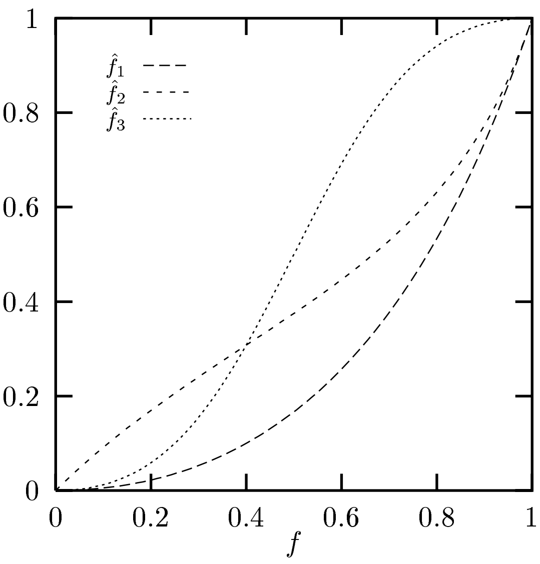
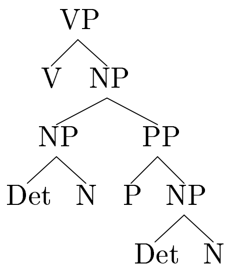


Figure 5: The estimated normalized frequency $\hat{f}$ of NP attachment using the PCFG models discussed in the text as a function of the relative frequency f of NP attachment in the training data.

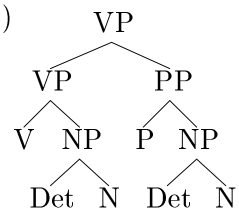
4.2 'Chomsky adjunction' representations

Now suppose that the corpus contains the following two trees (A_2) and (B_2), which are the 'Chomsky adjunction' representations of NP and VP attached PP's respectively, with relative frequencies f and $(1-f)$ as before.

(A_2)



(B_2)



The counts C_2 and the non-unit production probability estimates $\widehat{\mathrm{P}}_2$ for the PCFG induced from this two-tree corpus are as follows:

R	$C_2(R)$	$\widehat{\mathrm{P}}_2(R)$
VP $\rightarrow$ V NP	1	$1/(2-f)$
VP $\rightarrow$ VP PP	$1-f$	$(1-f)/(2-f)$
NP $\rightarrow$ Det N	2	$2/(2+f)$
NP $\rightarrow$ NP PP	f	$f/(2+f)$

The estimated likelihoods using $\widehat{\mathrm{P}}_2$ of the tree structures (A_2) and (B_2) are:

$$\begin{aligned} \widehat{\mathrm{P}}_2(\mathrm{A}_2) &= \frac{4f}{(4-f^2)(2+f)^2} \\ \widehat{\mathrm{P}}_2(\mathrm{B}_2) &= \frac{4\,(1-f)}{(4-f^2)^2} \end{aligned}$$

As in the previous subsection $\widehat{\mathrm{P}}_2(\mathrm{A}_2) < f$ and $\widehat{\mathrm{P}}_2(\mathrm{B}_2) < (1-f)$ because the PCFG assigns non-zero probability to trees not in the training corpus. Again, we calculate the estimated relative frequencies of (A_2) and (B_2) under $\widehat{\mathrm{P}}_2$.

$$\begin{aligned} \hat{f}_2 &= \widehat{\mathrm{P}}_2(\tau = \mathrm{A}_2 : \tau \in \{\mathrm{A}_2, \mathrm{B}_2\}) \\ &= \frac{f^2 - 2f}{2f^2 - f - 2} \end{aligned}$$

The relationship between f and $\hat{f}_2$ is plotted in Figure 5. The value of $\hat{f}_2$ can diverge from f, although not as widely as $\hat{f}_1$. For example, at $f = 0.48$ $\hat{f}_2 = 0.36$. Thus the precise tree structure representations used to train a PCFG can have a marked effect on its performance.

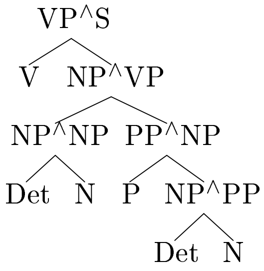
4.3 Penn II representations with parent annotation

One of the weaknesses of a PCFG is that it is insensitive to non-local relationships between nodes. If these relationships are significant then a PCFG will be a poor language model. Indeed, the sense in which the set of trees generated by a CFG is "context free" is precisely that the label on a node completely characterizes the relationships between the subtree dominated by the node and the set of nodes that properly dominate this subtree.

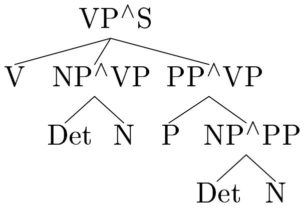
Thus one way of relaxing the independence assumptions implicit in a PCFG model is to systematically encode more information in node labels about their context. This subsection explores a particularly simple kind of contextual encoding: the label of the parent of each non-root nonpreterminal node is appended to that node's label. The labels of the root node and the terminal and preterminal nodes are left unchanged.

For example, assuming that the Penn II format trees (A_1) and (B_1) of subsection 4.1 are immediately dominated by a node labelled S, this relabelling applied to those trees produces the trees (A_3) and (B_3) below.

(A_3)



(B_3)



We can perform the same theoretical analysis on this two tree corpus that we applied to the previous corpora to investigate the effect of this relabelling on the PCFG modelling of PP attachment structures.

The counts C_3 and the non-unit production probability estimates $\widehat{\mathrm{P}}_3$ for the PCFG induced from this two-tree corpus are as follows:

R	$C_3(R)$	$\widehat{\mathrm{P}}_3(R)$
VP^S $\rightarrow$ V NP^VP	f	f
VP^S $\rightarrow$ V NP^VP PP^VP	$1-f$	$1-f$
NP^VP $\rightarrow$ Det N	$1-f$	$1-f$
NP^VP $\rightarrow$ NP^NP PP^NP	f	f

The estimated likelihoods using $\widehat{\mathrm{P}}_3$ of the tree structures (A_3) and (B_3) are:

$$\begin{aligned} \widehat{\mathrm{P}}_3(\mathrm{A}_3) &= f^2 \\ \widehat{\mathrm{P}}_3(\mathrm{B}_3) &= (1-f)^2 \end{aligned}$$

As in the previous subsection $\widehat{\mathrm{P}}_3(\mathrm{A}_3) < f$ and $\widehat{\mathrm{P}}_3(\mathrm{B}_3) < (1-f)$. Again, we calculate the estimated relative frequencies of (A_2) and (B_2) under $\widehat{\mathrm{P}}_2$.

$$\hat{f}_3 = \widehat{\mathrm{P}}_3(\tau = \mathrm{A}_3 : \tau \in \{\mathrm{A}_3, \mathrm{B}_3\})$$

$$= \frac{f^2}{f^2 + (1-f)^2}$$

The relationship between f and $\hat{f}_3$ is plotted in Figure 5. The value of $\hat{f}_3$ can diverge from f, just like the other estimates. For example, at $f = 0.48$ $\hat{f}_3 = 0.46$. Thus as expected, increased context information in the form of an enriched node labelling scheme can markedly change PCFG modelling performance.

5 Tree transformations

The last section presented simplified theoretical analyses of the effect of variation in tree representation and node labelling on PCFG modelling of PP attachment preferences. This section reports the results of an empirical investigation into the effect of changes in tree representation. These experiments were conducted by:

1. systematically transforming the trees in the training corpus F2-21 by applying a tree transform X,
2. inducing a PCFG G_X from the transformed F2-21 trees,
3. finding the maximum likelihood parses $Y(\tilde{\tau})_X$ of the yield of each sentence in the F22 corpus with respect to the PCFG G_X,
4. applying the inverse transform X^{-1} to these maximum likelihood parse trees $Y(\tilde{\tau})_X$ to yield a sequence of 'detransformed' trees $X^{-1}(Y(\tilde{\tau})_X)$ using (approximately) the same representational system as the tree bank itself, and
5. evaluating the detransformed trees $X^{-1}(Y(\tilde{\tau})_X)$ with the standard labelled precision and recall measures.

Statistics were also collected on the properties of the grammar G_X and its detransformed maximum likelihood parses $X^{-1}(Y(\tilde{\tau})_X)$; the full results are presented in Table 1.

The columns of that table correspond to different sequences of trees as follows.

F22: the trees from the F22 subcorpus of the Penn II tree bank,

F22 Id: the maximum likelihood parses of the yields of the F22 subcorpus using the PCFG estimated from the F22 subcorpus itself,

Id: the maximum likelihood parses of the yields of the F22 subcorpus using the PCFG estimated from the F2-21 subcorpus (i.e., this corresponds to applying an identity transform),

Parent: as above, except that the parent annotation transform described in subsection 4.3 was used in training and evaluation,

VP: as in **Id**, except that the flat VP structures used in the Penn II tree bank were transformed into recursive Chomsky adjunction structures as described below,

NP: as above, except that the one-level NP structures used in the Penn II tree bank were transformed into recursive Chomsky adjunction structures, and

VP-NP: as above, except that both NP and VP structures were transformed into recursive Chomsky adjunction structures.

The F22 tree sequence column provides information on the distribution of subtrees in the test tree sequence itself. The F22 Id PCFG gives data on the case where the PCFG is trained on the same data that it is evaluated on, namely the F22 subcorpus. This column is included because it is often assumed that the performance of such a model is a reasonable upper bound on what can be expected from models induced from training data distinct from the test data.

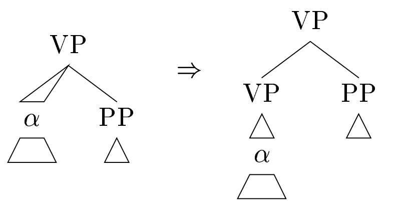
The remaining columns describe PCFGs induced from versions of the F2-21 subcorpora obtained by applying tree transformations in the manner described above.

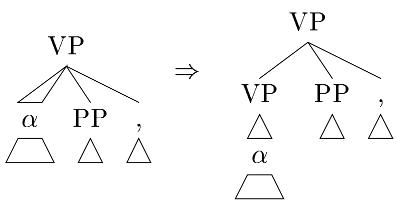
The VP transform is the result of exhaustively applying the tree transforms below. The first transform transforms VP expansions with

	F22	F22 Id	Id	Parent	VP	NP	VP-NP
No. of rules		2,269	14,962	22,773	14,393	14,866	14,297
Precision	1	0.772	0.735	0.801	0.722	0.738	0.730
Recall	1	0.728	0.696	0.793	0.677	0.698	0.705
NP attachments	279	0	67	217	0	51	329
VP attachments	299	424	384	350	240	427	0
NP* attachments	339	3	67	234	3	61	401
VP* attachments	412	668	663	461	493	650	151

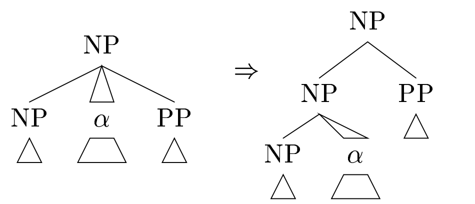
Table 1: The results of an empirical study of the effect of tree structure on PCFG models. Each column corresponds to the sequence of trees, either consisting of the F22 subcorpus or transforms of the maximum likelihood parses of the yields of the F22 subcorpus with respect to different PCFGs, as explained in the text. The first row reports the number of productions in these PCFGs, and the next two rows give the labelled precision and recall of these sequences of trees. The last four rows report the number of times particular kinds of subtrees appear in these sequences of trees, as explained in the text.

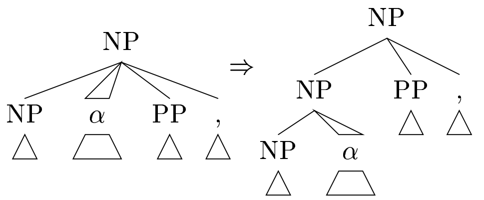
final PPs into Chomsky adjunction structures, and the second transform adjoins final PPs with a following comma punctuation into Chomsky adjunction structures. In both cases it is required that the 'lowered' sequence of subtrees α be of length 2 or greater. This ensures that the transforms will only apply a finite number of times. These two rules have the effect of converting VP final PPs into Chomsky adjunction structures.





The NP transform is similiar to the VP transform. It too is the result of exhaustively applying two tree transformation rules. These have the effect of converting NP final PPs into Chomsky adjunction structures. In this case, we require that α be of length 1 or greater.





The NP-VP transform is the result of applying all four of the above tree transforms.

The rows of Table 1 provide descriptions of these tree sequences (after 'untransformation', as described above) and, if appropriate, the PCFGs that generated them.

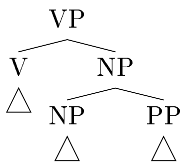
The labelled precision and recall figures are obtained by regarding a sequence of trees $\tilde{\tau}$ as a multiset or bag $E(\tilde{\tau})$ of edges, i.e., triples $\langle N, l, r \rangle$ where N is a nonterminal label and l and r are left and right string positions in yield of the entire corpus. (Root nodes and preterminal nodes are ignored in these edge sets, as they are given as input to the parser). Relative to a 'test sequence' of trees $\tilde{\tau}'$ (here the F22 subcorpus) the labelled precision and recall of a sequence of trees $\tilde{\tau}$ with the same yield as $\tilde{\tau}'$

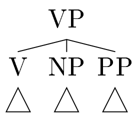
are calculated as follows:

$$\begin{aligned}\text{Precision}(\tilde{\tau}) &= \frac{|E(\tilde{\tau}) \cap E(\tilde{\tau}')|}{|E(\tilde{\tau})|} \\ \text{Recall}(\tilde{\tau}) &= \frac{|E(\tilde{\tau}) \cap E(\tilde{\tau}')|}{|E(\tilde{\tau}')|}\end{aligned}$$

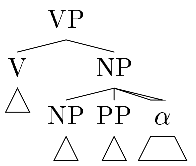
(The '$\cap$' operation above refers to multiset intersection). Precision is the fraction of edges in the tree sequence to be evaluated which also appear in the test sequence, and recall is the fraction of edges in the test sequence which also appear in sequence to be evaluated.

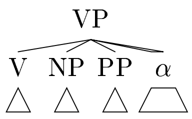
The rows labelled NP attachments and VP attachments provide the number of times the following tree schema, which represent a single PP attachment, match the tree sequence.[2] In these schema, V can be instantiated by any of the verbal preterminal tags used in the Penn II corpus.





The rows labelled NP* attachments and VP* attachments provide the number of times that the following more relaxed schema match the tree sequence. Here α can be instantiated by any sequence of trees, and V can be instantiated by the same range of preterminal tags as above.





As expected, the PCFG based on the Parent transformation, which copies the label of each parent node onto those of its children, outperforms all other PCFGs in terms of labelled precision and recall.

[2] The Penn II markup scheme permits a 'pseudo-attachment' notation for indicating ambiguous attachment. However, this is only used relatively infrequently—the pseudo-attachment markup only appears 27 times in the entire Penn II tree bank—and was ignored here. Pseudo-attachment structures count as VP attachment structures here.

The various adjunction transformations only had minimal effect on labelled precision and recall. Perhaps this is because PP attachment ambiguities, despite their important role in linguistic and parsing theory, are just one source of ambiguity among many in real language, and the effect of the alternative representations has only minor effect.

Indeed, in some cases moving to the purportedly linguistically more realistic tree Chomsky adjunction representations actually decreased performance on these measures. On reflection, perhaps this should not be surprising. The Chomsky adjunction representations are motivated within the theoretical framework of Transformational Grammar, which explicitly argues for nonlocal, indeed, non context free, dependencies. Thus its poor performance when used as input to a statistical model which is insensitive to such dependencies is to be expected. Indeed, it might be the case that the additional adjunction nodes inserted in the tree transformations above have the effect of converting a local dependency (which can be described by a PCFG) into a nonlocal dependency (which cannot).

Another initially surprising property of the tree sequences produced by the PCFGs is that they do not reflect at all well the frequency of the different kinds of PP attachment found in the Penn II corpus. This is in fact to be expected, since the sequences consist of *maximum likelihood* parses. To see this, consider any of the examples analysed in section 4. In all of these cases, the corpora contained two tree structures, and the induced PCFG associates each with an estimated likelihood. If these likelihoods differ, then a maximum likelihood parser will always return the same maximum likelihood tree structure each time it is presented with its yield, and will never return the tree structure with lower likelihood, even though the PCFG assigns it a nonzero likelihood.

Thus the surprising fact is that these PCFG parsers ever produce a nonzero number of NP attachments and VP attachments in the same tree sequence. This is possible because the node label V in the attachment schema above abbreviates several different preterminal labels (i.e., the set of all verbal tags). Further investigation shows that once the V label in NP and VP attachment schemas is instantiated with a particular verbal tag, only either the relevant NP attachment schema or the VP attachment schema appears in the tree sequence. For instance, in the Id tree sequence (i.e., produced by the standard tree bank grammar) the 67 NP attachments all occured with the V label instantiated to the verbal tag AUX.[3]

6 Subsumed rules in tree bank grammars

It was mentioned in subsection 4.1 that it is possible for the PCFG induced from a tree bank to generate trees that are not meaningful representations with respect to the original tree bank representational scheme. The PCFG induced from the F2-21 subcorpus contains the following two productions:

$$\begin{aligned}\widehat{\mathrm{P}}(\mathrm{NP} \rightarrow \mathrm{NP\ PP}) &= 0.112 \\ \widehat{\mathrm{P}}(\mathrm{NP} \rightarrow \mathrm{NP\ PP\ PP}) &= 0.006\end{aligned}$$

These productions generate the Penn II representations of one and two PP adjunctions to NP, as explained above. However, the second of these productions will never be used in a maximum likelihood parse, as the parse of the sequence NP PP PP involving two applications of the first rule has a higher estimated likelihood.

In fact, *all* of the productions of the form NP $\rightarrow$ NP PP^n where $n > 1$ in the PCFG induced from the F2-21 subcorpus are subsumed by the NP $\rightarrow$ NP PP production in this way. Thus PP adjunction to NP in the maximum likelihood parses using this PCFG always appear as Chomsky adjunctions, even though the original tree bank did not use this representational scheme for adjunction!

In fact, a large number of productions in the PCFG induced from the F2-21 subcorpus are subsumed in this way. Of the 14,962 productions in the PCFG, 1,327, or just under 9%, are subsumed by combinations of two or more productions. Since these productions are never used to construct a maximum likelihood parse, they can be ignored if only maximum likelihood parses are required.

[3] This tag was introduced by (Charniak, 1996) to distinguish auxiliary verbs from main verbs.

7 Conclusion

There may be several ways of representing a particular linguistic construction as a tree. Because of the independence assumptions implicit in a PCFG, the kind of tree representation employed can have a dramatic impact on the quality of the PCFG model induced. This paper introduces a new methodology for examining these effects utilitizing tree transformations, and showed that one transformation, which copies the label of a parent node onto the labels of its children, can dramatically improve the performance of a PCFG model in terms of labelled precision and recall. It also pointed out that if only maximum likelihood parses are of interest then many productions can be ignored, since they are subsumed by combinations of other productions in the grammar.

References

Bies, Ann, Mark Ferguson, Karen Katz, and Robert MacIntyre, 1995. *Bracketting Guideliness for Treebank II style Penn Treebank Project*. Linguistic Data Consortium.

Charniak, Eugene. 1996. Tree-bank grammars. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 1031–1036, Menlo Park. AAAI Press/MIT Press.

Charniak, Eugene. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, Menlo Park. AAAI Press/MIT Press.

Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *The Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, San Francisco. Morgan Kaufmann.

Shieber, Stuart M. 1985. Evidence against the Context-Freeness of natural language. *Linguistics and Philosophy*, 8(3):333–344.