A Corpus-based Evaluation of Syntactic Locality in TAGs

Fei Xia and Tonia Bleam
Institute for Research in Cognitive Science
University of Pennsylvania
Philadelphia, PA 19104, USA
{fxia/tbleam}@linc.cis.upenn.edu

Abstract
This paper presents a new methodology for examining cases of non-locality. The algorithm presented here allows us to extract from a large annotated corpus sentences that appear to require non-local MCTAG. We examine one such case, extraposition from NP, and argue that the dependency involved is not syntactic and therefore does not require non-local MCTAG.

1. Introduction
Much important work has been done to investigate the adequacy of local TAGs to account for various linguistic phenomena, see, e.g., (Heycock, 1987; Becker et al., 1992; Abeillé, 1994; Bleam, 1994; Kulick, 1998; Joshi et al., 2000). This paper presents a new methodology for doing this kind of research. The algorithm presented here allows us to extract from a large annotated corpus (the Penn Treebank) constructions that seem to require non-local1 derivations. We propose that, in fact, these non-local dependencies should not be represented syntactically, and therefore do not constitute a problem for maintaining tree-local MCTAG.

2. Extracting MC sets from the Treebank
Extracting multi-component (MC) tree sets from Treebanks is one of the tasks performed by a grammar development system named LexTract, whose structure is shown in Figure 1, with the components relevant to the MC extraction task marked in bold. There are three main steps in the MC extraction procedure: first, a bracketed structure in a Treebank (ttree) is decomposed into a set of elementary trees (etrees); second, a derivation tree is built to show how the etrees are combined; third, any pair of etrees that contain co-indexed components are placed in a trees set with the etrees that connect them in the derivation tree. If the size of the set is more than three, the relation between the co-indexed components is not tree-local, assuming the correctness of Treebank annotations. For lack of space, we will use an example to demonstrate these main steps without going into the details of the algorithms (see (Xia, 1999) for details).

2.1. The extracted grammar
To ensure that the extracted etrees are compact and linguistically sound, we require that each etree in the grammar fall into one of three types determined by the relations between the anchor of the etree and other nodes in the tree, as shown in Figure 2:

---

1 We would like to thank Aravind Joshi, Jeff Lidz, Anoop Sarkar and the XTAG research group for their help and suggestions. This work was supported by NSF Grant SBR 8920230.

1 By non-local, we mean non-tree-local.
In LTAGs, on the other hand, arguments and adjuncts are distinguished. To overcome this difference in notation, the algorithm first fully brackets trees by adding intermediate nodes so that etrees express one of three relations: a predicate-argument relation, a modification relation, or a coordination relation.

The next step is to extract the component etrees from a fully bracketed tree. Recursive structures become mod-etrees or conj-etrees, and the remaining structures become spine-etrees. For instance, in the fully bracketed tree in Figure 4, along the path $S \rightarrow VP_1 \rightarrow VP_2 \rightarrow VP_3 \rightarrow PP \rightarrow IN$, three adjuncts (the relative clause, the NP yesterday and the auxiliary verb were) are factored out and each forms a mod-etree (#13, #11 and #3 resp.), while the remaining structures become a spine etree #4. The whole tree yields the fifteen etrees shown in Figure 5.

Some nodes in the tree are numbered and split into the top and bottom pairs. Recall that when a pair of etrees are combined during parsing, the root of one etree is merged with a node in the other etree. Splitting nodes into top and bottom pairs during the decomposition of the fully bracketed tree is the reverse process of merging nodes during parsing. For the sake of simplicity, we show the top and the bottom parts of a node only when the two parts will end up in different etrees.

2.2. Extracting etrees from trees

The first step of the MC extraction procedure is to extract etrees from trees. A tree from Penn English Treebank is shown in Figure 3, where reference indices (e.g. -1 and -2) mark co-indexed constituents.

```plaintext
S (NP-SBJ (NN supply) (NNS troubles))
   (VP (VBD were)
    (PP-LOC-PRD (IN on)
     (NP (NP (DT the) (NN minds)))
     (PP (IN of)
      (NP (NP (NNP Treasury) (NNS investors))
       (SBAR (-NONE- (WHNP-1 (WP who))))
       (S (NP-SBJ (-NONE- (WHNP-2 (WP who))))
        (VP (VBD were)
         (PP-LOC-PRD (IN about)
          (NP (DT the) (NN flood)))))))))
```

The trees in the Treebank are partially bracketed in a way that does not explicitly distinguish arguments from adjuncts.

Figure 1: The structure of LexTract

Figure 2: Forms of extracted etrees

Figure 3: An example from the Treebank

Figure 4: The etree set is a decomposition of the fully bracketed tree.
2.3. Building derivation trees
Having extracted the etree from a ttree, the next step for MC extraction is to build the derivation tree. Under the assumptions that no adjunctions are allowed at the foot nodes and at most one adjunction at any one node, and given the etree, the mapping between the fully bracketed ttree and the derivation tree is one-to-one. The derivation tree for the ttree in Figure 4 is shown in Figure 6.

2.4. Building MC tree sets
We construct MC sets using the derivation trees and the reference indices in the ttrees. Given a pair of constituents that are co-indexed in a ttree, let $e_i$ and $e_j$ be the two etree that the two constituents belong to. There exists a unique path that connects the two etrees in the derivation tree. The etrees on the path form a tree set. If the size of the set is more than three, the relation between the co-indexed components is not tree-local, assuming the correctness of Treebank annotations. In our example, the relation between WHNP-1 and *T*-1 (both are in tree #13) is tree-local, whereas the relation between *ICH*-2 (in tree #10) and SBAR-2 (in tree #13) is not.

3. Experiments
We ran the algorithm on the Penn Treebank II (Marcus et al., 1994). Table 1 gives the breakdown of MC sets by size. Out of 3151 MC sets, 999 sets (31.7%) had more than three etrees and were thus not tree-local. Table 2 shows the classifications of these non-local sets.

(1) That is [a skill], Sony badly needs $t_i$ and Warner is loath to lose $t_i$.

(2) It $t_i$ would be my inclination [to advise clients not to sell].

(3) Federal Express goes further $t_i$ in this respect [than any company].

(4) [Of all the ethnic tensions in America], which $t_i$ is the most troublesome right now?

(5) [JMB officials are expected to be hired to represent the pension fund on the Santa Fe Pacific Realty board, Mr Roulac said $t_i$, to insulate the fund from potential liability problems].

(6) The Diet doesn't normally even debate bills because the opposition parties are so often $t_i$ opposed to whatever LDP does [that it would be a waste of time].

---

3 Notice if a list etrees $E_i$ all modify the same etree $E$, $E_i$ will form a chain in the derivation tree, as circled in Figure 6. Those intermediate modified etrees are not included in the MC tree set.
Table 1: Numbers of extended MC sets and their frequencies in PTB

| size of MC sets | ≤ 3 (tree-local sets) | 4 | 5 | 6 | 7 | 8 | subtotal | total |
|-----------------|------------------------|---|---|---|---|---|----------|-------|
| # of MC sets (type) | 2152(68.3%) | 874 | 94 | 26 | 4 | 1 | 999(31.7%) | 3151 |
| # of MC sets (token) | 19994(91.3%) | 1772 | 102 | 26 | 4 | 1 | 1905(8.7%) | 21899 |

Table 2: Classification of 999 extended MC sets that look non-local

In each of these "non-local" cases, the Treebank notation establishes a dependence between two elements, as shown in (1) - (6). We suggest that, in fact, in all of the cases, the dependence is not syntactic, and so these examples do not constitute cases where non-local MCTAG would be required. Due to space considerations, however, we cannot address each case independently. Instead, we focus on one construction, that of Extraposition (EXP) from NPs, both because this was the most common type of "non-local" example found by the algorithm and because it is potentially the strongest case against tree-locality. We will show that even for this difficult case, tree-locality can be maintained.

4. Extraposition

One example of EXP was discussed in Section 2 (cf. Figure 3-6). Further examples are illustrated in (7) and (8), where the bracketed prepositional phrase is construed as an argument (7) or a modifier (8) of the NP in bold.4

(7) Younkers rang up sales in 1988 [of $313 million].

(8) The company gave us discounts all last year [on their premium brands].

Most generative analyses of this phenomenon associate the extraposed phrase (EXP phrase) with a gap in the NP

---

4 Adjunct status was determined using two tests: one-substitution and wh-extraction.
A Corpus-based Evaluation of Syntactic Locality in TAGs

9.5 This approach works for argument EXP, but it faces two problems when applied to adjunct EXP. The first problem is that, given current assumptions, the derivation of even the simplest cases requires non-local MCTAG (Weir, 1988). The trees required to derive (8) are given in Figures 9 and 10.

In this derivation \(\beta(c)\) adjoins to the NP of \(\alpha\) (discounts), and \(\beta(on)\) adjoins to the VP node in \(\alpha\) (give).

A second problem with the gap analysis is pointed out by (Abeille, 1994) citing (Gunnarson, 1982). Extraposited adjunct phrases (9) allow pronominalization of the head noun, something that is not allowed if the adjunct phrase is not extraposited (10). This is clear evidence that there is no movement since the putative underlying representation is impossible.

(9) John makes lists every day [with names of people who owe us money], and I make them every day [with names of people who we owe money to].

(Abeillé, 1994) thus proposes that the relationship between adjunct extraposition and the head noun should be a semantic one rather than a syntactic one. These "base-generated" cases are handled using synchronous TAG (S-TAG), where the syntax and semantics are represented by parallel TAG derivations. Representing the semantics with a TAG allows Abeillé to preserve the locality effects that we find in argument EXP, which do require a syntactic dependence. We refer to this locality property as \textit{etree boundedness} (ETB). As Abeillé notes, her analysis predicts that EXP is NP-bounded; that is, the extraposed element "has to be a complement of the top N, and cannot be a dependent of an embedded N". While ETB holds of argument EXP, we have found that adjunct EXP does not obey this condition, and hence cannot be accounted for in the S-TAG analysis. (11) and (12) are examples from the Treebank of non-NP-bounded EXP. In (11), the extraposed relative clause who worried... is not associated with an argument of the \(\textit{etrees}\) to which it attaches, but rather to a more deeply embedded NP, thus violating ETB.

(11) Supply troubles were on the minds of \textit{Treasury investors} yesterday [who worried about the flood of new government securities].

(12) Major rivals have been following a policy of continuous and deep discounting for at least the past 18 months [on their premium brands].

These examples show that the S-TAG analysis of the semantic dependence is too re-
strictive for adjunct EXP. Instead, we propose that the semantic dependency must be calculated post-derivationally, as, for example, in (Joshi & Vijay-Shanker, 1999), where the semantic representation is read off the derivation tree. The process of calculating this dependency must make reference to structure, but it does not adhere to the strict locality that the S-TAG analysis requires.

5. Conclusions
We have presented an algorithm to extract from the Penn Treebank constructions that seem to require non-local MCTAG. We propose that all these non-local dependencies should not be represented syntactically, and therefore do not require non-local MCTAG. One such example is NP-EXP, which has been previously argued to be a locally-bounded dependency. Our algorithm has revealed that adjunct EXP does not obey the locality constraints previously posited by linguists. If these examples are to be derived syntactically, they would require an LTAG more powerful than Tree-local MCTAG. We show, however, that the dependency between the head noun and the EXP phrase is not a syntactic one, but a semantic one. We conclude that extraposition does not constitute a case for using non-local MCTAG; tree-locality can be maintained.

References
ABEILLE A. (1994). Syntax or Semantics? Handling Nonlocal Dependencies with MCTAGs or Synchronous TAGs. Computational Intelligence, 10, 471-485.

ANDREWS A. (1975). Studies in the Syntax of Relative and Comparative Clauses. PhD thesis, MIT.

BALTIN M. (1981). Strict Bounding. In C. BAKER & J. McCARTHY, Eds., The Logical Problem of Language Acquisition. MIT Press.

BECKER T., RAMBOW O. & NIV M. (1992). The Derivational Generative Power, or, Scrambling is Beyond LCFRS. Technical Report IRCS-92-38, University of Pennsylvania.

BLEAM T. (1994). Clitic Climbing in TAG: A GB Perspective. In Proc. of TAG+3.

CULICOVER P. & ROCHEMONT M. (1990). Extraposition and the Complement Principle. Linguistic Inquiry, 21, 23-47.

GUERON J. (1980). On the Syntax and Semantics of PP Extraposition. Linguistic Inquiry, 11, 637-678.

GUNNARSON A. K. (1982). Trois constructions à dépendance entre sujet et pp. Lingvisticae Investigationes.

HEYCOCK C. (1987). The Structure of the Japanese Causative. University of Pennsylvania.

JOSHI A., BECKER T. & RAMBOW O. (2000). The complexity of scrambling and the competence performance distinction. In A. ABEILLE & O. RAMBOW, Eds., Tree Adjoining Grammar: Formalism, Computation, Applications. CSLI Publications.

JOSHI A. & VIJAY-SHANKER K. (1999). Compositional Semantics with LTAG: How Much Underspecification is Necessary? In Proc of 3nd International Workshop on Computational Semantics.

KROCH A. S. & JOSHI A. K. (1987). Analyzing Extraposition in a Tree Adjoining Grammar. In G. HUCK & A. OJEDA, Eds., Discontinuous Constituents, Syntax and Semantics, volume 20. Academic Press.

KULICK S. (1998). TAG and Clitic Climbing in Romance. In Proc. of TAG+4.

MARCUS M., KIM G., MARCINKIEWICZ M. A. et al. (1994). The Penn Treebank: annotating predicate argument structure. In Proc of ARPA speech and natural language workshop.

POLLARD C. & SAG I. A. (1994). Head-Driven Phrase Structure Grammar. The University of Chicago Press.

WEIR D. (1988). Characterizing mildly context-sensitive grammar formalisms. PhD thesis, University of Pennsylvania.

XIA F. (1999). Extracting tree adjoining grammars from bracketed corpora. In Proc. of NLPR5-99, Beijing, China.