Wh-islands in TAG and Related Formalisms

Owen Rambow* and K. Vijay-Shanker†
* CoGenTex, Inc. (owen@cogentex.com)
† University of Delaware (vijay@udel.edu)

1 Introduction: TAG and *wh*-Movement

The analysis of *wh*-movement given within TAG is a very convincing argument for the use of a constrained tree-rewriting formalism in syntax, since *wh*-movement does not require any special mechanism in TAG. *wh*-movement can be localized to elementary trees, and island effects are obtained naturally. This situation contrasts with approaches based on string-rewriting formalisms such as CFG, which require extensions (mathematical or at any rate definitional) to the basic mathematical formalism (resulting in theories such as GPSG, HPSG, LFG, or transformational grammar).

However, the question arises how other *tree-rewriting* formalisms such as D-Tree Grammar (Rambow et al., 1995) can handle *wh*-movement. Specifically, the question arises whether an equally elegant solution to the problem of *wh*-movement can be found. In this paper, we propose to study exactly which what features of the formal (mathematical) definition of TAG contribute to the correct analysis of *wh*-movement (in English). We will mainly concentrate on TAG, but occasionally mention tree-local MC-TAG.

The paper is structured as follows. In Section 2, we present the relevant elements of the definition of TAG. We then proceed to discuss specific island types and how these can be expressed in TAG: relative clause and other adjunct islands in Section 3, sentential subject islands in Section 4, and *wh*-islands in Section 5.

2 Elements of the Definition of TAG

In this paper, we will distinguish the following elements of the definition of TAG. (For a full mathematical definition, see (Vijay-Shanker, 1987).)

- The **extended domain of locality** (EDL). In TAG, the elementary structures are trees (rather than strings), so we can state extensive linguistically motivated restrictions on the shape of the elementary trees of a grammar. In fact, any such linguistic restriction on the shape of elementary structures exploits EDL.
- The **geometry of adjunction** (GA). By this term, we mean the specific, mathematical definition of the adjunction operation in TAG and, especially, the shape of the resulting derived tree. Specifically, an auxiliary tree β has a designated footnode; when β is adjoined in a tree α at node ν, it is inserted in its entirety into α. In the process, β remains intact, but α is divided in two subtrees at node ν, with β now attached at ν and the subtree formerly rooted in ν now attached to the footnode of β.
- The **factoring of recursion** (FR). By definition, in an auxiliary tree β, the footnode and the root node must have the same label, A. Furthermore, β can only be adjoined at a node labeled A. We observe that this aspect of the definition of TAG is not essential in the sense that the restrictions could be lifted without affecting the remainder of the definition, in particular the geometry of adjunction. The crucial part for the geometry of adjunction is the presence of a footnode; its label does not *a priori* matter.

We observer that tree-local MC-TAG has the same notion of EDL as TAG, and it has it own

notion of GA. FR is limited to those cases in which adjunction of on eof the component trees takes place.

By definition, any other tree-rewriting system[1] will also have EDL, while GA and FR are specific to TAG. Thus, we are in particular interested in the extent to which GA and FR are used in deriving island constraints, since such use would not necessarily carry over to other tree-rewriting systems.

In the following, we will be making an important assumption. Because of the EDL of the elementary structures of TAG, it is possible to *lexicalize* TAG in a straightforward manner (Schabes, 1990), meaning that each elementary tree in a grammar is associated with exactly one lexical item. Furthermore, we can require that each tree corresponding to a lexical item has positions (substitution nodes or a footnode) corresponding to each syntactic argument of that lexical item, and that the derivation thus reflects the syntactic relation between the lexical items involved (Rambow and Joshi, 1996) (the "lexical derivation constraint"). In this paper, we will only be interested in lexicalized grammars and in derivations that conform to the lexical derivation constraint.

3 Relative Clause Islands and Other Adjunct Islands

Sentence-initial extraction from certain adjuncts such as relative clauses modifying non-fronted object NPs or VP sentential adjuncts is ruled out simply by GA (in conjunction with the lexical derivation constraint). It is simply impossible to adjoin (or substitute) a tree into a (non-fronted) object, or adjoin a tree at a VP node (in a tree which has a subject NP to the left of the VP node), and obtain a derived tree in which some part of the adjoined tree is now in sentence-initial position.

In contrast, it is quite possible to adjoin a relative clause to a subject or adjoin an S-adjunct to a clausal tree (i.e., and adjunct phrase rooted in S), and obtain a *wh*-extraction to sentence-initial position. A sample auxiliary tree that would result in illicit extraction is shown in Figure 1. This tree can be ruled out in several different ways resorting to linguistic arguments. For example, one could exclude the tree by saying that extraction beyond the root node of an adjunct is impossible since the root node is not part of the projection of the lexeme anchoring the adjunct, or one could say that the tree in Figure 1 is illicit because of independently formulated constraints on node labels. In any case, one would be exploiting the EDL to express linguistically motivated constraints on the shape of elementary structures in the grammar. But, crucially, these constraints would carry over to the case of the relative clause modifying an object NP, and to the case of the VP-adjunct: it is not plausible that the linguistic constraints would be formulated in such a way that they only apply to subject relative clauses (or S adjuncts), but not to object relative clauses (or VP adjuncts). Thus, these cases are redundantly rules out by GA.

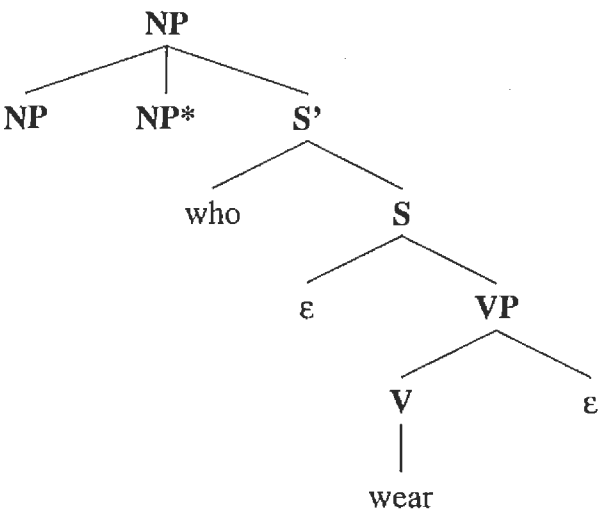

Figure 1: Relative clause with *wh*-moved element

Furthermore, there is a point that is easily overlooked. While object relative clauses with sentence-initial fronting are ruled out by GA, we *also* need to rule out non-initial fronting:

(1) *I saw what$_i$the man who was wearing t_i

While these kinds of sentences may be pathologically bad, they still need to be ruled out in a TAG grammar

[1]We include in this category systems which operate on tree-like structures.

We close by observing that if we are using tree-local MC-TAG, an argument very similar to the one above can be made to demonstrate that any predictive power obtained from the geometry of tree-local multicomponent adjunction is redundant with respect to independently required linguistic restrictions on the shape of the elementary tree sets. We omit the details.

4 Sentential Subjects

It would be possible to derive extraction from sentential subjects in the same manner that we derive extraction from sentential objects, namely by adjoining a matrix clause of the type shown in Figure 2 into the subordinate clause. In order to exclude such a derivation, we must say that the subject position, even when labeled S, cannot be a footnode. Thus, simply saying that we have factoring of recursion does not limit the extraction patterns: we must, in addition, make a linguistically motivated choice among possible footnodes. Designating a footnode is equivalent to allowing extraction from that position.

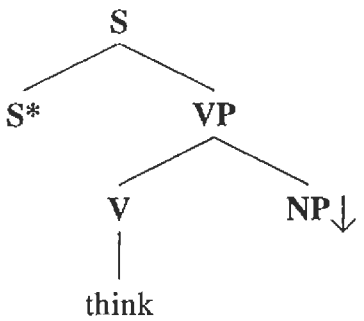

Figure 2: Matrix clause with sentential subject

However, the designation of the footnode is not sufficient. This is because of a well-known asymmetry in extraction from picture-NPs: while extraction from certain object NPs is possible, extraction from subject NPs never is.

(2) a. What$_i$ did you buy a picture of t_i?

b. * What$_i$ did a picture of t_i fall on your head?

Thus, if we use tree-local multicomponent MC-TAG to derive picture-NP extraction by substituting the main NP and substituting or adjoining the extracted *wh*-element, we must still specifically rule out extraction from subject position in some manner.[2] Furthermore, the same problem arises when we want to distinguish between verbs that allow picture-NP extraction and those that do not (as readily). Therefore, we will need some formal device (say, a feature EXTRACT on frontier nodes which regulates multicomponent derivations across them) for blocking extraction from certain positions *in addition to the choice of footnodes.* (This will also exclude extraction from sentential subjects if these are analyzed as projecting to NP.) The use of the device will need to be linguistically motivated. Some sort of equivalent device with similar linguistic motivation for its use can be used in tree rewriting systems which do not have FR or GA.

5 *Wh*-Islands

In English, we can exclude some wh-islands by restricting the shape of elementary trees.

(3) *What$_i$ do you know whom$_j$ Mary gave t_j t_i?

(3) is excluded because the elementary tree for give, which would need t have two *wh*-moved elements, is already excluded (we never have multiple *wh*-movement in English elementary trees). This analysis exploits the EDL and transfers to other tree-rewriting formalisms.

But that does not cover all cases of *wh*-islands.

(4) *What$_i$ do you know whom$_j$ Mary told t_j that she had bought t_i?

In (4), there is only one *wh*-extraction per elementary verbal tree. These cases can be excluded in several ways, but they all use FR. We

[2]Kroch (1989) suggests instead that the traces of picture-NP extractions are found in the elementary structures of the main verb. They are not licensed in the verbal tree because not bound; an index is adjoined through multi-component adjunction (along with the *wh*-element), which provides the binding. However, unlike traces in object position, traces in subject position are never licensed to begin with. This analysis exploits the EDL and could be expressed in other tree-rewriting formalisms as well.

take (Frank, 1992) as the most advanced example. There, trees in which *wh*-extraction from below takes place are footed in C′ and (hence by FR) are rooted in C′, while those without *wh*-extraction from below are both footed and rooted in CP. This ensures that if there is a *wh* element below (and assuming *wh* elements are always in SPEC(CP)), then the tree below must project to CP, and then the foot node must be CP, and hence the root node as well. Therefore, there is no room for a further *wh* element up front that would come from below. Note that if there is a single *wh*-movement at any depth of embedding, then because of the recursion part of FR, all trees above it must be CP-footed-and-rooted as well.

Frank's analysis makes use of several linguistic constraints on elementary structures (exploiting EDL), among which:

1. In an elementary tree, a C′ may never dominate a CP.
2. An elementary tree may not have two CP nodes one immediately dominating the other (the "anti-CP-recursion stipulation").
3. Each tree can only contain a single lexical item and its projection and (crucially) *no part of a different lexical item's projection.* Otherwise, we could have *(did) (john) wonder whether* in one tree which is rooted and footed in C′. Such a tree would allow sentences such as **Who did John wonder whether Sue saw?.*

Given these linguistic constraints as well as FR, it is impossible to obtain a node labeled CP immediately dominating a *wh*-element on the path separating a "moved" *wh*-element from the rest of its tree.

In tree rewriting systems that do not have FR, it will be necessary to derive the path constraint in some other manner. In DTG, it is possible to include path constraints explicitly in the elementary structures. In such an approach, the linguistic restrictions can be relaxed; it is not necessary to assume the anti-CP-recursion constraint, for example, and it would even be possible to allow an inversion of CP and C′.

6 Conclusion

In conclusion, we have seen that for relative clause islands and clausal adjunct islands, and for sentential subject islands, the TAG analysis exploits EDL but not GA or FR. These analyses would therefore carry over to other tree-rewriting systems. In the case of *wh*-islands, FR is exploited in conjunction with several linguistic EDL-type constraints in order to limited the occurrence of certain nodes on the path of *wh*-"movement". While this can not be replicated exactly in a system without FR, any other device to restrict the path has the same effect.

References

Frank, Robert (1992). *Syntactic Locality and Tree Adjoining Grammar: Grammatical, Acquisition and Processing Perspectives.* PhD thesis, Department of Computer and Information Science, University of Pennsylvania.

Kroch, Anthony (1989). Asymmetries in long distance extraction in a Tree Adjoining Grammar. In Baltin, Mark and Kroch, Anthony, editors, *Alternative Conceptions of Phrase Structure*, pages 66–98. University of Chicago Press.

Rambow, Owen and Joshi, Aravind (1996). A formal look at dependency grammars and phrase-structure grammars, with special consideration of word-order phenomena. In Wanner, Leo, editor, *Current Issues in Meaning-Text Theory.* Pinter, London.

Rambow, Owen; Vijay-Shanker, K.; and Weir, David (1995). D-Tree Grammars. In *33rd Meeting of the Association for Computational Linguistics (ACL'95)*, pages 151–158. ACL.

Schabes, Yves (1990). *Mathematical and Computational Aspects of Lexicalized Grammars.* PhD thesis, Department of Computer and Information Science, University of Pennsylvania.

Vijay-Shanker, K. (1987). *A study of Tree Adjoining Grammars.* PhD thesis, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA.