Extraction Phenomena in Synchronous TAG Syntax and Semantics

Rebecca Nesson and Stuart M. Shieber
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
{nesson,shieber}@deas.harvard.edu

Abstract

We present a proposal for the structure of noun phrases in Synchronous Tree-Adjoining Grammar (STAG) syntax and semantics that permits an elegant and uniform analysis of a variety of phenomena, including quantifier scope and extraction phenomena such as wh-questions with both moved and in-place wh-words, pied-piping, stranding of prepositions, and topicalization. The tight coupling between syntax and semantics enforced by the STAG helps to illuminate the critical relationships and filter out analyses that may be appealing for either syntax or semantics alone but do not allow for a meaningful relationship between them.

1 Introduction

Nesson and Shieber (2006) showed how a now-standard variant of the tree-adjoining grammar (TAG) formalism (multi-component, multiple adjunction, finite-feature-based TAG), when synchronized, leads to a natural analysis of the syntax-semantics relation, including handling of syntactic movement phenomena such as wh questions and relativization, semantic "movement" phenomena such as quantification, quantifier scope ambiguity, and even their interactions as found in pied-piped relative clauses.[1] phenomena were previously viewed as problematic for TAG analyses, leading to the hypothesizing of various extensions to the TAG formalism (Kallmeyer and Romero, 2004, and work cited therein). Independently, Han (2006a) developed a similar synchronous TAG analysis of pied-piping, providing evidence for the naturalness of the analysis.

Here, we update the analyses of noun phrases found in the previous works in one simple way, again with no additional formal TAG innovations, and show that it allows a further coverage of extraction and quantification phenomena as well as in-situ wh-phrases and topicalization. We emphasize that no novel formal devices are postulated to achieve this increased coverage — just a simple, natural and uniform change to the canonical structure of NPs and their semantics.

A word may be useful on the pertinence of this work in a workshop on "syntax and structure in machine translation", above and beyond the intrinsic importance of exploring the "applications of [synchronous/transduction grammars] to related areas including... formal semantics" underlying the workshop. Tree-structured mappings are advocated for machine translation systems because they allow for the expression of generalizations about relationships between languages more accurately and effectively. Evidence for this benefit ought to be found in the ability of the formalisms to characterize the primitive linguistic relationships as well, in particular, the form-meaning relationship for a natural language. The present work is part of a general program to explore the suitability of synchronous grammars for expressing this primary linguistic relationship. Inso-

[1] This work was supported in part by grant IIS-0329089 from the National Science Foundation.

Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation, pages 9–16,
Rochester, New York, April 2007.

far as it is successful, it lends credence to the use of these formal tools for a variety of language processing tasks, including MT. Insofar as it reveals insufficiencies in the formalism, it may lead to insights in the design or deployment of alternative systems.

We present a proposal for the structure of noun phrases in Synchronous Tree-Adjoining Grammar (STAG) syntax and semantics that permits an elegant and uniform analysis of a variety of phenomena, including quantifier scope and extraction phenomena such as wh-questions with both moved and in-situ wh-words, pied-piping, stranding of prepositions, and topicalization. Furthermore, the tight coupling between syntax and semantics enforced by grammar synchronization helps to illuminate the critical relationships and filter out analyses that may be appealing for either syntax or semantics alone but do not allow for a meaningful relationship between them.

We begin in Section 2 with a brief review of synchronous TAG and its application to English syntax and semantics. In Section 3, we present an analysis of quantifier scope that elucidates the relationship between the syntactic and semantic structures and explains an anomaly of previously proposed analyses. We apply the underlying idea from Section 3 to wh-questions in Section 4, showing that an alteration of the standard TAG syntax analysis of wh-questions produces the same derived trees while also elegantly modeling in-place wh-words. In Section 5 we present a challenging case for STAG syntax and semantics, the stranding of prepositions. This case is particularly difficult because the syntactic analyses suggested by previous work in STAG syntax do not encapsulate the relationships that appear necessary for the semantics. Our proposed analysis falls out naturally from the revision to the syntax of wh-words and respects both Frank's Condition on Elementary Tree Minimality (CETM) and the semantic relationships in the construction. In Section 6 we give an analysis of topicalization that also follows from the underlying ideas of the earlier analyses. We summarize the main ideas of the analysis in Section 7.

2 Introduction to Synchronous TAG

A tree-adjoining grammar (TAG) consists of a set of elementary tree structures of arbitrary depth,

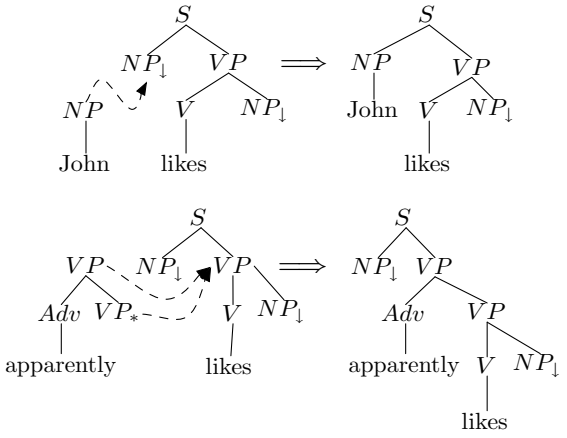

Figure 1: Example TAG substitution and adjunction.

which are combined with two operations, substitution and adjunction. Internal nodes in the elementary trees are labeled with a nonterminal symbol. Frontier nodes may be labeled with either terminal symbols or nonterminal symbols annotated with one of the diacritics $\downarrow$ or $*$. The $\downarrow$ diacritic marks a frontier nonterminal node as a *substitution node*, the target of the substitution operation. The *substitution* operation occurs when an elementary tree rooted in a nonterminal symbol A replaces a substitution node with the same nonterminal symbol.

Auxiliary trees are elementary trees in which the root and a frontier node, called the *foot node* and distinguished by the diacritic $*$, are labeled with the same nonterminal A. The *adjunction* operation involves splicing an auxiliary tree in at an internal node in an elementary tree also labeled with nonterminal A. Trees without a foot node, intended for substitution rather than adjunction into other trees, are called *initial trees*. Examples of the substitution and adjunction operations on sample elementary trees are shown in Figure 1. For further information, refer to Joshi and Schabes (1997).

Synchronous TAG (Shieber, 1994; Shieber and Schabes, 1990) extends TAG by taking the elementary structures to be pairs of TAG trees with links between particular nodes in those trees. Derivation proceeds as in TAG except that all operations must be paired. That is, a tree can only be substituted or adjoined at a node if its pair is simultaneously substituted or adjoined at a linked node. We notate the links by using boxed indices $\boxed{i}$ marking linked nodes.

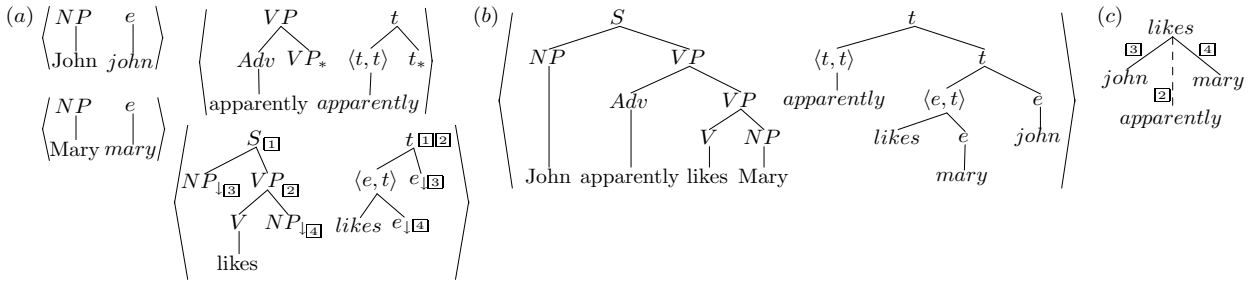

Figure 2: An English syntax/semantics STAG fragment (a), derived tree pair (b), and derivation tree (c) for the sentence "John apparently likes Mary."

As first described by Shieber and Schabes (1990), STAG can be used to provide a semantics for a TAG syntactic analysis by taking the tree pairs to represent a syntactic analysis synchronized with a semantic analysis.

For example, Figure 2(a) contains a sample English syntax/semantics grammar fragment that can be used to analyze the sentence "John apparently likes Mary". The node labels we use in the semantics correspond to the semantic types of the phrases they dominate.

Figure 2(c) shows the derivation tree for the sentence. Substitutions are notated with a solid line and adjunctions are notated with a dashed line. Each link in the derivation tree specifies a link number in the elementary tree pair, providing the location at which the operations take place. In this case, the tree pairs for the noun phrases *John* and *Mary* substitute into the *likes* tree pair at links 3 and 4, respectively. The word *apparently* adjoins at link 2. The tree pair so derived is shown in Figure 2(b). The resulting semantic representation can be read off the right-hand derived tree by treating the leftmost child of a node as a functor and its siblings as its arguments. Our sample sentence thus results in the semantic representation $apparently(likes(john, mary))$.

3 Quantifier Scope

We start by reviewing the prior approach to quantifier semantics in synchronous TAG. Consider the sentence "Everyone likes someone." We would like to allow both the reading where *some* takes scope over *every* and the reading where *every* takes scope over *some*. We start with the proposal of Shieber and Schabes (1990), which used multi-component TAG for the semantic portion of a synchronous TAG. Each quantified noun phrase has a two-component tree set as its semantics. One component introduces the variable quantified over in the scope of the quantifier; the other adjoins over the scope to provide the quantifier and restriction. Williford (1993) explored the use of multiple adjunction (Schabes and Shieber, 1993) to achieve scope ambiguity. Since the scope components of subject and object noun phrases adjoin at the same location in the semantic tree, they give rise to a systematic ambiguity as to which dominates the other in the derived tree, reflecting the semantic scope ambiguity of the sentence; the derivation tree itself is therefore a scope neutral representation. Previous work by Han (2006a; 2006b) and Nesson and Shieber (2006) describe this approach in detail, showing its applicability to a range of semantic phenomena.

A range of research has proceeded in an alternative line of using complex-feature-based TAG — rather than synchronous TAG — for TAG semantics (Kallmeyer and Romero, 2004, and work cited therein). Semantic representations are carried in features associated with nodes. Nonetheless, multi-component TAG with separate trees for bound position and scope is used here too. However, the two trees are syntactic trees, the quantified NP tree and a vestigial S tree, respectively. (An example is shown in Figure 6.) In such analyses, the single-node auxiliary S tree is used for the scope part of the syntax in order to get the desired relationship between the quantifier and the quantified expression in features threaded through the derivation tree and hence in the semantics.

The present analysis marries these two ap-

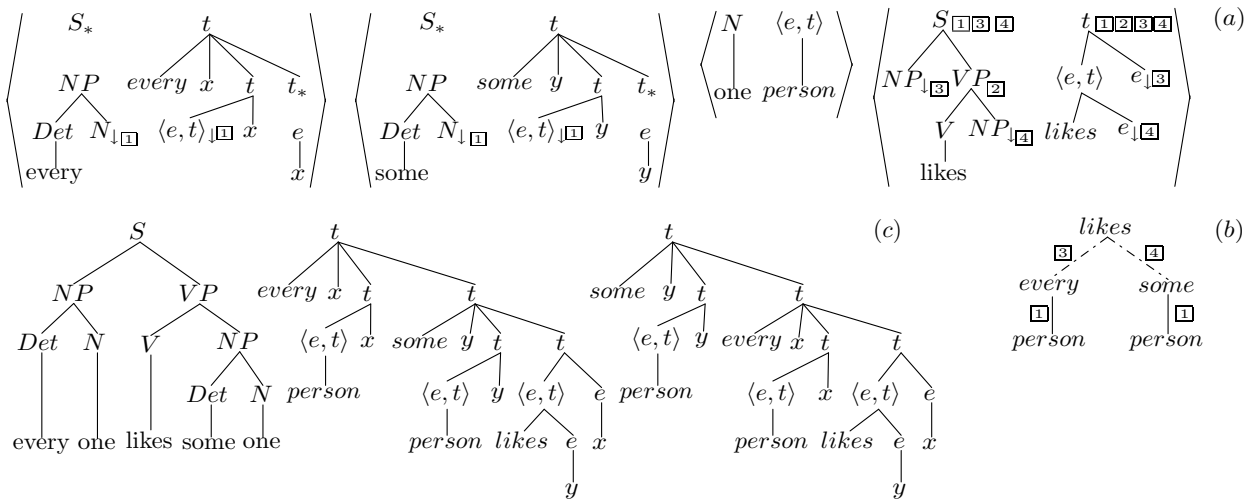

Figure 3: The elementary tree pairs (a), derivation tree (b), and derived trees (c) for the sentence "Everyone likes someone". Note that the derivation tree is a scope neutral representation: depending on whether *every* or *some* adjoins higher, we obtain different semantic derived trees and scope orderings.

proaches. Like the previous STAG work, we propose a solution in which a multi-component tree set provides semantics for quantified phrases, with multiple adjunction providing scope ambiguity. Like the complex-feature-based approach, we reflect the multi-component structure *in the syntax as well*. It is this single change in the analysis that makes possible the coverage of the wide range of phenomena we describe here.

Combining these two approaches, we give both the syntactic and semantic trees for quantifiers two parts, as depicted in Figure 3(a). In the semantics, the top part corresponds to the scope of the quantifier and attaches where the quantifier takes scope. The bottom part corresponds to the bound variable of the quantifier. By multiply adjoining the scope parts of the semantic trees of the quantifiers at the same location in the *likes* tree, we generate both available scope readings of the sentence.[2] Correspondingly on the syntax side, an NP tree provides the content of the noun phrase with a vestigial S tree available as well. Prior to the analyses given in this paper, the use of two trees in the quantifier syntax was an arbitrary stipulation used to make the semantic analysis possible. The pairing of the upper tree in the syntax with the scope tree in the semantics explicitly demonstrates their relationship and leads naturally to the exploration of non-degenerate upper trees in the syntax that we explore in this paper.

In order to use these multi-component quantifiers, we change the links in the elementary trees for verbs to allow a single link to indicate two positions in the syntax and semantics where a tree pair can adjoin, as shown in Figure 3(a). We add four-way links and drop the two-way links used by the unquantified noun phrases in the first example. This choice forces all noun phrase tree pairs to be multi-component in the syntax and semantics. Essentially, all noun phrases are "lifted" à la Montague. We explore the consequences of this in Section 6.

We turn now to the ramifications of this new syntactico-semantic STAG representation, showing its utility for a range of phenomena.

4 Wh-questions

The structure we propose for quantifiers suggests a new possibility for the TAG analysis of wh-words. We propose to simply treat wh-words as regular noun phrases by making them a multi-component tree set with an auxiliary tree that adjoins at the root of the verb tree and contains the lexical content and an initial tree with an empty frontier that substitutes at the argument position. This syntactic tree set can

[2] Nesson and Shieber (2006) provide a more in-depth explanation of the multiple-adjunction-driven approach to scope neutrality in STAG.

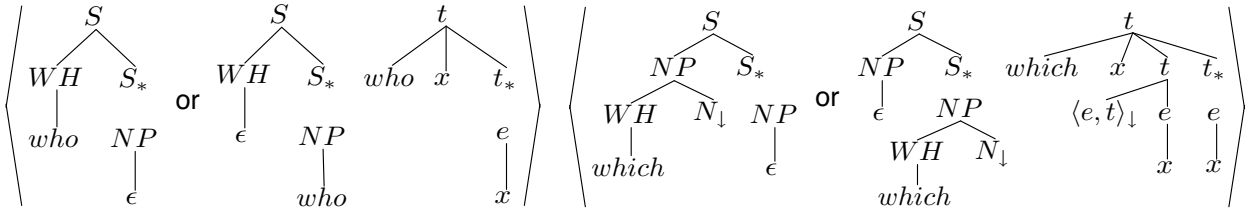

Figure 4: Elementary tree pairs for *who* and *which*. The left and middle tree sets are the syntactic alternatives used to model wh-movement and in-situ wh-words. The tree sets on the right provide the semantics.

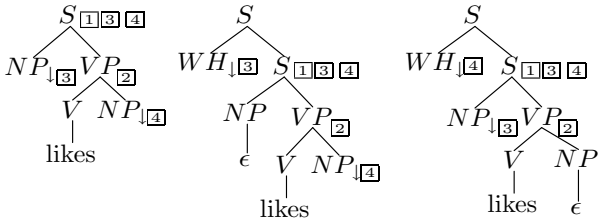

Figure 5: Traditional elementary trees for the verb *likes*. Using a revised, elementary syntax tree set for wh-words like *who*, only the left tree is necessary.

be paired with a multi-component semantic tree set that has an auxiliary tree containing the scope part and an initial tree that contains the bound variable. Wh-questions with the wh-word in place can be elegantly modeled with an alternative syntactic tree set in which the auxiliary tree has no lexical content and the wh-word is on the frontier of the initial tree that substitutes into the argument position. The semantic tree sets for both syntactic variations is the same. These trees are shown in Figure 4.

Besides the incorporation of a semantics, the basic analyses for wh-questions familiar from TAG syntax are otherwise unchanged because the top piece of the syntax tree set still ends up at the root of the main verb in sentences such as the following:

(1) Who likes Mary?
$who(x, likes(mary, x))$

(2) Which person does John like?[3]
$which(x, person(x), likes(x, john))$

(3) Which person does Bill think John likes?
$which(x, person(x), thinks(bill, likes(x, john)))$

(4) Who does each person like?
$who(x, each(y, person(y), likes(x, y)))$
$each(y, person(y), who(x, likes(x, y)))$

Note that in Sentence 3 *thinks* is not constrained to appear to the right of *who* in the syntax, because *thinks* and *who* both adjoin at the same location in the syntax. However, we can use a feature to force embedding verbs to adjoin lower than wh-words. The same situation exists in Sentence 4, though only in the semantics; the order of words in the syntax is well-defined but the multiple adjunction of the scope of *who* and the scope of *each* underspecifies the scope ordering between them. Both scope orderings are indeed arguably valid. Again, the preferences for certain orderings can be regulated using a feature. These issues highlight the many open questions about how to combine quantification and wh-terms, but also provides a first step towards their analysis within a concise STAG construction.

Our approach has several distinct advantages. First, it allows wh-words to be analyzed in a way that is uniform with the analysis of other noun phrases and allows us to simplify the lexical entries for verbs. In the traditional TAG analysis, wh-words substitute into specialized lexical trees for verbs that add an additional frontier node for the wh-word and abstract over one of the arguments of the verb by adding an empty terminal node at the frontier. Our revision to the elementary trees for wh-words allows us to remove several tree pairs from the elementary tree sets for verbs such as *like*. Instead of requiring an elementary tree pair for declarative sentences and an additional elementary tree for each argument

[3]The presence of do-support in wh-questions can be handled independently using a feature on the NP node into which the bottom part of the wh-word tree pair substitutes that governs whether and where a *do* tree adjoins.

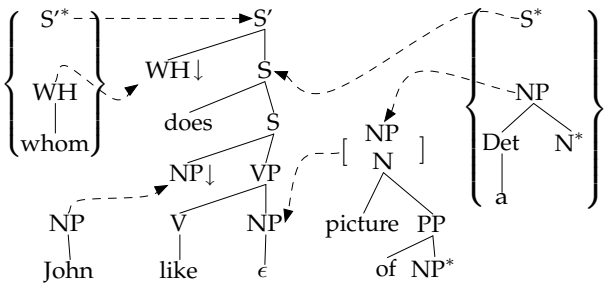

Figure 6: Kallmeyer and Scheffler's syntactic analysis for Sentence 6.

that can be replaced by a fronted wh-word to form a question (as shown in Figure 5), we can use just the single declarative sentence elementary tree.

Second, it provides a simple and elegant characterization of the syntax and semantics of wh-movement and the relationship between fronted and in-place wh-words. Using the alternative syntax tree set given in Figure 4 we model in-place use of wh-words as in Sentence 5 while still maintaining the usual semantic analysis:

(5) John likes who?
who(x, likes(x, john))

5 Stranded Prepositions

Sentence 6 presents a particularly challenging case for TAG semantics. The problem arises because *who* must contribute its bound variable, x, to the noun phrase "a picture of x". However, in the standard syntactic analysis *who* substitutes into the *likes* tree, and in any reasonable semantic analysis, *who* takes scope at the root of the *likes* tree.

(6) Who does John like a picture of?
$who(x, a(y,\ and(picture(y), of(x,y)),$
$likes(john, y)))$

Kallmeyer and Scheffler (2004) propose a syntactic analysis in which "a picture of" adjoins into the syntactic tree for "likes". The syntax for this analysis is shown for comparison in Figure 6. Associated with the syntactic analysis is a semantic analysis, which differs from ours in that all of the semantic computation is accomplished by use of a flexible set of features that are associated with nodes in the syntactic trees. This analysis maintains Frank's Constraint on Elementary Tree Minimality (CETM) if one analyzes the prepositional phrase as a complement of *picture* but it does so at the expense of a straightforward compositional semantics.[4] The source of the problem is that *who* contributes its bound variable to *likes* to form an intermediate semantics $who(x, likes(john, x))$, then *a picture of* combines non-compositionally to form the complete semantics given in Sentence 6.

Kroch (1989) describes the intuition eschewing this analysis: "The problem is that under such a derivation, the preposed wh-phrase changes its thematic role with each adjunction and the interpretation of the derived tree is not a simple function of the interpretations of its component elementary trees." When we consider the semantics of the two sentences, the anomaly of this analysis becomes apparent. In the first sentence the entity liked by John is referred to by the variable contributed by *who*. In the second sentence John likes an entirely different entity: the entity referred to by the variable contributed by *a*. Kallmeyer and Scheffler obtain the correct semantics by making use of non-local TAG operations to have the scope part of *a* adjoin into *likes* to capture the semantics of the likes proposition and employing a feature-based mechanism for swapping the variables as necessary.

Our revision to the syntax of wh-words provides an alternative way of maintaining the CETM that offers a much simpler semantic analysis. The details of the analysis are given in Figure 7. We adjoin *who* into the preposition *of* at link 1 where it contributes both variable and scope. The tree pair for *of* attaches to *a* at link 1, thus allowing the scope parts of the quantifier *a* and the wh-word *who* to end up taking scope over the main verb as in the analysis of prepositional phrases given by Nesson and Shieber (2006). It also places all the bound variables in the correct propositions without use of non-local operations or additional manipulation. A diagram of the derived syntax and semantics is given in Figure 8.

[4] In addition to suggesting a non-compositional semantics, their syntactic analysis makes use of non-local multi-component TAG in order to achieve the necessary semantic relationships. Although their use of non-local TAG may be benign in terms of complexity, our analysis is set-local. Our proposal therefore simplifies the syntactic analysis while also bringing it in line with a straightforward, compositional semantics.

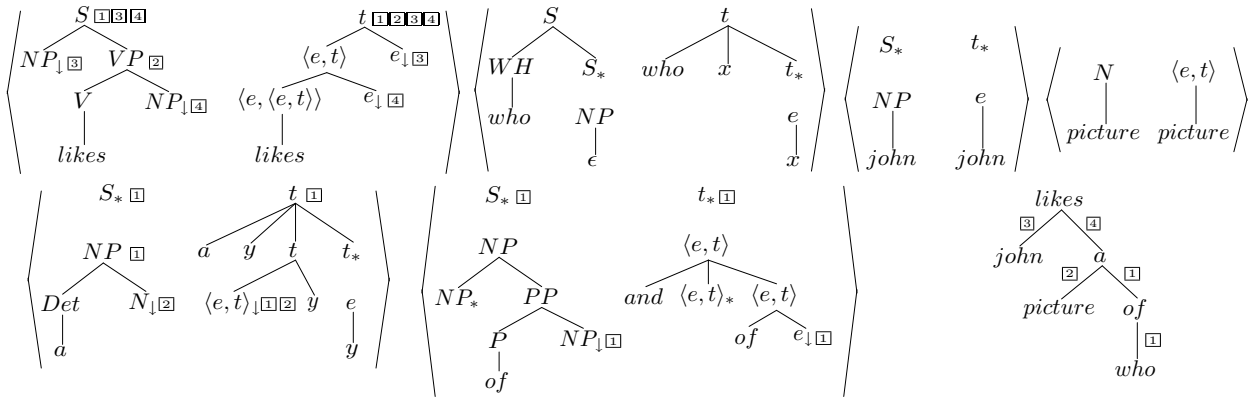

Figure 7: The elementary tree pairs and derivation tree for the sentence "Who does John like a picture of?".

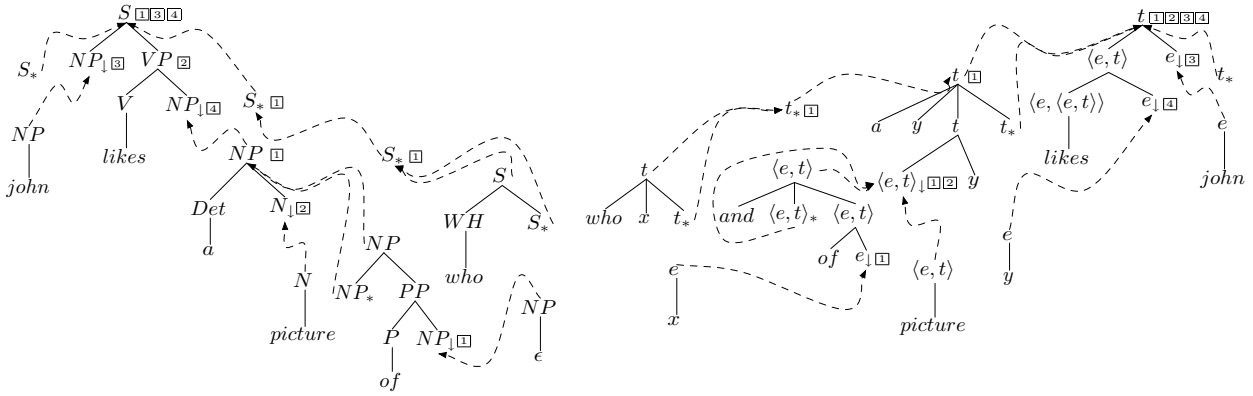

Figure 8: The derived syntax and semantics for Sentence 6.

6 Topicalization

The insight that allows us to model in-place wh-words extends to an elegant analysis of topicalization as well. The vestigial S_* tree that we added to the tree set for the syntax of every noun phrase need not always be contentless. Just as we moved the wh-word *who* from the top tree in its set to the bottom tree to model in-situ wh-words, we can move the lexical content of noun phrases to the top tree in their sets to model topicalization. For instance, the alternative tree pair for *Mary* shown in Figure 9 provides for an analysis of the sentence

(7) Mary, John likes.
likes(mary,john)

The analysis interacts properly with that for preposition stranding, so that the sentence

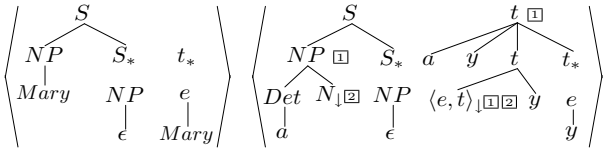

Figure 9: Alternative tree pairs for *Mary* and *a* that model topicalization.

(8) A picture of Mary, John likes.
a(x, and(picture(x), of(mary,x)), likes(x,john))

follows from the tree pair for *a* in the same figure.

7 Conclusion

In this paper we have proposed a uniform change to the structure of noun phrases in the STAG syntactico-semantic grammar. The formal tools we avail ourselves of comprise synchronous TAG with

set-local multicomponent adjunction and multiple adjunction. Nothing more is required.

All noun phrases now have a uniform multi-component structure in both the syntax and the semantics. In the semantics the top part corresponds to the scope-giving piece provided by the noun phrase and the bottom part to the bound variable or simple noun-phrase meaning. In the syntax, the top part corresponds to the lexical material that should appear moved to the edge of the sentence or clause; the bottom part corresponds to the lexical material that will fill an argument position of some head. By moving lexical material among the pieces of the multi-component set in the syntax, we can simply model phenomena like in-place wh-words and topicalization.

Making the top parts of wh-word tree sets into auxiliary trees allows them to adjoin not just to the main verb but also to heads of modifying clauses, such as prepositional phrases. This allows us to handle more complex sentences like Sentence 6 without violating either the CETM or going beyond simple compositional semantics. In order to allow the scope-giving part of the wh-word to percolate up to the root of the semantics of the main verb, each tree set that it adjoins into on its way must also have a scope part in the semantics to which it can adjoin. Scope carriers, such as prepositions, are therefore also multi-component in the semantics with a top node to which scope-givers can adjoin. One nice property of this analysis is that it predicts the observed facts about disallowed scope orderings in sentences that have three quantifiers, one of which is in a modifying clause. The scope part of the quantifier of the modified clause and the scope part of the quantifier of the modifying clause form an indivisible set as the derivation proceeds so that when they adjoin multiply with the scope part of the unmodified clause, that quantifier cannot intervene between them.

Our synchronous grammar treatment of the syntax-semantic relation with TAG is at least as simple and arguably more accurate than previous TAG proposals, offering treatments of such phenomena as in-situ wh-words, stranded prepositions, and topicalization.

References

Chung-Hye Han. 2006a. Pied-piping in relative clauses: Syntax and compositional semantics based on synchronous tree adjoining grammar. In *Proceedings of the 8th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+ 8)*, pages 41–48, Sydney, Australia.

Chung-Hye Han. 2006b. A tree adjoining grammar analysis of the syntax and semantics of it-clefts. In *Proceedings of the 8th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+ 8)*, pages 33–40, Sydney, Australia.

Aravind K. Joshi and Yves Schabes. 1997. Tree-adjoining grammars. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, pages 69–124. Springer.

Laura Kallmeyer and Maribel Romero. 2004. LTAG semantics with semantic unification. In *Proceedings of TAG+7*, pages 155–162, Vancouver, May.

Laura Kallmeyer and Tatjana Scheffler. 2004. LTAG analysis for pied-piping and stranding of wh-phrases. In *Proceedings of TAG+7*, pages 32–39, Vancouver, May.

Anthony Kroch. 1989. Asymmetries in long distance extraction in a tree adjoining grammar. In Mark Baltin and Anthony Kroch, editors, *Alternative Conceptions of Phrase Structure*. University of Chicago Press.

Rebecca Nesson and Stuart M. Shieber. 2006. Simpler TAG semantics through synchronization. In *Proceedings of the 11th Conference on Formal Grammar*, Malaga, Spain, 29–30 July.

Yves Schabes and Stuart M. Shieber. 1993. An alternative conception of tree-adjoining derivation. *Computational Linguistics*, 20(1):91–124.

Stuart M. Shieber and Yves Schabes. 1990. Synchronous tree-adjoining grammars. In *Proceedings of the 13th International Conference on Computational Linguistics*, volume 3, pages 253–258, Helsinki.

Stuart M. Shieber. 1994. Restricting the weak-generative capacity of synchronous tree-adjoining grammars. *Computational Intelligence*, 10(4):371–385, November.

Sean Williford. 1993. Application of synchronous tree-adjoining grammar to quantifier scoping phenomena in English. Undergraduate Thesis, Harvard College.