Some remarks on an extension of synchronous TAG*

David Chiang†, William Schuler†, and Mark Dras‡

†University of Pennsylvania
Dept of Computer and Information Science
200 S 33rd St
Philadelphia PA 19104 USA
{dchiang,schuler}@linc.cis.upenn.edu

‡University of Pennsylvania
Inst for Research in Cognitive Science
Suite 400A, 3401 Walnut St
Philadelphia PA 19104 USA
madras@linc.cis.upenn.edu

Abstract

We explore some properties of the synchronous formalism introduced in Dras (1999), showing that it handles an interaction, noted in Schuler (1999), between bridge and raising verbs which is problematic for synchronous TAG. We also show that it has greater formal power than synchronous TAG and discuss its computational complexity.

1. Introduction

Synchronous TAG (S-TAG), as defined by Shieber (1994), defines relations between languages by assembling paired elementary structures into isomorphic derivations. This isomorphism requirement is formally and computationally attractive, but for practical applications somewhat too strict. For this reason, Shieber suggests relaxing this requirement by treating bounded subderivations as elementary, but there are a few cases which remain problematic because they involve unbounded non-isomorphisms.

One such case is described by Schuler (1999). If a predicate is analyzed as a VP-adjunct in one language but an S-adjunct in another, then an unbounded non-isomorphism will arise when this predicate interacts with other VP-adjuncts. Consider the following sentences from English and Portuguese:

(1) X is supposed to (be going to ...) have to fly.

(2) É pressuposto que X (vai ...) tem/ter que voar.

We might analyze these sentences with the trees in Figure 1, but the resulting derivations for (1) and (2) would be non-isomorphic (see Figure 2).

Shieber (1994) describes this situation as "elimination of dominance"; in this case the non-isomorphism is potentially unbounded because the tree for *supposed to* adjoins into the lowest VP-adjunct on the derivation tree in English, but into the highest tree (that is, the initial tree) in Portuguese.

Schuler (1999) describes a solution to this problem based on a compositional semantics for TAG (Joshi & Vijay-Shanker, 1999) which relies on a mapping of contiguous ranges of scope in source and target derivations, but because it does not map subderivations in the source to subderivations in the target, this solution can only be used on individual

*This research is partially supported by ARO AASERT grant N00014-97-1-0603, ARO grant DAAG55971-0228, and NSF grant SBR-89-20230-15.

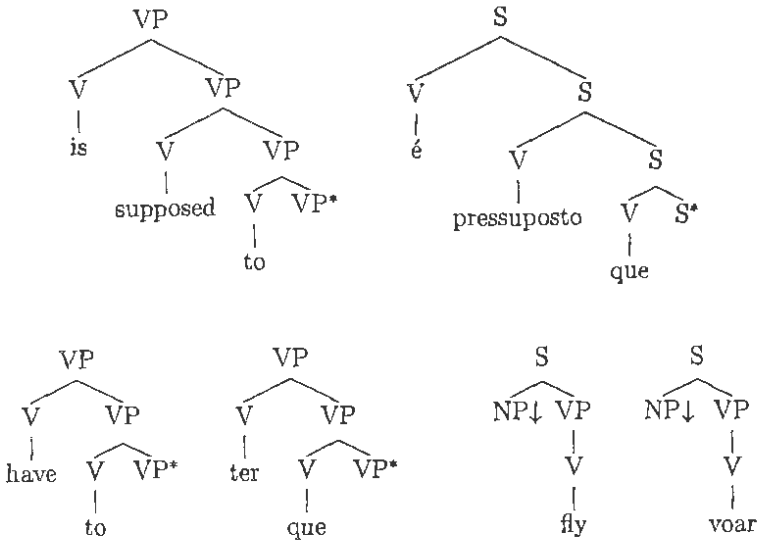

Figure 1: Elementary trees for sentences (1) and (2).

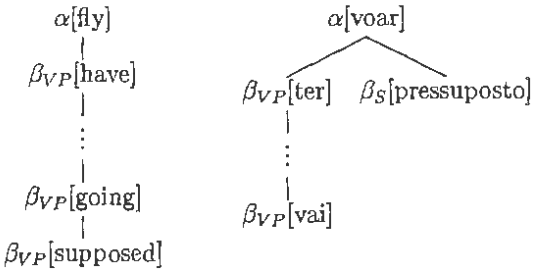

Figure 2: Derivation trees demonstrating *supposed/pressuposto* non-isomorphism.

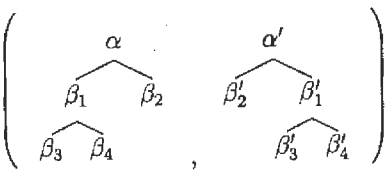

Figure 3: Paired derivation trees

derivation trees and not (tractably) on entire shared forests of possible derivations (Vijay-Shanker & Weir, 1993). Thus, for example, it is not directly possible to parse a natural language question and prune the chart using constraints on a semantic target.[1]
This paper shows that Schuler's example of unbounded non-isomorphism can be handled by the use of a meta-grammar, as in Dras (1999); specifically, by using a TAG meta-grammar in the regular form of Rogers (1994). (We will refer to this formalism as RF-2L(evel)TAG.) In addition, this paper explores how synchronous RF-2LTAG is more powerful than S-TAG: even though the weak generative capacity of the component TAGs is not altered by the synchronisation, the extra strong generative capacity of synchronous RF-2LTAG (that is, the extra structural descriptions it can produce) enables it to describe more *relations* between languages (that is, languages of pairs of strings). We also discuss the computational complexity of this formalism.

2. Using a meta-grammar

Dras (1999) describes what is in effect a relaxation of the requirement in the standard definition of S-TAG that paired derivation trees be isomorphic (as unordered trees). Since TAG derivation trees can be thought of as generated by context-free rules (Weir, 1988), we can likewise think of isomorphic derivation trees as generated by paired context-free rules (Aho & Ullman, 1969). For example, the derivation trees of Figure 3 would be generated by the following:

$$\langle \alpha \rightarrow \beta_1 \boxed{1}\ \beta_2 \boxed{2} \quad , \quad \alpha' \rightarrow \beta'_2 \boxed{2}\ \beta'_1 \boxed{1} \rangle$$
$$\langle \beta_1 \rightarrow \beta_3 \boxed{1}\ \beta_4 \boxed{2} \quad , \quad \beta'_1 \rightarrow \beta'_3 \boxed{1}\ \beta'_4 \boxed{2} \rangle$$

The relaxation proposed by Dras (1999) is to allow some other type of grammar to specify the pairings,[2] namely, TAG: with its greater domain of locality than CFGs, it can specify relationships between nodes of a derivation tree pair which are arbitrarily far apart. A meta-grammar thus pairs substructures in the derivation tree, rather than individual nodes; there is consequently an isomorphism between the trees representing the derivations of the derivations (the 'meta-derivations').
If the TAG meta-grammar is in the regnlar form of Rogers (1994), then the set of derivation trees is recognizable, and the weak generative capacity of the formalism is unchanged (Dras, 1999). Nevertheless, the additional strong generative capacity allows more mappings to be specified.
For example, a TAG meta-grammar can resolve the English-Portuguese mismatch noted above. If we use the same elementary tree pairs from Figure 1, the resulting derivation tree

[1] Ordinary synchronous TAG could use semantic target expressions to filter parse forests, but only if the target grammar were designed to accommodate a particular source grammar, with artificial notions of 'bridge' and 'raising' logical forms.

[2] Shieber's suggestion of treating bounded subderivations as elementary would be analogous to using a tree substitution grammar instead of a CFG to specify the pairings.

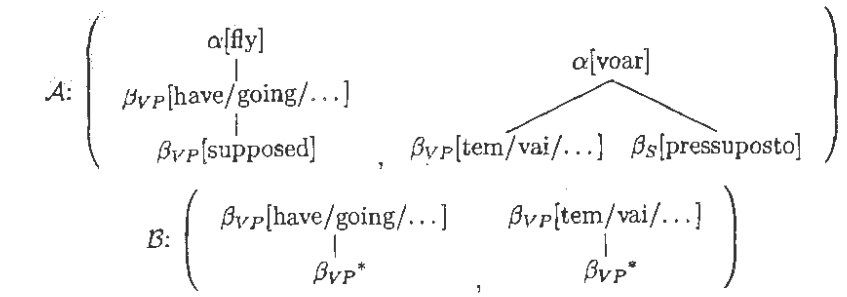

Figure 4: One possible meta-grammar for the *supposed/pressuposto* translation

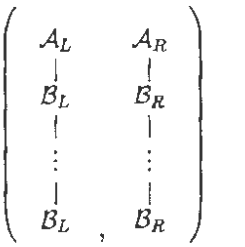

Figure 5: Meta-derivation trees.

structures (Figure 2) are non-isomorphic: in the English case, β[*fly*] and β[*supposed*] get stretched apart by an unbounded number of raising verbs, whereas in the Portuguese case, β[*pressuposto*] attaches directly to α[*voar*] and does not get stretched away. A TAG meta-grammar can be used to factor out the recursive material with pairs of auxiliary trees, like the pair $\mathcal{B}$ in Figure 4. An initial tree pair $\mathcal{A}$ specifies the difference between the English 'linear' derivation structure versus the Portuguese 'branching' derivation structure. The meta-derivation trees are as in Figure 5, with $\mathcal{A}_L$ and $\mathcal{A}_R$ being the left and right projections respectively of $\mathcal{A}$, and similarly for $\mathcal{B}$; they are clearly isomorphic, as desired.

3. Formal properties

Synchronous RF-2LTAG has the weak language preservation property (Rambow & Satta, 1996)—that is, the left and right projection languages of synchronous RF-2LTAGs are all TALs. However, as we have suggested, synchronous RF-2LTAG can specify relations between TALs which synchronous TAG cannot, as the following two claims show:

Claim (synchronous pumping lemma). If L is a language of pairs defined by a synchronous TAG, then there is a constant n such that if $\langle z, z' \rangle \in L$ and $|z| \geq n$ and $|z'| \geq n$, then $\langle z, z' \rangle$ may be written as $\langle u_1 v_1 w_1 v_2 u_2 v_3 w_2 v_4 u_3, u'_1 v'_1 w'_1 v'_2 u'_2 v'_3 w'_2 v'_4 u'_3 \rangle$, with $|v_1 v_2 v_3 v_4 v'_1 v'_2 v'_3 v'_4| > 0$, $|v_1 w_1 v_2 v_3 w_2 v_4| \leq n$, $|v'_1 w'_1 v'_2 v'_3 w'_2 v'_4| \leq n$, such that for all $i \geq 0$, $\langle u_1 v_1^i w_1 v_2^i u_2 v_3^i w_2 v_4^i u_3, u'_1 v'^{i}_1 w'_1 v'^{i}_2 u'_2 v'^{i}_3 w'_2 v'^{i}_4 u'_3 \rangle \in L$.

The proof is similar to that of the normal pumping lemma for TALs (Vijay-Shanker, 1987). The intuition is that the pumping lemma for local sets is applied to the derivation trees, and since paired derivation trees are isomorphic, the pumping constant can be chosen so that the pumping lemma holds for both sides simultaneously.

Claim. $L = \{\langle \mathrm{a}^i 1^j 2^j \mathrm{b}^i \mathrm{c}^i 3^j 4^j \mathrm{d}^i, 1^j \mathrm{a}^i \mathrm{b}^i 2^j 3^j \mathrm{c}^i \mathrm{d}^i 4^j \rangle \mid i, j \geq 0\}$ is not definable by a syn-

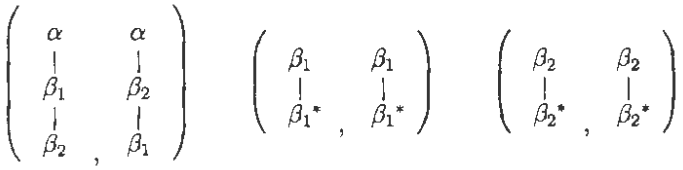

Figure 6: TAG meta-grammar for defining L

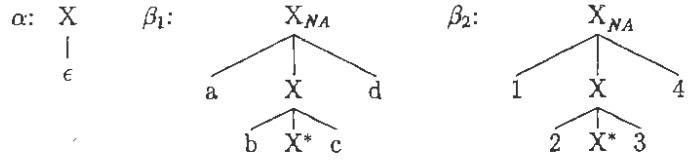

Figure 7: Object level trees for defining L

chronous TAG.

Proof. Assume that L is definable by a synchronous TAG. If n is the constant given by the pumping lemma, let $\langle z, z' \rangle = \langle \mathrm{a}^n 1^n 2^n \mathrm{b}^n \mathrm{c}^n 3^n 4^n \mathrm{d}^n, 1^n \mathrm{a}^n \mathrm{b}^n 2^n 3^n \mathrm{c}^n \mathrm{d}^n 4^n \rangle$ Then z and z' have to be written so that the v_i and v_i' are all letters or all numerals, or else the "pumped" pairs will not be in L. But if they are all letters, then $|v_1 w_1 v_2 v_3 w_2 v_4| > n$; if they are all numerals, then $|v_1' w_1' v_2' v_3' w_2' v_4'| > n$. Since $\langle z, z' \rangle$ cannot be rewritten in the manner indicated by the pumping lemma, L must not be definable by a synchronous TAG.

L can, however, be defined by the synchronous RF-2LTAG in Figure 6, where α, β_1, and β_2 are the same for both sides, shown in Figure 7.

So synchronous RF-2LTAG is more powerful than synchronous TAG; however, just as RF-TAG can be parsed in $\mathcal{O}(n^3)$ time like CFG, RF-2LTAG can be parsed in $\mathcal{O}(n^6)$ time like TAG. We can do this by keeping track of meta-adjunctions using stacks inside the chart items (Rogers, 1994). Because of the regular-form condition, the stacks will have bounded depth.

If we wish to transfer entire shared forests of derivations (Vijay-Shanker & Weir, 1993) rather than single parses, we may incur additional complexity, but this problem can still be solved in polynomial time, because there is a subderivation in the target grammar for every subderivation in the source. In contrast, the method of (Schuler, 1999) would require exponential time because it is defined only on completed parses.

One remaining question is, is it sufficient to use a TAG as a meta-grammar? For any k, define a language over the alphabet $\{\mathrm{a}_1, \mathrm{a}_2, \ldots, \mathrm{a}_k\}$: SEPARATE-$k$ $= \{\langle w, \mathrm{a}_1^{i_1} \mathrm{a}_2^{i_2} \cdots \mathrm{a}_k^{i_k} \rangle \mid$ w has exactly i_j occurrences of $\mathrm{a}_j\}$. SEPARATE-8 can be generated by a synchronous RF-2LTAG (the grammar is not complicated, but large), but SEPARATE-9 cannot. This can be seen by left-intersecting with $(\mathrm{a}_1 \mathrm{a}_2 \cdots \mathrm{a}_9)^*$ (this can be done without disrupting the synchronization): the right projection of the result will be $\{\mathrm{a}_1^n \mathrm{a}_2^n \cdots \mathrm{a}_9^n\}$, which is not generable by any 2LTAG.

More generally, SEPARATE-2^{k+1} can be generated by a synchronous k-level TAG, but SEPARATE-$(2^{k+1}+1)$ cannot. These are all well-behaved relations between *regular* languages; thus the weak language preservation property does not provide a natural ceiling on how powerful a meta-grammar can be. It remains to be seen what kinds of meta-grammars are actually practically useful, and what bounds can be placed on their computational

complexity.

4. Conclusion

In the future we hope to explore the possibility of using meta-level structures for linguistic description (in particular, shifting the Condition on Extended Tree Minimality (Frank, 1992) to meta-level elementary trees); in such an approach it becomes possible to eliminate the *supposed/pressuposto* non-isomorphism entirely.

Under the present approach, however, we have shown that a synchronous TAG meta-grammar provides the extra strong generative capacity needed to localize certain unbounded non-isomorphisms, overcoming some of the limitations of standard synchronous TAG while preserving the essential idea of local synchronization and its attendant advantages.

References

AHO A. V. & ULLMAN J. D. (1969). Syntax directed translations and the pushdown assembler. *J. Comp. Syst. Sci.*, **3** (1 p.), 37–56.

DRAS M. (1999). A meta-level grammar: redefining synchronous TAG for translation and paraphrase. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99)*.

FRANK R. (1992). *Syntactic locality and tree adjoining grammar: grammatical acquisition and processing perspectives*. PhD thesis, Computer Science Department, University of Pennsylvania.

JOSHI A. & VIJAY-SHANKER K. (1999). Compositional Semantics with Lexicalized Tree-Adjoining Grammar (LTAG): How Much Underspecification is Necessary? In *Proceedings of the 2nd International Workshop on Computational Semantics*.

RAMBOW O. & SATTA G. (1996). Synchronous Models of Language. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL '96)*.

ROGERS J. (1994). Capturing CFLs with tree adjoining grammars. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL '94)*.

SCHULER W. (1999). Preserving semantic dependencies in synchronous tree adjoining grammar. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99)*.

SHIEBER S. M. (1994). Restricting the weak-generative capability of synchronous tree adjoining grammars. *Computational Intelligence*, **10** (4 p.).

VIJAY-SHANKER K. (1987). *A study of tree adjoining grammars*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania.

VIJAY-SHANKER K. & WEIR D. (1993). The use of shared forests in tree adjoining grammar parsing. In *Proceedings of EACL '93*, p. 384–393.

WEIR D. (1988). *Characterizing Mildly Context-Sensitive Grammar Formalisms*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania.