Lexicalized non-local MCTAG with dominance links is NP-complete

Lucas Champollion

July 19, 2007

1 Introduction

An NP-hardness proof for nonlocal multicomponent tree-adjoining grammar (MCTAG, Joshi, 1985; Weir, 1988)[1] is extended to some linguistically relevant restrictions of that formalism. It is found that there are NP-hard languages among the languages described by nonlocal MCTAGs even if the following restrictions are imposed: every (or alternatively at least one) tree in every tree set has a lexical anchor; every tree set may contain at most two trees; in every such tree set, there is a dominance link between the foot node of one tree and the root node of the other tree and this dominance link must be obeyed in the derived tree. This is the version of MCTAG used in Becker et al. (1991).

While standard TAGs are closed under lexicalization (Schabes, 1990), it is not known whether this also applies to nonlocal MCTAG. So it would be conceivable that *lexicalized* nonlocal MCTAG are mildly context-sensitive. However, we show that lexicalized nonlocal MCTAG in fact contains languages that are NP-complete. Moreover, even if both restrictions (dominance links and lexicalization) are applied to nonlocal MCTAG at the same time, it still remains NP-complete. (For reasons of space, only NP-hardness but not membership in NP is discussed in this abstract.)

The restriction of the proof to the lexicalized and dominance-link conditions is mathematically straightforward. It is linguistically significant, however, because it has been argued (Becker et al., 1991) that phenomena such as German scrambling put natural language outside of the class LCFRS, a characterization of the class of *mildly context-sensitive languages* (Weir, 1988). This would put natural language outside of standard TAG, and even outside of set-local MCTAG, which is equivalent to LCFRS. (Joshi, 1985; Weir, 1988).

It should be noted that there exist alternative views on the complexity of scrambling. The data that would put it outside of LCFRS is based on the assumption that scrambling is grammatical for an unbounded number of levels of embedding, no matter what the order of scrambled arguments is. However, this assumption is hard to check empirically, because beyond three levels of embedding it is near impossible to obtain reliable grammaticality judgments from speakers due to the processing load. It is exactly the judgments about sentences beyond three levels of embedding that would be necessary in order to choose among grammar classes inside LCFRS and those outside of it. For example, tree-local MCTAG is inside LCFRS, and it derives some though not all of the sentences with more than tree levels of embedding (Aravind Joshi, p.c.). See also Joshi et al. (2002) for discussion. We are *not* pursuing the interesting empirical question of choosing among these perspectives here. Our enterprise is merely mapping out the boundary between those linguistically relevant grammar classes which are polynomially recognizable and those which, assuming $P \neq NP$, are not.

Since there are polynomially recognizable languages outside of the class LCFRS (Boullier, 1998), the hope is that a suitable grammar class can be found that contains all natural languages and whose members are still polynomially recognizable. The contribution of the present work to that search is that nonlocal MCTAG is, unfortunately, not a candidate for such a class even if linguistically plausible restrictions – i.e. lexicalization and/or dominance links – are applied.

[1] In an MCTAG, instead of auxiliary trees being single trees like in standard TAG (Joshi and Schabes, 1997) we have auxiliary sets, where a set consists of one or more (but still a fixed number of) auxiliary trees. Adjunction is defined as the simultaneous adjunction of all trees in a set to different nodes. In a *tree-local* MCTAG, all trees from one set S must be simultaneously adjoined into the same elementary tree T. In a *set-local* MCTAG, all trees from one set S must be simultaneously adjoined into trees that all belong to the same set S_2. If this requirement is dropped altogether, we obtain *non-local* MCTAG.

2 A nonlocal MCTAG that generates an NP-hard language

This section presents a proof of the NP-hardness of standard (i.e. nonlexicalized, non-dominance-links) nonlocal MCTAG with adjunction constraints. This is essentially the proof that was reported by Dahlhaus and Warmuth (1986) for scattered grammars. It was noted by Rambow and Satta (1992) and Rambow (1994) that the proof carries over to certain nonlocal MCTAGs in principle, but they do not actually perform the construction of the NP-hard grammar. We flesh out the proof in detail here, as we are going to need it later where it will be modified for the restricted cases.

Intuitively, the main property of nonlocal MCTAG that is underlying this proof is the following: We make use of the fact that nonlocal MCTAG allows us to introduce pairs of terminals into the derivation at two different (indeed arbitrarily distant) places in the tree, but requires us to introduce them at the same time. This allows us to build a grammar that counts up to the same arbitrary number in two places of the derivation. In the final string, each of these numbers is expressed in unary as a block of identical terminals. In designing our grammar, we may either choose to delimit these blocks from each other by special separator symbols, or simulate addition by leaving out these separators. In this case, since the string contains no record of the derivation, a recognizer only sees the sum and not the summands, and must in effect guess which summands have been chosen.

We now present a polynomial reduction from the strongly NP-complete problem *3-Partition* to a specific MCTAG.

3-Partition.

Instance. A set of $3k$ natural numbers n_i, and a bound B.

Question. Can the numbers be partitioned into k subsets of cardinality 3, each of which sums to B?

To simplify the construction, assume that 3-Partition is restricted in the way that there are at least three numbers n_i (i.e. that $k \geq 1$) and that each of the numbers n_i is greater or equal to two.

An instance of 3-Partition can be described as the sequence $\langle n_1, \ldots, n_{3k}, B\rangle$, or equivalently as the string $xa^{n_1}xa^{n_2}\ldots xa^{n_{3k}}(yb^B)^k$ where a, b, x, y are arbitrary symbols. (In this string, x and y are only used as separators. It will be seen later why the end of the string was chosen to be repeated k times.) In Figure 1, we provide a nonlocal MCTAG G_1 that has the property that $\langle n_1, \ldots, n_{3k}, B\rangle$ is an instance of 3-Partition if and only if the string $xa^{n_1}xa^{n_2}\ldots xa^{n_{3k}}(yb^B)^k$ is accepted by G_1. (Ignore the dominance links in Figure 1 for now.) This grammar is based on the growing scattered grammar G in Dahlhaus and Warmuth (1986).

G_1 has one initial tree, α_{start}; one single auxiliary tree, $\beta_{create-triple}$; and five auxiliary tree sets. We indicate obligatory adjunction sites with OA and null-adjunction sites with NA. Foot nodes are always null-adjunction sites and therefore not explicitly marked as such. There are no substitution sites in G_1.

To get an idea of how the grammar works, look at Figure 1. All terminals are introduced to the left of the spine of their auxiliary tree, so whatever is introduced towards the top of the derived tree will appear towards the left of the string.

Call a non-terminal node *saturated* iff it has a null-adjunction (NA) constraint, and *unsaturated* iff it has an obligatory adjunction (OA) constraint. In the derived trees produced by G_1, every non-terminal node is either saturated or unsaturated. All non-terminal nodes that are introduced into the derivation, except root and foot nodes, are unsaturated – they must be adjoined into at some point. Because the root (as well as the foot) nodes of every auxiliary tree have null-adjunction constraints, as soon as an auxiliary tree is adjoined into a node, that node is saturated – it is replaced by a null-adjunction node. Most trees contain exactly one unsaturated node and therefore adjoining them keeps the number of unsaturated nodes in the derivation constant. The exceptions are the singleton trees α_{start} and $\beta_{create-triple}$, which introduce more than one unsaturated node, and the trees $\beta_{close-triple.2}, \beta_{end.1}$ and $\beta_{end.2}$, which introduce none.

All derivations must start with the initial tree α_{start}, which introduces an X and a triple $\langle Y, \hat{Y}, \hat{Y}\rangle$. Subsequent steps in the derivation may use $\beta_{create-triple}$ nondeterministically to introduce any number of additional triples $\langle Y, \hat{Y}, \hat{Y}\rangle$. For clarity of exposition, we can assume that these triples are all introduced as early as possible and that there are k of them, corresponding to the number of sets created by the partition in the instance of 3-Partition that is to be recognized.

At all times there is at most one of $\{X, \overline{X}\}$ in the derivation. So after α_{start} and $\beta_{create-triple}$ have produced the original X and some number of $\langle Y, \hat{Y}, \hat{Y}\rangle$, any derivation can only proceed as follows:

1. Pick the X and some Y (resp. $\hat{Y}$) and use $\beta_{consume-y}$ (resp. $\beta_{consume-\hat{y}}$) to generate xa on the left and yb (resp. b) on the right. This introduces $\overline{X}$ on the left and $\overline{Y}$ on the right.
2. Optionally use $\beta_{fill-triple}$ to add an equal number of a's and b's to the left and right of the string.
3. Finally replace $\overline{X}$ by aa and $\overline{Y}$ by bb. Either $\beta_{close-triple}$ or β_{end} can be used for this. The only difference consists in whether another X is introduced. But there is no real choice here: If there are any Y's or $\hat{Y}$'s left on the right, they need to be consumed by introducing an X on the left and then going through steps 1 to 3 again with that X. If not, no X can be introduced or the derivation would get stuck.

This way, the grammar produces a sequence of blocks of a's followed by a sequence of blocks of b's. The sizes of the blocks of a's correspond to the numbers n_i. While X is deriving xa^{n_i} followed by X, either some Y derives yb^{n_i} or some $\hat{Y}$ derives b^{n_i}. There is a block of b's for each n, but the blocks of b's are permuted and grouped in threes. While the grammar produces more words than the ones that correspond to solutions of 3-Partition, those words in which each group of three sums to B are exactly the ones that correspond to some solution.

Proof.[2] Suppose we are given a solution of the instance of 3-Partition, i.e. disjoint sets $A_1, \ldots, A_k$, each of which contains 3 n_i's that add to B. It will be shown that the word $w = xa^{n_1}xa^{n_2}\ldots xa^{n_{3k}}(yb^B)^k$ that describes the instance of 3-Partition is in $L(G_1)$.

For any derived MCTAG tree t, do a left-to-right preorder traversal of t concatenating all the node labels and skipping any saturated non-terminals, and call the resulting string the *unsaturated yield of* t. Define a relation "$\Rightarrow$" ("is rewritten to") as holding between two strings s_1 and s_2 wrt. an MCTAG G iff there exist trees t_1, t_2 with unsaturated yields s_1, s_2 such that t_2 can be obtained from t_1 in a single (possibly multicomponent) substitution or adjunction step. We write $G \Rightarrow s$ iff G contains an initial tree t rooted in the start symbol of G such that there is a string s_t that is the unsaturated yield of t and $s_t \Rightarrow s$. As usual, we write $\stackrel{*}{\Rightarrow}$ for the reflexive and transitive closure of $\Rightarrow$. Obviously, for all $w \in \Sigma^*$, G derives w iff $G \stackrel{*}{\Rightarrow} w$.

Clearly $G_1 \stackrel{*}{\Rightarrow} X(Y\hat{Y}\hat{Y})^k$. Associate each set $A_q, 1 \leq q \leq k$, with the qth group $Y\hat{Y}\hat{Y}$ and associate each of the three elements of the set with one of the three symbols Y, $\hat{Y}$, and $\hat{Y}$, respectively, in the group. The association within each group is arbitrary. The derivation $X(Y\hat{Y}\hat{Y})^k \stackrel{*}{\Rightarrow} w$ is organized in $3k$ phases. In the jth phase, for $1 \leq j < 3k$, X is rewritten to $xa^{n_j}X$ and in parallel the Y-symbol (resp. $\hat{Y}$-symbol) that is associated with n_j is rewritten to yb^{n_j} (resp. b^{n_j}). In the $3k$th phase X is rewritten to $xa^{n_{3k}}$ and in parallel the Y-symbol (resp. $\hat{Y}$-symbol) that is associated with n_{3k} is rewritten to $yb^{n_{3k}}$ (resp. $b^{n_{3k}}$). Since the numbers of A_q add to B, each group $Y\hat{Y}\hat{Y}$ derives yb^B.

For the opposite direction, assume now that $G_1 \stackrel{*}{\Rightarrow} w$, where $w = xa^{n_1}xa^{n_2}\ldots xa^{n_{3k}}(yb^B)^k$. Normalize the derivation by adjoining all instances of $\beta_{create-triple}$ as early as possible within the derivation of w. The normalized derivation has the form:

$$G_1 \stackrel{*}{\Rightarrow} X(Y\hat{Y}\hat{Y})^k \stackrel{*}{\Rightarrow} w$$

The symbol X is rewritten to $\overline{X}$ and after a number of steps to X again. More exactly, X produces $xa^{n_i}X$ at the jth phase, for $1 \leq j < 3k$, and $xa^{n_{3k}}$ in the last phase. Furthermore, in the ith phase, for $1 \leq i \leq 3k$, a particular Y (resp. $\hat{Y}$) is rewritten to yb^{n_i} (resp. b^{n_i}). Observe that each non-terminal Y is responsible for a terminal y in w and the Y's produce exactly B b's. Each group thus corresponds to a different set of three numbers that adds to B and there are k such sets. □

3 Restriction to dominance links

The above proof can be easily restricted to nonlocal MCTAG with dominance links (MCTAG-DL) by adding dominance links to G_1 as in Figure 1 to produce a strongly equivalent MCTAG-DL G_2. Since the two

[2] From Dahlhaus and Warmuth (1986), with a few extensions.

grammars have the same language, it follows that there exist MCTAG-DL with NP-hard languages.

Proof. Call any element of $\{X, \overline{X}\}$ an *X-like symbol* and any element of $\{Y, \overline{Y}, \hat{Y}\}$ a *Y-like symbol*. Observe that in the tree α_{start} in the original grammar G_1, and vacuously in all the other trees of the grammar, any X-like symbol dominates any Y-like symbol. Call any elementary or derived tree with this property an *X-over-Y tree*. Moreover, in every tree set the tree with the X-like foot node contains only X-like non-terminals and the tree with the Y-like root node contains only Y-like non-terminals. By straightforward induction, every derived tree generated by G_1 can be shown to be X-over-Y. We can now add dominance links as shown in Figure 1 in a way such that a derived tree that violates any of these dominance links would have a Y-like root node dominate an X-like foot node and would therefore not be X-over-Y. Thus the dominance links are in effect redundant and void in the sense that adding them has no effect on the generated string set. It follows that G_2 is NP-hard. □

4 Restriction to lexicalized grammars

The grammar G_1 can be modified to get a lexicalized grammar G_3 that accepts a slightly different language than G_1 does. It can be shown that this language is NP-hard as well. The lexicalization constraint gives us NP-completeness. The proof is straightforward and is omitted here for reasons of space.

Since both restrictions just presented can be applied to G_1 at the same time and do not interact, there obtains the main result of this paper:

Corollary. There exist lexicalized nonlocal MCTAGs with dominance links that generate NP-complete languages. □

References

Becker, T., Joshi, A. K., and Rambow, O. (1991). Long-distance scrambling and Tree Adjoining Grammars. In *EACL*, pages 21–26.

Boullier, P. (1998). A generalization of mildly context-sensitive formalisms. In *Proceedings of the fourth international workshop on Tree Adjoining Grammars and Related Frameworks (TAG+4)*, pages 17–20. University of Pennsylvania.

Dahlhaus and Warmuth (1986). Membership for Growing Context-Sensitive Grammars is polynomial. *JCSS: Journal of Computer and System Sciences*, 33.

Joshi, A. K. (1985). Tree Adjoining Grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In Dowty, L. K. D. R. and Zwicky, A. M., editors, *Natural Language Parsing*. Cambridge University Press, Cambridge.

Joshi, A. K., Becker, T., and Rambow, O. C. (2002). Complexity of scrambling: A new twist to the competence-performance distinction. In Abeille, A. and Rambow, O. C., editors, *Tree-adjoining grammars*. Stanford: CSLI.

Joshi, A. K. and Schabes, Y. (1997). Tree-adjoining grammars. In *Handbook of formal languages and automata*. Springer, Berlin.

Rambow, O. and Satta, G. (1992). Formal properties of non-locality. In *1st Int. Workshop on Tree Adjoining Grammars. 1992*.

Rambow, O. C. (1994). *Formal and computational aspects of natural language syntax*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA.

Schabes, Y. (1990). *Mathematical and computational aspects of lexicalized grammars*. PhD thesis, University of Pennsylvania.

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

$G_2 = (NT, \Sigma, S, I, A)$ where

$$
\begin{aligned}
NT &= \{X, \overline{X}, Y, \overline{Y}, \hat{Y}\} \\
\Sigma &= \{a, b, x, y\} \\
I &= \{\alpha_{start}\} \\
A &= \{\beta_{create-triple}, \beta_{consume-y}, \beta_{consume-\hat{y}}, \beta_{fill-triple}, \beta_{close-triple}, \beta_{end}\}
\end{aligned}
$$

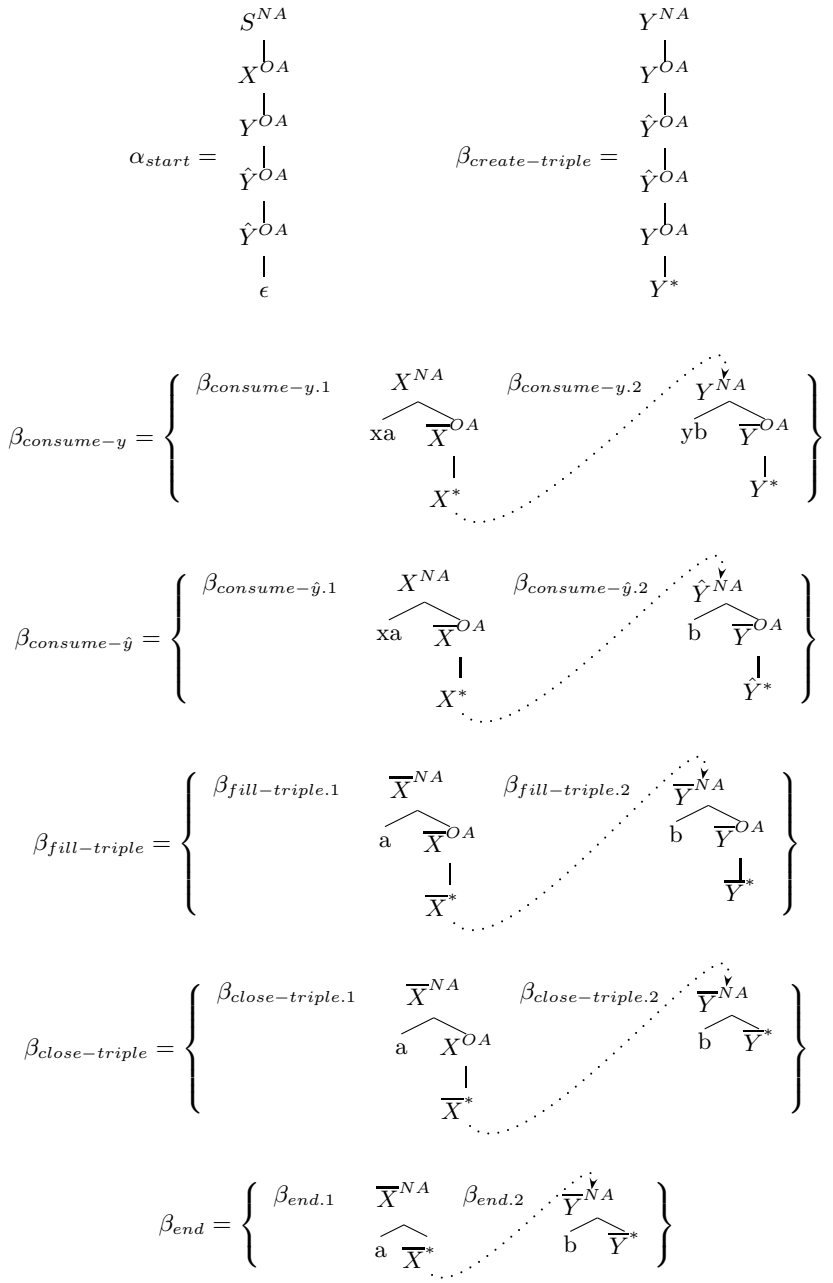

Figure 1: The MCTAG with dominance links G_2. (The MCTAG G_1 is obtained by ignoring the dominance links, represented as dashed lines.)