# Lexicalized nonlocal MCTAG with dominance links is NP-complete

Lucas Champollion

July 31, 2007

## 1 Introduction

(1) In this talk it is argued that nonlocal MCTAG is not a suitable candidate for the representation of natural language syntax, since (unless P=NP) it is not a polynomially parsable grammar class *even if we only consider linguistically relevant subclasses.*

(2) For this purpose, we take the following result from Rambow and Satta (1992); Rambow (1994):

**Theorem 1.** *There exist languages generated by nonlocal MCTAGs for which the membership problem is NP-hard.*

and show that a simple argument extends it to the following cases:

(i.) Restriction to languages generated by lexicalized languages

(ii.) Restriction to languages generated by grammars with dominance links

(3) These cases are linguistically more relevant than arbitrary MCTAGs because:

- It is generally accepted that only the lexicalized variants of TAGs are suitable candidates for encoding natural language.
- It is an open question whether lexicalization restricts the weak generative power of nonlocal MCTAG, therefore Theorem 1 does not trivially extend to lexicalized cases.
- The restrictions on nonlocal MCTAG (i.) and (ii.) above specifically have been proposed by Becker et al. (1991) for long-distance scrambling.
  - In German (and in many other SOV languages: Korean, Hindi, Japanese...) a constituent of an embedded clause may be moved from that clause into a higher clause.
  - More than one constituent may undergo movement into higher clauses. In this case, the scrambled constituents need not retain their original relative order to each other after scrambling.

## 2 Concepts: A Reminder

(4) A *Tree adjoining Grammar* (TAG) is a kind of tree-rewriting system. The operations, substitution and adjunction, are schematically shown in Figures 1 and 2.

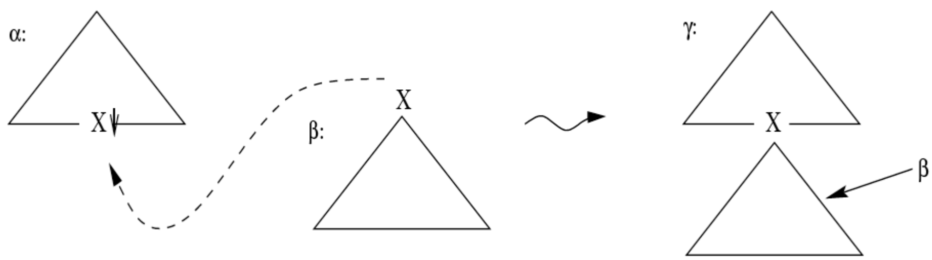


Figure 1: Substitution.

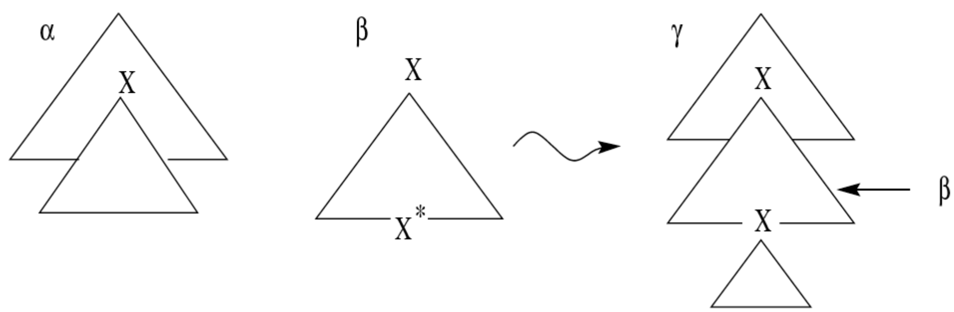


Figure 2: Adjoining.

(5) In a *multi-component TAG* (MCTAG) (Joshi, 1985; Weir, 1988), instead of auxiliary trees being single trees we have auxiliary sequences, where a sequence 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 sequence to different nodes.

- In a *tree-local* MCTAG, all trees from one sequence $S$ must be simultaneously adjoined into the same elementary tree $T$.
- In a *set-local* MCTAG, all trees from one sequence $S$ must be simultaneously adjoined into trees that all belong to the same sequence $S_2$.
- If this requirement is dropped altogether, we obtain *non-local* MCTAG.

(6) An MCTAG is *lexicalized* iff each elementary tree contains at least one terminal.

(7) In an MCTAG *with dominance links* each auxiliary sequence contains two trees. In the final derived tree, the foot node of one of the trees of each sequence has to dominate the root node of the other tree of the sequence. We may impose additional constraints between the two nodes such as c-command.

| | |
|---|---|
| 0 | Unrestricted |
| 1 | Context-sensitive |
| 1.2? | Non-local MCTAG, general case |
| **1.25** | **Lexicalized non-local MCTAG** |
| 1.5 | Set-local MCTAG, LCFRS, etc. (mildly context-sensitive) |
| 1.75 | TAG proper, tree-local MCTAG *(only weakly equivalent)* |
| 2 | Context free |
| 3 | Fnite state |
| 4 | Non-counting |
| 5 | Fnite |

Figure 3: The Chomsky hierarchy (extended, following Kornai).

## 2.1 Outline of the talk

- Sketch of the original proof by Dahlhaus and Warmuth (1986); Rambow and Satta (1992).
- Restriction to dominance links.
- Restriction to lexicalization.
- Discussion of the linguistic implications: What is the status of German scrambling?

# 3 NP-hardness of membership for nonlocal MCTAG

(8) Reduction from *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 $B$?

(9) 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.

(10) The MCTAG $G_1$ (see figure 4) 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 $G_1$.

(11) All derivations allowed by $G_1$ follow the same general pattern:

`step 1` Initialize the derivation by $\alpha_{start}$.

`step 2` Create $k$ triples by using $\beta_{create-triple}$ as many times as needed.

`step 3` 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.

`step 4` Optionally use $\beta_{fill-triple}$ to add an equal number of $a$'s and $b$'s to the left and right.

`step 5` Finally replace $\overline{X}$ by $a$ and $\overline{Y}$ by $b$. Either $\beta_{close-triple}$ or $\beta_{end}$ 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 `step 3` through `step 5` again with that $X$. If not, no $X$ can be introduced or the derivation would get stuck.

(12) The MCTAG $G_1$ in "multicomponent CFG" (USCG) format:

| | |
|---|---|
| start | $S \rightarrow XY\hat{Y}\hat{Y}$ |
| create-triple | $Y \rightarrow Y\hat{Y}\hat{Y}Y$ |
| consume-$y$ | $X \rightarrow xa\overline{X},\ Y \rightarrow yb\overline{Y}$ |
| consume-$\hat{y}$ | $X \rightarrow xa\overline{X},\ \hat{Y} \rightarrow b\overline{Y}$ |
| fill-triple | $\overline{X} \rightarrow a\overline{X},\ \overline{Y} \rightarrow b\overline{Y}$ |
| close-triple | $\overline{X} \rightarrow aX,\ \overline{Y} \rightarrow b$ |
| end | $\overline{X} \rightarrow a,\ \overline{Y} \rightarrow b$ |

(13) Sample derivation of the 3-partition instance: $\langle 4, 1, 3, 2, 5, 3; B = 9 \rangle$

| | | | | |
|---|---|---|---|---|
| | *init* | $S$ | | |
| `step 1` | start | $X$ | $Y\hat{Y}\hat{Y}$ | |
| `step 2` | create-triple | $X$ | $Y\hat{Y}\hat{Y}$ | $Y\hat{Y}\hat{Y}$ |
| `step 3` | consume-$y$ | $xa\overline{X}$ | $Y\hat{Y}\hat{Y}$ | $yb\overline{X}\hat{Y}\hat{Y}$ |
| `step 4` | fill-triple | $xaa\overline{X}$ | $Y\hat{Y}\hat{Y}$ | $ybb\overline{Y}\hat{Y}\hat{Y}$ |
| `step 4` | fill-triple | $xaaa\overline{X}$ | $Y\hat{Y}\hat{Y}$ | $ybbb\overline{Y}\hat{Y}\hat{Y}$ |
| `step 5` | close-triple | $xaaaaX$ | $Y\hat{Y}\hat{Y}$ | $ybbbb\hat{Y}\hat{Y}$ |
| `step 3` | consume-$\hat{y}$ | $xaaaa\ xa\overline{X}$ | $Yb\overline{Y}\hat{Y}$ | $ybbbb\hat{Y}\hat{Y}$ |
| `step 5` | close-triple | $xaaaa\ xaX$ | $Yb\hat{Y}$ | $ybbbb\hat{Y}\hat{Y}$ |
| `step 3` | consume-$\hat{y}$ | $xaaaa\ xa\ xa\overline{X}$ | $Yb\hat{Y}$ | $ybbbb\hat{Y}b\overline{Y}$ |
| `step 4` | fill-triple | $xaaaa\ xa\ xaa\overline{X}$ | $Yb\hat{Y}$ | $ybbbb\hat{Y}bb\overline{Y}$ |
| `step 5` | close-triple | $xaaaa\ xa\ xaaaX$ | $Yb\hat{Y}$ | $ybbbb\hat{Y}bbb$ |
| `step 3` | consume-$\hat{y}$ | $xaaaa\ xa\ xaaa\ xa\overline{X}$ | $Yb\hat{Y}$ | $ybbbbb\overline{Y}bbb$ |
| `step 5` | close-triple | $xaaaa\ xa\ xaaa\ xaaX$ | $Yb\hat{Y}$ | $ybbbbbbbbb$ |
| `step 3` | consume-$y$ | $xaaaa\ xa\ xaaa\ xaa\ xa\overline{X}$ | $yb\overline{Y}b\hat{Y}$ | $ybbbbbbbbb$ |
| `step 4` | fill-triple | $xaaaa\ xa\ xaaa\ xaa\ xaa\overline{X}$ | $ybb\overline{Y}b\hat{Y}$ | $ybbbbbbbbb$ |
| `step 4` | fill-triple | $xaaaa\ xa\ xaaa\ xaa\ xaaa\overline{X}$ | $ybbb\overline{Y}b\hat{Y}$ | $ybbbbbbbbb$ |
| `step 4` | fill-triple | $xaaaa\ xa\ xaaa\ xaa\ xaaaa\overline{X}$ | $ybbbb\overline{Y}b\hat{Y}$ | $ybbbbbbbbb$ |
| `step 5` | close-triple | $xaaaa\ xa\ xaaa\ xaa\ xaaaaaX$ | $ybbbbbb\hat{Y}$ | $ybbbbbbbbb$ |
| `step 3` | consume-$\hat{y}$ | $xaaaa\ xa\ xaaa\ xaa\ xaaaaa\ xa\overline{X}$ | $ybbbbbbb\overline{Y}$ | $ybbbbbbbbb$ |
| `step 4` | fill-triple | $xaaaa\ xa\ xaaa\ xaa\ xaaaaa\ xaa\overline{X}$ | $ybbbbbbbb\overline{Y}$ | $ybbbbbbbbb$ |
| `step 5` | end | $xaaaa\ xa\ xaaa\ xaa\ xaaaaa\ xaaa$ | $ybbbbbbbbb$ | $ybbbbbbbbb$ |

# 4 Restriction to dominance links

**Theorem 2.** *There exist languages generated by nonlocal MCTAGs* with dominance links *for which the membership problem is NP-hard.*

*Proof.* We simply add vacuous dominance links to all the tree sequences in $G_1$ in the manner shown in $G_2$ (Figure 5). None of these dominance links will ever rule out a derivation because they only require that some X-like symbol (i.e. $X, \overline{X}$) dominate some Y-like symbol (i.e. $Y, \overline{Y}, \hat{Y}$). But in fact even without the dominance links it can never be the case that a Y-like symbol dominates an X-like symbol. In every tree set in $G_1$, 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 tree derived by $G_1$ can be shown to have the property that all X-like symbols dominate all Y-like symbols. Therefore the links in $G_2$ are vacuous and the languages generated by $G_1$ and $G_2$ are identical. □

# 5 Restriction to lexicalization

**Theorem 3.** *There exist languages generated by nonlocal* lexicalized *MCTAGs (with or without dominance links) for which the membership problem is NP-hard.*

*Proof.* The grammar $G_1$ (or $G_2$) can also be modified to get a lexicalized grammar $G_3$ that accepts a slightly different language than $G_1$ and $G_2$ do (with an obvious polynomial-time mapping between the two). The resulting grammar $G_3$ is shown in Figure 6. □

# 6 NP-completeness

**Theorem 4.** *Any lexicalized nonlocal MCTAG is at most NP-complete.*

*Proof.* In a lexicalized grammar, every derivation step introduces terminals to the derivation. So it always takes at most $|w|$ steps to derive $w$. □

# 7 Implications

(14) Unless P=NP, MCTAG with dominance links cannot be parsed in polynomial time and is therefore outside LCFRS.

(15) The conjecture by Rambow (1994) that dominance links do not decrease the weak generative power of MCTAG is confirmed.

(16) The proposal by Becker et al. (1991) to model German scrambling by nonlocal MCTAG with dominance links is undermined.

(17) On the other side, it should be noted that there exist alternative views on the complexity of scrambling.

(18) Becker et al. (1991) assumed that any number n of verbal arguments can be scrambled at once and that all scrambling orders are possible.

(19) Recent results (Chen-Main and Joshi, 2007) about *tree-local MCTAG* indicate the following:

(a) for certain classes of scramblings (permutations) all patterns are possible for all n (number of arguments)

(b) for all other classes of scramblings (permutations) not all derivations are possible for $n > 3$.

(20) These combinatorial properties of tree-local MCTAG are relevant because judgments about scramblings beyond 4 arguments (some even at the level of 4 arguments) are impossible in general, although some special cases (such as no permutation or an end-around permutation, for example) are much easier to judge positively for all n.

(21) Thus, it may be that the only data that would discriminate between a polynomial-time and an NP-complete grammar class is unavailable for judgments, presumably due to processing costs.

# 8 Appendix

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

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

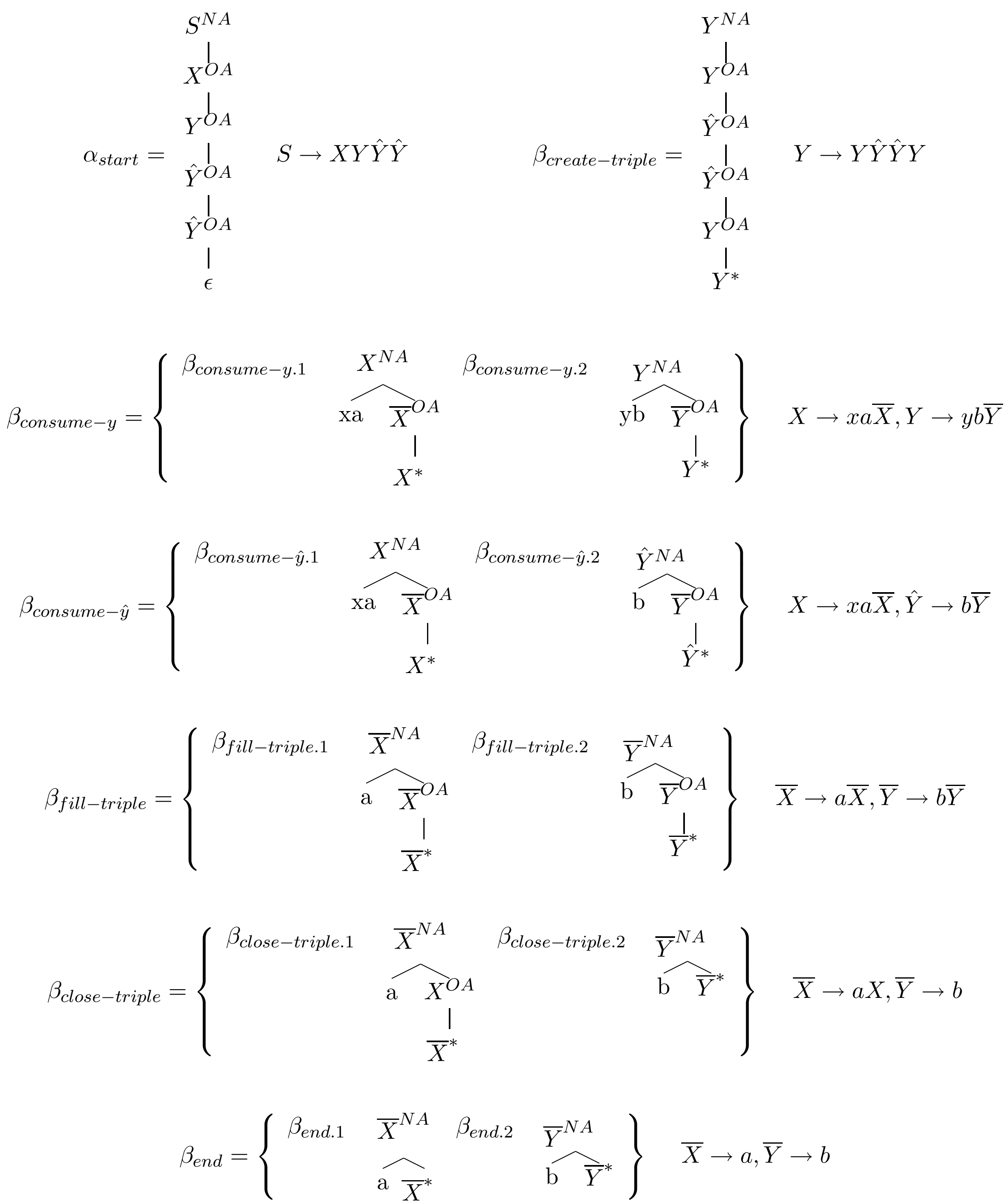


Figure 4: The MCTAG $G_1$ with its corresponding CFG rules.

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

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

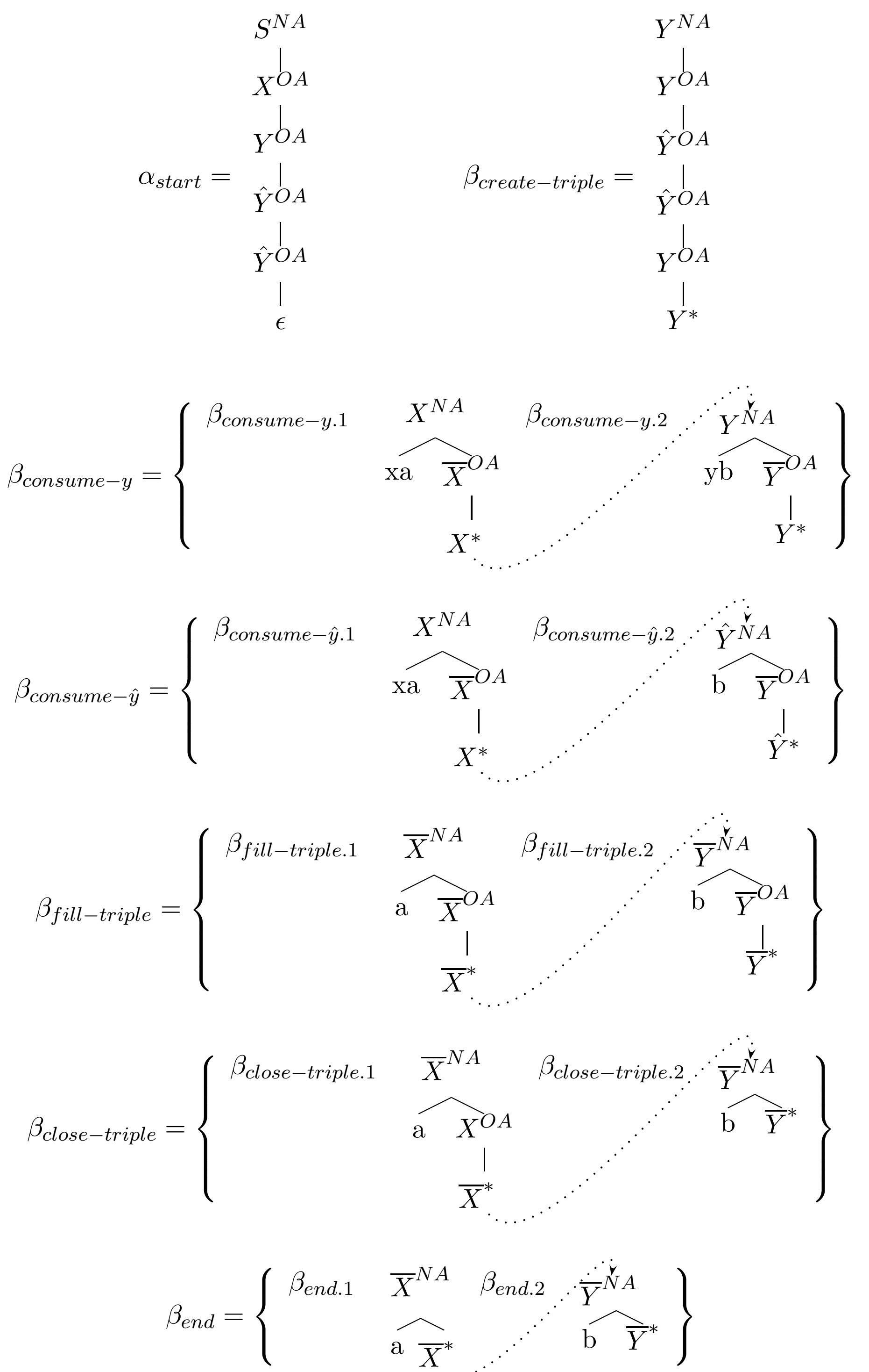

Figure 5: The MCTAG with dominance links $G_2$ (Identical to $G_1$ except for the dominance links.)

$G_3 = (NT, \Sigma, S, I, A)$ where

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

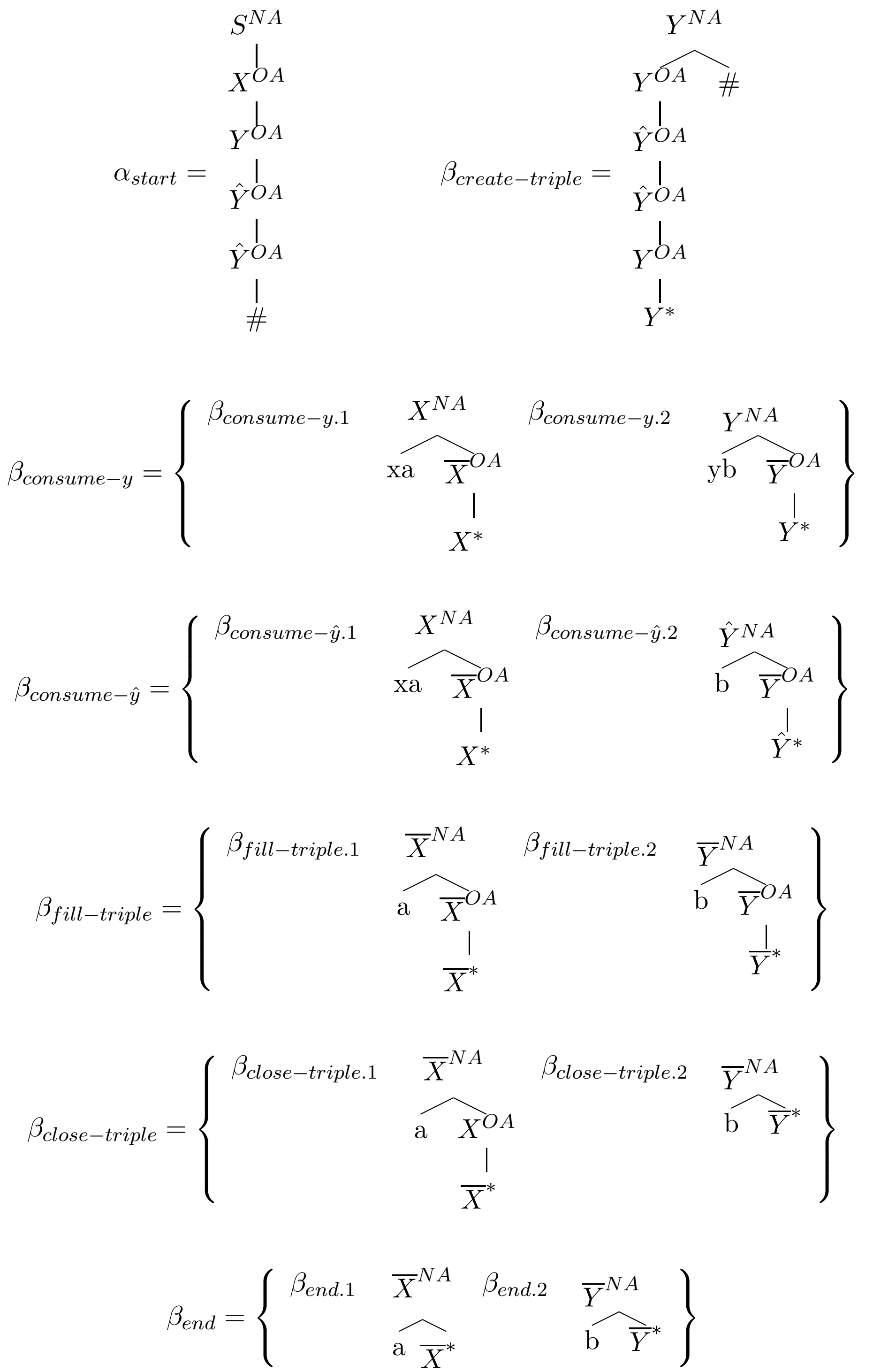


Figure 6: The lexicalized MCTAG $G_3$. (Identical to $G_1$ except that new terminals have been added to $\alpha_{start}$ and to $\beta_{create-triple}$.)

# References


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

Chen-Main, J. and Joshi, A. K. (2007). Some observations on a "graphical" model-theoretical approach and generative models. In *ESSLLI 2007 workshop on Model Theoretic Syntax*.

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.

Rambow, O. and Satta, G. (1992). Formal properties of non-locality. In *$1^{st}$ 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.

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