Prefix Lexicalization of Synchronous CFGs using Synchronous TAG: Supplementary Material

Logan Born and **Anoop Sarkar**
Simon Fraser University
School of Computing Science
{loborn,anoop}@sfu.ca

This is a supplementary document containing the complete proof for lemma 1. We repeat Figure 3 from the paper as figure 1 for ease of reference.

Lemma 1. *G_{XA} generates the language $L_{XA} = \{\langle u, v\rangle | \langle X\boxed{1}, A\boxed{1}\rangle \Rightarrow^*_{TTLD} \langle u, v\rangle\}$.*

Proof. We prove this lemma by induction over derivations of increasing length. We show first that every TTLD starting from $\langle X\boxed{1}, A\boxed{1}\rangle$ corresponds to a unique derivation in G_{XA}; we then show the other direction, that each derivation in G_{XA} corresponds to a TTLD starting from $\langle X\boxed{1}, A\boxed{1}\rangle$.

1 TTLD to STAG

We first consider the direction from TTLDs in G to derivations in G_{XA}. For the sake of brevity, the rest of this section uses TTLD as shorthand for "TTLD starting from $\langle X\boxed{1}, A\boxed{1}\rangle$". We show that the last n steps of every TTLD correspond to some derivation over n trees from G_{XA}. We show as well that whenever the derivation in G_{XA} is complete (there are no open substitution sites left) it generates the same string as the TTLD.

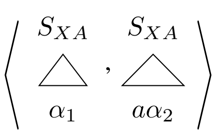
Base Cases As a base case, consider a TTLD of length 1, as in (1):

$$\langle X\boxed{1}, A\boxed{1}\rangle \Rightarrow \langle \alpha_1, a\alpha_2\rangle \quad (1)$$

where $a \in \Sigma$, and $\alpha_i \in (N \cup \Sigma)^*$. Such a derivation involves the application of one rule which must be of the form in (2):

$$\langle X \rightarrow \alpha_1, A \rightarrow a\alpha_2\rangle \quad (2)$$

By construction, we know that if such a rule exists in G, then G_{XA} must contain a corresponding tree pair of the shape depicted in Figure 1(a). This implies that the following is a valid derived tree in G_{XA}:



(3)

This derived tree produces the same string pair as the TTLD in (1). Thus we see that for every single-step TTLD in G there exists a (unique) derivation in G_{XA} which produces the same sentential form.

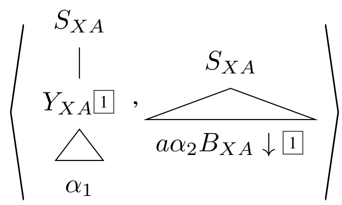
As a second base case, consider a TTLD of length > 1. This will be a derivation of the form in (4)

$$\langle X\boxed{1}, A\boxed{1}\rangle \Rightarrow^*_{TTLD} \langle uY\boxed{1}v, B\boxed{1}w\rangle \Rightarrow \langle u\alpha_1 v, a\alpha_2 w\rangle \quad (4)$$

where $Y, B \in N \setminus \{S\}$, $a \in \Sigma$, and $u, v, w, \alpha_i \in (N \cup \Sigma)^*$. Now the last step of this TTLD must involve the application of some rule of the form in (5)

$$\langle Y \rightarrow \alpha_1, B \rightarrow a\alpha_2\rangle \quad (5)$$

By construction, we know that if such a rule exists in G, then G_{XA} must contain a corresponding tree pair of the shape in Figure 1(b). This implies that the following is a valid derivation in G_{XA}:

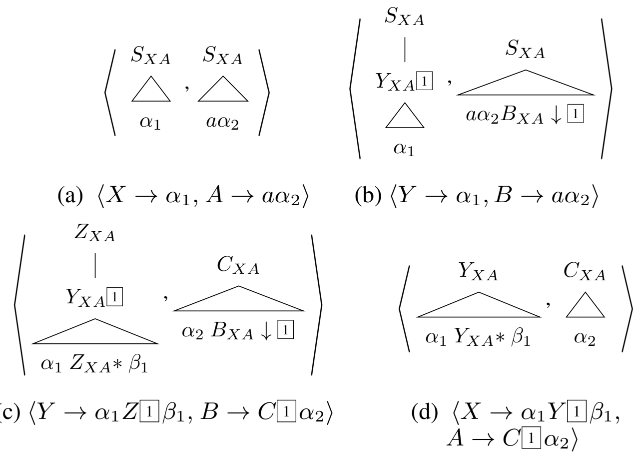


(6)

This is a derivation over a single tree pair; it is not a complete derivation, however, as there remains an open substitution site in the target-side tree. This derivation produces the sentential form $\langle \alpha_1, a\alpha_2 B_{XA}\rangle$, which is the same form produced by the last step of the TTLD in question, up to the addition of a B_{XA} in the target string. Finally, note that this derivation contains an open Y_{XA} adjunction site on the source side linked to an open B_{XA} substitution site on the target side.

Taken together, these base cases show the following:

- Every TTLD of length 1 has a corresponding derivation in G_{XA} which produces the same sentential form as that TTLD.
- For every TTLD of length > 1, the last step of that TTLD corresponds to some single-tree derivation in G_{XA}. This correspondence satisfies the following:
  - the last step of the TTLD produces the same sentential form as the derivation in G_{XA}, up to the addition of some nonterminal in the target string;



(a) $\langle X \to \alpha_1, A \to a\alpha_2 \rangle$ (b) $\langle Y \to \alpha_1, B \to a\alpha_2 \rangle$

(c) $\langle Y \to \alpha_1 Z\boxed{1}\beta_1, B \to C\boxed{1}\alpha_2 \rangle$ (d) $\langle X \to \alpha_1 Y\boxed{1}\beta_1,$ $A \to C\boxed{1}\alpha_2 \rangle$

Figure 1: Tree-pairs in G_{XA} and the rules in G from which they derive.

– if the last step of the TTLD involves overwriting some pair of nonterminals $\langle Y\boxed{1}, B\boxed{1}\rangle$, then the derivation in G_{XA} contains a Y_{XA} adjunction site in the source tree linked to a B_{XA} substitution site in the target tree.

Inductive Step Assume that the following inductive hypotheses are true for some n:

For every TTLD of length $> n$, the last n steps of that TTLD correspond to some derivation in G_{XA} over n trees. This correspondence satisfies the following:

- the last n steps of the TTLD produce the same sentential form as the derivation in G_{XA}, up to the addition of some nonterminal in the target string;
- if the nth-from-last step of the TTLD involves overwriting some pair of nonterminals $\langle Y\boxed{1}, B\boxed{1}\rangle$, then the derivation in G_{XA} contains an Y_{XA} adjunction site in the source tree linked to a B_A substitution site in the target tree.

Also, for every TTLD of length exactly n, that TTLD corresponds to some derivation in G_{XA} over n trees. This correspondence satisfies the following:

- the TTLD produces the same sentential form as the derivation in G_{XA}.
- the derivation in G_{XA} contains no open adjunction or substitution sites.

We now prove that if these hypotheses hold for some n, then they must also hold for $n+1$. There are two cases to consider: either a TTLD contains more than $n+1$ steps, or it contains exactly $n+1$ steps.

First Case: TTLD of length $> n+1$ Consider the last $n+1$ steps of such a TTLD, as shown in (7)

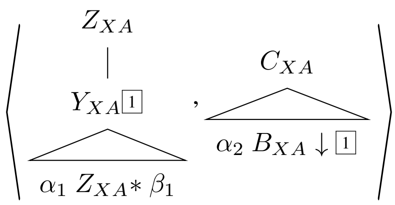
$$\langle Y\boxed{1}, B\boxed{1}\rangle \Rightarrow \langle \alpha_1 Z\boxed{1}\beta_1, C\boxed{1}\gamma_1\rangle \Rightarrow^*_{TTLD} \langle \alpha_1\alpha_2\beta_1, a\gamma_2\gamma_1\rangle \quad (7)$$

where $Y, Z, B, C \in N \setminus \{S\}$, $a \in \Sigma$, and $\alpha_i, \beta_i, \gamma_i \in (N \cup \Sigma)^*$. By the first inductive hypothesis, the last n of these steps correspond to some derivation over n trees in G_{XA}. Since the first of these n steps must involve rewriting the C which is at the left edge of the target string, the inductive hypothesis implies that the derivation in G_{XA} contains a C_{XA} substitution site linked to a Z_{XA} adjunction site. Furthermore, by the inductive hypothesis this derivation produces the same sentential form as the last n steps of the TTLD, up to the addition of a C_{XA} at the edge of the target string.

Now, from (7) we also see that the step $n+1$ operations before the end of the TTLD involves a rule of the form

$$\langle Y \to \alpha_1 Z\boxed{1}\beta_1, B \to C\boxed{1}\gamma_1\rangle \quad (8)$$

By construction, the existence of this rule in G implies that G_{XA} contains a tree pair of the shape in Figure 1(c), repeated here as (9)



(9)

This tree pair can be added to the n-tree derivation which the inductive hypothesis tells us must exist: the source tree can adjoin to the open Z_{XA} adjunction site, and the target tree can substitute into the C_{XA} substitution site.

The result will be a new $n+1$ tree derivation which satisfies the following:

- it produces the same sentential form as the last $n+1$ steps of the TTLD. This can be verified by observing that all adjunction sites in G_{XA} are near the root of the tree, so that when the new source tree adjoins it must necessarily wrap α_1 and β_1 to either side of the existing source string, to produce the required form; on the target side, the new tree will overwrite the C_{XA} node at the right edge of the string so that α_2 will also be in the correct position.

- it contains an open Y_{XA} adjunction site on the source side and a B_{XA} substitution site on the target side, as can be seen by inspection of (9)

Therefore we see that the first inductive hypothesis will also hold for a derivation of length $n+1$ given that it holds for a derivation of length n.

Second Case: TTLD of length $n+1$ Consider a completed TTLD of length $n+1$, as shown in (10)

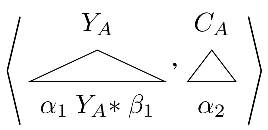
$$\langle X\boxed{1}, A\boxed{1}\rangle \Rightarrow \langle \alpha_1 Y\boxed{1}\beta_1, C\boxed{1}\gamma_1\rangle \Rightarrow^*_{TTLD} \langle \alpha_1\alpha_2\beta_1, a\gamma_2\gamma_1\rangle \tag{10}$$

where $Y, C \in N \setminus \{S\}$, $a \in \Sigma$, and $\alpha_i, \beta_i, \gamma_i \in (N \cup \Sigma)^*$. By the first inductive hypothesis, the last n steps of this TTLD correspond to some derivation over n trees in G_{XA}. Since the first of these n steps must involve rewriting the C which is at the left edge of the target string, the derivation in G_{XA} must contain a C_{XA} substitution site linked to a Y_{XA} adjunction site. Furthermore, this derivation must produce the same string as the last n steps of the TTLD, up to the addition of C_{XA} at the right edge of the target string.

Now, from (10) we also see that the first step of the derivation involves a rule of the form

$$\langle X \rightarrow \alpha_1 Y\boxed{1}\beta_1, A \rightarrow C\boxed{1}\gamma_1\rangle \tag{11}$$

By construction, the existence of this rule in G implies that G_{XA} contains a tree pair of the shape in Figure 1(d), repeated here as (12)



(12)

This tree pair can be added to the n-tree derivation which the inductive hypothesis tells us must exist: the source tree can adjoin to the open Y_{XA} adjunction site, and the target tree can substitute into the C_{XA} substitution site.

The result will be a new $n+1$ tree derivation which satisfies the following:

- it produces the same sentential form as the entire $n+1$ step TTLD. This can be verified by observing that all adjunction sites in G_{XA} are near the root of the tree, so that when the new source tree adjoins it will wrap α_1 and β_1 to either side of the existing source string to produce the required form; on the target side, the new tree will overwrite the C_A node at the right edge of the string so that α_2 will also be in the correct position.

- it is a completed derivation, as there are no open adjunction or substitution sites.

Therefore it follows that the second inductive hypothesis also holds for $n+1$ given that the first hypothesis holds for n.

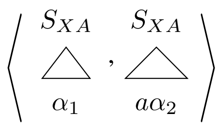
Conclusion Taken together, the preceding two cases show that there is a derivation in G_{XA} corresponding to every TTLD starting from $\langle X\boxed{1}, A\boxed{1}\rangle$. To obtain a one-to-one correspondence, we now prove the other direction, that for every derivation in G_{XA} there exists a corresponding TTLD in G.

2 STAG to TTLD

We now show that the first n steps of every derivation in G_{XA} correspond to the last n steps of a TTLD in G, and every complete derivation in G_{XA} corresponds to a TTLD starting from $\langle X\boxed{1}, A\boxed{1}\rangle$.

Preliminaries In TAG, derivations are generally assumed to be unordered, and all operations are taken to occur at once. In the case of a grammar like G_{XA}, however, we may talk about the "first" and "last" operations, because every tree has rank at most 1. Concretely, we shall say that the first tree pair in a derivation is the one rooted in the start symbol S_{XA}. Then the second tree pair in that derivation is the one which substitutes or adjoins to the first; the third tree pair substitutes or adjoins to the second; and so on.

Base Cases As a base case, consider a derivation in G_{XA} comprising a single tree pair of the shape given in Figure 1(a), repeated here:



(13)

where $a \in \Sigma$, and $\alpha_i \in (N \cup \Sigma)^*$. By construction, we know that this tree pair must have been added to G_{XA} on the basis of some rule in G. In particular, there must be a corresponding rule in G of the shape in (14)

$$\langle X \rightarrow \alpha_1, A \rightarrow a\alpha_2\rangle \tag{14}$$

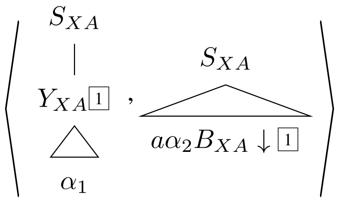
where $a \in \Sigma$, and $\alpha_i \in (N \cup \Sigma)^*$.

Using (14), we may construct a TTLD of length 1, shown in (15):

$$\langle X\boxed{1}, A\boxed{1}\rangle \Rightarrow \langle \alpha_1, a\alpha_2\rangle \tag{15}$$

This is a completed TTLD which generates the same string pair as the derivation in G_{XA}. Thus we see that for every completed single-tree derivation in G_{XA}, there exists a corresponding TTLD in G which produces the same string.

As a second base case, consider a derivation in G_{XA} comprising more than one tree pair. This derivation must start with some tree pair rooted in S_{XA}; furthermore, since it includes more than one tree pair in total, it cannot start with a pair of the shape in 1(a), because such a pair has no open substitution or adjunction sites. The only remaining possibility is for the derivation to start with a tree pair of the shape in 1(b), repeated below:



(16)

where $Y, B \in N \setminus \{S\}$, $a \in \Sigma$, and $\alpha_i \in (N \cup \Sigma)^*$. By construction, we know that this tree pair must have been added to G_A on the basis of some rule in G. In particular, there must be a corresponding rule in G of the shape in (17)

$$\langle Y \rightarrow \alpha_1, B \rightarrow a\alpha_2 \rangle \quad (17)$$

where $Y, B \in N \setminus \{S\}$, $a \in \Sigma$, and $\alpha_i \in (N \cup \Sigma)^*$.

Using (17), we may construct the derivation in (18):

$$\langle Y\boxed{1}, B\boxed{1} \rangle \Rightarrow \langle \alpha_1, a\alpha_2 \rangle \quad (18)$$

This is valid TTLD; furthermore this derivation produces the string pair $\langle \alpha_1, a\alpha_2 \rangle$, which is the same pair produced by the first tree in the derivation in G_{XA}, up to the removal of B_{XA} from the right edge of the target string. Note that (18) starts by rewriting the pair $\langle Y\boxed{1}, B\boxed{1} \rangle$, and that (16) likewise contains a Y_{XA} adjunction site linked to a B_{XA} substitution site.

Taken together, the two base cases show the following:

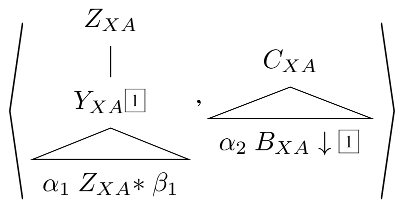
- Every completed, single-tree-pair derivation in G_{XA} has a corresponding TTLD in G which produces the same sentential form as that derivation.
- For every derivation in G_{XA} comprising more than one tree pair, the first tree pair in that derivation corresponds to the end of some TTLD in G. This correspondence satisfies the following:
  - the last step of the TTLD produces the same sentential form as the first tree pair of the derivation in G_{XA}, up to the removal of some nonterminal from the target string;
  - if the first tree pair in the derivation in G_{XA} contains a Y_{XA} adjunction site in the source tree linked to a B_{XA} substitution site in the target tree, then the last step of the TTLD involves overwriting the pair of nonterminals $\langle Y\boxed{1}, B\boxed{1} \rangle$.

Inductive Step Assume that the following inductive hypotheses are true for some n:

> For every derivation in G_{XA} comprising $> n$ tree pairs, the first n tree pairs in that derivation correspond to some TTLD in G involving n rule applications. This correspondence satisfies the following:
>
> - the first n tree pairs produce the same sentential form as the TTLD, up to the removal of some nonterminal from the right edge of the target string;
> - if the nth tree pair of the derivation in G_{XA} contains a Y_{XA} adjunction site in the source tree linked to a B_{XA} substitution site in the target tree, then the first step of the TTLD involves overwriting the pair of nonterminals $\langle Y\boxed{1}, B\boxed{1} \rangle$.
>
> Also, for every derivation in G_{XA} of length exactly n, that derivation corresponds to some TTLD involving n rule applications. This correspondence satisfies the following:
>
> - the TTLD produces the same sentential form as the derivation in G_{XA}.
> - the TTLD starts from $\langle X\boxed{1}, A\boxed{1} \rangle$.

We now prove that if these hypotheses hold for some n, then they must also hold for $n+1$. There are two cases to consider: either a derivation in G_{XA} involves more than $n+1$ tree pairs, or it involves exactly $n+1$ pairs.

First Case: $> n+1$ tree pairs Consider the $n+1$th tree pair in such a derivation. This must be of the shape in Figure 1(c), repeated below as (19). This is because this is the only kind of tree pair in G_{XA} which both (i) contains open substitution/adjunction sites to perpetuate the derivation (since we assume it is longer than $n+1$ operations) and (ii) is not rooted in S_{XA}, and is therefore able to appear in the middle of a derivation.



(19)

Since the $n+1$th pair must compose with the nth pair, the nth pair must contain an open adjunction site labeled Z_{XA} linked to a substitution site labeled C_{XA}, where Y_{XA} and C_{XA} are the nonterminals at the root of the $n+1$th pair's source and target trees respectively.

Furthermore, by the first inductive hypothesis, the first n tree pairs in this derivation must correspond to some n-step TTLD in G. Since the nth pair has open Z_{XA} and C_{XA} sites, we know by the same hypothesis that the corresponding TTLD starts from $\langle Z\boxed{1}, C\boxed{1} \rangle$, as in (20):

$$\langle Z\boxed{1}, C\boxed{1} \rangle \Rightarrow^*_{TTLD} \langle \alpha_2, a\gamma_2 \rangle \quad (20)$$

Now, by construction we know that if G_{XA} contains a tree pair of the shape in (19), then G must contain a production of the shape in (21):

$$\langle Y \rightarrow \alpha_1 Z\boxed{1}\beta_1, B \rightarrow C\boxed{1}\gamma_1 \rangle \quad (21)$$

By applying the rule in (21), followed by the rest of the derivation in (20), we obtain a new $n+1$-step TTLD shown in (22):

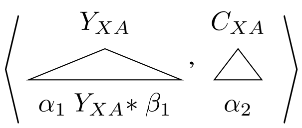
$$\langle Y\boxed{1}, B\boxed{1}\rangle \Rightarrow \langle \alpha_1 Z\boxed{1}\beta_1, C\boxed{1}\gamma_1\rangle \\ \Rightarrow^{*}_{TTLD} \langle \alpha_1\alpha_2\beta_1, a\gamma_2\gamma_1\rangle \quad (22)$$

The new TTLD in (22) satisfies the following:

- it produces the same sentential form as the first $n+1$ tree pairs of the derivation in G_{XA}, up to the removal of a nonterminal from the right edge of the target string. This can be verified by observing that prepending the new production to the existing TTLD wraps α_1 and β_1 around the existing source string in the same way that adjoining the $n+1$th source tree wraps α_1 and β_1 around the rest of the tree; on the target side, γ_2 is appended to the right edge in the same position that the $n+1$th target tree appends $\gamma_2 B_A$.
- it starts from the pair $\langle Y\boxed{1}, B\boxed{1}\rangle$, where Y_{XA} and B_{XA} are the labels on the adjunction and substitution sites in the $n+1$th tree pair.

Therefore we see that the first inductive hypothesis holds for derivations of length $n+1$ given that it holds for derivations of length n. In other words, we have so far proven that for every derivation in G_{XA}, every step up to the last step of the derivation corresponds to some TTLD in G. We now prove the final case, which shows that the last step of the derivations also correspond.

Second Case: exactly $n+1$ tree pairs Consider a completed derivation in G_{XA} containing $n+1$ tree pairs. The last tree pair must be of the shape in Figure 1(d), repeated below as (23), because this is the only tree pair which can compose with a derivation without introducing any new adjunction or substitution sites.



(23)

Since the $n+1$th tree pair must compose with the nth pair, the nth pair must contain an open adjunction site labeled Y_{XA} linked to a substitution site labeled C_{XA}, where Y_{XA} and C_{XA} are the nonterminals at the root of the $n+1$th pair's source and target trees respectively.

Furthermore, by the first inductive hypothesis, the first n tree pairs in this derivation must correspond to some n-step TTLD in G. Since the nth pair has open Y_{XA} and C_{XA} sites, we know by the same hypothesis that the corresponding TTLD starts from $\langle Y\boxed{1}, C\boxed{1}\rangle$, as in (24):

$$\langle Y\boxed{1}, C\boxed{1}\rangle \Rightarrow^{*}_{TTLD} \langle \alpha_2, a\gamma_2\rangle \quad (24)$$

Now, by construction we know that if the $n+1$th tree pair is of the shape in (23), then G must contain a production of the shape in (25):

$$\langle X \rightarrow \alpha_1 Y\boxed{1}\beta_1, A \rightarrow C\boxed{1}\gamma_1\rangle \quad (25)$$

By applying the rule in (25), followed by the rest of the derivation in (24), we obtain a new $n+1$-step TTLD shown in (26):

$$\langle X\boxed{1}, A\boxed{1}\rangle \Rightarrow \langle \alpha_1 Y\boxed{1}\beta_1, C\boxed{1}\gamma_1\rangle \\ \Rightarrow^{*}_{TTLD} \langle \alpha_1\alpha_2\beta_1, a\gamma_2\gamma_1\rangle \quad (26)$$

The new TTLD in (26) satisfies the following:

- it produces the same sentential form as the first $n+1$ tree pairs of the derivation in G_{XA}. This can be verified by observing that prepending the new production to the existing TTLD wraps α_1 and β_1 around the existing source string in the same way that adjoining the $n+1$th source tree wraps α_1 and β_1 around the rest of the tree; on the target side, γ_2 is appended to the right edge in the same position that the $n+1$th target tree appends γ_2.
- it starts from the pair $\langle X\boxed{1}, A\boxed{1}\rangle$.

Therefore we see that the second inductive hypothesis holds for derivations of length $n+1$ given that the first hypothesis holds for derivations of length n.

Conclusion Taken together, the preceding two cases show that there is a TTLD in G corresponding to every derivation in G_{XA}. Furthermore every completed derivation in G_{XA} corresponds to a TTLD which starts from the pair $\langle X\boxed{1}, A\boxed{1}\rangle$.

Combining the results in both of the preceding sections, we see that there is a one-to-one correspondence between completed derivations in G_{XA} and TTLDs in G which start from $\langle X\boxed{1}, A\boxed{1}\rangle$. By extension, we have shown that G_{XA} generates precisely the language $L_{XA} = \{\langle u, v\rangle | \langle X\boxed{1}, A\boxed{1}\rangle \Rightarrow^{*}_{TTLD} \langle u, v\rangle\}$.

□