Linking Theorems for Tree Transducers

Zoltán Fülöp[a,1,2], Andreas Maletti[b,2,*]

[a] *Department of Foundations of Computer Science, University of Szeged*
Árpád tér 2, H-6720 Szeged, Hungary
[b] *Institute of Computer Science, Universität Leipzig*
Augustusplatz 10–11, 04109 Leipzig, Germany

Abstract

Linear extended multi bottom-up tree transducers are presented in the framework of synchronous grammars, in which the input and the output tree develop in parallel by rewriting linked nonterminals (or states). These links are typically transient and disappear once the linked nonterminals are rewritten. They are promoted to primary objects here, preserved in the semantics, and carefully studied. It is demonstrated that the links computed during the derivation of an input and output tree pair are hierarchically organized and that the distance between (input and output) link targets is bounded. Based on these properties, two linking theorems are developed that postulate the existence of certain natural links in each derivation for a given input and output tree pair. These linking theorems allow easy, high-level proofs that certain tree translations cannot be implemented by (compositions of) linear extended multi bottom-up tree transducers.

1. Introduction

The notion of a multi bottom-up tree transducer was originally introduced and studied in [4, 26], albeit under different names. The deterministic variant was rediscovered in [13, 14], where the name "multi bottom-up tree transducer" was coined. Quite recently, it was established [11, 28] that the (weighted) linear extended variant has very nice algorithmic properties. It was thus further developed into a formal model for tree-to-tree translation [29, 31], which is a sub-discipline in syntax-based statistical machine translation [23]. An open-source implementation of a statistical machine translation system based on shallow linear extended multi bottom-up tree transducers [6] inside the Moses framework [24] is available and has been evaluated on an English-to-German translation task.

Here we consider linear extended multi bottom-up tree transducers (for short: MBOT) and present them in the form of synchronous grammars [7]. In such grammars, the nonterminals (or states) occurring in sentential forms are linked, and the linked nonterminals are replaced at the same time. Consequently, productions (or rules) of synchronous grammars often have at least two components: the input side and the output side. An MBOT rule might have even more than two components because it contains a vector of output trees. More formally, an MBOT is a finite-state tree transducer, in which the rules are of the form $\langle \ell, q, \vec{r} \rangle$, where the left-hand side ℓ is a tree that is linear in the states (i.e., each state can occur at most once), q is a state, and the right-hand side $\vec{r}$ is a vector of trees, in which states can also occur. It is required that all states that occur in $\vec{r}$ also appear in ℓ. In contrast to other presentations [11] we do not use ranked alphabets — neither for the input and output symbols nor for the states. It is easy to see that our

*Corresponding author

[1] Supported by the program TÁMOP-424B / 2-11 / 1-2012-0001, grant B2 / 2R / 3350.

[2] Supported by the German Research Foundation (DFG) grant MA / 4959 / 1-1 and the German Academic Exchange Service (DAAD) and Hungarian Scholarship Board Office (MÖB) exchange project "Theory and Applications of Automata" (grant 5567).

(rank-free) formalization (syntactically) includes all ranked versions including those that permit different ranked alphabets for the input and output symbols. Our model thus becomes slightly more powerful than traditional MBOT since it allows the same symbol to occur with different ranks in the input or output trees.

The semantics of our MBOT is defined by means of synchronous rewriting or, more generally, with the help of a derivation relation over sentential forms. In the synchronous rewriting approach, several parts of the sentential form develop (via the rules) at the same time. Typically, the left-hand side of the rule contributes to the input tree of the sentential form and the right-hand side contributes at the same time to the output tree of the sentential form. For MBOT, the right-hand side consists of a vector of trees, so it can act simultaneously at several positions in the output tree. The input and output positions that are supposed to develop in parallel are recorded by links (v, w), which (in our case) relate a position v in the input tree to a position w in the output tree. Consequently, a form of an MBOT is a tuple $\langle \xi, A, I, \zeta \rangle$ consisting of a (partial) input tree ξ, two sets $A, I \subseteq \mathrm{pos}(\xi) \times \mathrm{pos}(\zeta)$ of links, and a (partial) output tree ζ. The elements of A and I are called active and inactive links, respectively. The active links fulfill the already mentioned purpose of recording which parts are supposed to develop in parallel, whereas inactive links simply record all links that have been active at some point during the derivation. In this way, we preserve all links and can later argue about their structure, which will allow us to prove properties about MBOT.

On these forms we now define our derivation steps. A form $\langle \xi, A, I, \zeta \rangle$ derives a form $\langle \xi', A', I', \zeta' \rangle$, written $\langle \xi, A, I, \zeta \rangle \Rightarrow \langle \xi', A', I', \zeta' \rangle$, if we can select a rule $\langle \ell, q, \vec{r} \rangle$ and an occurrence $v \in \mathrm{pos}(\xi)$ of q in the input tree ξ such that $\vec{r}$ has as many components as there are output positions $A(v) = \{w \mid (v, w) \in A\}$ that are actively linked to v and

- ξ' is obtained from ξ by replacing the occurrence v of q by the tree ℓ,
- A' is obtained from A by removing the used links $\{(v, w) \mid w \in A(v)\}$ and adding the links induced by the rule $\langle \ell, q, \vec{r} \rangle$,
- I' is obtained from I by simply adding the used links $\{(v, w) \mid w \in A(v)\}$, and
- ζ' is obtained from ζ by replacing the linked subtrees $A(v)$ in ζ by $\vec{r}$ (in lexicographic order).

As usual, we apply derivation steps until no active links and no occurrences of states remain. The initial sentential form consists simply of two actively linked initial states (and no disabled links). Since we are interested in the links encountered during the derivation, the set of computed dependencies consists of all $\langle t, I, u \rangle$ such that $\langle q_0, \{(\varepsilon, \varepsilon)\}, \emptyset, q_0 \rangle \Rightarrow^* \langle t, \emptyset, I, u \rangle$, where q_0 is an initial state and t and u are trees, in which no state occurs.

Making additional information from the derivation process (like the links) visible in the computed output has been explored already. In particular, the origin [34, 35], which associates to output positions the input position from which this output fragment was created, has been intensively studied. For example, in [12] origin information is used to characterize those macro tree transducers that are MSO definable, and in [25] it is used to get a Myhill-Nerode characterization of deterministic top-down tree transducers. Information on the type of transition used (in a push-down automaton) yields the visibly push-down languages [2], and in the recent work [5] origin information is built into the semantics of a string transducer. However, for us the links are purely a tool, so the computed tree relation is obtained from the computed dependencies simply by projecting onto the input and output trees (i.e., removing the links) as usual. This computed tree relation coincides with the tree relation computed by means of other semantics [4, 11, 13].

Our goal is to provide generic linking theorems (see Theorems 5 and 6), which given a tree relation with particular properties (essentially it must contain a specific tree relation) predict certain natural links that must be present in a dependency containing a specific tree pair. Roughly speaking, the linking theorems will establish that whenever we preserve an input subtree in the output (i.e., we copy this part of the input tree verbatim to the output), then these occurrences must be linked. This conventional and intuitive wisdom had to be essentially reproved for each particular tree relation under investigation because the typical arguments used (e.g., the fooling technique) require negative information (i.e., information about tree pairs that are *not* in the desired tree relation), which often means that the proof cannot be reused in similar scenarios. Our linking theorems only use positive information (i.e., only the knowledge that certain tree pairs are in the tree relation), so they readily transfer to similar scenarios. We can then use these established links and general properties of dependencies computed by MBOT to show very easily that certain (classic) tree relations cannot be computed by (compositions of) certain subclasses of MBOT.

Before we start with the investigation of the properties of dependencies computed by MBOT, we show an example of a proof utilizing the classic fooling technique. This example proof from [33] shall demonstrate how the traditional proof technique works, so that it can be compared to our later solution using the linking theorems. However, for the linking theorems to be useful, we first need to establish some basic properties that we can use to reason with links. It turns out that the links in each dependency are organized hierarchically [25, 30]. More precisely, a set L of links is input hierarchical if for all $(v_1, w_1), (v_2, w_2) \in L$ the condition $v_1 \prec v_2$ implies that $w_2 \not\preceq w_1$ and that there exists a $(v_1, w_1') \in L$ such that $w_1' \preceq w_2$, where $\preceq$ is the prefix order. Moreover, it is strictly input hierarchical if $v_1 \prec v_2$ implies $w_1 \preceq w_2$, and $v_1 = v_2$ implies $w_1 \preceq w_2$ or $w_2 \preceq w_1$. Trivially, if L is strictly input hierarchical, then it is also input hierarchical. Finally, L is (strictly) output hierarchical if L^{-1} is (strictly) input hierarchical. We prove that the links in the dependencies computed by an MBOT are always input hierarchical and strictly output hierarchical (see Theorem 2). If the vector $\vec{r}$ has at most one component for every rule $\langle \ell, q, \vec{r} \rangle$ of an MBOT M, then M is actually a (linear) extended top-down tree transducer with regular look-ahead [33] (for short: XTOP$^{\mathrm{R}}$). The links in the dependencies computed by an XTOP$^{\mathrm{R}}$ are even strictly input hierarchical (see Corollary 1). In addition, the shape of the MBOT rules and the derivation process guarantee that there cannot be large "gaps" between two "adjacent" link positions (i.e., positions that are the source or target of a link). This property is again true for both the input as well as the output side of each dependency computed by an MBOT. More precisely, for every MBOT, there is an integer b that limits the distance between input link positions. For the output side, even the stronger statement requiring a link position every b positions is true along each path (see Theorem 3). If the vector $\vec{r}$ has exactly one component for every rule $\langle \ell, q, \vec{r} \rangle$ of an MBOT M, then M is actually a (linear) nondeleting extended top-down tree transducer [33] (for short: n-XTOP), for which the stricter distance property is also true for the input side (see Corollary 2).

Next, we prove our two linking theorems for ε-free MBOT, which are MBOT in which no rule has a left-hand side that is just a state. The first linking theorem (see Theorem 5) concerns arbitrary compositions of ε-free XTOP$^{\mathrm{R}}$, whereas the second linking theorem concerns a single ε-free MBOT (see Theorem 6). In both cases, we assume that the computed tree relation contains a sub-relation that is obtained by plugging trees from a simple, yet infinite tree language into an input-output context pair. Finally, we demonstrate how to apply these linking theorems in Section 7. In particular, we show in Theorem 7 and in a reproof of Theorem 1 that the counterexample tree relations of [4] and [33] cannot be computed by any ε-free XTOP$^{\mathrm{R}}$. Additionally, the linking theorems have been used in [16, 32] and the main theorems of [32] are proved in detail here. More precisely, Theorem 8 shows that the inverse of abstract topicalization [1, 8, 20] cannot be computed by any ε-free MBOT, and in addition, Theorem 9 shows that abstract topicalization itself cannot be computed by any composition of ε-free XTOP$^{\mathrm{R}}$.

2. Preliminaries

The sets of all nonnegative integers and all positive integers are $\mathbb{N}$ and $\mathbb{N}_+$, respectively (i.e., $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$). We let $[k] = \{i \in \mathbb{N}_+ \mid i \leq k\}$ for all $k \in \mathbb{N}$, which yields $[0] = \emptyset$. A relation ρ from a set S to a set T is a subset $\rho \subseteq S \times T$. Given a relation $\rho \subseteq S \times T$ and $S' \subseteq S$, the inverse $\rho^{-1} \subseteq T \times S$ is the relation $\rho^{-1} = \{(t, s) \mid (s, t) \in \rho\}$, and the image $\rho(S') \subseteq T$ of S' via ρ is $\rho(S') = \{t \mid \exists s \in S' \colon (s, t) \in \rho\}$. For every $s \in S$, we let $\rho(s) = \rho(\{s\})$. The composition $\rho \,;\, \tau$ of ρ with the relation $\tau \subseteq T \times U$ is

$$\rho \,;\, \tau = \{(s, u) \in S \times U \mid \rho(s) \cap \tau^{-1}(u) \neq \emptyset\} \ .$$

The set of all finite words (or sequences) over S is S^*, where $\varepsilon \in S^*$ is the empty word. The concatenation of the words $v, w \in S^*$ is $v.w$ or simply vw. The length of a word $w \in S^*$ is denoted by $|w|$. Given a word (or sequence) $w \in \Sigma^*$ of length $k = |w|$ and $i \in [k]$, we write w_i for the i-th letter (or component) in w. An alphabet Σ is a nonempty and finite set, of which the elements are called symbols. In the following, let Σ be an alphabet and S be a set with $\Sigma \cap S = \emptyset$.

The set $T_\Sigma(S)$ of Σ-trees indexed by S is the smallest set T such that $S \subseteq T$ and $\sigma(t_1, \ldots, t_k) \in T$ for every $k \in \mathbb{N}$, $\sigma \in \Sigma$, and $t_1, \ldots, t_k \in T$. Since $\Sigma \cap S = \emptyset$, we can safely write α instead of $\alpha()$

for every $\alpha \in \Sigma$. Moreover, for all $t \in T_\Sigma(S)$, $\gamma \in \Sigma$, and $n \in \mathbb{N}$, we denote $\gamma(\gamma(\cdots\gamma(t)\cdots))$ containing n occurrences of γ in the abbreviated list of γ-symbols by $\gamma^n(t)$, and we let $T_\Sigma = T_\Sigma(\emptyset)$. The set $\mathrm{pos}(t) \subseteq \mathbb{N}_+^*$ of positions of $t \in T_\Sigma(S)$ is inductively defined by $\mathrm{pos}(s) = \{\varepsilon\}$ for every $s \in S$ and $\mathrm{pos}(\sigma(t_1, \ldots, t_k)) = \{\varepsilon\} \cup \{i.v \mid i \in [k], v \in \mathrm{pos}(t_i)\}$ for every $k \in \mathbb{N}$, $\sigma \in \Sigma$, and $t_1, \ldots, t_k \in T_\Sigma(S)$. We denote the usual linear lexicographic order on $\mathbb{N}_+^*$ by $\leq$, and the usual partial prefix order on $\mathbb{N}_+^*$ by $\preceq$. Given a finite set $P \subseteq \mathbb{N}_+^*$ of positions, we let $\mathrm{seq}(P) = (w_1, \ldots, w_k)$ be the sequence of the positions of P in lexicographic order; i.e., $P = \{w_1, \ldots, w_k\}$ and $w_1 < \cdots < w_k$. The size $|t| \in \mathbb{N}_+$ and the height $\mathrm{ht}(t) \in \mathbb{N}$ of t are $|t| = |\mathrm{pos}(t)|$ and $\mathrm{ht}(t) = \max\{|w| \mid w \in \mathrm{pos}(t)\}$, respectively.

Let $t \in T_\Sigma(S)$ and $w \in \mathrm{pos}(t)$. The label of t at w is $t(w) \in \Sigma \cup S$, and the w-rooted subtree of t is $t|_w \in T_\Sigma(S)$. Formally, $s(\varepsilon) = s|_\varepsilon = s$ for every $s \in S$, and for every $k \in \mathbb{N}$, $\sigma \in \Sigma$, and $t_1, \ldots, t_k \in T_\Sigma(S)$

$$\big(\sigma(t_1, \ldots, t_k)\big)(w) = \begin{cases} \sigma & \text{if } w = \varepsilon \\ t_i(v) & \text{if } w = i.v \text{ with } i \in [k] \text{ and } v \in \mathrm{pos}(t_i) \end{cases}$$

$$\sigma(t_1, \ldots, t_k)|_w = \begin{cases} \sigma(t_1, \ldots, t_k) & \text{if } w = \varepsilon \\ t_i|_v & \text{if } w = i.v \text{ with } i \in [k] \text{ and } v \in \mathrm{pos}(t_i) \end{cases} .$$

For every $S' \subseteq S$, we let $\mathrm{pos}_{S'}(t) = \{w \in \mathrm{pos}(t) \mid t(w) \in S'\}$ be those leaf positions whose label is in S', and $\mathrm{pos}_s(t) = \mathrm{pos}_{\{s\}}(t)$ for every $s \in S$. If $|\mathrm{pos}_s(t)| \leq 1$ for every $s \in S$, then the tree t is linear, and we denote the set of all linear trees of $T_\Sigma(S)$ by $T_\Sigma^{\mathrm{lin}}(S)$. Moreover, we let $\mathrm{idx}(t) = \{s \in S \mid \mathrm{pos}_s(t) \neq \emptyset\}$ be the indices that occur in t. For all $u \in T_\Sigma(S)$, the expression $t[u]_w$ denotes the tree that is obtained from t by replacing the subtree $t|_w$ at w by u. Formally, $s[u]_\varepsilon = u$ for all $s \in S$ and

$$\sigma(t_1, \ldots, t_k)[u]_w = \begin{cases} u & \text{if } w = \varepsilon \\ \sigma(t_1, \ldots, t_{i-1}, t_i[u]_v, t_{i+1}, \ldots, t_k) & \text{if } w = i.v \text{ with } i \in [k] \text{ and } v \in \mathrm{pos}(t_i) \end{cases}$$

for all $k \in \mathbb{N}$, $\sigma \in \Sigma$, and $t_1, \ldots, t_k \in T_\Sigma(S)$. For every nonnegative integer $n \in \mathbb{N}$, we extend this notation to sequences $\vec{u} = (u_1, \ldots, u_n)$ of trees $u_1, \ldots, u_n \in T_\Sigma(S)$ and sequences $\vec{w} = (w_1, \ldots, w_n)$ of positions $w_1, \ldots, w_n \in \mathrm{pos}(t)$ such that the positions of $\vec{w}$ are pairwise incomparable with respect to $\prec$ (i.e., $w_i \not\preceq w_j$ for all $i, j \in [n]$ with $i \neq j$).[3] Then $t[\vec{u}]_{\vec{w}}$ denotes the tree obtained from t by replacing in parallel, for every $i \in [n]$, the subtree $t|_{w_i}$ at w_i by u_i. Formally, $t[\vec{u}]_{\vec{w}} = (\cdots(t[u_1]_{w_1})\cdots)[u_n]_{w_n}$.

Finally, let us recall contexts. We reserve the set $X = \{x_i \mid i \in \mathbb{N}_+\}$ of special symbols, which are called variables. For every $n \in \mathbb{N}$, we let $X_n = \{x_i \mid i \in [n]\}$. A tree $t \in T_\Sigma^{\mathrm{lin}}(X_n)$ is an n-context over Σ if $\mathrm{idx}(t) = X_n$. Alternatively, we can require that the set $\mathrm{pos}_{x_i}(t)$ is a singleton for every $i \in [n]$. The set of all n-contexts over Σ is denoted by $C_\Sigma(X_n)$. Let $c \in C_\Sigma(X_n)$. For all $i \in [n]$ we identify $\mathrm{pos}_{x_i}(c)$ with its (unique) element. Given $u_1, \ldots, u_n \in T_\Sigma$, we write $c[u_1, \ldots, u_n]$ for $c[\vec{u}]_{\vec{w}}$, where $\vec{u} = (u_1, \ldots, u_n)$ and $\vec{w} = (\mathrm{pos}_{x_1}(c), \ldots, \mathrm{pos}_{x_n}(c))$. A more detailed introduction to trees and automaton models working on trees can be found in [17, 18].

3. The transformational model

We select a variant of the (linear extended) multi bottom-up tree transducer (for short: MBOT), which was introduced and investigated in [4, 11, 13, 14, 26, 28, 29]. Roughly speaking, our version of MBOT, which is slightly more expressive than the traditional linear extended MBOT[4], is a synchronous grammar formalism, in which the output side might be discontinuous. As in all synchronous grammars [7] the input and the output trees develop (via the rules) at the same time, so each rule specifies part of the input and part of the

[3] This incomparability is needed to ensure that the parallel substitutions do not affect each other.

[4] The additional expressive power comes from the possibility to use the same symbol with different ranks. For example, our model can use a symbol in the output tree as a binary, unary, and even nullary symbol, whereas traditional models only allow one rank for each symbol.

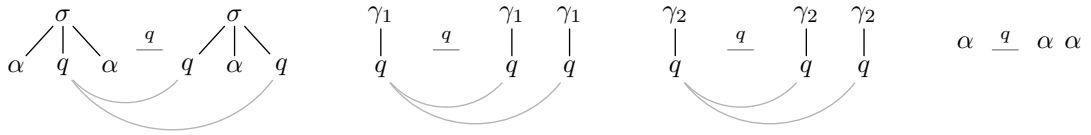

Figure 1: Rules of the MBOT M_{ex} in Example 1.

output tree. Since MBOT offer discontinuity on the output side, we can in fact specify several parts of the output tree in each rule. Note that we decided to use only one alphabet for both the input and the output symbols. Any rule of a traditional linear extended MBOT using input symbols from Σ and output symbols from Δ is a valid rule for our MBOT using the set $\Sigma \cup \Delta$ of symbols.[5]

Definition 1 (cf. [29]) A *multi bottom-up tree transducer* (for short: MBOT) is a tuple $M = (Q, \Sigma, Q_0, R)$, where

- Q is the alphabet of *states* and $Q_0 \subseteq Q$ contains the *initial states*,
- Σ is the alphabet of *input* and *output symbols* such that $\Sigma \cap Q = \emptyset$, and
- $R \subseteq T^{\text{lin}}_{\Sigma}(Q) \times Q \times T_{\Sigma}(Q)^*$ is the nonempty, finite set of *rules* such that $\bigcup_{i=1}^{n} \operatorname{idx}(r_i) \subseteq \operatorname{idx}(\ell)$ for all $\langle \ell, q, (r_1, \dots, r_n) \rangle \in R$.

If additionally $|\vec{r}| \leq 1$ for all $\langle \ell, q, \vec{r} \rangle \in R$, then M is a (linear) *extended top-down tree transducer with regular look-ahead* [3, 9, 10, 21, 22, 27, 33] (for short: XTOP$^{\text{R}}$), and if additionally $|\vec{r}| = 1$ for all $\langle \ell, q, \vec{r} \rangle \in R$, then M is a (linear) *nondeleting extended top-down tree transducer* (for short: n-XTOP). Finally, the MBOT M is *ε-free* if $\ell \notin Q$ for all $\langle \ell, q, \vec{r} \rangle \in R$. ◻

For the remaining discussion, let $M = (Q, \Sigma, Q_0, R)$ be an MBOT. As usual, we call ℓ and $\vec{r}$ of a rule $\langle \ell, q, \vec{r} \rangle \in R$ the *left-* and *right-hand side* of the rule, respectively. We also write $\ell \xrightarrow{q} \vec{r}$ instead of $\langle \ell, q, \vec{r} \rangle$.

Example 1 Let us consider the MBOT $M_{\text{ex}} = (\{q\}, \Sigma, \{q\}, R)$ with $\Sigma = \{\sigma, \gamma_1, \gamma_2, \alpha\}$ and

$$R = \big\{ \sigma(\alpha, q, \alpha) \xrightarrow{q} \sigma(q, \alpha, q), \quad \gamma_1(q) \xrightarrow{q} \gamma_1(q) \,.\, \gamma_1(q), \quad \gamma_2(q) \xrightarrow{q} \gamma_2(q) \,.\, \gamma_2(q), \quad \alpha \xrightarrow{q} \alpha \,.\, \alpha \big\} \ .$$

In Figure 1 we display those rules, where the gray splines show that each state occurring in the right-hand side also occurs in the left-hand side. It can easily be verified that M_{ex} is an ε-free MBOT, but neither an XTOP$^{\text{R}}$ nor a n-XTOP. ◻

Next, we prepare the definition of the semantics of an MBOT. In this definition and in the rest of the paper, the concept of a link plays a central role. A *link* is just an element $(v, w) \in \mathbb{N}_+^* \times \mathbb{N}_+^*$. However, we will mostly be interested in links for which v and w are positions in an input and an output tree, respectively. We let

$$\mathcal{F}(Q, \Sigma) = \{ \langle \xi, A, I, \zeta \rangle \mid \xi, \zeta \in T_{\Sigma}(Q),\ A, I \subseteq \operatorname{pos}(\xi) \times \operatorname{pos}(\zeta) \}$$

be the *forms over Q and Σ*. Such a form $\langle \xi, A, I, \zeta \rangle \in \mathcal{F}(Q, \Sigma)$ consists of a partial input tree ξ, a partial output tree ζ, and two sets A and I of links. Elements in A and I are called *active* and *inactive* links, respectively. These links essentially stem from the rules. Let $\ell \xrightarrow{q} \vec{r} \in R$ be a rule. Roughly speaking, each occurrence w' of a state $p \in Q$ in the trees of the right-hand side $\vec{r}$ is linked to the (unique) occurrence v' of p in the left-hand side ℓ. In this fashion a link (v', w') is formed. The gray splines in Figure 1 indicate exactly those links. For technical convenience, we prefix the positions by other positions in the next definition.

Definition 2 Let $\rho = \ell \xrightarrow{q} \vec{r} \in R$ be a rule, $n = |\vec{r}|$ be the number of fragments in the right-hand side, and let $v, w_1, \dots, w_n \in \mathbb{N}_+^*$. The set $\operatorname{links}_{v,\vec{w}}(\rho) \subseteq \mathbb{N}_+^* \times \mathbb{N}_+^*$ of *links induced by ρ, v, and $\vec{w} = (w_1, \dots, w_n)$* is

$$\operatorname{links}_{v,\vec{w}}(\rho) = \bigcup_{q' \in Q} \left(\bigcup_{i=1}^{n} \{ (v.v', w_i.w') \mid v' \in \operatorname{pos}_{q'}(\ell),\ w' \in \operatorname{pos}_{q'}(r_i) \} \right) \ . \qquad \text{◻}$$

[5] Since we do not restrict ourselves to ranked alphabets [17, 18], this union can always be taken.

For example, $\{(1.2.2, 2.1), (1.2.2, 2.3)\}$ are the links induced by the left-most rule ρ of Figure 1 , $v = 1.2$, and $\vec{w} = (2)$. We graphically represent links as splines.

The semantics of MBOT is presented using synchronous substitution. Since each rule has one partial tree in the left-hand side and potentially several partial trees in the right-hand side, each rule acts on one position of the input tree and potentially several positions in the output tree of a form. More precisely, the application of a rule $\ell \xrightarrow{q} \vec{r}$ replaces, at the same time, one occurrence v of the state q in ξ and potentially several occurrences of q in ζ of a form $\langle \xi, A, I, \zeta \rangle$. In fact, all actively linked occurrences (i.e., all positions that are linked to v in A) are replaced in the output tree ζ.

Definition 3 Let $\langle \xi, A, I, \zeta \rangle, \langle \xi', A', I', \zeta' \rangle \in \mathcal{F}(Q, \Sigma)$ be forms. We write $\langle \xi, A, I, \zeta \rangle \Rightarrow_M \langle \xi', A', I', \zeta' \rangle$ if there exist a rule $\ell \xrightarrow{q} \vec{r} \in R$ and an occurrence $v \in \mathrm{pos}_q(\xi)$ of q in ξ such that

- $|\vec{r}| = |A(v)|$,
- $\xi' = \xi[\ell]_v$ and $\zeta' = \zeta[\vec{r}]_{\vec{w}}$ with $\vec{w} = \mathrm{seq}(A(v))$, and
- $I' = I \cup U$ and $A' = (A \setminus U) \cup \mathrm{links}_{v,\vec{w}}(\ell \xrightarrow{q} \vec{r})$ with $U = \{(v, w) \mid w \in A(v)\}$.

As usual $\Rightarrow_M^*$ is the reflexive and transitive closure of $\Rightarrow_M$. The set $\mathcal{SF}(M) \subseteq \mathcal{F}(Q, \Sigma)$ of *sentential forms computed by* M is

$$\mathcal{SF}(M) = \bigcup_{q_0 \in Q_0} \{\langle \xi, A, I, \zeta \rangle \mid \langle q_0, \{(\varepsilon, \varepsilon)\}, \emptyset, q_0 \rangle \Rightarrow_M^* \langle \xi, A, I, \zeta \rangle\} \ .$$

Additionally, the MBOT M computes the set $\mathcal{D}(M)$ of *dependencies* given by

$$\mathcal{D}(M) = \{\langle t, I, u \rangle \mid \langle t, \emptyset, I, u \rangle \in \mathcal{SF}(M) \text{ and } t, u \in T_\Sigma\}$$

and the *(tree) relation* $M \subseteq T_\Sigma \times T_\Sigma$ given by $M = \{\langle t, u \rangle \mid \exists I \text{ such that } \langle t, I, u \rangle \in \mathcal{D}(M)\}$. □

Roughly speaking, in a derivation step we select a rule $\ell \xrightarrow{q} \vec{r} \in R$ and an occurrence v of the state q in the partial input tree ξ. Next, we select all positions $A(v)$ of the partial output tree ζ that are actively linked to v. We test whether $|A(v)| = |\vec{r}|$; i.e., whether v has as many active link partners as there are trees in the right-hand side $\vec{r}$. If this test is not successful, then the rule cannot be applied at v. Otherwise, we replace the occurrence v of q in ξ by the left-hand side ℓ, and similarly, we replace all subtrees at $A(v)$ in lexicographic order $\mathrm{seq}(A(v))$ in ζ by the trees in the right-hand side $\vec{r}$ (in the order that they are listed in the rule). In addition, the set $U = \{(v, w) \mid w \in A(v)\}$, which collects the links used in the derivation step, is subtracted from the active links A and added to the inactive links I. Finally, the links induced by the rule and the positions at which it is applied are added to the active links.

In the literature [7, 19] the used links U are often simply removed during a derivation step, but we want to investigate and reason about those links as in [30], so we preserve them in the set I of inactive links. More precisely, we even output them as part of the computed dependencies, which are essentially sentential forms of state-free input and output trees without any active links. At this point, it is also clear that each rule $\ell \xrightarrow{q} \varepsilon$ is a *look-ahead rule* because it only produces part of the input tree (or alternatively: it only checks the input tree). Such look-ahead rules can be used to check whether an input tree belongs to a certain regular tree language [17, 18].

Example 2 A short derivation using the MBOT M_{ex} of Example 1 is shown in Figure 2. Since the final sentential form of that derivation contains no states and has no active links, the MBOT M_{ex} can compute the dependency $\langle t, I, u \rangle \in \mathcal{D}(M_{\mathrm{ex}})$ with

$$I = \{(\varepsilon, \varepsilon), (2, 1), (2, 3), (2.1, 1.1), (2.1, 3.1), (2.1.1, 1.1.1), (2.1.1, 3.1.1)\} \ ,$$

$t = \sigma(\alpha, \gamma_1(\gamma_2(\alpha)), \alpha)$, and $u = \sigma(\gamma_1(\gamma_2(\alpha)), \alpha, \gamma_1(\gamma_2(\alpha)))$. In general, M_{ex} computes the tree relation $\{\langle \sigma(\alpha, t, \alpha) \, , \, \sigma(t, \alpha, t) \rangle \mid t \in T\}$, where T is the smallest tree language such that $\alpha \in T$ and $\{\gamma_1(t), \gamma_2(t)\} \subseteq T$ for all $t \in T$. □

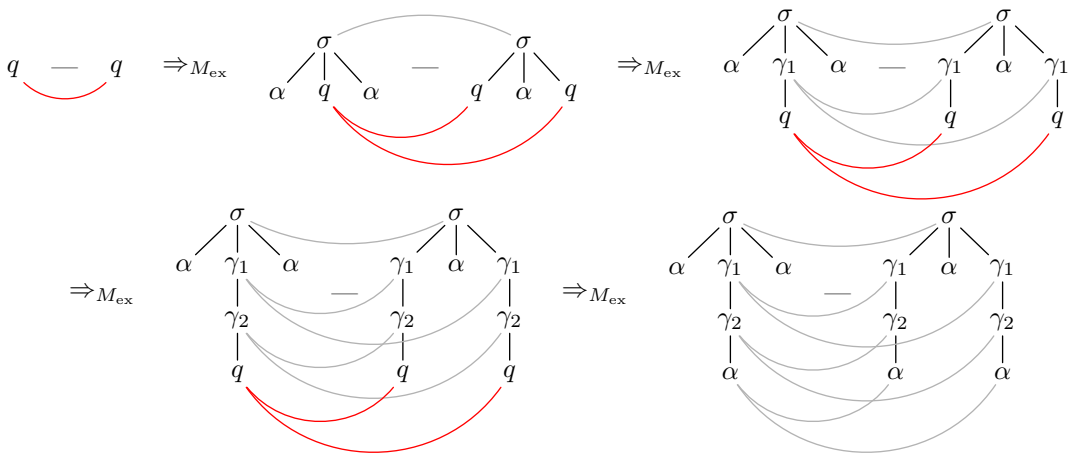

Figure 2: Derivation of the MBOT M_{ex} of Example 1. The active links are clearly marked, whereas inactive links are light gray.

Figure 3: Tree relation used in Theorem 1, where $s, t, u \in T$.

4. A traditional proof

Next, we want to illustrate a standard technique, coined fooling technique, for proving that a given tree relation cannot be computed by a given model. We chose an instance of [33, Theorem 5.2], which demonstrates that a given tree relation cannot be computed by any XTOP$^{\text{R}}$. We additionally restrict ourself to ε-free XTOP$^{\text{R}}$ for reasons mentioned later (see Section 6). Finally, we adjust the proof of [33] to our notation.

Theorem 1 (see [33, Theorem 5.2]) *The tree relation τ over $\Sigma = \{\delta, \gamma, \alpha\}$ given by*

$$\tau = \left\{ \left\langle \delta(\gamma^n(\delta(s,t)), u)\,,\, \delta(s, \delta(t,u)) \right\rangle \mid n \in \mathbb{N},\, s,t,u \in T \right\} \qquad \text{with} \qquad T = \{\gamma^k(\alpha) \mid k \in \mathbb{N}\}\ ,$$

which is illustrated in Figure 3, cannot be computed by any ε-free XTOP$^{\text{R}}$. □

Proof For the sake of a contradiction, suppose that the tree relation τ can be computed by some ε-free XTOP$^{\text{R}}$ $M = (Q, \Sigma, Q_0, R)$. Let $m > |Q|$ and $n > \max\{\max(\text{ht}(\ell), \text{ht}(r_1)) \mid \ell \xrightarrow{q} \vec{r} \in R\}$ be a constant larger than the height of all left- and right-hand sides. We select the contexts $c = \delta(\gamma^n(\delta(x_1, x_2)), x_3)$ and $c' = \delta(x_1, \delta(x_2, x_3))$. Moreover, we select a tree $t \in T$ such that $\text{ht}(t) > m + n$. Clearly, such a tree t exists and $\langle c[t,t,t],\, c'[t,t,t]\rangle \in \tau = M$. Consequently, there exists a dependency $\langle c[t,t,t], I,\, c'[t,t,t]\rangle \in \mathcal{D}(M)$ and hence $\langle c[t,t,t], \emptyset, I, c'[t,t,t]\rangle \in \mathcal{SF}(M)$. From the latter we conclude that there exists an initial state $q_0 \in Q_0$ such that $\langle q_0, \{(\varepsilon,\varepsilon)\}, \emptyset, q_0\rangle \Rightarrow_M^* \langle c[t,t,t], \emptyset, I, c'[t,t,t]\rangle$. Obviously, at least one derivation step needs to be applied, so let

$$\langle q_0, \{(\varepsilon,\varepsilon)\}, \emptyset, q_0\rangle \Rightarrow_M \langle u, A', I', u'\rangle \Rightarrow_M^* \langle c[t,t,t], \emptyset, I, c'[t,t,t]\rangle$$

for some $\langle u, A', I', u'\rangle \in \mathcal{SF}(M)$. Let $\rho = \ell \xrightarrow{q_0} \vec{r} \in R$ be the rule that is used in the first derivation step. We now distinguish several cases for ρ.

- First, suppose that $r_1 \in Q$. By Definition 1 we have $|\mathrm{pos}_{r_1}(\ell)| = 1$, so let $w \in \mathrm{pos}_{r_1}(\ell)$ be the unique occurrence of the state r_1 in ℓ. Note that $w \neq \varepsilon$ because M is ε-free. Hence $w = i.w'$ for some $i \in \{1, 2\}$ and $w' \in \mathbb{N}_+^*$. We now distinguish two subcases.
  - Let $i = 1$. Then $w = 1^k$ for some $1 \leq k < n$ because $\mathrm{ht}(\ell) < n$. Since $\mathrm{ht}(t) > n > \mathrm{ht}(\ell)$, another state q must occur in $\ell|_2$ (i.e., $\mathrm{idx}(\ell|_2) \neq \emptyset$). This state q can only develop using look-ahead rules since it does not occur in r_1. We will argue a bit informally here since we want to avoid introducing more material. Since the tree t is tall enough, the pumping lemma for regular tree languages (see [17]) permits us to select a tree $t' \in T$ with $t \neq t'$, which the look-ahead rules can also create. Then
    $$\langle q_0, \{(\varepsilon, \varepsilon)\}, \emptyset, q_0 \rangle \Rightarrow_M \langle u, A', I', u' \rangle \Rightarrow_M^* \langle c[t, t, t'], \emptyset, I, c'[t, t, t] \rangle \ ,$$
    which yields $\langle c[t, t, t'], c'[t, t, t] \rangle \in M = \tau$ contradicting the definition of τ because $t \neq t'$.
  - The case $i = 2$ can be handled in much the same manner.
- Alternatively, we have $r_1 = \delta(t_1, t_2)$ for some $t_1, t_2 \in T_\Sigma(Q)$. We can assume that r_1 is linear since M is an $\textsc{xtop}^{\mathrm{R}}$ (we omit a proof of this claim). Moreover, since $\mathrm{ht}(t) > n > \mathrm{ht}(r_1)$, both t_1 and t_2 must contain a state (i.e., $\mathrm{idx}(t_1) \neq \emptyset \neq \mathrm{idx}(t_2)$. Clearly, these two states $q_1, q_2 \in Q$ are different because r_1 is linear. By Definition 1 the left-hand side ℓ must contain both those states q_1 and q_2, so $\ell = \delta(\gamma^{k_1}(q_1), \gamma^{k_2}(q_2))$ for some $k_1, k_2 \in \mathbb{N}$ with $k_1, k_2 < n$. Now we again need to distinguish two subcases.
  - First, suppose that $q_1 \in \mathrm{idx}(t_1)$. Clearly, there are no further states in r_1 because there cannot be any further state occurrences in ℓ. Again, we switch to a rather informal description here. We must have $r_1 = \delta(t_1, q_2)$ because $\mathrm{ht}(r_1) < n < \mathrm{ht}(t)$. Intuitively speaking, the root of t_2 can only be δ or q_2. If it were δ, then one of its subtrees would not contain a state, which contradicts the generated tree because both subtrees of that occurrence of δ are higher than n. Hence the state q_2 must develop into a subtree of t in the input and into $\delta(t, t)$ in the output. In that subderivation we consider the rule that produces the δ in the output. This rule can only contain one state by the shape of the input subtree that it generates, but then it cannot generate $\delta(t, t)$ in the output because $\mathrm{ht}(t)$ is strictly larger than the height of the right-hand side of that rule.
  - The remaining case, in which $q_1 \in \mathrm{idx}(t_2)$, can be contradicted in a similar manner.

All cases are contradictory. Hence an ε-free $\textsc{xtop}^{\mathrm{R}}$ computing τ cannot exist. ∎

To recapture the main points of the previous proof, we suppose that we can compute the tree relation and then cleverly select a particular pair in the relation, for which we must have a derivation. Then we inspect the rules used in that derivation and show that (i) such rules cannot exist [as in the first subcase in the second case] or (ii) the rules permit additional undesired derivations [as in the first subcase of the first case]. For the latter, we need knowledge about pairs of trees that are not in the desired tree relation.

We also note that the proof is feasible because we know both the input and the output tree, which allows us to restrict the shape of the considered rules [as demonstrated in all cases]. However, if we want to prove that a particular tree relation cannot be computed by a composition of $\textsc{xtop}^{\mathrm{R}}$, then we would need to argue over (potentially) several unknown intermediate trees, which would yield many additional potential cases for the rules, which often makes this proof approach rather unappealing.

5. Basic properties of links

In this section, we introduce some important properties, which were already discussed in [25, 30], for sets of links [such as A and I in a form $\langle \xi, A, I, \zeta \rangle \in \mathcal{F}(Q, \Sigma)$] and the dependencies $\mathcal{D}(M)$. Let us start with the properties that relate links in a set to each other. We generally only define the properties for the input side, but assume that the same properties are also defined (in the same manner) for the output side.

Definition 4 (see [30, Definition 8]) A set $L \subseteq \mathbb{N}_+^* \times \mathbb{N}_+^*$ of links is

- *input hierarchical* if for all links $(v_1, w_1), (v_2, w_2) \in L$
  - (i) $v_1 \prec v_2$ implies $w_2 \not\prec w_1$ and
  - (ii) $v_1 \prec v_2$ implies that there exists a link $(v_1, w_1') \in L$ with $w_1' \preceq w_2$, and
- *strictly input hierarchical* if for all links $(v_1, w_1), (v_2, w_2) \in L$
  - (i') $v_1 \prec v_2$ implies $w_1 \preceq w_2$ and
  - (ii') $v_1 = v_2$ implies $w_1 \preceq w_2$ or $w_2 \preceq w_1$.

A form $\langle \xi, A, I, \zeta \rangle \in \mathcal{F}(Q, \Sigma)$ is (strictly) input hierarchical whenever $A \cup I$ is. Finally, $\mathcal{D}(M)$ has those properties if for each $\langle t, I, u \rangle \in \mathcal{D}(M)$ the corresponding form $\langle t, \emptyset, I, u \rangle$ has them (i.e., I has them). □

Trivially, any strictly input hierarchical set L is also input hierarchical because (i') implies (i) and (ii). Roughly speaking, input hierarchical sets of links have no crossing links, which are links $(v_1, w_1), (v_2, w_2) \in L$ such that $v_1 \prec v_2$ and $w_2 \prec w_1$ [contradicting (i)]. The property *(strictly) output hierarchical* [for sets L of links, forms $\langle \xi, A, I, \zeta \rangle \in \mathcal{F}(Q, \Sigma)$, and $\mathcal{D}(M)$] is defined by requiring the corresponding input-side property for the inverted set L^{-1} of links, the inverted form $\langle \zeta, A^{-1}, I^{-1}, \xi \rangle$, and the inverted dependencies $\mathcal{D}(M)^{-1} = \{\langle u, I^{-1}, t \rangle \mid \langle t, I, u \rangle \in \mathcal{D}(M)\}$. Consequently, L is strictly output hierarchical if and only if L^{-1} is strictly input hierarchical.

Example 3 The links I of Example 2 (those of the final sentential form in Figure 2) are input hierarchical. They are not strictly input hierarchical because the links $(2, 1), (2.1, 3.1) \in I$ violate (i'). Given the same two links, we can select the link $(2, 3) \in I$ to fulfill (ii). However, I is strictly output hierarchical. □

The properties (input hierarchical and strictly output hierarchical) mentioned in Example 3 are not accidental, but rather they are true for the dependencies $\mathcal{D}(M)$ computed by each MBOT M [30, Lemma 22] as we show next.

Theorem 2 *For every* MBOT M*, the set* $\mathcal{D}(M)$ *is input hierarchical and strictly output hierarchical.* □

PROOF Let $M = (Q, \Sigma, Q_0, R)$. We prove the more general statement that $\langle \xi, A, I, \zeta \rangle$ [i.e., the set $A \cup I$ of links] is input hierarchical and strictly output hierarchical for every sentential form $\langle \xi, A, I, \zeta \rangle \in \mathcal{SF}(M)$. Let us prove these properties by induction on the length of the derivation $\langle q_0, \{(\varepsilon, \varepsilon)\}, \emptyset, q_0 \rangle \Rightarrow_M^* \langle \xi, A, I, \zeta \rangle$, where $q_0 \in Q_0$.

The properties are clearly true for the initial sentential form $\langle q_0, \{(\varepsilon, \varepsilon)\}, \emptyset, q_0 \rangle$ with $q_0 \in Q_0$, so we covered the induction base. In the induction step, we assume that they hold for some $\langle \xi, A, I, \zeta \rangle \in \mathcal{SF}(M)$, and prove them for all sentential forms $\langle \xi', A', I', \zeta' \rangle \in \mathcal{SF}(M)$ such that $\langle \xi, A, I, \zeta \rangle \Rightarrow_M \langle \xi', A', I', \zeta' \rangle$. By the definition of $\Rightarrow_M$ (see Definition 3) there exist a rule $\ell \xrightarrow{q} \vec{r} \in R$ and an occurrence $v \in \mathrm{pos}_q(\xi)$ of q in ξ such that (i) $|\vec{r}| = |A(v)|$, (ii) $\xi' = \xi[\ell]_v$ and $\zeta' = \zeta[\vec{r}]_{\vec{w}}$ with $\vec{w} = \mathrm{seq}(A(v))$, and (iii) $I' = I \cup U$ and $A' = (A \setminus U) \cup L$ with $U = \{(v, w) \mid w \in A(v)\}$ and $L = \mathrm{links}_{v,\vec{w}}(\ell \xrightarrow{q} \vec{r})$. In the following, let $m = |\vec{r}|$.

Let us first prove that $A' \cup I'$ is input hierarchical. To this end, let $(v_1', w_1'), (v_2', w_2') \in A' \cup I'$ with $v_1' \prec v_2'$. Note that $A' \cup I' = A \cup I \cup L$. If $(v_2', w_2') \in A \cup I$, then $(v_1', w_1') \in A \cup I$ (because $v \preceq v'$ for all $(v', w') \in L$ and $v \not\prec v_2'$ since v is a leaf position in ξ) and thus the required properties (i) and (ii) are trivially true by the induction hypothesis and $A \cup I \subseteq A' \cup I'$. It remains to investigate the case where $(v_2', w_2') \in L$, which yields $(v_1', w_1') \in A \cup I$ because v' is a leaf position of ξ' for every link $(v', w') \in L$ and thus $v' \not\prec v_2'$. Consequently, the link (v_2', w_2') points to $v_2' = vv'$ and $w_2' = ww'$ for some $v' \in \mathrm{pos}(\ell)$, $w \in A(v)$, and $w' \in \mathbb{N}_+^*$ by the definition of L. Hence we can establish (i) [i.e., $w_2' \not\prec w_1'$] immediately because $w \preceq w_2'$ but $w \not\prec w_1'$ as w is a leaf position in $\mathrm{pos}(\zeta)$ and $w_1' \in \mathrm{pos}(\zeta)$. For (ii) we observe that $v_1' \preceq v$ because v is a leaf position of ξ, $v_1' \prec vv' = v_2'$, and $v_1' \in \mathrm{pos}(\xi)$. If $v_1' = v$, then $(v, w) \in A$ is a link required for (ii) because $w \preceq w_2'$ and $A \subseteq A' \cup I'$. Finally, if $v_1' \prec v$, then by the induction hypothesis there exists a link $(v_1', w_1'') \in A \cup I$ with $w_1'' \preceq w$. Clearly, this link (v_1', w_1'') is also in $A' \cup I'$ and fulfills $w_1'' \preceq w \preceq w_2'$. This establishes that $A' \cup I'$, and thus $\langle \xi', A', I', \zeta' \rangle$ is input hierarchical.

Secondly, we need to show that $A' \cup I'$ is strictly output hierarchical. To this end, we assume two links $(v_1', w_1'), (v_2', w_2') \in A' \cup I'$ such that $w_1' \preceq w_2'$. If $v_1' = v_2'$, then both (i') and (ii') are trivially fulfilled. Hence assume that $v_1' \neq v_2'$. Again if $(v_1', w_1'), (v_2', w_2') \in A \cup I$, then the required properties (i') and (ii')

are true by the induction hypothesis. Next we consider the case when $w'_1 = w'_2$ and $(v'_2, w'_2) \notin A \cup I$. Then $v'_2 = vv'$ and $w'_2 = ww'$ for some $v' \in \mathrm{pos}(\ell)$, $w \in A(v)$, and $w' \in \mathbb{N}_+^*$ by the definition of L. Since each output position links to at most one input position in L and $v'_1 \neq v'_2$, we obtain that $(v'_1, w'_1) \in A \cup I$. Consequently, $w' = \varepsilon$ because otherwise $w'_1 = w'_2 \notin \mathrm{pos}(\zeta)$. So in summary we have $w'_1 = w'_2 = w$ and two links $(v, w), (v'_1, w) \in A \cup I$. From the latter we can conclude that $v \preceq v'_1$ or $v'_1 \preceq v$ by the induction hypothesis. Since v is a leaf position in ξ and $v'_1 \in \mathrm{pos}(\xi)$, we know that $v'_1 \preceq v$ must be true. Together with $v \preceq v'_2$ we obtain $v'_1 \preceq v \preceq v'_2$ as required for (ii'). Analogous arguments also prove the case $w'_1 = w'_2$ and $(v'_1, w'_1) \notin A \cup I$. In the remaining case, we have $w'_1 \prec w'_2$, $(v'_2, w'_2) \in L$, and $(v'_1, w'_1) \in A \cup I$ as $w'_1 \prec w'_2$ yields that the links cannot be both in L and we already covered the case in which both links are in $A \cup I$. Consequently, $v'_2 = vv'$ and $w'_2 = ww'$ for some $v' \in \mathrm{pos}(\ell)$, $w \in A(v)$, and $w' \in \mathbb{N}_+^*$ by the definition of L. Since w is a leaf position of ζ, $w'_1 \prec ww' = w'_2$, and $w'_1 \in \mathrm{pos}(\zeta)$, we obtain that $w'_1 \preceq w$. Since (v, w) and (v'_1, w'_1) are both in $A \cup I$ and $w \preceq w'_1$, we can apply the induction hypothesis to them and distinguish two subcases. If $w'_1 \prec w$, then $v'_1 \preceq v$ by the induction hypothesis, which together with $v \preceq v'_2$ yields $v'_1 \preceq v \preceq v'_2$ as required for (i'). Similarly, if $w'_1 = w$, then $v'_1 \preceq v$ or $v \preceq v'_1$ by the induction hypothesis, from which we can conclude that $v'_1 \preceq v$ because v is a leaf position in ξ and $v'_1 \in \mathrm{pos}(\xi)$. Consequently, we again obtain $v'_1 \preceq v \preceq v'_2$ as required for (i').

Since all sentential forms of M are input hierarchical and strictly output hierarchical, these properties also hold for $\mathcal{D}(M)$. ■

Corollary 1 (of the proof of Theorem 2) *For every* XTOP$^{\mathrm{R}}$ M*, the set* $\mathcal{D}(M)$ *is strictly input and strictly output hierarchical.* □

PROOF Again, let $M = (Q, \Sigma, Q_0, R)$. Theorem 2 shows that $\mathcal{D}(M)$ is strictly output hierarchical. The proof of that property can also be applied to the input side in an XTOP$^{\mathrm{R}}$ because for each rule $\rho = \ell \xrightarrow{q} \vec{r} \in R$ and suitable positions v and $\vec{w}$ as in that proof, each input position links to at most one output position in $L = \mathrm{links}_{v,\vec{w}}(\rho)$.[6] Following the same reasoning as for the output side, we obtain that $\mathcal{D}(M)$ is also strictly input hierarchical. ■

Now we have established the interrelations between the links in the dependencies computed by MBOT and XTOP$^{\mathrm{R}}$. However, this property by itself is not yet very useful because in order to apply it we first need to establish the existence of links with certain properties [to fulfill the preconditions of the properties (i), (ii), (i'), or (ii')]. Consequently, we also need to establish the existence of "enough" links to which we can then apply Theorem 2 and Corollary 1. Fortunately, the derivation process guarantees that there cannot be large "gaps" between two "adjacent" links. In other words, for each MBOT there should be an integer b that limits the distance between links. Note that due to the presence of look-ahead rules, we cannot, in general, require the stricter variant that each path in the input tree should have a link every b steps. We present this phenomenon in more detail in Example 4.

Definition 5 (cf. [30, Definition 10]) Let $b \in \mathbb{N}$. A form $\langle \xi, A, I, \zeta \rangle \in \mathcal{F}(Q, \Sigma)$ has

- *link distance b in the input* if for all links $(v_1, w_1), (v_2, w_2) \in A \cup I$ with $v_1 \prec v_2$ and $|v_2| - |v_1| > b$, there exists a link $(v, w) \in A \cup I$ such that $v_1 \prec v \prec v_2$ and $|v| - |v_1| \leq b$, and
- *strict link distance b in the input* if for all positions $v_1, v_2 \in \mathrm{pos}(\xi)$ with $v_1 \prec v_2$ and $|v_2| - |v_1| > b$, there exists a link $(v, w) \in A \cup I$ such that $v_1 \prec v \prec v_2$ and $|v| - |v_1| \leq b$.

The set $\mathcal{D}(M)$ of dependencies has those properties if for each $\langle t, I, u \rangle \in \mathcal{D}(M)$ the corresponding sentential form $\langle t, \emptyset, I, u \rangle \in \mathcal{SF}(M)$ has them. Moreover, $\mathcal{D}(M)$ is *(strictly) link-distance bounded in the input* if there exists an integer $b \in \mathbb{N}$ such that it has (strict) link distance b in the input. □

Clearly, a form $\langle \xi, A, I, \zeta \rangle$ with strict link distance b in the input has link distance b in the input because the strict property again trivially implies the non-strict property. As before, we assume that (strict) link distance is also defined for the output side, so a form $\langle \xi, A, I, \zeta \rangle$ and $\mathcal{D}(M)$ have *(strict) link distance b in the output* if $\langle \zeta, A^{-1}, I^{-1}, \xi \rangle$ and $\mathcal{D}(M)^{-1} = \{\langle u, I^{-1}, t \rangle \mid \langle t, I, u \rangle \in \mathcal{D}(M)\}$ have (strict) link distance b in the input, respectively. Let us illustrate the difference between the two 'link distance' notions.

[6] Note that $|\vec{w}| \leq 1$ because M is an XTOP$^{\mathrm{R}}$ and thus $|\vec{r}| \leq 1$.

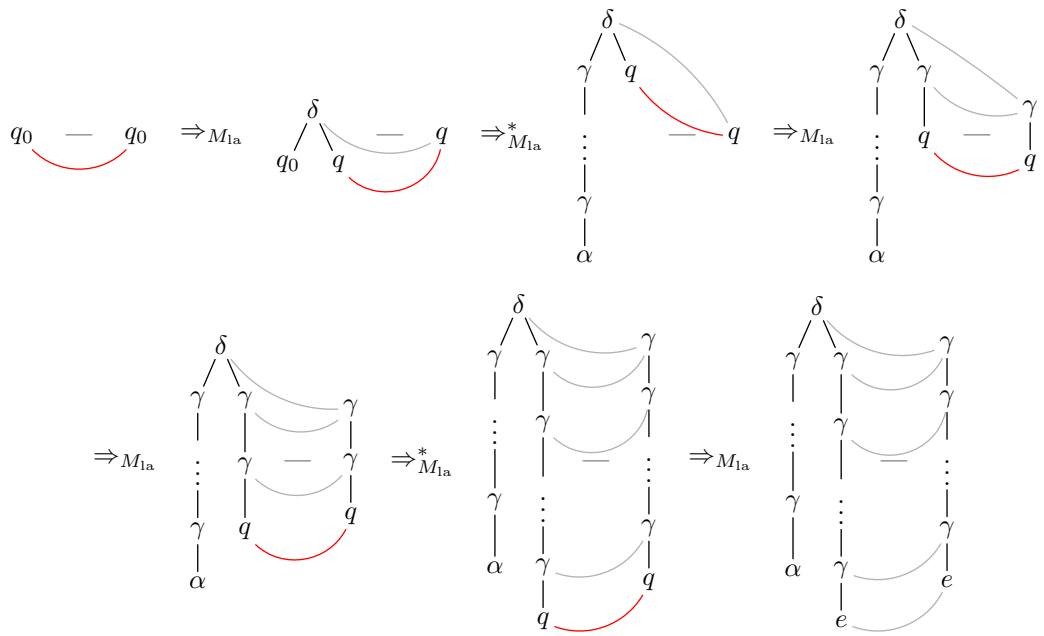

Figure 4: Illustration of the dependencies (with $e \in \{\alpha, \beta\}$) computed by the XTOP$^{\mathrm{R}}$ M_{la} of Example 4. Inactive links are shown in light gray.

Example 4 Let us consider the ε-free XTOP$^{\mathrm{R}}$ $M_{\mathrm{la}} = (\{q_0, q\}, \Sigma, \{q_0\}, R)$ with $\Sigma = \{\delta, \gamma, \alpha, \beta\}$ and exactly the following rules in R:

$$\delta(q_0, q) \xrightarrow{q_0} q \qquad \gamma(q_0) \xrightarrow{q_0} \varepsilon \qquad \alpha \xrightarrow{q_0} \varepsilon$$
$$\beta \xrightarrow{q} \beta \qquad \gamma(q) \xrightarrow{q} \gamma(q) \qquad \alpha \xrightarrow{q} \alpha \ .$$

Clearly, M_{la} computes the dependencies

$$\mathcal{D}(M_{\mathrm{la}}) = \{\langle \delta(\gamma^m(\alpha), \gamma^n(e)) \,,\, I_n \,,\, \gamma^n(e) \rangle \mid e \in \{\alpha, \beta\},\, m, n \in \mathbb{N}\} \qquad \text{with}$$
$$I_n = \{(\varepsilon, \varepsilon), (2, \varepsilon)\} \cup \{(21^i, 1^i) \mid i \in [n]\} \ .$$

We illustrate the computed dependencies in the last sentential form of Figure 4. The set $\mathcal{D}(M_{\mathrm{la}})$ has link distance 1 in both the input and the output. In addition, it has strict link distance 1 in the output, but it is not strictly link-distance bounded in the input because for a given $b \in \mathbb{N}$ we can select the dependency $\langle \delta(\gamma^b(\alpha), \gamma(\alpha)), I_1, \gamma(\alpha) \rangle$ and the input positions $\varepsilon, 1^{b+1} \in \mathrm{pos}(\delta(\gamma^b(\alpha), \gamma(\alpha)))$, the latter of which points to the α-leaf in the left branch. Then $\varepsilon \prec 1^{b+1}$ and $|1^{b+1}| - |\varepsilon| = b + 1 > b$, but there is no link $(v, w) \in I_1$ such that $\varepsilon \prec v \prec 1^{b+1}$. □

As before, the properties observed in Example 4 are not accidental. For each MBOT M, the set $\mathcal{D}(M)$ of dependencies is link-distance bounded in the input and strictly link-distance bounded in the output [30, Lemma 22]. As before, the output side automatically fulfills the stricter variant, but in contrast to the hierarchical properties, Example 4 already demonstrates that this distinction remains true even for XTOP$^{\mathrm{R}}$.

Theorem 3 *For every* MBOT M*, the set* $\mathcal{D}(M)$ *is link-distance bounded in the input and strictly link-distance bounded in the output.* □

Proof Let $M = (Q, \Sigma, Q_0, R)$,

$$a = \max\{\mathrm{ht}(\ell) \mid \ell \xrightarrow{q} \vec{r} \in R\} \quad \text{and} \quad b = \max\{\mathrm{ht}(r_i) \mid \ell \xrightarrow{q} (r_1, \ldots, r_n) \in R,\, i \in [n]\} \ .$$

We show that every sentential form $\langle \xi, A, I, \zeta \rangle \in \mathcal{SF}(M)$ has link distance a in the input and strict link distance b in the output. Again, we prove the statement by induction on the length of the derivation $\langle q_0, \{(\varepsilon, \varepsilon)\}, \emptyset, q_0 \rangle \Rightarrow_M^* \langle \xi, A, I, \zeta \rangle$, where $q_0 \in Q_0$.

The two link distance properties are trivially true for the initial sentential form $\langle q_0, \{(\varepsilon, \varepsilon)\}, \emptyset, q_0 \rangle$ with $q_0 \in Q_0$ (i.e., induction base), in which we simply cannot select two different links [there is only the link $(\varepsilon, \varepsilon)$] nor two different input tree positions (since $\mathrm{pos}(q_0) = \{\varepsilon\}$). In the induction step, we assume that the link distance properties hold for some $\langle \xi, A, I, \zeta \rangle \in \mathcal{SF}(M)$, and we need to prove them for all sentential forms $\langle \xi', A', I', \zeta' \rangle \in \mathcal{SF}(M)$ such that $\langle \xi, A, I, \zeta \rangle \Rightarrow_M \langle \xi', A', I', \zeta' \rangle$. By the definition of $\Rightarrow_M$ (see Definition 3) there exist a rule $\ell \xrightarrow{q} \vec{r} \in R$ and an occurrence $v \in \mathrm{pos}_q(\xi)$ of q in ξ such that (i) $|\vec{r}| = |A(v)|$, (ii) $\xi' = \xi[\ell]_v$ and $\zeta' = \zeta[\vec{r}]_{\vec{w}}$ with $\vec{w} = \mathrm{seq}(A(v))$, and (iii) $I' = I \cup U$ and $A' = (A \setminus U) \cup L$ with $U = \{(v, w) \mid w \in A(v)\}$ and $L = \mathrm{links}_{v,\vec{w}}(\ell \xrightarrow{q} \vec{r})$.

Let $m = |\vec{r}|$, and we start with link distance a in the input. To this end, let $(v_1', w_1'), (v_2', w_2') \in A' \cup I'$ be such that $v_1' \prec v_2'$ and $|v_2'| - |v_1'| > a$. By the induction hypothesis and $A \cup I \subseteq A' \cup I'$, there exists a link $(v', w') \in A' \cup I'$ with $v_1' \prec v' \prec w_2'$ and $|v'| - |v_1'| \leq a$ provided that $(v_1', w_1'), (v_2', w_2') \in A \cup I$. Consequently, we consider the case that at least one of the two links (v_1', w_1') and (v_2', w_2') is in L. It is clear from the definition of L that both links cannot be in L because $\mathrm{ht}(\ell) \leq a$ (i.e., the longest path in ℓ is at most of length a) but $|v_2'| - |v_1'| > a$. Similarly, the case $(v_1', w_1') \in L$ is impossible because it implies $v \preceq v_1'$ and $(v_2', w_2') \in A \cup I$. These two statements in turn yield $v_2' \in \mathrm{pos}(\xi)$ and $v \preceq v_1' \prec v_2'$, but v is a leaf position in ξ, which contradicts $v_2' \in \mathrm{pos}(\xi)$. Only the case $(v_2', w_2') \in L$ and $(v_1', w_1') \in A \cup I$ remains. Since $v_1' \prec v_2'$ we can conclude $v_1' \preceq v \preceq v_2'$ and there exists $w \in A(v)$ such that $w \preceq w_2'$. Note that $v_1' = v$ is impossible because $v_1' = v$ yields $|v_2'| - |v_1'| \leq a$ contradicting $|v_2'| - |v_1'| > a$. Hence $v_1' \prec v \preceq v_2'$. If $|v| - |v_1'| \leq a$, then $(v, w) \in A' \cup I'$ is a suitable link because $v_1' \prec v \prec v_2'$, where $v = v_2'$ is excluded by the conditions $|v| - |v_1'| \leq a$ and $|v_2'| - |v_1'| > a$. Otherwise, $|v| - |v_1'| > a$ and since $(v_1', w_1'), (v, w) \in A \cup I$ with $v_1' \prec v$ we can apply the induction hypothesis to obtain a link $(v', w') \in A \cup I$ with $v_1' \prec v' \prec v$ and $|v'| - |v_1'| \leq a$. This link also fulfills $v_1' \prec v' \prec v \prec v_2'$ and is in $A' \cup I'$ because $A \cup I \subseteq A' \cup I'$, so we completed the link distance a in the input.

Next, we move to the output side, for which we show the strict link distance b. Let $w_1', w_2' \in \mathrm{pos}(\zeta')$ such that $w_1' \prec w_2'$ and $|w_2'| - |w_1'| > b$. Again, the property is satisfied by the induction hypothesis and $A \cup I \subseteq A' \cup I'$ whenever $w_1', w_2' \in \mathrm{pos}(\zeta)$. We observe that $w_1' \notin \mathrm{pos}(\zeta') \setminus \mathrm{pos}(\zeta)$ because $\mathrm{ht}(r_i) \leq b$ for all $i \in [m]$ (i.e., the longest path in any r_i with $i \in [m]$ is at most of length b) and $|w_2'| - |w_1'| > b$ by assumption. Thus, we only need to investigate the case, in which $w_1' \in \mathrm{pos}(\zeta)$ and $w_2' \in \mathrm{pos}(\zeta') \setminus \mathrm{pos}(\zeta)$. In this case, there exists $w \in A(v)$ such that $w_1' \preceq w \prec w_2'$. The case $w_1' = w$ is impossible because $w_1' = w$ yields $|w_2'| - |w_1'| \leq b$ contradicting $|w_2'| - |w_1'| > b$. Hence $w_1' \prec w \prec w_2'$. Now we again distinguish two cases. If $|w| - |w_1'| \leq b$, then $(v, w) \in A' \cup I'$ is a suitable link because $w_1' \prec w \prec w_2'$. Otherwise, $|w| - |w_1'| > b$ and we can apply the induction hypothesis to the positions $w_1', w \in \mathrm{pos}(\zeta)$ because $w_1' \prec w_i$. In this way, we obtain a link $(v', w') \in A \cup I$ with $w_1' \prec w' \prec w$ and $|w'| - |w_1'| \leq b$. This link also fulfills $w_1' \prec w' \prec w \prec w_2'$ and is in $A' \cup I'$, so we completed the induction.

Consequently, $\mathcal{D}(M)$ is link-distance bounded in the input and strictly link-distance bounded in the output, which establishes the main statement. ∎

Corollary 2 (of the proof of Theorem 3) *For every n-XTOP M, the set $\mathcal{D}(M)$ of computed dependencies is strictly link-distance bounded in both the input and the output.* □

PROOF Theorem 3 shows that $\mathcal{D}(M)$ is strictly link-distance bounded in the output. The proof of this property can also be used to prove the strict variant for the input side given an n-XTOP because for each derivation step (as in that proof) we have that the set $A(v)$ of output positions is never empty.[7] Following the same reasoning as for the output side, we obtain that $\mathcal{D}(M)$ is also strictly link-distance bounded in the input. ∎

[7] We have $|\vec{r}| = 1$ for every rule $\ell \xrightarrow{q} \vec{r} \in R$ because M is an n-XTOP.

Model \ Property	hierarchical input	hierarchical output	link distance bounded in the input	link distance bounded in the output
n-XTOP	strictly	strictly	strictly	strictly
XTOP^{R}	strictly	strictly	✓	strictly
MBOT	✓	strictly	✓	strictly

Table 1: Summary of the properties of the computed dependencies.

Now we know that there must be sufficiently many links [due to the (strict) link distance] and how those links interrelate [via the (strictly) hierarchical properties] in each dependency $\langle t, I, u\rangle \in \mathcal{D}(M)$. Table 1 summarizes the properties of the dependencies computed by various transducer models.

For the special case of ε-free MBOT, for which we will develop our linking theorems,[8] we now limit the number of output positions that are linked to a given input position. Trivially, given a rule $\rho = \ell \xrightarrow{q} \vec{r} \in R$ of an ε-free MBOT M (i.e., $\ell \notin Q$) and positions v and $\vec{w}$ we observe that there is no link $(v, w') \in \text{links}_{v,\vec{w}}(\rho)$. In other words, each derivation step of the ε-free MBOT M only adds links to "new" input positions. The next lemma captures this property formally. In particular, in each ε-free XTOP^{R} each input position can be used in at most one link. To simplify the notation, let $\text{rk}(M) = \max\{n \mid \ell \xrightarrow{q} (r_1, \ldots, r_n) \in R\}$. Clearly, $\text{rk}(M) \leq 1$ for each XTOP^{R} M. Moreover, let the output degree $\deg^{\text{o}}(M)$ of M be the smallest $k \in \mathbb{N}$ such that $\text{pos}(r_i) \subseteq [k]^*$ for all rules $\ell \xrightarrow{q} (r_1, \ldots, r_n) \in R$ and $i \in [n]$. In other words, each node in a tree in the right-hand side of the rules of M has at most $\deg^{\text{o}}(M)$ children.

Lemma 1 *For every ε-free* MBOT $M = (Q, \Sigma, Q_0, R)$*, we have $|I(v)| \leq \text{rk}(M)$ and the elements of $I(v)$ are pairwise incomparable with respect to the strict prefix order $\prec$ for all $\langle t, I, u\rangle \in \mathcal{D}(M)$ and $v \in \text{pos}(t)$.* □

PROOF A straightforward induction on the length of the derivation can be used to prove $|(A \cup I)(v)| \leq \text{rk}(M)$ and that the elements of $(A \cup I)(v)$ are pairwise incomparable with respect to $\prec$ for all sentential forms $\langle \xi, A, I, \zeta\rangle \in \mathcal{SF}(M)$ and $v \in \text{pos}(\xi)$. We leave the proof details to the reader. ■

We complete this section with results that limit the size and height of an output tree based on the size and height of the corresponding input tree. Several such results on size and height relations are folklore (e.g., for n-XTOP and XTOP^{R}) or well-known (e.g., [14, Lemma 3.7] for deterministic MBOT). Although it is well-known [11, Lemma 8] that MBOT can only compute tree relations, in which the size and height of the output tree is linearly related to the size and height of the input tree, respectively, we reprove this result here. It serves as a first example of how the links can be used to derive (in this case positive) useful results. Although the result is essentially proved in the same way in the literature, we can now obtain a more general result by establishing these relations between linked subtrees. In particular, we will derive bounds on the number of links below a position because these bounds will be an essential tool later on. On a first reading the details of these results can safely be skipped.

Lemma 2 *Let $M = (Q, \Sigma, Q_0, R)$ be an ε-free* MBOT*, and let $a, b \in \mathbb{N}$ be such that $a \geq \max(2, \deg^{\text{o}}(M))$ and that $\mathcal{D}(M)$ has strict link distance b in the output.*[9] *Then*

$$\left\lceil \frac{|u|_w|}{a^{b+1}} \right\rceil \leq |L| \leq \text{rk}(M) \cdot |t|_v|$$

with $L = \{(v', w') \in I \mid v \preceq v',\ w \preceq w'\}$ for every dependency $\langle t, I, u\rangle \in \mathcal{D}(M)$ and link $(v, w) \in I$. □

PROOF Since $\mathcal{D}(M)$ has strict link distance b in the output, tree fragments without links can only have height b. The size of each such fragment is strictly smaller than a^{b+1} because $a \geq \max(2, \deg^{\text{o}}(M))$. Consequently, the set $L' = \{(v', w') \in I \mid w \prec w'\}$ of links pointing to positions strictly below w in the

[8] We defer a discussion of why ε-freeness is essential in our approach to the next section.

[9] Such an integer b exists by Theorem 3.

output contains at least $\lceil \frac{|u|_w|}{a^{b+1}} \rceil - 1$ links, where we subtract the link (v, w), which is not in L'. Moreover, by Theorem 2 we know that I is strictly output hierarchical, so for each link $(v', w') \in L'$ we actually know that $v \preceq v'$ (i.e., they all point to positions below v in the input) and thus $(v', w') \in L$ because $w \prec w'$. Hence $L' \subseteq L$ and $(v, w) \in L \setminus L'$. This establishes the first inequality of the lemma. Finally, by Lemma 1, which is the only part that requires ε-freeness, each input position occurs in at most $\mathrm{rk}(M)$ links. Putting all the pieces together we obtain

$$\lceil \tfrac{|u|_w|}{a^{b+1}} \rceil \leq |L'| + 1 \leq |L| \leq \mathrm{rk}(M) \cdot |\mathrm{pos}(t|_v)| = \mathrm{rk}(M) \cdot |t|_v| \ ,$$

which proves the statement. ∎

Theorem 4 *For every ε-free* MBOT *M, there exist integers $z, b \in \mathbb{N}$ such that*
1. $|u|_w| \leq z \cdot |t|_v|$ *and*
2. $\mathrm{ht}(u|_w) \leq b \cdot \big(\mathrm{ht}(t|_v) + 2\big)$

for every dependency $\langle t, I, u\rangle \in \mathcal{D}(M)$ and link $(v, w) \in I$.[10] □

Proof Let $M = (Q, \Sigma, Q_0, R)$ and $a, b \in \mathbb{N}$ be such that $a \geq \max(2, \deg^{\mathrm{o}}(M))$ and that $\mathcal{D}(M)$ has strict link distance b in the output by Theorem 3. With the help of Lemma 2 we obtain

$$\frac{|u|_w|}{a^{b+1}} \leq \left\lceil \frac{|u|_w|}{a^{b+1}} \right\rceil \leq \mathrm{rk}(M) \cdot |t|_v|$$

and thus $|u|_w| \leq a^{b+1} \cdot \mathrm{rk}(M) \cdot |t|_v|$, which proves the first item for $z = a^{b+1} \cdot \mathrm{rk}(M)$.

We continue with the second item. A set $W \subseteq \mathbb{N}_+^*$ of positions is a prefix chain if all elements of W are pairwise comparable with respect to the prefix order $\preceq$ [i.e., for all $w', w'' \in W$ we have $w' \preceq w''$ or $w'' \preceq w'$]. A prefix chain $W \subseteq \mathrm{pos}(u)$ is an output-link chain of I if $I^{-1}(w) \neq \emptyset$ for every $w' \in W$. Obviously,

$$\begin{aligned}
\mathrm{ht}(u|_w) &= \max\{|w'| \mid w' \in \mathrm{pos}(u|_w)\} \\
&= \max\{|W| \mid W \subseteq \mathrm{pos}(u|_w),\ W \text{ is a prefix chain}\} - 1 \\
&\leq b \cdot \max\{|W| \mid W \subseteq \mathrm{pos}(u|_w),\ \{w\} \cdot W \text{ is an output-link chain}\} \\
&\leq b \cdot \Big(\max\{|W| \mid W \subseteq (\mathrm{pos}(u|_w) \setminus \{\varepsilon\}),\ \{w\} \cdot W \text{ is an output-link chain}\} + 1\Big) \qquad (\dagger)
\end{aligned}$$

because along a prefix chain of output positions we must have a link every b positions by the strict link distance b in the output. For each set W as in (†), the set $V_W = I^{-1}(\{w\} \cdot W)$ is again a prefix chain and $v \preceq v'$ for every $v' \in V_W$ because trivially $w \prec w'$ for all $w' \in \{w\} \cdot W$ and the fact that I is strictly output hierarchical according to Theorem 2. Finally, $|W| \leq |V_W|$ because for each $v' \in V_W$ there exists exactly one $w' \in W$ such that $(v', ww') \in I$ by Lemma 1. Consequently, $|W| \leq |V_W| \leq \mathrm{ht}(t|_v) + 1$. In total, we obtain $\mathrm{ht}(u|_w) \leq b \cdot \big(\mathrm{ht}(t|_v) + 2\big)$ as required. ∎

Corollary 3 (of Theorem 4) *For every ε-free* MBOT *M, there exist integers $z, b \in \mathbb{N}$ such that $|u| \leq z \cdot |t|$ and $\mathrm{ht}(u) \leq b \cdot (\mathrm{ht}(t) + 2)$ for every $\langle t, u\rangle \in M$.* □

Proof If $\langle t, u\rangle \in M$, then there exists $\langle t, I, u\rangle \in \mathcal{D}(M)$. Since $(\varepsilon, \varepsilon) \in I$, we obtain the statements from Theorem 4. ∎

6. Linking theorems

In this section, we develop our linking theorems for compositions of ε-free MBOT. The ε-freeness is unfortunately essential for our approach as we will use the linear size and height approximations of Theorem 4 as a

[10] The additive component in the second item is necessary since $\mathrm{ht}(t|_v) = 0$ [i.e., $t|_v$ is just a leaf] does not imply $\mathrm{ht}(u|_w) = 0$.

key component in our proofs. Size and height bounds for the output tree do not exist for general MBOT M because for an input tree $t \in T_\Sigma$ the set $\{u \mid \langle t, u\rangle \in M\}$ can contain arbitrarily large and high output trees.

Roughly speaking, our linking theorems establish the existence of certain interrelated links, which are forced simply by a subset of the tree relation computed by the composition. More precisely, we simply use the widely accepted approach of requiring the composition to reproduce parts of the input tree t exactly. It is generally assumed that for an (ε-free) MBOT to produce an exact copy in the output tree u, it needs to process the original (i.e., in other words, the reproduction must be a translation of the original). We can now make this widely used intuition precise by showing that such reproductions indeed force links between the original and the reproduction. Such relations are typically established using the fooling technique, wherein we replace one input subtree such that the overall surrounding computation is not affected. Then we observe an undesired effect in the output tree computed for this modified input tree unless the desired linking structure is in place. More precisely, it is usually demonstrated that by such an exchange an undesired tree pair becomes part of the computed tree relation. In this sense, negative information (knowledge that a certain tree pair is *not* in the computed tree relation) is usually necessary for the fooling technique. On the contrary our linking theorems only require positive information (certain tree pairs are in the computed tree relation) and then conclude the existence of certain interrelated links L. These links L represent an underspecification of the expected dependency $\langle t, I, u\rangle$ (i.e., $L \subseteq I$). These links L can then be used to prove that certain tree relations cannot be computed by compositions of MBOT as demonstrated in Section 7.

The absence of negative information in our linking theorems makes them applicable to a wide array of (similar) tree relations. On the contrary, whenever negative information is required, the tree relation typically has to be completely specified. This yields that proofs made for those tree relations do not easily (or automatically) generalize to similar tree relations. Our approach does not require negative information and can thus be applied to all tree relations that include our required tree pairs. Naturally, additional techniques such as the fooling technique might still be necessary to conclude the desired statement from the expected links, but our linking theorems establish the expected links, which relieves us from the effort to individually establish them in each proof. In this sense, our linking theorems allow us to focus on the high-level argument and allow for very nice and intuitive proofs of negative results about the expressive power of compositions of MBOT (as demonstrated in Section 7).

We start with simple utility definitions and statements. The definitions establish two simple properties of tree languages, which we will require in our linking theorems. The utility statements observe simple properties of general trees. In particular, we show that given a large tree there exists a large subtree at a certain depth.

Definition 6 A tree $t \in T_\Sigma$ is a *chain* (or unary tree) if $\mathrm{pos}(t) \subseteq [1]^*$, and t is a *binary* tree if $\mathrm{pos}(t) \subseteq [2]^*$.[11] A tree language $T \subseteq T_\Sigma$ is

- *unary shape-complete* if for every chain $t \in T_\Sigma$ there exists a tree $t' \in T$ with $\mathrm{pos}(t') = \mathrm{pos}(t)$, and
- *binary shape-complete* if for every binary tree $t \in T_\Sigma$ there exists a tree $t' \in T$ with $\mathrm{pos}(t') = \mathrm{pos}(t)$.

□

A unary shape-complete tree language T contains chains of any length, thus for any chain $t \in T_\Sigma$ we can find a chain $t' \in T$ such that $|t'| = |t| = \mathrm{ht}(t) + 1 = \mathrm{ht}(t') + 1$. In a binary shape-complete tree language T the situation is similar because for every binary tree $t \in T_\Sigma$ it contains a tree $t' \in T$ with $\mathrm{pos}(t') = \mathrm{pos}(t)$, which also yields $|t'| = |t|$ and $\mathrm{ht}(t') = \mathrm{ht}(t)$. These two properties will be essential in our linking theorems. Next, we recall two simple properties of trees.

Lemma 3 *Let $a \in \mathbb{N}_+$ and $t \in T_\Sigma$ be such that $\mathrm{pos}(t) \subseteq [a]^*$. For every $i \in \mathbb{N}$ with $i \leq \mathrm{ht}(t)$ there exists a position $w \in \mathrm{pos}(t)$ such that $|w| = i$ and $|t|_w| \geq \frac{|t|}{a^i} - 1$.* □

Proof Let $i \in \mathbb{N}$ with $i \leq \mathrm{ht}(t)$, and let $W' = [a]^i \cap \mathrm{pos}(t)$ be the set of positions of t of length i. Clearly, $|W'| \leq |[a]^i| = a^i$. Similarly, $\sum_{j=0}^{i-1} |[a]^j| = \sum_{j=0}^{i-1} a^j \leq a^i$, so there are at most a^i positions of length strictly

[11] In other words, t is a chain if and only if $\mathrm{pos}(t)$ is a prefix chain.

smaller than i. For a contradiction, suppose that $|t|_{w'}| \leq \frac{|t|}{a^i} - 2$ for all $w' \in W'$. It follows that

$$|t| = |\mathrm{pos}(t)| = \left(\sum_{j=0}^{i-1} |\mathrm{pos}(t) \cap [a]^j|\right) + \sum_{w' \in W'} |\mathrm{pos}(t|_{w'})| \leq \left(\sum_{j=0}^{i-1} |[a]^j|\right) + \sum_{w' \in W'} |t|_{w'}| \leq a^i + \sum_{w' \in W'} \left(\frac{|t|}{a^i} - 2\right)$$
$$\leq a^i + a^i \cdot \left(\frac{|t|}{a^i} - 2\right) = a^i \cdot \left(\frac{|t|}{a^i} - 1\right) = |t| - a^i \leq |t| - 1 \ ,$$

which is clearly a contradiction. Hence there exists $w \in W'$ such that $|t|_w| \geq \frac{|t|}{a^i} - 1$ as required. ∎

Corollary 4 (of Lemma 3) *Let $a \in \mathbb{N}_+$ and $t \in T_\Sigma$ be such that $a \geq 2$ and $\mathrm{pos}(t) \subseteq [a]^*$. For every $i \in \mathbb{N}$ with $i \leq \mathrm{ht}(t)$, if $|t| \geq a^{i+1}$, then there exists a position $w \in \mathrm{pos}(t)$ such that $|w| = i$ and $|t|_w| \geq \frac{|t|}{a^{i+1}}$.* □

Proof By Lemma 3 there exists a position $w \in \mathrm{pos}(t)$ such that $|w| = i$ and $|t|_w| \geq \frac{|t|}{a^i} - 1$. Consequently,

$$|t|_w| \geq \frac{|t|}{a^i} - 1 \geq \frac{|t|}{a^i} - \frac{|t|}{a^{i+1}} = \frac{|t|}{a^i}\left(1 - \frac{1}{a}\right) \geq \frac{|t|}{a^{i+1}}$$

using the assumption $|t| \geq a^{i+1}$, which yields $\frac{|t|}{a^{i+1}} \geq 1$, and $a \geq 2$, which yields $1 - \frac{1}{a} \geq \frac{1}{a}$. ∎

6.1. Linking theorem for ε-free XTOP$^{\mathrm{R}}$

We now start with the linking theorem for ε-free XTOP$^{\mathrm{R}}$. It is known that ε-free XTOP$^{\mathrm{R}}$ (and many important subclasses) are not closed under composition [33] and that their composition hierarchy collapses at the third power [16]. However, since some important subclasses (e.g., ε-free n-XTOP) have an infinite composition hierarchy [15], we state and prove our linking theorem for the composition of arbitrarily many ε-free XTOP$^{\mathrm{R}}$. Our linking theorem is only applicable to tree relations, which contain a sub-relation that is obtained with the help of an input and an output context into which we can plug trees from a unary shape-complete tree language. If such a tree relation τ is computed by a composition $\tau = M_1;\cdots;M_k$ of ε-free XTOP$^{\mathrm{R}}$ $M_1, \ldots, M_k$, then we can deduce dependencies with the natural links relating the corresponding subtrees of the contexts.

Theorem 5 *Let Σ be an alphabet, $k, n \in \mathbb{N}_+$, and $M_1, \ldots, M_k$ be ε-free* XTOP$^{\mathrm{R}}$ *over Σ such that*

$$\{\langle c[t_1, \ldots, t_n], c'[t_1, \ldots, t_n]\rangle \mid t_1 \in T_1, \ldots, t_n \in T_n\} \subseteq M_1 ; \cdots ; M_k$$

for some $c, c' \in C_\Sigma(X_n)$ and unary shape-complete tree languages $T_1, \ldots, T_n \subseteq T_\Sigma$. Then there exist trees $t_1 \in T_1, \ldots, t_n \in T_n$, dependencies $\langle u_0, I_1, u_1\rangle \in \mathcal{D}(M_1)$, $\langle u_1, I_2, u_2\rangle \in \mathcal{D}(M_2)$, …, $\langle u_{k-1}, I_k, u_k\rangle \in \mathcal{D}(M_k)$ with $u_0 = c[t_1, \ldots, t_n]$ and $u_k = c'[t_1, \ldots, t_n]$, and a family of links $(v_{ij}, w_{ij}) \in I_j$ for $i \in [n]$ and $j \in [k]$ such that for all $i \in [n]$

- $\mathrm{pos}_{x_i}(c') \preceq w_{ik}$,
- $v_{i(j+1)} \preceq w_{ij}$ *for all* $j \in [k-1]$, *and*
- $\mathrm{pos}_{x_i}(c) \preceq v_{i1}$. □

Let us attempt a quick proof overview before we present the actual proof. Since we only use the prefix order $\preceq$ in Theorem 5 and the link properties (Definitions 4 and 5), we will sometimes omit it in the following discussion and proof. Consequently, when positions are incomparable, we mean to say that they are incomparable with respect to $\prec$.

We chose the plugged trees $t_1, \ldots, t_n$ to be long chains, but such that all of them differ markedly in length (and thus also in size). Due to their size and the strict link distance in the output, we can conclude a link induced by the last XTOP$^{\mathrm{R}}$ M_k pointing into each of them in the output. Since those output positions are naturally pairwise incomparable, also the corresponding linked input positions are pairwise incomparable because the dependencies of the XTOP$^{\mathrm{R}}$ M_k are strictly input hierarchical by Corollary 1. Using Corollary 4 and Theorem 4 we can then estimate the size of the linked input subtree. Then we repeat these arguments until we established links into the input tree of the first XTOP$^{\mathrm{R}}$ M_1 and approximated those input subtree

sizes. These positions in the input tree are again pairwise incomparable, and the approximated sizes of the subtrees are still larger than the input context c, which yields that each such input position v is comparable to an occurrence of a variable in the input context c. The markedly different sizes still exist after all the approximations, so we can correctly associate the subtrees to the correct variables. If the largest subtree was plugged for x_i in c', then the occurrence of x_i in c must be comparable to the input position of the corresponding link because all other subtrees are simply too small. In this way we establish all the links relating the reproduced plugged trees. However, some of those links might still point to a prefix of the correct variable position. For such a link we use the size approximation once more to conclude that there must be more links in the corresponding subtrees. If the input position of this newly obtained link is still a prefix of the correct variable occurrence in the input, then we repeat the procedure until it no longer is. The approximated size is suitably large to allow enough iterations of this procedure. We note that we intentionally use different constants to illustrate their influence. In the second linking theorem (see Theorem 6) we will use a universal constant.

Proof (of Theorem 5) Let $m > \max_{j\in[k]} \deg^{\mathrm{o}}(M_j)$ be the maximal output degree of all the $\textsc{xtop}^{\mathrm{R}}$. Note that $m \geq 2$. Let $b > 1$ be such that $\mathcal{D}(M_j)$ has strict link distance b in the output (see Theorem 3) for all $j \in [k]$. Theorem 4 shows that for each of the ε-free $\textsc{xtop}^{\mathrm{R}}$ $M_1, \ldots, M_k$ the size of an output tree is linearly bounded by the size of the linked input tree. Let $a > 1$ be an upper bound for all these linear factors such that $z > |c|$, where $z = a \cdot m^{b+2}$ is our major constant.

Now we are ready to select the trees $t_1 \in T_1, \ldots, t_n \in T_n$. For each $i \in [n]$, let $t_i \in T_i$ be a chain such that $|t_i| = z^{2ki+k}$. Since $T_1, \ldots, T_n$ are unary shape-complete, such trees exist. We obtain the input tree $u_0 = c[t_1, \ldots, t_n]$ and the output tree $u_k = c'[t_1, \ldots, t_n]$. Since $\langle u_0, u_k\rangle \in M_1 \,;\cdots;\, M_k$ by assumption, there exist dependencies $\langle u_0, I_1, u_1\rangle \in \mathcal{D}(M_1)$, $\ldots$, $\langle u_{k-1}, I_k, u_k\rangle \in \mathcal{D}(M_k)$.

Next, we prove an *auxiliary statement*, which states that given $j \in [k]$, $\ell \in \mathbb{N}_+$, and $w' \in \mathrm{pos}(u_j)$ such that $|u_j|_{w'}| \geq z^{\ell+j}$, there exists a link $(v, w) \in I_j$ such that $w' \preceq w$ and $|u_{j-1}|_v| \geq z^{\ell+j-1}$. Clearly, $\mathrm{ht}(u_j|_{w'}) > b$ because $|u_j|_{w'}| \geq z \geq m^{b+2}$. By Corollary 4 (applied for $a \leftarrow m$, $t \leftarrow u_j|_{w'}$, and $i \leftarrow b+1$), which is applicable because $m^{b+2} \leq z \leq z^{\ell+j} \leq |u_j|_{w'}|$, there exists a position $w'' \in \mathrm{pos}(u_j|_{w'})$ such that $|w''| = b+1$ and

$$|u_j|_{w'w''}| \geq \frac{|u_j|_{w'}|}{m^{b+2}} \geq \frac{z^{\ell+j}}{m^{b+2}} = \frac{a^{\ell+j} \cdot (m^{b+2})^{\ell+j}}{m^{b+2}} = a^{\ell+j} \cdot (m^{b+2})^{\ell+j-1} \quad .$$

Thus, we have the dependency $\langle u_{j-1}, I_j, u_j\rangle \in \mathcal{D}(M_j)$ and two positions $w', w'w'' \in \mathrm{pos}(u_j)$ with $w' \prec w'w''$ and $|w'w''| - |w'| = |w''| = b+1$. Since $\mathcal{D}(M_j)$ has strict link distance b in the output, there exists a link $(v, w) \in I_j$ such that $w' \prec w \prec w'w''$, which yields that $|u_j|_w| \geq |u_j|_{w'w''}| \geq a^{\ell+j} \cdot (m^{b+2})^{\ell+j-1}$. By Theorem 4, which establishes the linear output size bound between linked subtrees, applied to the dependency $\langle u_{j-1}, I_j, u_j\rangle \in \mathcal{D}(M_j)$ and the link $(v, w) \in I_j$ we obtain that

$$|u_{j-1}|_v| \geq \frac{|u_j|_w|}{a} \geq \frac{a^{\ell+j} \cdot (m^{b+2})^{\ell+j-1}}{a} = z^{\ell+j-1} \quad ,$$

which proves the auxiliary statement.

Now we repeatedly apply the auxiliary statement to prove the *main statement*, which states that for all $i \in [n]$ and $j \in [k]$ there exists a link $(v_{ij}, w_{ij}) \in I_j$ with

- $\mathrm{pos}_{x_i}(c') \preceq w_{ik}$,
- $v_{i(j+1)} \preceq w_{ij}$ if $j \in [k-1]$,
- the positions $v_{1j}, \ldots, v_{nj}$ are pairwise incomparable, and
- $|u_{j-1}|_{v_{ij}}| \geq z^{2ki+j-1}$

by downward induction on j starting at k. In the induction base, we have $j = k$, and for every $i \in [n]$ we select the position $w_i = \mathrm{pos}_{x_i}(c')$. Clearly, the positions $w_1, \ldots, w_n$ are pairwise incomparable. For each $i \in [n]$ we have $|u_k|_{w_i}| = |t_i| = z^{2ki+k}$ as required in the auxiliary statement (that we apply with $w' \leftarrow w_i$, $j \leftarrow k$, and $\ell \leftarrow 2ki$), so we conclude that there exists a link $(v_{ik}, w_{ik}) \in I_k$ such that $\mathrm{pos}_{x_i}(c') = w_i \preceq w_{ik}$ and $|u_{k-1}|_{v_{ik}}| \geq z^{2ki+k-1}$ as required for the main statement. Moreover, the positions $v_{1k}, \ldots, v_{nk}$ are

pairwise incomparable since I_k is strictly input hierarchical by Corollary 1 and $w_{1k}, \ldots, w_{1n}$ are pairwise incomparable, which establishes the induction base because the precondition of the second item $k \in [k-1]$ is clearly not satisfied. Now we assume that the statement is true for $j+1 \in [k]$ and prove it for j. By the induction hypothesis we have that $|u_j|_{v_{i(j+1)}}| \geq z^{2ki+j}$ for all $i \in [n]$. Thus, for every $i \in [n]$ we can again apply the auxiliary statement (with $w' \leftarrow v_{i(j+1)}$, $j \leftarrow j$, and $\ell \leftarrow 2ki$) to obtain that there exists a link $(v_{ij}, w_{ij}) \in I_j$ such that $v_{i(j+1)} \preceq w_{ij}$, thereby establishing the second item, and $|u_{j-1}|_{v_{ij}}| \geq z^{2ki+j-1}$. Since I_j is strictly input hierarchical by Corollary 1 and the positions $v_{1(j+1)}, \ldots, v_{n(j+1)}$ are pairwise incomparable by the induction hypothesis, we obtain that also $v_{1j}, \ldots, v_{nj}$ are pairwise incomparable, which proves the main statement.

Consequently, we already established the first two items of the theorem. In addition, we have that $|u_0|_{v_{i1}}| \geq z^{2ki}$ for every $i \in [n]$ and that the positions $v_{11}, \ldots, v_{n1}$ are pairwise incomparable. Since for each $i \in [n]$ the subtree $u_0|_{v_{i1}}$ has size at least $z > |c|$, it follows that each of the positions $v_{11}, \ldots, v_{n1}$ is comparable to at least one of the positions $\mathrm{pos}_{x_1}(c), \ldots, \mathrm{pos}_{x_n}(c)$ because the context c itself is too small to contain $u_0|_{v_{i1}}$ completely. For each $i \in [n]$, all elements of $W_i = \{w \in \mathrm{pos}(u_0) \mid w \preceq \mathrm{pos}_{x_i}(c) \text{ or } \mathrm{pos}_{x_i}(c) \preceq w\}$ are pairwise comparable (i.e., W_i is a prefix chain) since $u_0|_{\mathrm{pos}_{x_i}(c)} = t_i$ is a chain.[12] Since the positions $v_{11}, \ldots, v_{n1}$ are pairwise incomparable, we know that each $v_{11}, \ldots, v_{n1}$ is comparable to exactly one position of $\mathrm{pos}_{x_1}(c), \ldots, \mathrm{pos}_{x_n}(c)$. Now we can use the size bounds to make the association precise. Let us show for every $i \in [n]$ that v_{i1}, for which we know that $|u_0|_{v_{i1}}| \geq z^{2ki}$, cannot be comparable to any of the positions $\mathrm{pos}_{x_1}(c), \ldots, \mathrm{pos}_{x_{i-1}}(c)$. If $i = 1$, then we are immediately done. Otherwise, $i > 1$, and if we suppose that v_{i1} is comparable to such a position, then

$$z^{2ki} \leq |u_0|_{v_{i1}}| \leq |t_{i-1}| + |c| < z^{2k(i-1)+k} + z = z^{2ki-k} + z \overset{(\dagger)}{\leq} z^{2ki-k+1}$$

because v_{i1} can only be comparable to one position of $\{\mathrm{pos}_{x_1}(c), \ldots, \mathrm{pos}_{x_{i-1}}(c)\}$ and $\mathrm{pos}_{x_{i-1}}(c)$ is the root of the largest such subtree to which we can potentially add part of c. The approximation marked (†) is valid because $2 \leq z \leq z^{k(2i-1)}$. Simplifying the inequality $z^{2ki} < z^{2ki-k+1}$ we obtain $k < 1$, which is a contradiction. Consequently, the position v_{i1} can only be comparable to one of the positions of $\{\mathrm{pos}_{x_i}(c), \ldots, \mathrm{pos}_{x_n}(c)\}$. Thus we know that for every $i \in [n]$ the position v_{i1} is comparable only to $\mathrm{pos}_{x_i}(c)$.

We are almost done now. For all $i \in [n]$ with $\mathrm{pos}_{x_i}(c) \preceq v_{i1}$ [see third item in theorem] we are already done. Thus, let us consider $i \in [n]$ such that $v_{i1} \prec \mathrm{pos}_{x_i}(c)$. Reconsidering (the first part of the proof of) the auxiliary statement (with $w' \leftarrow v_{i2}$, $j \leftarrow 1$, $\ell \leftarrow 2ki$) that we used to obtain the link $(v_{i1}, w_{i1}) \in I_1$, which is applicable because $|u_1|_{v_{i2}}| \geq z^{2ki+1}$ by the main statement, we know that $|u_1|_{w_{i1}}| \geq a^{2ki+1} \cdot (m^{b+2})^{2ki}$. Consequently, $|u_1|_{w_{i1}}| \geq z^{2ki}$. Applying Lemma 2 (with $a \leftarrow m$ and $b \leftarrow b$) to the dependency $\langle u_0, I_1, u_1 \rangle \in \mathcal{D}(M_1)$ and the link $(v_{i1}, w_{i1}) \in I_1$ we can conclude that

$$|L| \geq \lceil \tfrac{|u_1|_{w_{i1}}|}{m^{b+1}} \rceil \geq \tfrac{|u_1|_{w_{i1}}|}{m^{b+1}} \geq \tfrac{z^{2ki}}{a \cdot m^{b+2}} = z^{2ki-1} \quad ,$$

where $L = \{(v', w') \in I_1 \mid v_{i1} \preceq v',\ w_{i1} \preceq w'\}$. Since M_1 is ε-free, for each position $v \in \mathrm{pos}(u_0)$ there is at most one link $(v, w) \in I_1$ by Lemma 1. However, L contains at least $z^{2ki-1} \geq z > |c|$ links, so it must have at least one link $(v'_{i1}, w'_{i1}) \in L$ such that $\mathrm{pos}_{x_i}(c) \preceq v'_{i1}$. By definition of L we have $v_{i1} \preceq v'_{i1}$ and $w_{i1} \preceq w'_{i1}$. Consequently, all requirements of the theorem are now satisfied by replacing, for every $i \in [n]$ with $v_{i1} \prec \mathrm{pos}_{x_i}(c)$, the link $(v_{i1}, w_{i1}) \in I_1$ by $(v'_{i1}, w'_{i1}) \in I_1$. ∎

6.2. Linking theorem for ε-free MBOT

Our second linking theorem concerns ε-free MBOT. Fortunately, ε-free MBOT and several relevant subclasses (different from $\mathrm{XTOP}^{\mathrm{R}}$ and its subclasses) are closed under composition [11], so we do not consider compositions of ε-free MBOT.[13] As in Theorem 5 our linking theorem applies to tree relations that contain a sub-relation induced by two contexts into which we can plug all trees of a certain type of tree language. However, this time we require binary shape-complete tree languages and a slightly different approach.

[12] Here we need the special chain shape of t_i. This property would no longer be true if we allow non-chains.

[13] However, we could, in principle, extend the technique (in the same way as in Theorem 5) to deal with compositions of ε-free MBOT. The construction of the trees $t_1, \ldots, t_n$ becomes quite a bit more difficult in the extension.

Theorem 6 *Let $n \in \mathbb{N}_+$ and $M = (Q, \Sigma, Q_0, R)$ be an ε-free* MBOT *such that*

$$\{\langle c[t_1, \ldots, t_n]\,,\, c'[t_1, \ldots, t_n]\rangle \mid t_1 \in T_1, \ldots, t_n \in T_n\} \subseteq M$$

for some $c, c' \in C_\Sigma(X_n)$ and binary shape-complete tree languages $T_1, \ldots, T_n \subseteq T_\Sigma$. Then there exist trees $t_1 \in T_1, \ldots, t_n \in T_n$, a dependency $\langle u, I, u'\rangle \in \mathcal{D}(M)$ with $u = c[t_1, \ldots, t_n]$ and $u' = c'[t_1, \ldots, t_n]$, and links $(v_1, w_1), \ldots, (v_n, w_n) \in I$ such that $\mathrm{pos}_{x_i}(c) \preceq v_i$ and $\mathrm{pos}_{x_i}(c') \preceq w_i$ for every $i \in [n]$. □

Let us illustrate the proof approach first. Again, we select the plugged trees $t_1, \ldots, t_n$ to be large and tall such that different trees have suitably different sizes and heights, but this time we also need another particularity. Namely, the largest tree is also the least tall and the smallest tree in size is the tallest tree of $\{t_1, \ldots, t_n\}$. In other words, we establish a reciprocal relation between the size and the height. To these trees we add a long chain to the top. In this way, we can conclude that there are suitably many links pointing into the long chain, which need to link to a position below a variable occurrence in the input context (using the fact that $\mathcal{D}(M)$ for an MBOT M is strictly link distance bounded and strictly output hierarchical). Next, we use the height and size of the plugged tree to establish an approximation for the linked subtree in the input using Theorem 4. It shows that only one plugged tree (the one pointed to in the output) fulfills the restrictions such obtained, which establishes the desired links. This time we use a universal constant $a \in \mathbb{N}_+$ for simplicity.

Proof (of Theorem 6) Theorem 4 shows that the size and height of an output tree are linearly bounded by the size and height of the linked input tree, respectively. Using Theorems 3 and 4 we can conclude that there exists $a \in \mathbb{N}_+$ such that

- $a > \deg^{\mathrm{o}}(M)$ and $a > \mathrm{ht}(c) + 1$,
- a is larger than the factor of the size relation (Item 1. in Theorem 4),
- a is larger than the factor of the height relation (Item 2. in Theorem 4), and
- $\mathcal{D}(M)$ has strict link distance a in the output (Theorem 3).

Note that $a > 2$. Finally, let $z \in \mathbb{N}_+$ be such that $z > \max(2a^2 + 1, (4n)^2)$. From $z > (4n)^2$ we conclude $\sqrt{z} > 4n$ and thus $z = \sqrt{z} \cdot \sqrt{z} > 4n \cdot \sqrt{z} > 4n \cdot \log_2(\sqrt{z}) = \log_2(z^{2n})$, which yields $2^z > z^{2n}$. Using these inequalities we obtain for all $i \in \mathbb{N}_+$

$$z^i > (a^2 + a^2 + 1) \cdot z^{i-1} > az^{i-1} + a^2 + z^{i-1} > az^{i-1} + a^2 \tag{†}$$

$$2^{(az^i+1)} > 2^{az} > 2^{a+z} = 2^a \cdot 2^z > a \cdot z^{2n} \geq az^{2n-i+1}\ . \tag{‡}$$

Now we prepare the selection of the trees $t_1, \ldots, t_n$. In the following, let $i \in [n]$ be arbitrary. We note that $az^i < az^{2n-i+1} < 2^{(az^i+1)}$ by (‡). Consequently, there exists a binary tree $t_i'' \in T_\Sigma$ such that $\mathrm{ht}(t_i'') = az^i$ and $|t_i''| = az^{2n-i+1}$ because $\mathrm{ht}(t_i'') < |t_i''| < 2^{\mathrm{ht}(t_i'')+1}$. Let $t_i' = \gamma^{(a^2)}(t_i'')$ for some arbitrary $\gamma \in \Sigma$. Since T_i is binary shape-complete, there also exists a tree $t_i \in T_i$ with $\mathrm{pos}(t_i) = \mathrm{pos}(t_i')$. Note that $\mathrm{ht}(t_i) = az^i + a^2$ and $|t_i| = az^{2n-i+1} + a^2$. Hence we have trees $t_1 \in T_1, \ldots, t_n \in T_n$, and we let $u = c[t_1, \ldots, t_n]$ be the input tree and $u' = c'[t_1, \ldots, t_n]$ be the output tree. By assumption, we know that $\langle u, u'\rangle \in M$, so there exists a dependency $\langle u, I, u'\rangle \in \mathcal{D}(M)$.

Now we start our analysis with the output tree u'. For each $i \in [n]$, let $p_i' = \mathrm{pos}_{x_i}(c')$ be the unique occurrence of x_i in the context c'. Then $\mathrm{ht}(u'|_{p_i'}) = \mathrm{ht}(t_i) = az^i + a^2 > a^2$ and $1^{(a^2)} \in \mathrm{pos}(u'|_{p_i'}) = \mathrm{pos}(t_i)$. Consequently, there exist a links $(v_{i1}, w_{i1}), \ldots, (v_{ia}, w_{ia}) \in I$ such that $p_i' \preceq w_{i1} \prec \cdots \prec w_{ia} \preceq p_i'.1^{(a^2)}$ because $\mathcal{D}(M)$ has strict link distance a in the output. Since $\mathcal{D}(M)$ is strictly output hierarchical, we additionally obtain that $v_{i1} \preceq \cdots \preceq v_{ia}$, and together with Lemma 1 we can even conclude that $v_{i1} \prec \cdots \prec v_{ia}$ because all elements of $\{w_{i1}, \ldots, w_{ia}\}$ are pairwise comparable. We select the link $(v_i, w_i) = (v_{ia}, w_{ia}) \in I$, so $\mathrm{pos}_{x_i}(c') = p_i' \prec w_{ia}$ as required. Moreover, $w_{ia} \preceq p_i'.1^{(a^2)}$ and thus

$$\mathrm{ht}(u'|_{w_i}) \geq \mathrm{ht}(t_i'') = az^i \qquad \text{and} \qquad |u'|_{w_i}| = |t_i''| = az^{2n-i+1}\ .$$

Next we apply Theorem 4 to the dependency $\langle u, I, u'\rangle \in \mathcal{D}(M)$ and the link $(v_i, w_i) \in I$ to obtain that

$$\mathrm{ht}(u|_{v_i}) \geq \frac{\mathrm{ht}(u'|_{w_i})}{a} - 2 \geq \frac{az^i}{a} - 2 = z^i - 2 \qquad \text{and} \qquad |u|_{v_i}| \geq \frac{|u'|_{w_i}|}{a} \geq \frac{az^{2n-i+1}}{a} = z^{2n-i+1}\ .$$

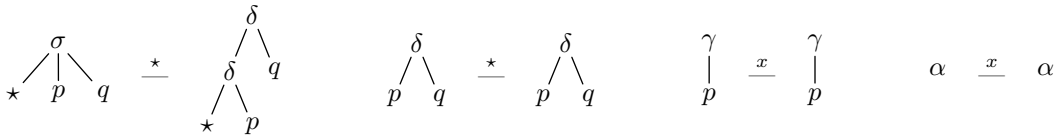

Figure 5: Rules of the n-XTOP M_1 of Example 5.

Since $v_{i1} \prec \cdots \prec v_{ia} = v_i$, we obtain that $|v_i| \geq a - 1 \geq \mathrm{ht}(c) + 1$. Consequently, $v_i \notin \mathrm{pos}(c)$ and thus there exists a (unique) $j \in [n]$ such that $\mathrm{pos}_{x_j}(c) \prec v_i$; i.e., a variable occurrence of c must be a strict prefix of the position v_i. The uniqueness is a simple consequence of the fact that the variable occurrences in c are pairwise incomparable. For every $j \in [n]$, let $p_j = \mathrm{pos}_{x_j}(c)$ be the unique occurrence of x_j in the source-side context c. Moreover, for every $i \in [n]$, let $j_i \in [n]$ be the unique integer such that $\mathrm{pos}_{x_{j_i}}(c) = p_{j_i} \prec v_i$. Next, we prove that $j_i \geq i$ using the height approximation. For the sake of a contradiction, suppose that $j_i < i$. Since $p_{j_i} \prec v_i$, we have $\mathrm{ht}(u|_{p_{j_i}}) > \mathrm{ht}(u|_{v_i}) \geq z^i - 2 > az^{i-1} + a^2$, where the last inequality is due to (†).[14] However,

$$\mathrm{ht}(u|_{p_{j_i}}) > az^{i-1} + a^2 \geq az^{j_i} + a^2 = \mathrm{ht}(t_{j_i}) = \mathrm{ht}(u|_{p_{j_i}}) \ ,$$

which is a contradiction. Thus $j_i \geq i$. Similarly, we will finally show that $j_i \leq i$ using the size approximation. Again, for the sake of a contradiction, suppose that $j_i > i$. Since we have $p_{j_i} \prec v_i$ we immediately obtain that $|u|_{p_{j_i}}| > |u|_{v_i}| \geq z^{2n-i+1} > az^{2n-i} + a^2$, where the last inequality is again due to (†). However,

$$|u|_{p_{j_i}}| > az^{2n-i} + a^2 \geq az^{2n-j_i+1} + a^2 = |t_{j_i}| = |u|_{p_{j_i}}| \ ,$$

which is again a contradiction. Thus $i \leq j_i \leq i$, which shows that $j_i = i$. Taking all the pieces together, the link $(v_i, w_i) \in I$ obeys $\mathrm{pos}_{x_i}(c) = p_i = p_{j_i} \prec v_i$ and $\mathrm{pos}_{x_i}(c') = p'_i \prec w_{ia} = w_i$ as required. ∎

7. Applications of the linking theorems

In this section, we present some applications of our linking theorems to existing results of the literature. We start with a classical result of [4]. Figure 6 illustrates the counterexample, which [4] use to show that the class of tree relations computed by XTOP$^{\mathrm{R}}$ (as well as those computed by n-XTOP) is not closed under composition. Our linking technique only applies to ε-free XTOP$^{\mathrm{R}}$, but offers a fully formalized, but still very intuitive proof of this claim. We start by recalling the particular tree relation τ, which can be computed by two ε-free n-XTOP.

Example 5 (see [4, Section 3.4]) Let $M_1 = (Q, \Sigma, \{\star\}, R_1)$ and $M_2 = (Q, \Sigma, \{\star\}, R_2)$ be the ε-free XTOP$^{\mathrm{R}}$ with $Q = \{\star, p, q, r\}$ and $\Sigma = \{\sigma, \delta, \gamma, \alpha\}$, where

- R_1 contains exactly the following rules for all $x \in \{p, q\}$:

$$\sigma(\star, p, q) \xrightarrow{\star} \delta(\delta(\star, p), q) \qquad \delta(p, q) \xrightarrow{\star} \delta(p, q) \qquad \gamma(p) \xrightarrow{x} \gamma(p) \qquad \alpha \xrightarrow{x} \alpha \ ,$$

- R_2 contains exactly the following rules for all $x \in \{q, p, r\}$:

$$\delta(r, p) \xrightarrow{\star} \delta(r, p) \qquad \delta(\delta(r, p), q) \xrightarrow{r} \sigma(r, p, q) \qquad \gamma(p) \xrightarrow{x} \gamma(p) \qquad \alpha \xrightarrow{x} \alpha \ .$$

Clearly, both M_1 and M_2 are ε-free n-XTOP. The rules of M_1 are illustrated in Figure 5. □

Theorem 7 (see [4, Section 3.4]) *The tree relation $\tau = M_1 \,;\, M_2$ of Example 5 (also illustrated in Figure 6) cannot be computed by any ε-free* XTOP$^{\mathrm{R}}$. □

[14] Note that the 3 strict inequalities in (†) justify $z^i - 2 > az^{i-1} + a^2$.

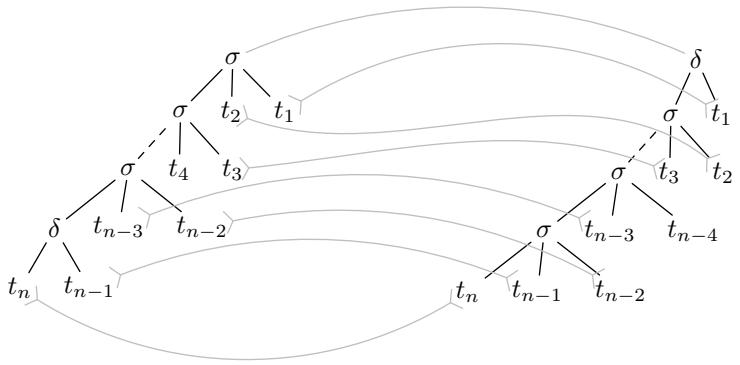

Figure 6: Counterexample relation of [4] with links, which we conclude from Theorem 5, where an inverse arrow head indicates that the link refers to a node (not necessarily the root) inside the subtree that the spline points to.

In [4] a more general version of this result for arbitrary XTOP$^{\mathrm{R}}$ is proved using as essential tool a lemma[15] establishing a normal form for XTOP$^{\mathrm{R}}$ that compute bijective tree relations [4, p. 47–50]. With this lemma the main proof [4, p. 53–54] is then rather straightforward. However, [4] also gives an intuitive explanation why a single XTOP$^{\mathrm{R}}$ cannot compute the tree relation τ [4, p. 52]. This intuitive explanation essentially delivers our proof approach since the linking theorem establishes the links required to formalize the presented intuition.

Proof (of Theorem 7) Suppose on the contrary that there exists an ε-free XTOP$^{\mathrm{R}}$ $M = (Q, \Sigma, Q_0, R)$ that computes τ. By Theorem 3, the set $\mathcal{D}(M)$ is link-distance bounded in the input, so let $b \in \mathbb{N}_+$ be such that $\mathcal{D}(M)$ has link distance b in the input. We let $n = 2b + 4$, and as in [4, p. 52], we select the contexts

$$c = \sigma(\sigma(\cdots\sigma(\delta(x_n, x_{n-1}), x_{n-2}, x_{n-3})\cdots, x_4, x_3), x_2, x_1)$$
$$c' = \delta(\sigma(\sigma(\cdots\sigma(x_n, x_{n-1}, x_{n-2})\cdots, x_5, x_4), x_3, x_2), x_1)$$

and the unary shape-complete tree languages $T_1 = \cdots = T_n = T$, where $T = \{\gamma^k(\alpha) \mid k \in \mathbb{N}\}$. Consequently, we meet the requirements of Theorem 5 and its application yields trees $t_1, \ldots, t_n \in T$, a dependency $\langle u, I, u' \rangle \in \mathcal{D}(M)$ with $u = c[t_1, \ldots, t_n]$ and $u' = c'[t_1, \ldots, t_n]$, and links $(v_1, w_1), \ldots, (v_n, w_n) \in I$ such that $\mathrm{pos}_{x_i}(c') \preceq w_i$ and $\mathrm{pos}_{x_i}(c) \preceq v_i$ for all $i \in [n]$. This situation is also depicted in Figure 6, where an inverse arrow head indicates that the link refers to some position inside the subtree that it (or better: its graphical representation) points at.

This was the preparation. Now we start the argument. We observe that $(\varepsilon, \varepsilon) \in I$ and $(v_n, w_n) \in I$. By the selection of c, we have $|v_n| > b$, and since $\mathcal{D}(M)$ has link distance b in the input, there exists another link $(v, w) \in I$ such that $\varepsilon \prec v \prec v_n$ and $|v| \leq b$. Consequently, $v = 1^m$ for some $m \in [b]$. Moreover, we observe that $v \prec v_{2m+2}$ and $v \prec v_{2m+1}$ because $v \prec \mathrm{pos}_{x_{2m+2}}(c)$ and $v \prec \mathrm{pos}_{x_{2m+1}}(c)$. Since I is strictly input hierarchical by Corollary 1, we obtain $w \preceq w_{2m+2}$ and $w \preceq w_{2m+1}$, which by the shape of c' yields that $w = 1^k$ for some $k \leq m$. However, this also yields that $w \prec \mathrm{pos}_{x_{2m}}(c') \prec w_{2m}$. Since I is also strictly output hierarchical by Corollary 1, we conclude that $1^m = v \preceq v_{2m}$, which contradicts the shape of c. Thus, we derived the required contradiction and can conclude that such an ε-free XTOP$^{\mathrm{R}}$ cannot exist. ∎

Corollary 5 (of Example 5 and Theorem 7) *The class of tree relations computable by ε-free* XTOP$^{\mathrm{R}}$ *(or ε-free n-*XTOP*) is not closed under composition.* □

Let us apply Theorem 5 again by providing an alternative proof for Theorem 1 that utilizes our linking theorem. This new proof follows exactly the intuition (see Figure 7). Roughly speaking, the depicted chain

[15] certainly of separate interest

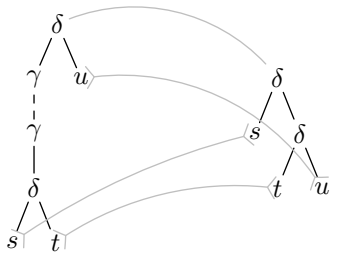

Figure 7: Counterexample relation of [33] with links, which we conclude from Theorem 5, where an inverse arrow head indicates that the link refers to a node (not necessarily the root) inside the subtree that the spline points to.

of γ-symbols in the input tree is, in general, too long to be covered by a single rule, so by a clever selection of the input tree there must be a link (v, w) in it. The input position v of that link dominates the subtrees s and t, so intuitively its output position w should also dominate the subtrees s and t in the output. However, only the root position dominates those two subtrees, so $w = \varepsilon$. But then w dominates the subtree u in the output, so intuitively it should also do so in the input, which allows us to conclude that $v = \varepsilon$ because that is the only position dominating also u in the input. But $v = \varepsilon$ contradicts that the link points to the γ-chain in the input. With the help of our linking theorem for XTOP^{R} we can make this informal argument formal as we will demonstrate next.

Proof (Alternative proof of Theorem 1) For the sake of a contradiction, we assume that there exists an ε-free XTOP^{R} $M = (Q, \Sigma, Q_0, R)$ that computes τ. By Theorem 3, the set $\mathcal{D}(M)$ is link-distance bounded in the input, so let $b \in \mathbb{N}_+$ be such that $\mathcal{D}(M)$ has link distance b in the input. We select the contexts

$$c = \delta(\gamma^b(\delta(x_1, x_2)), x_3) \qquad \text{and} \qquad c' = \delta(x_1, \delta(x_2, x_3))$$

and the unary shape-complete tree languages $T_1 = T_2 = T_3 = T$, where T is given in Theorem 1. Consequently, we meet the requirements of Theorem 5 and its application yields trees $t_1, t_2, t_3 \in T$, a dependency $\langle u, I, u' \rangle \in \mathcal{D}(M)$ with $u = c[t_1, t_2, t_3]$ and $u' = c'[t_1, t_2, t_3]$, and links $(v_1, w_1), (v_2, w_2), (v_3, w_3) \in I$ such that $\text{pos}_{x_j}(c') \preceq w_j$ and $\text{pos}_{x_j}(c) \preceq v_j$ for all $j \in \{1, 2, 3\}$. These links are already indicated in Figure 7, where an inverse arrow head indicates that the link refers to some position inside the subtree that it points at.

Now we can again derive a contradiction using those links. First, we observe that for $(v_1, w_1) \in I$ we have $|v_1| > b$ because $1^{b+2} = \text{pos}_{x_1}(c) \preceq v_1$. Since $(\varepsilon, \varepsilon), (v_1, w_1) \in I$ with $|v_1| > b$ and $\mathcal{D}(M)$ has link distance b in the input, there exists another link $(v, w) \in I$ such that $\varepsilon \prec v \prec v_1$ and $|v| \leq b$. Consequently, $v = 1^m$ for some $m \in [b]$, and thus $v \prec v_1$ and $v \prec v_2$ because $v = 1^m \prec 1^{b+1}.2 = \text{pos}_{x_2}(c)$. Since I is strictly input hierarchical by Corollary 1, we obtain that $w \preceq w_1$ and $w \preceq w_2$, which by the shape of c' yields that $w = \varepsilon$. However, this also yields that $w = \varepsilon \prec \text{pos}_{x_3}(c') \prec w_3$. Since I is also strictly output hierarchical by Corollary 1, we conclude that $v = 1^m \preceq 2 = \text{pos}_{x_3}(c) \preceq v_3$, which is a contradiction because $m \geq 1$. Thus, we derived the required contradiction and can conclude that such an ε-free XTOP^{R} cannot exist. ∎

In addition, Theorem 5 has been used in [16] to prove results about the composition closure of ε-free XTOP^{R}. Now let us also apply our linking theorem for ε-free MBOT (see Theorem 6) to an interesting counterexample, which is the inverse of abstract topicalization [1, 8, 20] and illustrated in Figure 9. This result is reported and discussed in [32]. Its consequences are discussed there as well. Here we are mostly interested in the ease of proving such results. In fact, the linking technique established here enables the proof of this result. We start by recalling abstract topicalization, which we present using the ε-free MBOT of Example 6.

Example 6 Let $M_{\text{tpc}} = (Q, \Sigma, \{\star\}, R)$ be the ε-free MBOT with $Q = \{\star, p, q, r\}$ and $\Sigma = \{\sigma, \delta, \gamma, \alpha\}$, where

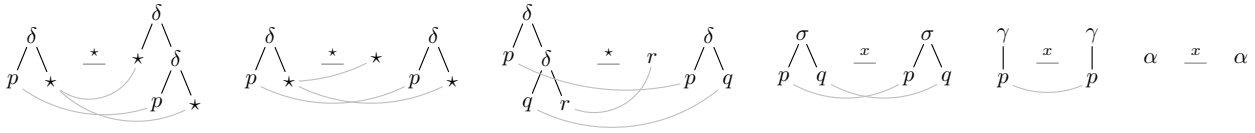

Figure 8: Rules of the MBOT M_{tpc} of Example 6.

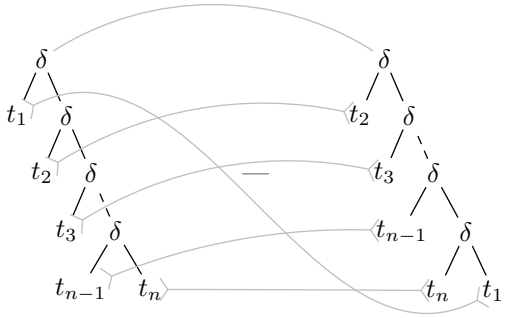

Figure 9: Counterexample relation M_{tpc}^{-1} of Example 6 with links, which we conclude from Theorem 6, where an inverse arrow head indicates that the link refers to a node (not necessarily the root) inside the subtree that the spline points to.

R contains exactly the following rules for every $x \in \{p, q, r\}$

$$\delta(p,\star) \xrightarrow{\star} \delta(\star,\delta(p,\star)) \qquad \delta(p,\star) \xrightarrow{\star} \star\,.\,\delta(p,\star) \qquad \delta(p,\delta(q,r)) \xrightarrow{\star} r\,.\,\delta(p,q)$$
$$\sigma(p,q) \xrightarrow{x} \sigma(p,q) \qquad \gamma(p) \xrightarrow{x} \gamma(p) \qquad \alpha \xrightarrow{x} \alpha\ .$$

The rules of M_{tpc} are illustrated in Figure 8, and the tree relation M_{tpc}^{-1} is illustrated in Figure 9. □

Theorem 8 (see [32, Theorem 8]) *The relation M_{tpc}^{-1} cannot be computed by any ε-free* MBOT. □

PROOF Again we suppose on the contrary that there exists an ε-free MBOT $M = (Q, \Sigma, Q_0, R)$ that computes M_{tpc}^{-1}. Let $b \in \mathbb{N}_+$ be such that $\mathcal{D}(M)$ has strict link distance b in the output (see Theorem 3). Moreover, let $n > b + 2$, and we select the contexts

$$c = \delta(x_1, \delta(x_2, \dots \delta(x_{n-1}, x_n) \cdots)) \qquad \text{and} \qquad c' = \delta(x_2, \delta(x_3, \dots \delta(x_{n-1}, \delta(x_n, x_1)) \cdots))$$

and the binary shape-complete tree languages $T_1 = \cdots = T_n = T$, where T is the smallest tree language such that $\alpha \in T$, $\gamma(t) \in T$ for all trees $t \in T$, and $\sigma(t_1, t_2) \in T$ for all trees $t_1, t_2 \in T$. At this point, we can apply Theorem 6 to obtain trees $t_1, \dots, t_n \in T$, a dependency $\langle u, I, u' \rangle \in \mathcal{D}(M)$ with $u = c[t_1, \dots, t_n]$ and $u' = c'[t_1, \dots, t_n]$, and a link $(v_i, w_i) \in I$ for every $i \in [n]$ such that $\text{pos}_{x_i}(c) \preceq v_i$ and $\text{pos}_{x_i}(c') \preceq w_i$. We depict this situation in Figure 9.

It remains to derive the contradiction. To this end, we first observe that $\varepsilon, 2^{b+1} \in \text{pos}(u')$ because $\text{pos}_{x_1}(c') = 2^{n-1} \in \text{pos}(u')$ and $n - 1 > b + 1$. Since $\mathcal{D}(M)$ has strict link distance b in the output, there exists a link $(v, w) \in I$ such that $\varepsilon \prec w \prec 2^{b+1}$. We also know that $w \prec \text{pos}_{x_n}(c') = 2^{n-2}.1 \preceq w_n$ and $w \prec \text{pos}_{x_1}(c') = 2^{n-1} \preceq w_1$ because $w \prec 2^{b+1} \preceq 2^{n-2}$. Now we can use that $\mathcal{D}(M)$ is strictly output hierarchical by Theorem 2, which applied to the links $(v, w), (v_1, w_1), (v_n, w_n) \in I$ yields that $v \preceq v_n$ and $v \preceq v_1$. However, $\text{pos}_{x_n}(c) \preceq v_n$ and $\text{pos}_{x_1}(c) \preceq v_1$, so the only possible selection for v is $v = \varepsilon$. Consequently, we have two links $(\varepsilon, \varepsilon), (\varepsilon, w) \in I$ with $\varepsilon \prec w$, which is contradicting Lemma 1. Thus, we derived the desired contradiction and such an ε-free MBOT computing M_{tpc}^{-1} cannot exist. ■

Corollary 6 (of Example 6 and Theorem 8) *The class of tree relations computable by ε-free* MBOT *is not closed under inverses.* □

Finally, let us present a more complex example to demonstrate the usefulness of the linking theorems. Namely we show that the relation M_{tpc} (abstract topicalization) cannot be computed by any composition of ε-free $\textsc{xtop}^{\mathrm{R}}$. This result is also reported in [32], and again relies on the linking technique presented here. An approach based on the fooling technique would be rather difficult (or hopeless in the eyes of the authors) in this case as we would need to argue over several (at least 2) unknown intermediate trees.

Theorem 9 (see [32, Theorem 6]) *The relation M_{tpc} cannot be computed by any chain of ε-free $\textsc{xtop}^{\mathrm{R}}$.*$\Box$

Proof Again we prove the statement by contradiction. Therefore, we assume that M_{tpc} is computed by a composition of several ε-free $\textsc{xtop}^{\mathrm{R}}$. By [16, Theorem 11] we know that 3 ε-free $\textsc{xtop}^{\mathrm{R}}$ suffice, so there are ε-free $\textsc{xtop}^{\mathrm{R}}$ M_1, M_2, and M_3 over Σ such that $M_{\mathrm{tpc}} = M_1 \,;\, M_2 \,;\, M_3$. Let $b \in \mathbb{N}_+$ be such that $\mathcal{D}(M_1)$, $\mathcal{D}(M_2)$, and $\mathcal{D}(M_3)$ have strict link distance b in the output (see Theorem 3). Moreover, let $n > (b+1)^3$, and (essentially as in the proof of Theorem 8) we select the contexts

$$c = \delta(x_2, \delta(x_3, \dots \delta(x_{n-1}, \delta(x_n, x_1)) \cdots)) \qquad \text{and} \qquad c' = \delta(x_1, \delta(x_2, \dots \delta(x_{n-1}, x_n) \cdots))$$

and the unary shape-complete tree languages $T_1 = \cdots = T_n = T$, where $T = \{\gamma^k(\alpha) \mid k \in \mathbb{N}\}$. Theorem 5 now yields trees $t_1, \dots, t_n \in T$ and dependencies $\langle u_0, I_1, u_1\rangle \in \mathcal{D}(M_1)$, $\langle u_1, I_2, u_2\rangle \in \mathcal{D}(M_2)$, and $\langle u_2, I_3, u_3\rangle \in \mathcal{D}(M_3)$ with $u_0 = c[t_1, \dots, t_n]$ and $u_3 = c'[t_1, \dots, t_n]$, and link $(v_{i1}, w_{i1}) \in I_1$, $(v_{i2}, w_{i2}) \in I_2$, and $(v_{i3}, w_{i3}) \in I_3$ for every $i \in [n]$ such that

- $\mathrm{pos}_{x_i}(c) \preceq v_{i1}$,
- $v_{i2} \preceq w_{i1}$ and $v_{i3} \preceq w_{i2}$, and
- $\mathrm{pos}_{x_i}(c') \preceq w_{i3}$.

We depict this situation in Figure 10.

Now we start the argumentation. Clearly, $\varepsilon, 2^{(b^3+2b^2+b+1)} \in \mathrm{pos}(u_3)$ because $\mathrm{pos}_{x_n}(c') = 2^{n-1}$ and $n > (b+1)^3 > b^3 + 2b^2 + b + 1$. We can now use the strict link distance b in the output to conclude the existence of $k' > (b+1)^2$ links $(v'_1, w'_1), \dots, (v'_{k'}, w'_{k'}) \in I_3$ such that $\varepsilon \prec w'_1 \prec \cdots \prec w'_{k'} \prec 2^{(b^3+2b^2+b+1)}$. Using the fact that $\mathcal{D}(M_3)$ is strictly output hierarchical by Theorem 2, we obtain $v'_1 \preceq \cdots \preceq v'_{k'}$ and with the help of Lemma 1 we can sharpen this relation to $v'_1 \prec \cdots \prec v'_{k'}$. These links are indicated as group (3) in Figure 10. Clearly, $v'_1, v'_{k'} \in \mathrm{pos}(u_2)$ with $v'_1 \prec v'_{k'}$ and $|v'_{k'}| - |v'_1| \geq (b+1)^2$. To these positions in u_2 we can now apply the knowledge that $\mathcal{D}(M_2)$ has strict link distance b in the output. Consequently, there are $k > b+1$ links $(v_1, w_1), \dots, (v_k, w_k) \in I_2$ such that $v'_1 \prec w_1 \prec \cdots \prec w_k \prec v'_{k'}$. Again, the fact that $\mathcal{D}(M_2)$ is strictly output hierarchical implies that $v_1 \preceq \cdots \preceq v_k$ and also this relation can be sharpened to $v_1 \prec \cdots \prec v_k$. These links are marked (2) in Figure 10. Since $v_1, v_k \in \mathrm{pos}(u_1)$ with $v_1 \prec v_k$ and $|v_k| - |v_1| > b$ and $\mathcal{D}(M_1)$ has strict link distance b in the output, we obtain another link $(v, w) \in I_1$ such that $v_1 \prec w \prec v_k$. This link is marked (1) in Figure 10. Now we established all the required links, and we refer the reader to Figure 10 for an overview of the links and their relation.

It remains to derive the contradiction. Since $w'_1 \prec \cdots \prec w'_{k'} \prec 2^{b^3+2b^2+b+1} \prec 2^{(b+1)^3-1} = \mathrm{pos}_{x_n}(c')$, we conclude that $w'_{k'} \prec w_{(n-1)3}$ and $w'_{k'} \prec w_{n3}$. Using that $\mathcal{D}(M_3)$ is strictly output hierarchical once again applied to the links $(v'_{k'}, w'_{k'}), (v_{(n-1)3}, w_{(n-1)3}), (v_{n3}, w_{n3}) \in I_3$, we obtain that $v'_{k'} \preceq v_{(n-1)3}$ and $v'_{k'} \preceq v_{n3}$. Moving to the $\textsc{xtop}^{\mathrm{R}}$ M_2, we observe that $w_1 \prec \cdots \prec w_k \prec v'_{k'} \preceq v_{(n-1)3} \preceq w_{(n-1)2}$ and $w_1 \prec \cdots \prec w_k \prec v'_{k'} \preceq v_{n3} \preceq w_{n2}$. Since $\mathcal{D}(M_2)$ is also strictly output hierarchical, the links $(v_k, w_k), (v_{(n-1)2}, w_{(n-1)2}), (v_{n2}, w_{n2}) \in I_2$ yield that $v_k \preceq v_{(n-1)2}$ and $v_k \preceq v_{n2}$. Finally, we need to iterate this process once more for the $\textsc{xtop}^{\mathrm{R}}$ M_1. We first observe that $w \prec v_k \preceq v_{(n-1)2} \preceq w_{(n-1)1}$ and $w \prec v_k \preceq v_{n2} \preceq w_{n1}$. Also the dependencies $\mathcal{D}(M_1)$ of M_1 are strictly output hierarchical, so applied to the links $(v, w), (v_{(n-1)1}, w_{(n-1)1}), (v_{n1}, w_{n1}) \in I_1$ we obtain $v \preceq v_{(n-1)1}$ and $v \preceq v_{n1}$. Together with $\mathrm{pos}_{x_{n-1}}(c) \preceq v_{(n-1)1}$ and $\mathrm{pos}_{x_n}(c) \preceq v_{n1}$ we obtain that $v \prec \mathrm{pos}_{x_1}(c') \preceq v_{11}$.

Now we start deriving another link relation. Since $\varepsilon \prec w'_1$, we have that $\mathrm{pos}_{x_1}(c')$ and w'_1 are incomparable, which is thus also true for the positions w_{13} and w'_1 because $\mathrm{pos}_{x_1}(c') \preceq w_{13}$. We know that $\mathcal{D}(M_3)$ is strictly input hierarchical by Corollary 1, so applied to the links $(v'_1, w'_1), (v_{13}, w_{13}) \in I_3$ we obtain that v'_1 and v_{13} are incomparable. Since $v'_1 \prec w_1$ and $v_{13} \preceq w_{12}$, also the positions w_1 and w_{12} are incomparable. Looking at the links $(v_1, w_1), (v_{12}, w_{12}) \in I_2$ we can conclude that v_1 and v_{12} are incomparable because also $\mathcal{D}(M_2)$ is strictly input hierarchical. Since $v_1 \prec w$ and $v_{12} \preceq w_{11}$ also the positions w and w_{11} are

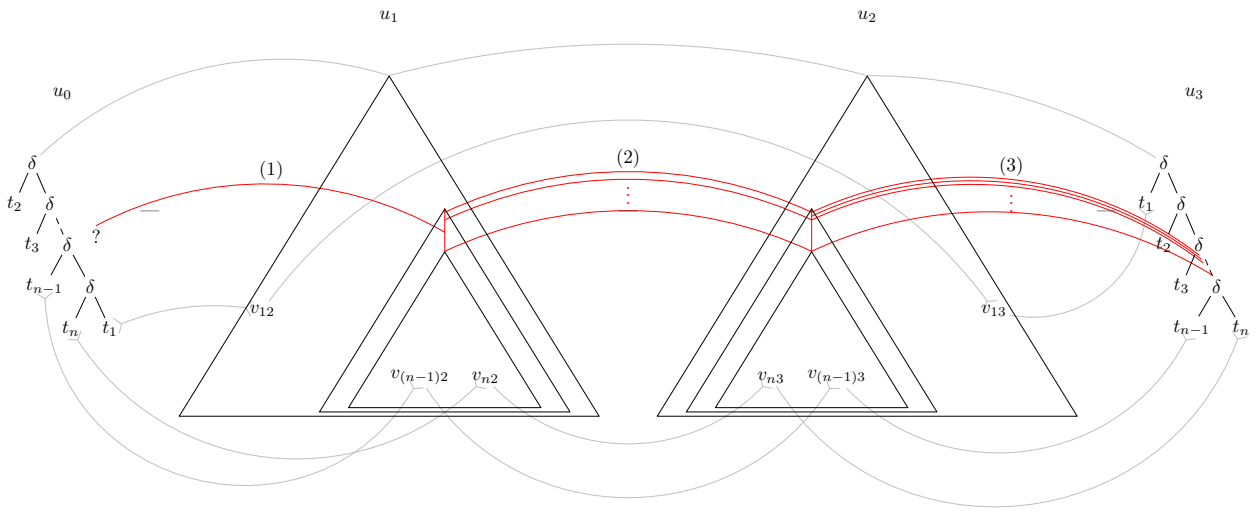

Figure 10: Illustration of the links discussed in the proof of Theorem 9. Inverted arrow heads indicate that the link points to a position below the one indicated by the spline.

incomparable. Applying that $\mathcal{D}(M_1)$ is strictly input hierarchical to the links $(v, w), (v_{11}, w_{11}) \in I_1$ we thus obtain that v and v_{11} are incomparable. However, this contradicts $v \prec v_{11}$, which we derived in the previous paragraph, so such ε-free XTOP$^{\mathrm{R}}$ M_1, M_2, and M_3 cannot exist. ∎

Summary and further research

We considered linear extended multi bottom-up tree transducers [4, 11, 26, 28], which are used, for example, as translation models in syntax-based statistical machine translation [6]. Following the tradition in that application area, we presented them in the form of synchronous grammars, in which the input tree and the output tree develop simultaneously in the derivation steps. During rule application, the left- and right-hand side contribute, respectively, to the input and to the output tree. To keep track of the input and output positions that are supposed to develop in parallel, links (v, w) are used, where v and w are, respectively, a position in the input and the output tree. In contrast to much of the literature, we collect all links created during the derivation of an input tree t and an output tree u in a set I. Roughly speaking, the links in I provide origin information; i.e., they record the parts of u that were generated due to a particular part of t. Making these links explicit, each MBOT computes a set of triples $\langle t, I, u\rangle$, each of which is called a dependency. We studied the structural properties of those dependencies for MBOT and its subclasses XTOP$^{\mathrm{R}}$ and n-XTOP. As expected, the links in these dependencies have a certain hierarchical organization (Theorem 2 and Corollary 1) and there exists a bound on the distances between linked positions both in the input and the output side (Theorem 3 and Corollary 2). Our main contributions are the linking theorems for compositions of ε-free XTOP$^{\mathrm{R}}$ (Theorem 5) and for ε-free MBOT (Theorem 6), which guarantee the existence of the natural links relating identical subtrees provided that a particular relation is a subset of the computed tree relation. To demonstrate the usefulness of those linking theorems, we use them to reprove classic results in an intuitive fashion and also showcase recent results that have been obtained with their help.

Our current linking theorems only apply to (compositions of) ε-free MBOT, for which the computed output tree is always in a linear relation to the input tree in terms of both size and height. This linear dependence is a key ingredient in the proofs of our linking theorems. The authors naturally expect that the essence of the linking theorems remains true also for (compositions of) arbitrary MBOT, but a relation similar to the linear size and height dependence would be required to establish those generalizations of our linking theorems (using our approach). Second, the authors believe that it will be interesting to see, which additional results can easily be reproved or even proved for the first time using our linking theorems.

Acknowledgements: The authors are indebted to Joost Engelfriet for his remarks on a draft version and to the reviewers for their constructive feedback.

References

[1] D. Adger, Core Syntax: A Minimalist Approach, Core Linguistics, Oxford University, 2003.
[2] R. Alur, P. Madhusudan, Adding nesting structure to words, J. ACM 56 (3) (2009) 1–43.
[3] A. Arnold, M. Dauchet, Bi-transductions de forêts, in: Proc. ICALP, Edinburgh University Press, 1976, pp. 74–86.
[4] A. Arnold, M. Dauchet, Morphismes et bimorphismes d'arbres, Theoret. Comput. Sci. 20 (1) (1982) 33–93.
[5] M. Bojańczyk, Transducers with origin information, in: Proc. ICALP, vol. 8573 of LNCS, Springer, 2014, pp. 26–37.
[6] F. Braune, A. Maletti, D. Quernheim, N. Seemann, Shallow local multi bottom-up tree transducers in statistical machine translation, in: Proc. ACL, Association for Computational Linguistics, 2013, pp. 811–821.
[7] D. Chiang, An introduction to synchronous grammars, Part of an ACL tutorial given with Kevin Knight (2006).
[8] N. Chomsky, The Minimalist Program, Current Studies in Linguistics, MIT Press, 1995.
[9] M. Dauchet, Transductions inversibles de forêts, Thèse 3ème cycle, Université de Lille (1975).
[10] J. Engelfriet, Top-down tree transducers with regular look-ahead, Math. Systems Theory 10 (1) (1977) 289–303.
[11] J. Engelfriet, E. Lilin, A. Maletti, Composition and decomposition of extended multi bottom-up tree transducers, Acta Inf. 46 (8) (2009) 561–590.
[12] J. Engelfriet, S. Maneth, Macro tree translations of linear size increase are MSO definable, SIAM J. Comput. 32 (4) (2003) 950–1006.
[13] Z. Fülöp, A. Kühnemann, H. Vogler, A bottom-up characterization of deterministic top-down tree transducers with regular look-ahead, Inf. Process. Lett. 91 (2) (2004) 57–67.
[14] Z. Fülöp, A. Kühnemann, H. Vogler, Linear deterministic multi bottom-up tree transducers, Theoret. Comput. Sci. 347 (1–2) (2005) 276–287.
[15] Z. Fülöp, A. Maletti, Composition closure of linear extended top-down tree transducers, Tech. Rep. 1301.1514, arXiv (2013).
[16] Z. Fülöp, A. Maletti, Composition closure of ε-free linear extended top-down tree transducers, in: Proc. DLT, vol. 7907 of LNCS, Springer, 2013, pp. 239–251.
[17] F. Gécseg, M. Steinby, Tree Automata, Akadémiai Kiadó, Budapest, 1984.
[18] F. Gécseg, M. Steinby, Tree languages, in: G. Rozenberg, A. Salomaa (eds.), Handbook of Formal Languages, vol. 3, chap. 1, Springer, 1997, pp. 1–68.
[19] D. Gildea, On the string translations produced by multi bottom-up tree transducers, Comput. Linguist. 38 (3) (2012) 673–693.
[20] T. Givón, Syntax: An Introduction, John Benjamins Publishing, Amsterdam, 2001.
[21] J. Graehl, K. Knight, J. May, Training tree transducers, Comput. Linguist. 34 (3) (2008) 391–427.
[22] K. Knight, J. Graehl, An overview of probabilistic tree transducers for natural language processing, in: Proc. CICLing, vol. 3406 of LNCS, Springer, 2005, pp. 1–24.
[23] P. Koehn, Statistical Machine Translation, Cambridge University Press, 2010.
[24] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, Moses: Open source toolkit for statistical machine translation, in: Proc. ACL, Association for Computational Linguistics, 2007, pp. 177–180.
[25] A. Lemay, S. Maneth, J. Niehren, A learning algorithm for top-down XML transformations, in: Proc. PODS, ACM, 2010, pp. 285–296.
[26] E. Lilin, Propriétés de clôture d'une extension de transducteurs d'arbres déterministes, in: Proc. CAAP, vol. 112 of LNCS, Springer, 1981, pp. 280–289.
[27] A. Maletti, Compositions of extended top-down tree transducers, Inform. and Comput. 206 (9–10) (2008) 1187–1196.
[28] A. Maletti, An alternative to synchronous tree substitution grammars, J. Natur. Lang. Engrg. 17 (2) (2011) 221–242.
[29] A. Maletti, How to train your multi bottom-up tree transducer, in: Proc. ACL, Association for Computational Linguistics, 2011, pp. 825–834.
[30] A. Maletti, Tree transformations and dependencies, in: Proc. MOL, vol. 6878 of LNAI, Springer, 2011, pp. 1–20.
[31] A. Maletti, Every sensible extended top-down tree transducer is a multi bottom-up tree transducer, in: Proc. NAACL, Association for Computational Linguistics, 2012, pp. 263–273.
[32] A. Maletti, The power of regularity-preserving multi bottom-up tree transducers, in: Proc. CIAA, vol. 8587 of LNCS, Springer, 2014, pp. 278–289.
[33] A. Maletti, J. Graehl, M. Hopkins, K. Knight, The power of extended top-down tree transducers, SIAM J. Comput. 39 (2) (2009) 410–430.
[34] A. van Deursen, P. Klint, F. Tip, Origin tracking, J. Symb. Comput. 15 (5–6) (1993) 523–545.
[35] A. van Deursen, P. Klint, F. Tip, Origin tracking and its applications, in: A. van Deursen, J. Heering, P. Klint (eds.), Language Prototyping: An Algebraic Specification Approach, vol. 5 of AMAST Series in Computing, World Scientific, 1996, pp. 249–294.