1 Introduction

Many recent variants of Tree Adjoining Grammars (TAG) allow an underspecification of the parent relation between nodes in a tree, i.e. they do not deal with fully specified trees as it is the case with TAGs. Such TAG variants are for example Description Tree Grammars (DTG) (Rambow, Vijay-Shanker and Weir 1995), Unordered Vector Grammars with Domiance Links (UVC-DL) (Rambow 1994a, 1994b), a definition of TAGs via so-called quasi-trees (Vijay-Shanker 1992), (Rogers and Vijay-Shanker 1994), (Rogers 1994) and (Local) Tree Description Grammars (TDG) (Kallmeyer 1997, 1998a). The last TAG variant, local TDGs, is an extension of TAG generating tree descriptions. Local TDGs even allow an underspecification of the dominance relation between node names and thereby provide the possibility to generate underspecified representations for structural ambiguities such as quantifier scope ambiguities.

This abstract deals with formal properties of local TDGs. A hierarchy of local TDGs is established together with a pumping lemma for local TDGs of a certain rank. With this pumping lemma one can prove that the class of local TDGs of a certain rank \( n \) contains the language \( L_n := \{a_1^{k_1} \cdots a_t^{k_t} \mid k \geq 0 \} \) if \( t \leq 2n \).

2 Local TDGs

Local TDGs, proposed in (Kallmeyer 1997), consist of tree descriptions, so-called elementary descriptions, and a specific start description. These tree descriptions are negation and disjunction free formulas in a quantifier-free first order logic. This logic allows the description of relations between node names \( k_1,k_2 \) such as parent relation (i.e. immediate dominance) \( k_1 \prec k_2 \), dominance (reflexive transitive closure of the parent relation) \( k_1 \prec^* k_2 \), linear precedence \( k_1 \prec k_2 \) and equality \( k_1 \equiv k_2 \). Furthermore, nodes are supposed to be labelled by terminals or by atomic feature structures. The labeling function is denoted by \( \delta \), and for a node name \( k \), \( \delta(k) \equiv t \) signifies that \( k \) has a terminal label \( t \), and \( \{ \delta(k) \} \equiv v \) signifies that \( k \) is labelled by a feature structure containing the attribute value pair \( (\alpha, v) \).

Tree descriptions in a local TDG are of a certain form, roughly speaking they consist of fully specified (sub)tree descriptions that are connected by dominance relations.\(^1\)

In an elementary description \( \psi \), some of the node names are marked (those in the set \( K_\psi \)); this is important for the derivation of descriptions. A sample local TDG is shown in Fig. 1 (in the graphical representations, some of the node names are omitted for reasons of readability). Conjuncts such as \( k_1 \prec^* k_2 \) in \( \phi_S \) that are not entailed by the other conjuncts, are called strong dominance.

Starting from the start description \( \phi_S \), local TDGs generate tree descriptions. In each derivation step, a derived \( \phi_1 \) and an elementary description \( \psi \) are combined to obtain a new description \( \phi_2 \). Roughly said, \( \phi_2 \) can be viewed as a conjunction of \( \phi_1 \) and \( \psi \) and new formulas \( k \prec k' \) or \( k \prec^* k' \) where \( k \) is a name from \( \phi_1 \) and \( k' \) a name from \( \psi \). This derivation step must be such that

1. for a node name \( k_\psi \) in \( \psi \), there is a new equivalence iff either \( k_\psi \) is marked or \( k_\psi \) is minimal (dominated by no other name, e.g. \( k_5 \) in \( \psi_1 \) and \( k_{11} \) in \( \psi_2 \) in Fig. 1),

2. a marked or minimal name \( k' \) in \( \psi \) that is not a leaf name (i.e. dominates other names) but does not dominate any other marked name must become equivalent to a leaf name in \( \phi_1 \),

3. the names \( k \) from \( \phi_1 \) that are used for the new equivalences must be part of one single element

\(^1\)Some of the conditions holding for descriptions in a local TDG are left aside here. For a formal definition of local TDGs see (Kallmeyer 1998a).
tary or start description, the so-called deriv-

tion description of this derivation step (first lo-

cality condition).

4. For each marked name $k_\psi$ in $\psi$ with a parent,

there must be a strong dominance $k_1 \approx k_2$ in $\phi_1$

such that $k_2 \approx k_\psi$ is added and the subdescription

between $k_2$ and the next marked or mini-

mal name dominating $k_\psi$ must be dominated

by $k_1$ (second locality condition).

5. and the result $\phi_2$ must be maximally underspec-

fied.

As the first condition shows, marked names are

comparable to foot nodes in an auxiliary tree in a

TAG since they specify those parts of an elementary
description $\psi$ that must be connected to a derived
description $\phi$ when adding $\psi$ to $\phi$ in a derivation

step.

The second condition describes a kind of substitu-
tion. Only leaf names in the old description can

come equivalent to names that do not dominate

other marked names.

Conditions 3 and 4. express the locality of the
derivations. All names in the old description that

are chosen for new equivalences must be part of

the derivation description, and furthermore a sub-
derscription between two minimal or marked names

must be "inserted" into a strong dominance where

the dominated name is part of the derivation de-
scription. These conditions can be compared to the

locality restriction of the derivation in a set-local

multicomponent TAG (MC-TAG) (Weir 1988). In

fact, for each set-local MC-TAG, an equivalent local

TDG can be constructed (Kallmeyer 1998a). How-

ever, local TDGs are more powerful than set-local

MC-TAGs because the locality condition restricts

only the derivation of descriptions but not the way

a minimal structure for a derived description is ob-
tained. This locality constitutes a crucial difference

between local TDGs and DTGs since derivations in

DTGs are non-local. Each subtree of a d-tree that

is added in a derivation step to a derived d-tree $\gamma$
can be inserted into any of the d-edges in $\gamma$.

If a marked name has no parent, then an under-
specification of the dominance relation can occur

in the result of a derivation step (see (Kallmeyer

1998b, Kallmeyer 1998a)). In this paper, such cases

are not considered, and for the examples mentioned

here, the fifth condition is of no consequence.

In Fig. 1 for example, a derivation step $\phi_5 \Rightarrow \phi_1$
is possible with $\phi_1 = \phi_5 \land \psi_2 \land k_1 \approx k_{11} \land k_2 \approx k_{12} \land k_4 \approx k_{22} \land k_5 \approx k_{19}$.

A local TDG generates a set of descriptions. Each

of these descriptions denotes infinitely many trees.
The trees in the tree language of a local TDG are

those trees that are "minimal" for one of the derived

descriptions. A minimal tree of a description $\phi$ is a

tree $\gamma$ satisfying $\phi$ in such a way that

1. all parent relations in $\gamma$ are described in $\phi$, and

2. if two different node names in $\phi$ denote the same

node in $\gamma$, then these two names neither have

both a parent in $\phi$ nor have both a daughter in

$\phi$.

The first condition makes sure that everything in

$\gamma$ is described in $\phi$, and with the second condi-
tion no parent relation in the tree is described more

than once in $\phi$.

For the local TDG in Fig. 1 for example, only

those descriptions have a minimal tree that are de-
viled by adding $\psi_1$ in the last derivation step.

The string language of a local TDG $G$ is the set of

all strings yielded by the trees in the tree language

of $G$.

TDGs allow "multicomponent" derivations and a

uniform complementation operation similar to sub-

scription in DTGs. Furthermore, they provide un-
derspecified representations for scope ambiguities

(Kallmeyer 1998b) since they allow the generation

of descriptions with underspecified dominance rela-

tions.

3 Rank of a Local TDG

For a given TAG, an equivalent local TDG with

at most one marked name per elementary de-
scription can be easily constructed. Obviously, the extra

power of local TDGs in contrast to TAGs arises from

the possibility of marking more than one node name

in an elementary description. In Fig. 1 for example,

$\psi_1$ and $\psi_2$ both contain two marked names. The

language generated by this local TDG is no TAL. This

suggests the definition of a hierarchy of local TDGs

depending on the maximal number of marked node

names in an elementary description.

Two kinds of marked names can be distinguished:

marked names where the part of the description

dominating this name can be put somewhere "in be-
tween" on the one hand (e.g. $k_{17}$ and $k_{20}$ in $\psi_5$ in Fig.

1), and on the other hand marked node names that

must be identified with a leaf name (e.g. $k_3$ and $k_4$

in $\psi_2$ in Fig. 2). Since there is a similarity between foot

nodes of auxiliary trees in TAGs and the first kind

of marked node names, these are called adjunct-

marked (a-marked). For similar reasons, the second
Start description:
\[
\phi_5 = k_1 \land k_2 \land k_3 <^* k_4 \land k_5 \land k_6 <^* k_7 \land k_8 < k_9 \land k_{10} < k_{11} \\
\land \text{cat}(\hat{\delta}(k_1)) = S \land \text{cat}(\hat{\delta}(k_2)) = T_1 \\
\land \text{cat}(\hat{\delta}(k_3)) = T_2 \land \text{cat}(\hat{\delta}(k_4)) = T_3 \land \delta(k_5) = \epsilon
\]

Elementary descriptions:
\[
\psi_1 = k_6 <^* k_7 \land k_1 \land k_8 <^* k_9 \land k_{10} < k_{11} \\
\land \text{cat}(\hat{\delta}(k_6)) = S \land \cdots \\
\psi_2 = k_{11} <^* k_7 \land k_{12} <^* k_{13} \land k_1 \land k_8 <^* k_9 \land k_{10} < k_{11} \\
\land k_{12} < k_{13} \land k_{14} < k_{15} < k_{16} \land k_4 <^* k_{14} \land k_{15} < k_{16} \land \cdots \\
\land \text{cat}(\hat{\delta}(k_{11})) = S \land \text{cat}(\hat{\delta}(k_{12})) = S \land \cdots \\
\land \delta(k_{16}) = a_7 \land \delta(k_{22}) = a_8
\]

Graphical representations:

(3-marked names with asterisk)

Figure 1: Local TDG for \(\{a_1a_2a_3a_4a_5|0 \leq n\}\) with two a-marked names in each elementary description

Figure 2: Local TDG for \(\{a_1a_2a_3a_4a_5|0 \leq n\}\) with two s-marked names in each elementary description

kind of marked names are called substitution-marked (s-marked).

Roughly speaking, in a derivation step, for each s-marked name in the new elementary description, there is one substring added to the yield of the description, and for each a-marked name, two substrings are added (e.g. \(a_1a_2a_3\) for \(k_3\) in Fig. 2, \(a_3a_4\) and \(a_5a_6\) for \(k_{12}\) in Fig. 1 and \(a_7a_8\) and \(a_9a_{10}\) for \(k_{22}\) in Fig. 1). Therefore, a-marked names count twice as much as s-marked names for the rank of a local TDG: a local TDG \(G\) is of rank \(n\) iff \(n = \max\{i | \text{there is an elementary } \psi \text{ in } G \text{ such that } i \text{ is twice the number of a-marked names in } \psi \}\).

For a given local TDG it is always possible to find a weakly equivalent local TDG with one more s-marked name per elementary description. Therefore, the class of languages generated by local TDGs of rank \(i\) forms a subset of the class of languages generated by local TDGs of rank \(i+1\) for \(i \geq 0\).

As shown in (Kallmeyer 1998a), the classes of local TDLs of rank 0 and 1 are equal, they are exactly the context-free languages. The class of local TDLs of rank 2 contains all TALs.

4 A pumping lemma

The idea of the pumping lemma for local TDGs of a certain rank \(n\) is similar to the one leading to the pumping lemma for TALs in (Vijay-Shanker 1987). As shown in (Kallmeyer 1997), the derivation process in a local TDG can be described by a context-free grammar \(G_{CF}\). For \(G_{CF}\), the pumping lemma for context-free languages holds. This means that in a derivation tree (of \(G_{CF}\)) from a certain tree height on, there is a subtree \(\gamma\) that can be iterated. For the corresponding local TDG, this signifies that an elementary \(\psi\) can be added twice such that: before adding \(\psi\) again we have the following situation for a string \(w\) yielded by the old description: \(w = x_10v_1 \cdot \cdot \cdot x_{m-1}v_m x_m\) where \(x_i \in T^+\), \(v_1 \cdot \cdot \cdot v_m\) is the string yielded by the subdescription derived from \(\psi\) (ordered by linear precedence). As a next derivation step, \(\psi\) is added again. If the grammar is of rank \(n\), then by adding \(\psi\), the string \(w\) can be split by inserting at most \(n\) new strings. Before the next adding of \(\psi\) (corresponding to another iteration) takes place, these substrings will be expanded to substrings \(w_1, \ldots, w_n\) with \(w_1 \cdot \cdot \cdot w_n = v_1 \cdot \cdot \cdot v_m\). These \(w_i\) may be split into several words (with other words in between) but the order of the letters is as

These two characterizations are not exclusive, for examples of node names that are both a-marked and s-marked see (Kallmeyer 1998a).
in \( v_1 \cdots v_m \). If this is repeated \( k \) times, \( k \geq 1 \), then one ends up with a word containing the letters of

\[ z_1 := z_1 \cdots z_{1m} \]

and \( k \) occurrences of all symbols of

\[ w_1 \cdots w_n \]

that are for each of these occurrences (from left to right) ordered as in \( w_1 \cdots w_n \). In the last steps (after the iterations of the derivation subtree \( \gamma \)), the symbols of some string \( z_2 \in \mathcal{T} \) are added.

Therefore the pumping lemma is as follows: for each word \( w \) in the string language of a local TDG of rank \( n \) with \(|w|\) greater than some constant \( c_0 \); after removing the letters of some words \( z_1 \) and \( z_2 \) from \( w \), the resulting word has the form \( w_1 \cdots w_n \). Then for each \( k \) there is a word \( u^{(k)} \) in the language containing also the letters of \( z_1 \) and \( z_2 \), such that: if these letters are removed from \( u^{(k)} \), the result \( \tilde{w}^{(k)} \) is a word that can be obtained by taking \( k \) occurrences of \( w_1 \cdots w_n \) and then, starting with \( \epsilon \), taking (in arbitrary order) always the left letter of one of these \( k \) words as the next letter in \( \tilde{w}^{(k)} \). Furthermore, \( \tilde{w}^{(k)} \) still contains as substrings one occurrence of each of the words \( w_1, \ldots, w_n \) (in this order).

For the language \( L_{2n} := \{a^{i_1}_{m_{1}} \cdots a^{i_m}_{m_{m}} | 0 \leq m \} \) for example the lemma for rank \( n \) holds with \( c_0 = 2n - 2 \); \( z_1 \) is \( z_2 \); \( \epsilon \). If \( w = a^{i_1}_{m_{1}} \cdots a^{i_m}_{m_{m}} \), then \( w_1 = a_{m_{1}}^{i_1} \cdots a_{m_{m}}^{i_m} \).

With the pumping lemma, it can be easily shown that for \( i > 2n \), \( L_i := \{a_{m_{1}}^{i_1} \cdots a_{m_{m}}^{i_m} | m \geq 0 \} \) does not satisfy the pumping lemma for TDGs of rank \( n \) and therefore cannot be generated by a local TDG of rank \( n \).

Consequently, for all \( n \geq 1 \), the string languages of TDGs of rank \( n \) form a proper subset of the string languages generated by TDGs of rank \( n + 1 \).

5 Conclusion

In this paper, the rank of a local TDG was defined based on the number of marked names in the elementary descriptions of the grammar. Two kinds of marked names are distinguished, namely s-marked and a-marked names. Since derivations in local TDGs can be described by a context-free grammar, the pumping lemma for context-free grammars can be applied to the derivation trees of a local TDG. This leads to the proof of a pumping lemma for local TDGs of a certain rank \( n \). Roughly said, according to this pumping lemma, to a derivation step, for each s-marked name in the new elementary description, one substring is added, and for each a-marked name, two substrings are added. With this pumping lemma one can show that for \( n \geq 1 \) the languages generated by local TDGs of rank \( n \) form a proper subset of languages generated by local TDGs of rank \( n + 1 \).

References

Kallmeyer, L.: 1997, Local Tree Description Grammars, Proceedings of the Fifth Meeting on Mathematics of Language, DFKI Research Report.

Kallmeyer, L.: 1998a, Tree Description Grammars and Underspecified Representations, PhD thesis, Universität Tübingen. To appear in Arbeitsspapier des SFB 340.

Kallmeyer, L.: 1998b, Underspecification in Tree Description Grammars, in H. P. Kolb and U. Mönnich (eds), The Mathematics of Syntactic Structures, Mouton de Gruyter. To appear.

Rambow, O.: 1994a, Formal and Computational Aspects of Natural Language Syntax, PhD thesis, University of Pennsylvania.

Rambow, O.: 1994b, Multiset-Valued Linear Index Grammars: Imposing dominance constraints on derivations, Proceedings of ACL.

Rambow, O., Vijay-Shanker, K. and Weir, D.: 1995, D-Tree Grammars, Proceedings of ACL.

Rogers, J.: 1994, Studies in the Logic of Trees with Applications to Grammar Formalisms, PhD thesis, University of Delaware.

Rogers, J. and Vijay-Shanker, K.: 1994, Obtaining trees from their descriptions: an application to Tree-Adjoining Grammars, Computational Intelligence 10(4), 401-421.

Vijay-Shanker, K.: 1987, A Study of Tree Adjoining Grammars, PhD thesis, University of Pennsylvania.

Vijay-Shanker, K.: 1992, Using descriptions of trees in a tree adjoining grammar, Computational Linguistics 18(4), 481-517.

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