A hierarchy of local TDGs

Laura Kallmeyer
Universität Tübingen
Seminar für Sprachwissenschaft
Wilhelmstr. 113
D-72074 Tübingen, Germany
lk@sfs.nphil.uni-tuebingen.de

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 Dominance Links (UVG-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 TDG, 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 $L_i := \{a_1^k \cdots a_i^k \mid k \geq 0\}$ iff $i \leq 2n$.

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) $k_1 \lhd k_2$, dominance (reflexive transitive closure of the parent relation) $k_1 \lhd^* k_2$, linear precedence $k_1 \prec k_2$ and equality $k_1 \approx k_2$. Furthermore, nodes are supposed to be labelled by terminals or by atomic feature structures. The labeling function is denoted by δ, and for a node name k, $\delta(k) \approx t$ signifies that k has a terminal label t, and $a(\delta(k)) \approx v$ signifies that k is labelled by a feature structure containing the attribute value pair $\langle a, v \rangle$.

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 ψ, some of the node names are *marked* (those in the set K_ψ); 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 $k_1 \lhd^* k_2$ in ϕ_S that are not entailed by the other conjuncts, are called *strong dominance*.

Starting from the start description ϕ_S, local TDGs generate tree descriptions. In each derivation step, a derived ϕ_1 and an elementary description ψ are combined to obtain a new description ϕ_2. Roughly said, ϕ_2 can be viewed as a conjunction of ϕ_1, ψ and new formulas $k \approx k'$ or $k \lhd^* k'$ where k is a name from ϕ_1 and k' a name from ψ. This derivation step must be such that

1. for a node name k_ψ in ψ, there is a new equivalence iff either k_ψ is marked or k_ψ is minimal (dominated by no other name, e.g. k_6 in ψ_1 and k_{11} in ψ_2 in Fig. 1),

2. a marked or minimal name k' in ψ 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 ϕ_1

3. the names k from ϕ_1 that are used for the new equivalences must be part of one single elemen-

[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 *derivation description* of this derivation step (first locality condition),

4. for each marked name k_ψ in ψ with a parent, there must be a strong dominance $k_1 \lhd^* k_2$ in ϕ_1 such that $k_2 \approx k_\psi$ is added and the subdescription between k_ψ and the next marked or minimal name dominating k_ψ must be dominated by k_1 (second locality condition),

5. and the result ϕ_2 must be maximally underspecified.

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 ψ that must be connected to a derived description ϕ when adding ψ to ϕ in a derivation step.

The second condition describes a kind of substitution. Only leaf names in the old description can become 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 subdescription between two minimal or marked names must be "inserted" into a strong dominance where the dominated name is part of the derivation description. 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). However, 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 obtained. 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 γ can be inserted into any of the d-edges in γ.

If a marked name has no parent, then an underspecification 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 $\phi_S \overset{\psi_2}{\Rightarrow} \phi_1$ is possible with $\phi_1 = \phi_S \wedge \psi_2 \wedge k_1 \approx k_{11} \wedge k_2 \approx k_{17} \wedge k_4 \approx k_{23} \wedge k_3 \lhd^* k_{18}$.

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 ϕ is a tree γ satisfying ϕ in such a way that

1. all parent relations in γ are described in ϕ, and

2. if two different node names in ϕ denote the same node in γ, then these two names neither have both a parent in ϕ nor have both a daughter in ϕ.

The first condition makes sure that everything in γ is described in ϕ, and with the second condition no parent relation in the tree is described more than once in ϕ.

For the local TDG in Fig. 1 for example, only those descriptions have a minimal tree that are derived by adding ψ_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 subsertion in DTGs. Furthermore, they provide underspecified representations for scope ambiguities (Kallmeyer 1998b) since they allow the generation of descriptions with underspecified dominance relations.

3 Rank of a local TDG

For a given TAG, an equivalent local TDG with at most one marked name per elementary description 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, ψ_1 and ψ_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 between" on the one hand (e.g. k_{17} and k_{23} in ψ_2 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 ψ_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 *adjunction-marked (a-marked)*. For similar reasons, the second

Start description:

$$\phi_S = k_1 \lhd^* k_2 \wedge k_2 \lhd k_3 \wedge k_3 \lhd^* k_4 \wedge k_4 \lhd k_5$$
$$\wedge\, cat(\delta(k_1)) \approx S \wedge cat(\delta(k_2)) \approx T_1$$
$$\wedge\, cat(\delta(k_3)) \approx T_2 \wedge cat(\delta(k_4)) \approx T_3 \wedge \delta(k_5) \approx \epsilon$$

Elementary descriptions:

$$\psi_1 = k_6 \lhd^* k_7 \wedge k_7 \lhd k_8 \wedge k_8 \lhd^* k_9 \wedge k_9 \lhd k_{10}$$
$$\wedge\, cat(\delta(k_6)) \approx S \wedge \cdots$$
$$\psi_2 = k_{11} \lhd^* k_{12} \wedge k_{12} \lhd k_{13} \wedge k_{12} \lhd k_{14} \wedge k_{12} \lhd k_{27}$$
$$\wedge\, k_{13} \prec k_{14} \wedge k_{14} \prec k_{27} \wedge k_{14} \lhd^* k_{15} \wedge \ldots$$
$$\ldots \wedge cat(\delta(k_{11})) \approx S \wedge cat(\delta(k_{12})) \approx S \wedge \ldots$$
$$\ldots \wedge \delta(k_{26}) \approx a_7 \wedge \delta(k_{27}) \approx a_8$$

$$K_{\psi_1} = \{k_8, k_{10}\}, K_{\psi_2} = \{k_{17}, k_{23}\}$$

Graphical representations:
(marked names with asterisk)

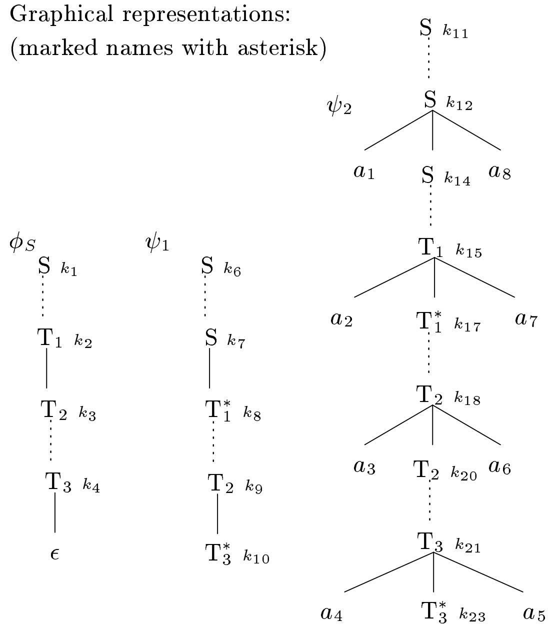

Figure 1: Local TDG for $\{a_1^n a_2^n a_3^n a_4^n a_5^n a_6^n a_7^n a_8^n \mid 0 \leq n\}$ with two a-marked names in each elementary description

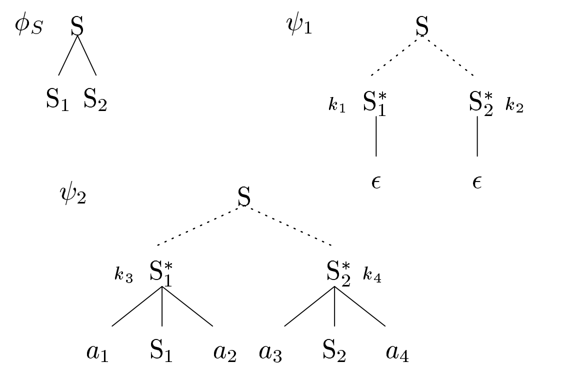

Figure 2: Local TDG for $\{a_1^n a_2^n a_3^n a_4^n \mid 0 \leq n\}$ with two s-marked names in each elementary description

kind of marked names are called *substitution-marked (s-marked)*.[2]

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_2 for k_3 in Fig. 2, a_1a_2 and a_7a_8 for k_{17} in Fig. 1 and a_3a_4 and a_5a_6 for k_{23} 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 $n = max\{i \mid$ there is an elementary ψ in G such that i is twice the number of a-marked names in ψ plus the number of s-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 γ that can be iterated. For the corresponding local TDG, this signifies that an elementary ψ can be added twice such that: before adding ψ again we have the following situation for a string w yielded by the old description: $w = x_{10}v_1 \cdots x_{1m-1}v_m x_{1m}$ where $x_{1i} \in T^*$, $v_1 \cdots v_m$ is the string yielded by the subdescription derived from ψ (ordered by linear precedence). As a next derivation step, ψ is added again. If the grammar is of rank n, then by adding ψ, the string w can be split by inserting at most n new strings. Before the next adding of ψ (corresponding to another iteration) takes place, these substrings will be expanded to substrings $w_1, \cdots, w_n$ with $w_1 \cdots w_n = v_1 \cdots v_m$. These w_i may be split into several words (with other words in between) but the order of the letters is as

[2] 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 contaning the letters of $x_1 := x_{10} \cdots x_{1m}$ 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 γ), the symbols of some string $x_2 \in T^*$ 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_G: after removing the letters of some words x_1 and x_2 from w, the resulting word has the form $w_1 \cdots w_n$. Then for each k there is a word $w^{(k)}$ in the language containing also the letters of x_1 and x_2, such that: if these letters are removed from $w^{(k)}$, the result $\hat{w}^{(k)}$ is a word that can be obtained by taking k occurrences of $w_1 \cdots w_n$ and then, starting with ϵ, taking (in arbitrary order) always the left letter of one of these k words as the next letter in $\hat{w}^{(k)}$. Furthermore, $\hat{w}^{(k)}$ still contains as substrings one occurrence of each of the words $w_1, \cdots, w_n$ (in this order).

For the language $L_{2n} := \{a_1^m \cdots a_{2n}^m \,|\, 0 \leq m\}$ for example the lemma for rank n holds with $c_G = 2n-1$, $x_1 = x_2 = \epsilon$: if $w = a_1^m \cdots a_{2n}^m$, then $w_i = a_{2i-1}^m a_{2i}^m$.

With the pumping lemma, it can be easily shown that for $i > 2n$, $L_i = \{a_1^m \cdots a_i^m \,|\, m \geq 0\}$ 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, in 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 Arbeitspapiere 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.