17

Well-Nested Drawings as Models of Syntactic Structure

Manuel Bodirsky, Marco Kuhlmann and Mathias Möhl

Abstract

This paper investigates *drawings* (totally ordered forests) as models of syntactic structure. It offers a new model-based perspective on lexicalised Tree Adjoining Grammar by characterising a class of drawings structurally equivalent to TAG derivations. The drawings in this class are distinguished by a restricted form of non-projectivity (*gap degree at most one*) and the absence of interleaving substructures (*well-nestedness*).

Keywords Model-theoretic syntax, Tree Adjoining Grammar

17.1 Introduction

There are two major approaches to formal accounts of the syntax of natural language, the proof-theoretic and the model-theoretic approach. Both aim at providing frameworks for answering the question whether a given natural language expression is grammatical. Their methodology, however, is rather different: In a proof-theoretic framework, one tries to set up a system of *derivation rules* (such as the rules in a context-free grammar) so that each well-formed natural language expression stands in correspondence with a derivation in that system. In contrast, in a model-theoretic framework, one attempts to specify a class of *models* for natural language expressions and a set of *constraints* on these models such that an expression is well-formed iff it has a model satisfying all the constraints. The main contribution of this

FG-MoL 2005.
James Rogers (ed.).
Copyright © 2009, CSLI Publications.

paper is the characterisation of a class of structures that provides a new model-based perspective on Tree Adjoining Grammar (TAG; Joshi and Schabes (1997)), a well-known proof-theoretic syntactic framework.

Every syntactic framework needs to account for at least two dimensions of syntactic structure: derivation structure and word order. The derivation structure captures linguistic notions such as dependency and constituency—the idea that a natural language expression can be composed of smaller expressions. Statements about word order are needed to account for the fact that not all permutations of the words of a grammatical sentence are neccessarily grammatical themselves.

One of the scales along which syntactic frameworks can vary is the flexibility they permit in the relationship between derivation structure and word order. Context-free grammars do not allow any flexibility at all; derivation structure determines word order completely. In mildly context-sensitive grammar formalisms like TAG or Combinatory Categorial Grammar (Steedman, 2001), certain forms of discontinuous derivations are permitted ("crossed-serial dependencies"). Other frameworks, such as non-projective dependency grammar (Plátek et al., 2001), allow for even more flexibility to account for languages with free word order.

In this paper we introduce *drawings*, a simple class of structures for which the relaxation of the relationship between derivation structure and word order can be easily measured (§ 17.2). There is a natural way in which TAG derivations can be understood as drawings (§ 17.3). We show that the class of drawings induced in this way can be identified by two structural properties: a restriction on the degree of word order flexibility and a global property called *well-nestedness*, which disallows interleaving subderivations. In combination, these two properties capture the "structural essence" of TAG (§ 17.4). The paper concludes with a discussion of the relevance of our results and an outlook on future research (§ 17.5). For further details and full formal proofs, we refer to the extended version of this paper (Bodirsky et al., 2005).

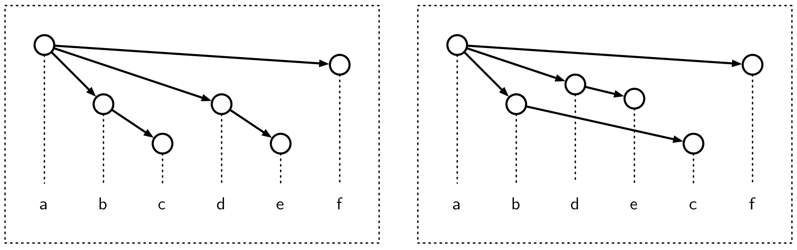
17.2 Drawings

We start by introducing some basic terminology.

17.2.1 Relational structures

A *relational structure* is a tuple whose first component is a non-empty, finite set V of *nodes*, and whose remaining components are (in this paper) binary relations on V. The notation Ru stands for the set of all nodes v such that $(u, v) \in R$. We use the standard notations for the transitive (R^+) and reflexive transitive (R^*) closure of binary relations.

In this paper, we are concerned with two types of relational struc-



(a) gap 0 (b) gap 1

FIGURE 1: Two drawings

tures in particular: *forests* and *total orders.* A relational structure $(V; \lhd)$ is called a *forest* iff $\lhd$ is acyclic and every node in V has at most one predecessor with respect to $\lhd$. Nodes in a forest with no $\lhd$-predecessors are called *roots.* A *tree* is a forest that has exactly one root. A *total order* is a relational structure $(V; \prec)$ in which $\prec$ is transitive and for all $v_1, v_2 \in V$, exactly one of the following three conditions holds: $v_1 \prec v_2$, $v_1 = v_2$, or $v_2 \prec v_1$. Given a total order, the *interval* between two nodes v_1 and v_2 is the set of all v such that $v_1 \preceq v \preceq v_2$. The *cover* (also known as *convex hull*) of a set $V' \subseteq V$, $\mathcal{C}(V')$, is the smallest interval containing V'. A set V' is *convex* iff it is equal to its cover. A *gap* in a set V' is a maximal, non-empty interval in $\mathcal{C}(V') - V'$. We call the number of gaps in a set the *gap degree* of that set and write $G_k(V)$ for the k-th gap in V (counted, say, from left to right).

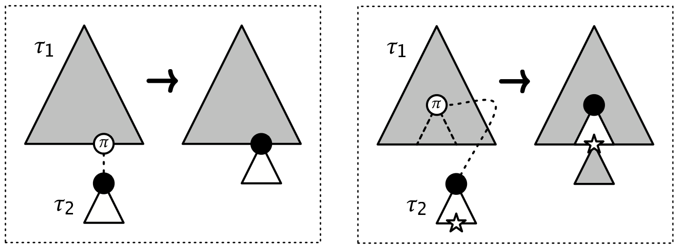
17.2.2 Drawings and gaps

Drawings are relational structures with two binary relations: a forest to model derivation structure, and a total order to model word order.

Definition 54 A *drawing* is a relational structure $(V; \lhd, \prec)$ where $(V; \lhd)$ forms a forest, and $(V; \prec)$ forms a total order. Drawings whose underlying forest is a tree will be called *T-drawings.*

Note that, in contrast to ordered forests (where order is defined on the direct successors of each node), order in drawings is *total.* By identifying each node v in a drawing with the set $(\lhd^*)v$ of nodes in the subtree rooted at v, we can lift the notions of cover and gap as follows: $\mathcal{C}(v) := \mathcal{C}((\lhd^*)v)$, $G_k(v) := G_k((\lhd^*)v)$. The gap degree of a drawing is the maximum among the gap degrees of its nodes.

Fig. 1 shows two drawings of the same underlying tree. The circles and solid arcs reflect the forest structure. The dotted lines mark the positions of the nodes with respect to the total order. The labels attached to the dotted lines give names to the nodes. Drawing 1a has gap degree



(a) substitution (b) adjunction

FIGURE 2: Combining tree structures in TAG

zero, since $\mathcal{C}(v) = (\triangleleft^*)v$ for all nodes v. In contrast, drawing 1b has gap degree one, since the set $\{d, e\} = \{b, d, e, c\} - \{b, c\} = \mathcal{C}(b) - (\triangleleft^*)b$ is a gap for node b, and no other node has a gap.

17.2.3 Related work

Our terminology can be seen as a model-based reconstruction of the terminology developed for non-projective dependency trees (Plátek et al., 2001), where gaps are defined with respect to tree structures generated by a grammar. The notion of gap degree is closely related to the notion of *fan-out* in work on (string-based) finite copying parallel rewriting systems (Rambow and Satta, 1999): fan-out measures the number of substrings that a sub-derivation *does* contribute to the complete yield of the derivation; dually, the gap degree measures the number of substrings that a sub-derivation *does not* contribute.

17.3 Drawings for TAG

Tree Adjoining Grammar (TAG) (Joshi and Schabes, 1997) is a proof-theoretic syntactic framework whose derivations manipulate tree structures. This section gives a brief overview of the formalism and shows how drawings model derivations in lexicalised TAGs.

17.3.1 Tree Adjoining Grammar

The building blocks of a TAG grammar are called *elementary trees;* they are successor-ordered trees in which each node has one of three types: *anchor* (or *terminal node*), *non-terminal node*, or *foot node.* Anchors and foot nodes must be leaves; non-terminal nodes may be either leaves or inner nodes. Each elementary tree can have at most one foot node. Elementary trees without a foot node are called *initial trees*; non-initial trees are called *auxiliary trees.* A TAG grammar is *strictly lexicalised*, if each of its elementary trees contains exactly one anchor.

Trees in TAG can be combined using two operations (Fig. 2): *Substitution* combines a tree structure τ_1 with an initial tree τ_2 by identifying a non-terminal leaf node π of τ_1 with the root node of τ_2 (Fig. 2a). *Adjunction* identifies an inner node π of a structure τ_1 with the root node of an auxiliary tree τ_2; the subtree of τ_1 rooted at π is excised from τ_1 and inserted below the foot node of τ_2 (Fig. 2b; the star marks the foot node). Combing operations are disallowed at root and foot nodes.

TAG *derivation trees* record information about how tree structures were combined during a derivation. Formally, they can be seen as unordered trees whose nodes are labelled with elementary trees, and whose edges are labelled with the nodes at which the combining operations took place. If v is a node in a derivation tree, we write $\ell(v)$ for the label of v. An edge $v_1 \;-\pi\!\rightarrow\; v_2$ signifies that the elementary tree $\ell(v_2)$ was substituted or adjoined into the tree $\ell(v_1)$ at node π.

TAG *derived trees* represent results of derivations; we write $\mathsf{drv}(D)$ for the derived tree corresponding to a derivation tree D. Derived trees are ordered trees made up from the accumulated material of the elementary trees participating in the derivation. In particular, each TAG derivation induces a mapping ρ that maps each node v in D to the root node of $\ell(v)$ in $\mathsf{drv}(D)$. In strictly lexicalised TAGs, a derivation also induces a mapping α that maps each node v in D to the anchor of $\ell(v)$ in $\mathsf{drv}(D)$.

For derivation trees D in strictly lexicalised TAGs, we define

$$\begin{aligned} \mathsf{derived}(v) &:= \{\, \alpha(u) \mid v \lhd^* u \text{ in } D \,\} \quad \text{and} \\ \mathsf{yield}(v) &:= \{\, \pi \mid \pi \text{ is an anchor and } \rho(v) \lhd^* \pi \text{ in } \mathsf{drv}(D) \,\}\,. \end{aligned}$$

The set $\mathsf{derived}(v)$ contains those anchors in $\mathsf{drv}(D)$ that are contributed by the partial derivation starting at $\ell(v)$; $\mathsf{yield}(v)$ contains those anchors that are dominated by the root node of $\ell(v)$. To give a concrete example: Fig. 3 shows a TAG derivation tree (3a) and its corresponding derived tree (3b). For this derivation, $\mathsf{derived}(\mathit{like}) = \{\mathit{what}, \mathit{Dan}, \mathit{like}\}$ and $\mathsf{yield}(\mathit{like}) = \mathsf{derived}(\mathit{like}) \cup \{\mathit{does}\}$.

17.3.2 TAG drawings

There is a natural relation between strictly lexicalised TAGs and drawings: given a TAG derivation, one obtains a drawing by ordering the nodes in the derivation tree according to the left-to-right order on their corresponding anchors in the derived tree.

Definition 55 Let D be a derivation tree for a strictly lexicalised TAG. A drawing $(V; \lhd, \prec)$ is a TAG-*drawing* iff (a) V is the set of nodes in D; (b) $v_1 \lhd v_2$ iff for some π, there is an edge $v_1 \;-\pi\!\rightarrow\; v_2$ in D; (c) $v_1 \prec v_2$ iff $\alpha(v_1)$ precedes $\alpha(v_2)$ with respect to the leaf order in $\mathsf{drv}(D)$.

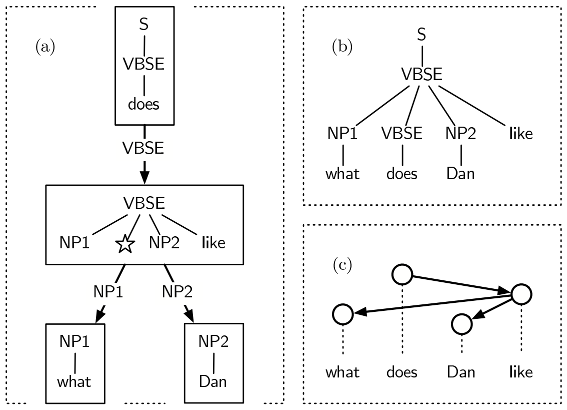


FIGURE 3: TAG derivation trees (a), derived trees (b), and drawings (c)

Fig. 3c shows the TAG drawing induced by the derivation in Figs. 3a–b.

17.4 The structural essence of TAG

Now that we have defined how TAG derivations induce drawings, we can ask whether *all* drawings (whose underlying forests are trees) are TAG drawings. The answer to this question is "no": TAG drawings form a proper subclass in the class of all drawings. As the major technical result of this paper, we will characterise the class of TAG drawings by two structural properties: a restriction on the gap degree and a property we call *well-nestedness* (Definition 56). The relevance of this result is that it provides a characterisation of "TAG-ness" that does not make reference to any specific grammar, but refers to purely structural properties: well-nested drawings with gap degree at most one are "just the right" models for TAG in the sense that every TAG derivation induces such a drawing, and for any such drawing we can construct a TAG grammar that allows for a derivation inducing that drawing.

17.4.1 TAG drawings are gap one

Gaps in TAG drawings correspond to adjunctions in TAG derivations: each adjunction may introduce material into the yield of a node that was not derived from that node. Since auxiliary trees have only one foot node, TAG drawings can have at most one gap.

Lemma 56 *Let D be a* TAG *derivation tree, and let v be a node in D. Then (a)* $\mathsf{derived}(v) \subseteq \mathsf{yield}(v)$*, (b)* $\mathsf{yield}(v) - \mathsf{derived}(v)$ *is convex, and (c)* $\mathsf{derived}(v)$ *contains at most one gap.*

Corollary 57 TAG *drawings have gap degree at most one.*

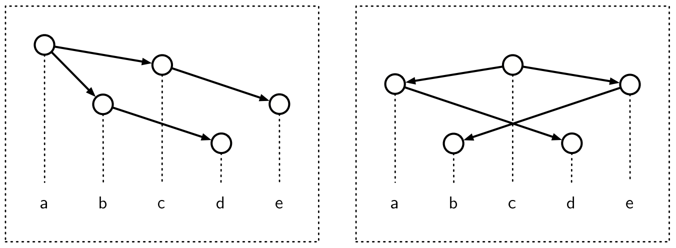


FIGURE 4: Two drawings that are not well-nested

17.4.2 TAG drawings are well-nested

The gap restriction alone is not sufficient to characterise TAG drawings: there are drawings with gap degree one that cannot be induced by a TAG. Fig. 4 shows two examples. To see why these drawings cannot be induced by a TAG notice that in both of them, the cover of two nodes overlap ($\mathcal{C}(b)$ and $\mathcal{C}(c)$ in the left drawing, $\mathcal{C}(a)$ and $\mathcal{C}(e)$ in the right one). Since each node in a drawing corresponds to a sub-derivation on the TAG side, this would require the overlap of two yields in the derived tree, which is impossible. The present section will make this statement precise.

Definition 56 Let T_1 and T_2 be disjoint subtrees in a drawing. We say that T_1 and T_2 *interleave* iff there are nodes $l_1, r_1 \in T_1$ and $l_2, r_2 \in T_2$ such that $l_1 \prec l_2 \prec r_1 \prec r_2$. A drawing is called *well-nested* iff it does not contain any interleaving subtrees.

Well-nestedness is a purely structural property: it does not make reference to any particular grammar at all. In this respect, it is similar to the condition of *planarity* (Yli-Jyrä, 2003). In fact, one obtains planarity instead of well-nestedness from Definition 56 if the disjointness condition is relaxed such that T_2 may also be a subtree of T_1, and l_1, r_1 are chosen from $T_1 - T_2$.

Lemma 58 TAG *drawings are well-nested.*

17.4.3 Constructing a TAG grammar for a drawing

To complete our characterisation of TAG drawings, we now present an algorithm that takes a well-nested drawing with gap degree at most one and constructs a TAG grammar whose only derivation induces the original drawing. Correctness of the algorithm establishes the following

Lemma 59 *Each well-nested T-drawing that has gap degree at most one is a* TAG *drawing.*

The algorithm (fully discussed in the extended version of this paper) performs a pre-order traversal of the tree structure underlying the drawing. For each node v, it constructs an elementary tree whose anchor is v and whose non-terminal nodes license exactly the combining operations required by the outgoing edges of v. Children w of v that do not have a gap induce sites for substitutions (non-terminal leaf nodes), other children induce sites for adjunctions (non-terminal inner nodes). If v itself has a gap, it also needs to include a foot node. The order and dominance relation on the nodes in the constructed elementary tree are determined by the order and nesting of the *scopes* of the nodes: the scope of an anchor is the anchor itself, the scope of a non-terminal node is the cover of the corresponding child node, and the scope of a foot node is the gap that triggered the inclusion of this foot node. Well-nestedness ensures that the nesting of the scopes can be translated into a tree relation between the nodes in the elementary tree.

The combination of Lemmata 56, 58 and 59 implies

Theorem 60 *A T-drawing is a* TAG *drawing iff it is well-nested and has gap degree at most one.*

17.5 Conclusion

This paper introduced *drawings* as models of syntactic structure and presented a novel perspective on lexicalised TAG by characterising a class of drawings structurally equivalent to TAG derivations. The drawings in this class—we called them TAG drawings—have two properties: they have a *gap degree* of at most one and are *well-nested.* TAG drawings are suitable structures for a model-theoretic treatment of TAG.

We believe that our results can provide a new perspective on the treatment of languages with free word order in TAG. Since TAG's ability to account for word order variations is extremely limited, various attempts have been made to move TAG into a description-based direction.[1] Drawings allow us to analyse these proposals with respect to the question how they extend the class of *models* of TAG, and what new *descriptive means* they offer to talk about these models. We feel that these issues were not clearly separated in previous work on model-theoretic TAG (Palm, 1996, Rogers, 2003).

A model-theoretic approach to natural language processing lends itself to constraint-based processing techniques. We have started to investigate the computational complexity of constraint satisfaction problems on TAG drawings by defining a relevant constraint language and formulating a constraint solver that decides in polynomial time whether

[1] Kallmeyer's dissertation (Kallmeyer, 1999) provides a comprehensive summary.

a formula in that language can be satisfied on a well-nested drawing (Bodirsky et al., 2005). This solver can be used as a propagator in a constraint-based processing framework for TAG descriptions.

Our immediate future work will be concerned with the further development of our processing techniques into a model-based parser for TAGs. The current constraint solver propagates information about structures that are already known; a full parser would need to construct these structures in the first place. In the longer term, we hope to characterise other proof-theoretic syntactic frameworks in terms of drawings, such as Multi-Component TAG and Combinatory Categorial Grammar.

Acknowledgements We are grateful to Alexander Koller, Guido Tack and an anonymous reviewer for useful comments on earlier versions of this paper. The work of Kuhlmann and Möhl is funded by the Collaborative Research Centre 378 "Resource-Adpative Cognitive Processes" of the Deutsche Forschungsgemeinschaft (DFG).

References

Bodirsky, Manuel, Marco Kuhlmann, and Mathias Möhl. 2005. Well-nested drawings as models of syntactic structure. Tech. rep., Saarland University. Electronic version available at `http://www.ps.uni-sb.de/Papers/`.

Joshi, Aravind and Yves Schabes. 1997. *Handbook of Formal Languages*, vol. 3, chap. Tree Adjoining Grammars, pages 69–123. Springer.

Kallmeyer, Laura. 1999. *Tree Description Grammars and Underspecified Representations*. Ph.D. thesis, Universität Tübingen.

Palm, Adi. 1996. From constraints to TAGs: A transformation from constraints into formal grammars. In *Second Conference on Formal Grammar*. Prague, Czech Republic.

Plátek, Martin, Tomáš Holan, and Vladislav Kuboň. 2001. On relax-ability of word-order by d-grammars. In C. Calude, M. Dinneen, and S. Sburlan, eds., *Combinatorics, Computability and Logic*, pages 159–174. Springer.

Rambow, Owen and Giorgio Satta. 1999. Independent parallelism in finite copying parallel rewriting systems. *Th. Computer Science* 223:87–120.

Rogers, James. 2003. Syntactic structures as multi-dimensional trees. *Research on Language and Computation* 1:265–305.

Steedman, Mark. 2001. *The Syntactic Process*. MIT Press.

Yli-Jyrä, Anssi. 2003. Multiplanarity – a model for dependency structures in treebanks. In *Second Workshop on Treebanks and Linguistic Theories*, pages 189–200. Växjö, Sweden.