A 'Tree Adjoining' Grammar without Adjoining
The case of scrambling in German

Gerard Kempen
Department of Psychology
Leiden University
PO Box 9555
NL-2300 RB Leiden
The Netherlands
kempen@rulfsw.fsw.leidenuniv.nl

Karin Harbusch
Computer Science Department
University of Koblenz-Landau
Rheinau 1
D-56075 Koblenz
Germany
harbusch@informatik.uni-koblenz.de

The psycholinguistically motivated grammar formalism of *Performance Grammar* (PG, [Kempen 97]) is similar to recent versions of *Tree Adjoining Grammar* (TAG; cf. [Joshi *et al*. 91]) in several important respects. It uses lexicalized initial trees; it generates derived trees synchronously linked to conceptual structures described in the same formalism (as in Synchronous TAGs [Shieber, Schabes 90]); and it factors dominance relationships and linear precedence in surface structure trees ([Joshi 87]).

PG differs from recent TAG versions in that the adjoining operation and auxiliary trees are absent. Adjunction is replaced by a combination of substitution—the only composition operation—and a special linearization component that takes care of ordering the branches of derived trees in a global manner without re-arranging the derived structures. PG has been worked out for substantial fragments of Dutch, including the well-known cross-serial dependencies in self-embedded clauses. Here we will outline how PG deals with scrambling phenomena in German without invoking adjunction. For TAG treatments of these phenomena we refer to [Becker *et al*. 91] and [Rambow 94].

PG's lexicalized initial trees, called *lexical frames*, are 3-tiered mobiles. The top layer of a frame consists of a single *phrasal* node (called the 'root'; e.g. S, NP, ADJP, PP), which is connected to one or more *functional* nodes in the second layer (e.g., SUBJect, HeaD, Direct OBJect, CoMPlement, MODifier). At most one exemplar of a functional node is allowed in the same frame, except for MOD nodes, which may occur several times (indicated by the Kleene star: MOD*). Every functional node dominates exactly one phrasal node in the third ('foot') layer, except for HD which immediately dominates a lexical (part-of-speech) node. Each lexical frame is 'anchored' to a lexical item—a 'lemma' printed below the lexical node serving as the frame's HeaD (Fig. 1).

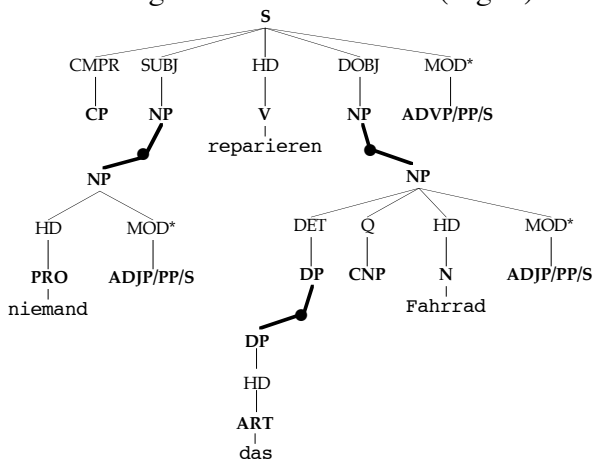


Fig. 1. Simplified examples of lexical frames. CP = Complementizer Phrase; CMPR = Complementizer; DP = Determiner Phrase. Left-to-right order of branches is arbitrary. The unifications (filled circles) correspond to German sentences such as Repariert niemand das Fahrrad? *or* Niemand repariert das Fahrrad *('Does nobody repair the bicycle?' or 'Nobody repairs the bicycle').*

Associated with nodes in the top and bottom layers are feature matrices (not discussed here), which can be *unified* with other matrices as part of the substitution process. Unification always involves one root and one foot node of two different lexical frames (see the filled circles in Fig. 1). Only non-recursive unification is used.

Left-to-right order of the branches of a lexical frame is determined by the 'linearizer' as-

sociated with a lexical frame. We assume that every lexical frame has a one-dimensional array specifying a fixed number of positions for foot nodes. For instance, verb frames (i.e., frames anchored to a verb) have an array whose positions can be occupied by a Subject NP, a Direct Object NP, the Head verb, etc. Fig. 2 shows 13 out of 14 slots where foot nodes of German verb frames can go. The positions numbered M1 through M11 belong to the Midfield (Ger. *Mittelfeld*); B1 and B2 make up the Backfield (*Nachfeld*). Not shown is the single Forefield (*Vorfeld*) slot F1, located to the left of M1. The annotations at the arcs denote possible fillers of the slots. For example, in a main clause the Head verb is assigned the first Midfield slot (M1); in a subordinate clause it goes to the last Midfield position (M11). Subject NPs that could not enter the Forefield (e.g. in subordinate clauses) are placed in M2 if its head is a personal pronoun, in M3 otherwise. (Note that frames anchored to other parts of speech than verbs (NP, PP) have their own specialized linearization array.)

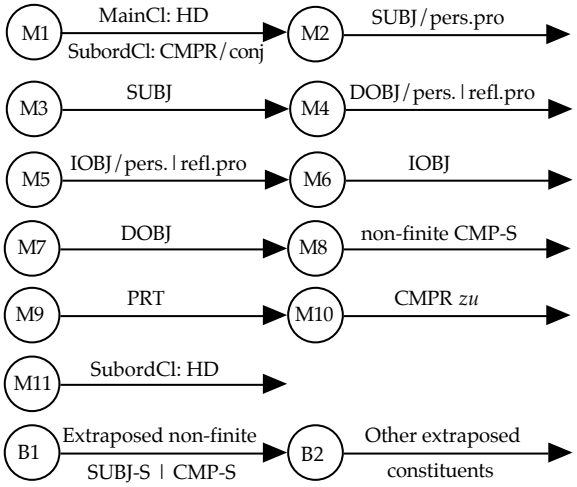


Fig. 2. Positions licensed to various types of constituents in the Midfield and Backfield of German clauses.

The fillers listed in slots M2 through M7 represent the unmarked order of verbal arguments (cf. [Uszkoreit 87]). They may be accompanied by additional constituents, in particular by modifiers and by arguments that, because of being in emphatic or contrastive focus, have been moved to the left (e.g. in *weil er ein Fahrrad den Kindern verspricht*; because-he-a-bike-the-children-promises). These companions are positioned after the 'standard' fillers (if any).

A key property of linearization in PG is that certain constituents may move out of their 'own' array and receive a position in an array located at a higher level. This is because, due to subcategorization features, a linearization array may be instantiated incompletely. For instance, if a verb takes a non-finite complement clause, then slots M1 through M3 are missing from the complement's array. If, in addition, the complement is subjected to 'clause union', slots M4 through M7 are absent as well. In such cases, verb arguments and adjuncts that need to be expressed overtly, look for a slot higher up in the hierarchy of verb frames and get hold of the first (i.e. lowest) slot that is within scope. E.g., in *daß sie den Lehrer das Fahrrad nicht reparieren sah* (that she didn't see the teacher repair the bike), *den Lehrer* and *das Fahrrad* occupy the same M7 slot, in order of increasing depth (Fig. 3).

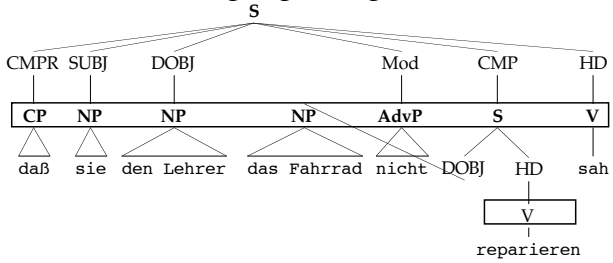
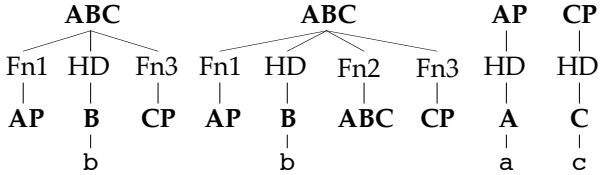


Fig. 3. The embedded DOBJ-NP has been lifted into the linearization array (rectangle) of the next higher verb frame. Due to a subcategorization feature of the lexical entry sehen *(to see), only slots M8-M10 of the complement clause have been instantiated. This causes* das Fahrrad *to land in the M7 slot of the matrix, joining* den Lehrer.



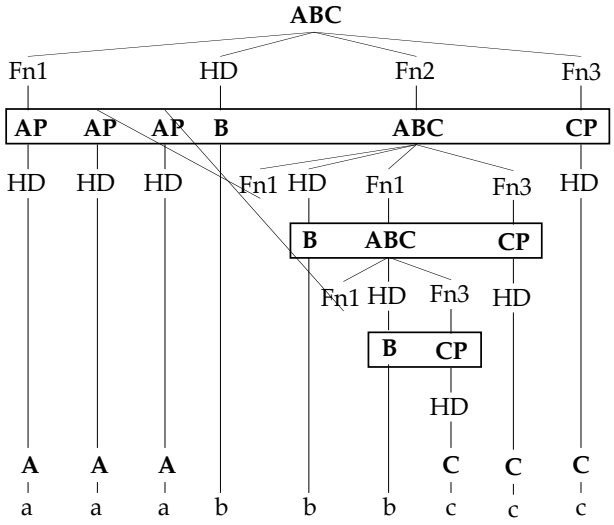


Fig. 4. Derivation of string $a^3b^3c^3$. (a) Initial lexical frames. (b) Derived tree. Notice that only the matrix linearization array is instantiated completely; the embedded ones are truncated, causing the A-phrases to be fronted.

The mechanism that controls the distribution of constituents over the slots of a linearization array, is modeled as a Finite-State Automaton (FSA). The FSA associated with a lexical frame traverses its array from left to right. At each slot, it inspects the set of constituents that are waiting for placement in the array, and inserts there any constituents meeting the placement conditions on that slot (see the labels on the edges of Fig. 2).

PG is capable of generating the mildly context-sensitive language $a^nb^nc^n$. Fig. 4b illustrates a possible derivation of $a^3b^3c^3$ based on the lexical frames in Fig. 4a. The linearization array associated with ABC frames contains four slots *S1...S4* to be filled, respectively, by constituents of type AP (any number, in arbitrary order), B, ABC, and CP. Furthermore, a subcategorization feature in the ABC foot node of the recursive ABC frame causes deletion of slot S1 of the embedded ABC linearization arrays.

Certain scrambling phenomena in German are interpretable as a consequence of PG's linearization scheme. Consider sentence (1), from [Rambow 94], with two non-finite clauses embedded in one another:

[S[S *das Fahrrad zu reparieren*] *zu versuchen*]

Rambow presents acceptability ratings for 30 scrambled versions of this sentence, viz. for all permutations in which the NPs precede the verbs they belong to. (Only five constituents are permutable: two NPs and three verbs.) See Table 1 for a selection from these data.

Table 1. Acceptability ratings for some scrambled version of sentence (1), based on judgments by several native speakers of German. Data from [Rambow 94].

6	weil das Fahrrad zu reparieren niemand zu versuchen verspricht	ok
20	weil niemand das Fahrrad zu reparieren verspricht zu versuchen	?
23	weil niemand zu versuchen verspricht, das Fahrrad zu reparieren	?
25	weil niemand das Fahrrad zu versuchen verspricht zu reparieren	*?
30	weil das Fahrrad zu versuchen niemand verspricht zu reparieren	*?
10	weil das Fahrrad zu versuchen niemand zu reparieren verspricht	*
24	weil niemand zu versuchen das Fahrrad verspricht zu reparieren	*

(1) weil niemand das Fahrrad zu reparieren
because nobody the bike to repair
zu versuchen verspricht
to try promises
'because nobody promises to try to repair the bike'

(2) weil niemand verspricht das Fahrrad zu reparieren zu versuchen

(3) weil niemand das Fahrrad verspricht zu reparieren zu versuchen

The verbs *versprechen* and *versuchen* can take several types of complement in addition to the one exemplified in (1). The non-finite complement clause may be extraposed, i.e. put behind the finite verb in subordinate clauses (as in (2)). Moreover, it allows the so-called "Third Construction" where only part of the non-finite complement clause, including the infinitival verb, is extraposed.

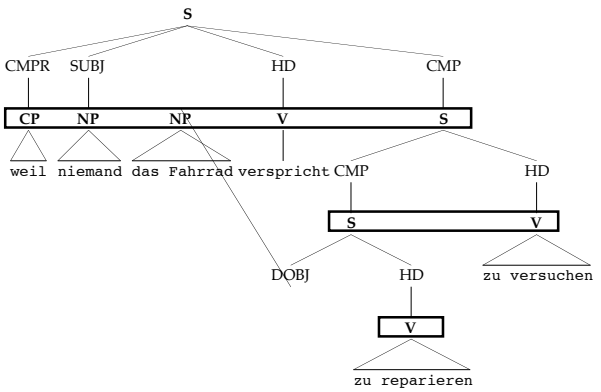


Fig 5. PG analysis of sentence (3).

In the PG treatment of these constructions (illustrated in Fig. 5), the linearization arrays play a crucial role. We assume that, in sentence (2), *reparieren*'s linearization array has been instantiated from slot M4 onward, and in sentence (3) only from slot M8 onward. Moreover, *versuchen*'s array has been truncated as well and only contains slots M8 through B2. This implies that, in (2), the direct object *das Fahrrad* could find a place in *reparieren*'s array, whereas it was moved upward into the finite clause in (3). As stated above, it is a subcategorization feature of a complement-taking verb that controls how the complement's linearization array will be instantiated.

Emphatic or contrastive focus is another factor causing a constituent to move upward. A focused constituent is assigned to early positions in a clause, e.g. M3 or M4. If that position is not available at the clause level it belongs to, it moves into the array of a higher clause.

The position of the two infinitives with respect to one another turns out to be the major source of variation in acceptability. In all fully or marginally acceptable versions ("ok" or "?"):

(A) the non-finite clauses are adjacent, or

(B) they are discontinuous, with the complement-taking infinitive (*zu versuchen*) following its complement (*zu reparieren*).

These properties are illustrated by the PG representations of Rambow's sentences (2) and (6) in Fig. 6.

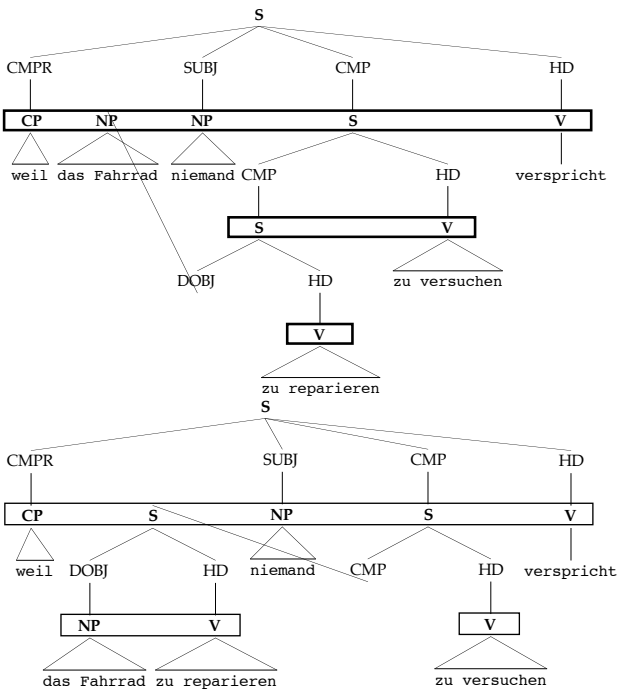


Fig. 6. PG analyses of two acceptable utterances in conformity with linearization rules. Top panel: both non-finite clauses occupy the standard position M8 in their respective arrays. NP das Fahrrad *is focused (slot M3 or M4). Bottom panel: CMP-S* versuchen *is in unmarked position M8; CMP-S* reparieren *is focused.*

On the other hand, in all unacceptable or bad versions ("*" or "*?"):

(A') the non-finite clauses are discontinuous,

(B') with the complement-taker preceding its complement.

Examples are Rambow's sentences (10) and (30), quasi-reconstructed here as Fig. 7.

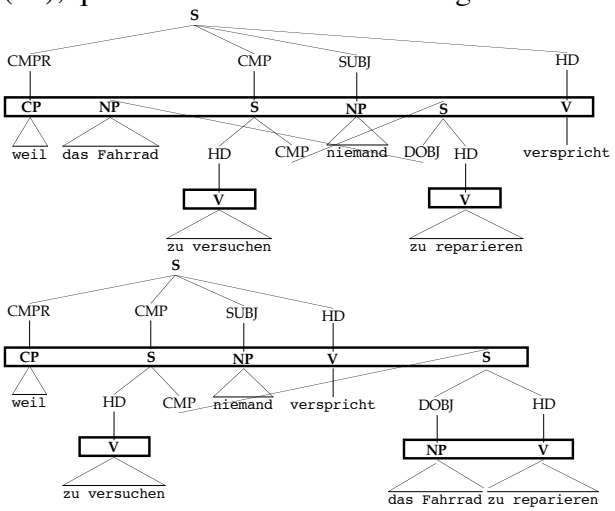


Fig. 7. Quasi-analyses of two unacceptable sentences.

The structures depicted in Fig. 7 violate PG's linearization scheme because of an illegal

attempt of *zu reparieren* to move into the finite clause: this CMP-S is not moving into a focus slot and therefore will be assigned a place at its own level, i.e. in slot M8 or B1 of *versuchen*'s array. All bad or unacceptable sentences in Table 1 suffer from this problem, while those rated good or marginal all adhere to PG's linearization scheme. Version (23), whose rating is relatively good although it manifests an illegal extraposition attempt, is the only exception.

We conclude that PG is capable of accounting for a considerable portion of the variance in the acceptability judgments reported by [Rambow 94]. This suggests that the combination of 'substitution + linearization FSA' in PG could serve as an alternative to 'adjunction + substitution' in TAG.

References

[Becker *et al*. 91] Becker, T., Joshi, A.K., Rambow, O. (1991). Long Distance Scrambling and Tree Adjoining Grammars. In: *Papers presented to EACL91*, Berlin.

[Joshi 87] Joshi, A.K. (1987). The relevance of Tree Adjoining Grammar to generation. In: Kempen, G. (Ed.), *Natural language generation*. Dordrecht: Kluwer.

[Joshi *et al*. 91] Joshi, A.K., Vijay-Shanker, K., Weir, D. (1991). The convergence of mildly context-sensitive grammatical formalisms. In Sells, P., Shieber, S.M., Wasow, T. (eds.), *The Mental Representation of Grammatical Relations*. Cambridge MA: MIT Press.

[Kempen 97] Kempen, G. (1997). Grammatical performance in human sentence production and comprehension. Ms, Leiden University.

[Rambow 94] Rambow, O. (1994). *Formal and computational aspects of natural language syntax*. PhD Thesis, University of Pennsylvania.

[Shieber, Schabes 90] Shieber, S.M., Schabes, Y. (1990). Synchronous Tree-Adjoining Grammars. In: Karlgren, H. (Ed.), *COLING-90*, Helsinki.

[Uszkoreit 87] Uszkoreit, H. (1987). *Word order and constituent structure in German*. Stanford CA: CSLI.