Introducing GerTT: A young TT-MCTAG for German

Laura Kallmeyer, Timm Lichte

University of Tübingen, Germany

NaTAL, 26.06.2008

Aims and scope

GerTT (German TT-MCTAG)

- Large-coverage TT-MCTAG for German
- Including semantics.
- Being implemented using XMG and TuLiPA.

Outline:

1. The formalism: (k-)TT-MCTAG, an extension of TAG
2. Sample analyses and limits of TT-MCTAG
3. Adding semantics to GerTT
4. The grammar implementation: state and coverage testing
5. The retrieval and integration of lexical resources

Aims and scope

GerTT (German TT-MCTAG)

- Large-coverage TT-MCTAG for German
- Including semantics.
- Being implemented using XMG and TuLiPA.

Outline:

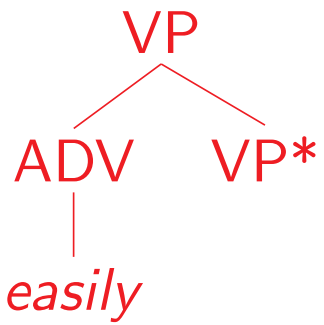
1. The formalism: (k-)TT-MCTAG, an extension of TAG
2. Sample analyses and limits of TT-MCTAG
3. Adding semantics to GerTT
4. The grammar implementation: state and coverage testing
5. The retrieval and integration of lexical resources

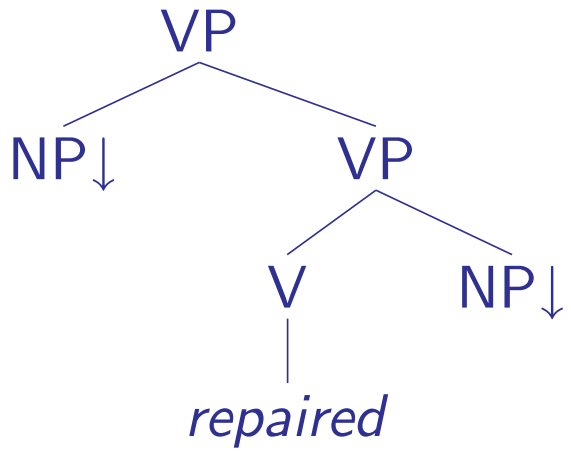
Tree-Adjoining Grammar - Basics

A **Tree Adjoining Grammar (TAG)** is a set of elementary trees:

- a finite set of initial trees
- a finite set of auxiliary trees

E.g.:

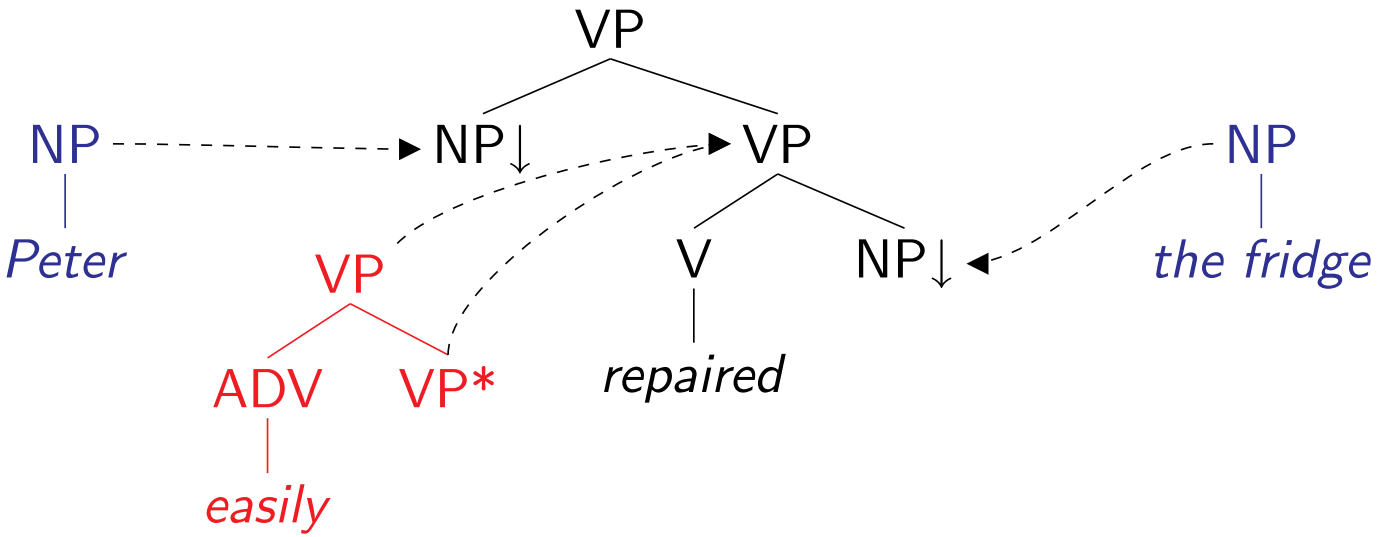




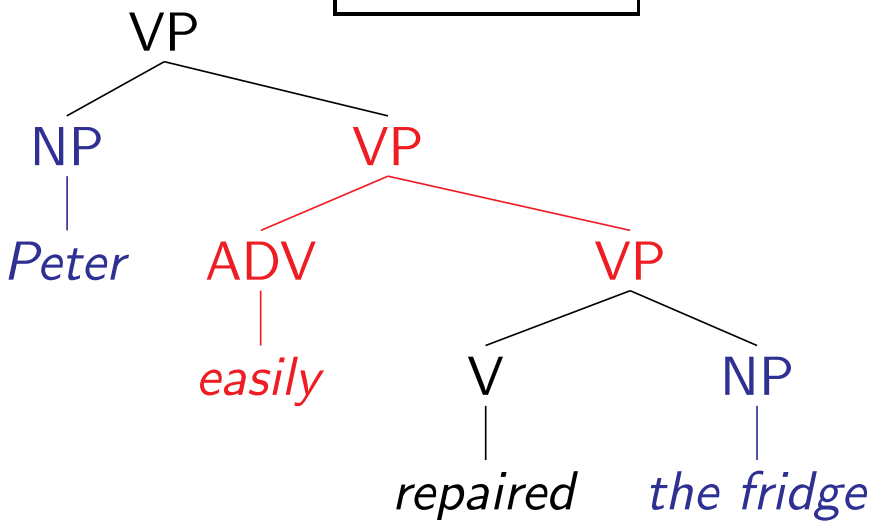
Combinatorial operations:

- substitution: replacing a non-terminal leaf with an initial tree
- adjunction: replacing an internal node with an auxiliary tree

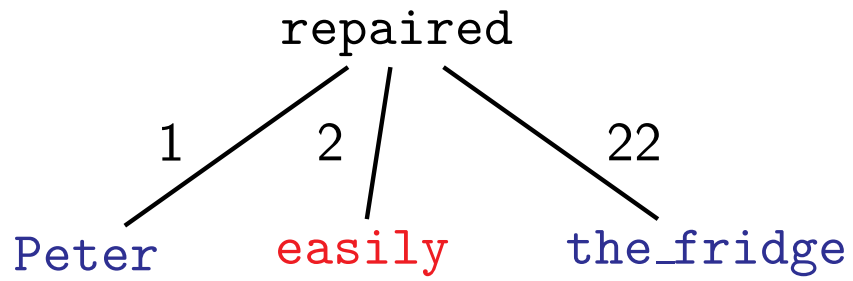
Tree-Adjoining Grammar - Example



derived tree



derivation tree



Tree-Adjoining Grammar - Basics

TAGs are **mildly context-sensitive:**

1. Polynomial time parsing complexity
2. Generation of limited crossing dependencies
3. Constant growth property (semilinearity)

Large TAG grammars:

- English and Korean (XTAG, UPenn)
- French TAG (Benoit Crabbé's PhD-thesis)
- ...

Why not TAG for German?

The order of complements (and adjuncts) of a verb is flexible.

(1) *Peter liebt Susi.*
1: `Peter loves Susi`
2: `Susi loves Peter`

(2) *dass Peter heute den Kühlschrank repariert hat*
dass den Kühlschrank heute Peter repariert hat
...
('that Peter has repaired the fridge today')

TAG is inappropriate for German, because it is:

- not powerful enough for some constructions (i.e., coherent constructions)
- not descriptively adequat (i.e., one elementary tree for each permutation)

Why not TAG for German?

The order of complements (and adjuncts) of a verb is flexible.

(1) *Peter liebt Susi.*
1: `Peter loves Susi`
2: `Susi loves Peter`

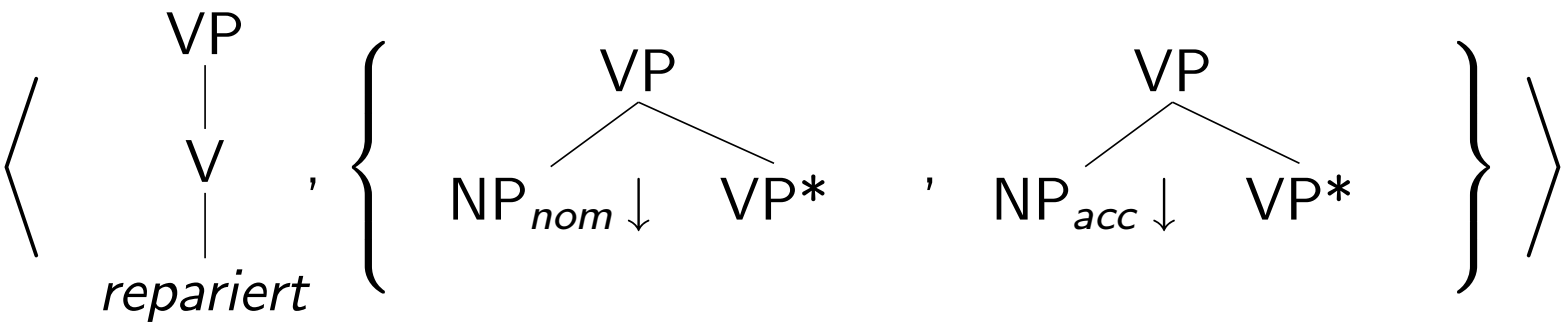
(2) *dass Peter heute den Kühlschrank repariert hat*
dass den Kühlschrank heute Peter repariert hat
...
('that Peter has repaired the fridge today')

TAG is inappropriate for German, because it is:

- not powerful enough for some constructions
  (i.e., coherent constructions)
- not descriptively adequat
  (i.e., one elementary tree for each permutation)

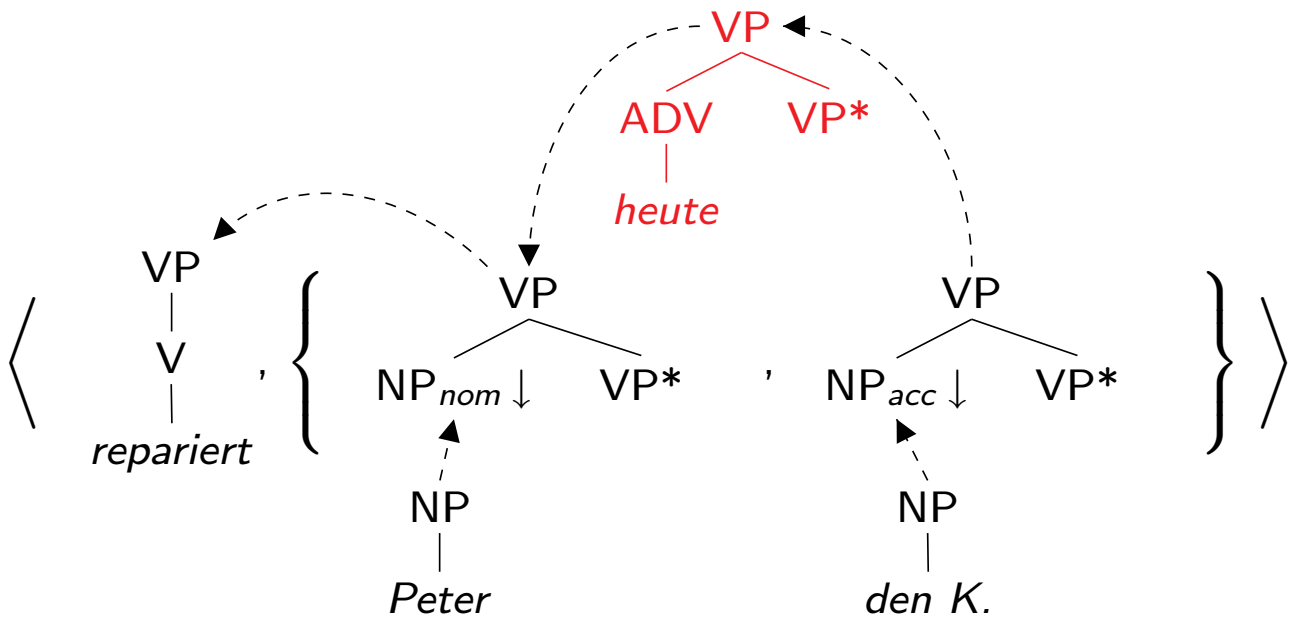
TT-MCTAG: a TAG-extension for German

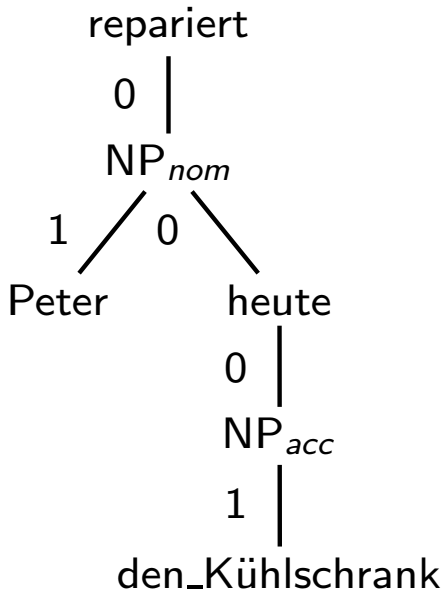
- Multi-Component TAG (MCTAG) with shared-nodes locality
- Elementary structures are **tuples** $\langle \gamma, \{\beta_1, ..., \beta_n\} \rangle$:
  - a lexicalized elementary tree γ (the head tree)
  - a tree set $\{\beta_1, ..., \beta_n\}$ (the complement trees)
- **Meaning of tree tuples:** During derivation, the β-trees have to attach to the γ-tree (via node sharing).
- **Node sharing:** In the derivation tree,
  1. a β-tree must either be the immediate daughter of its γ-tree,
  2. or the β-tree must be connected to the daughter of the γ-tree via a chain of root adjunctions.



TT-MCTAG example

(3) *dass den Kühlschrank heute Peter repariert*
("that Peter repairs the fridge today")





TT-MCTAG: Complexity issues

Is TT-MCTAG polynomially tractable?

Søgaard et al. (2007):

- The MIX language can be generated.
- The universal recognition problem is NP-complete.

k-TT-MCTAG:

- In each derivation step, the number of pending β trees is at most k.
- Mild context-sensitivity has been proven (Kallmeyer and Parmentier, 2008).

What k also means: maximal size of the complement set
What k does not mean: k-gap degree, arity from SN-MCTAG

TT-MCTAG: Complexity issues

Is TT-MCTAG polynomially tractable?

Søgaard et al. (2007):

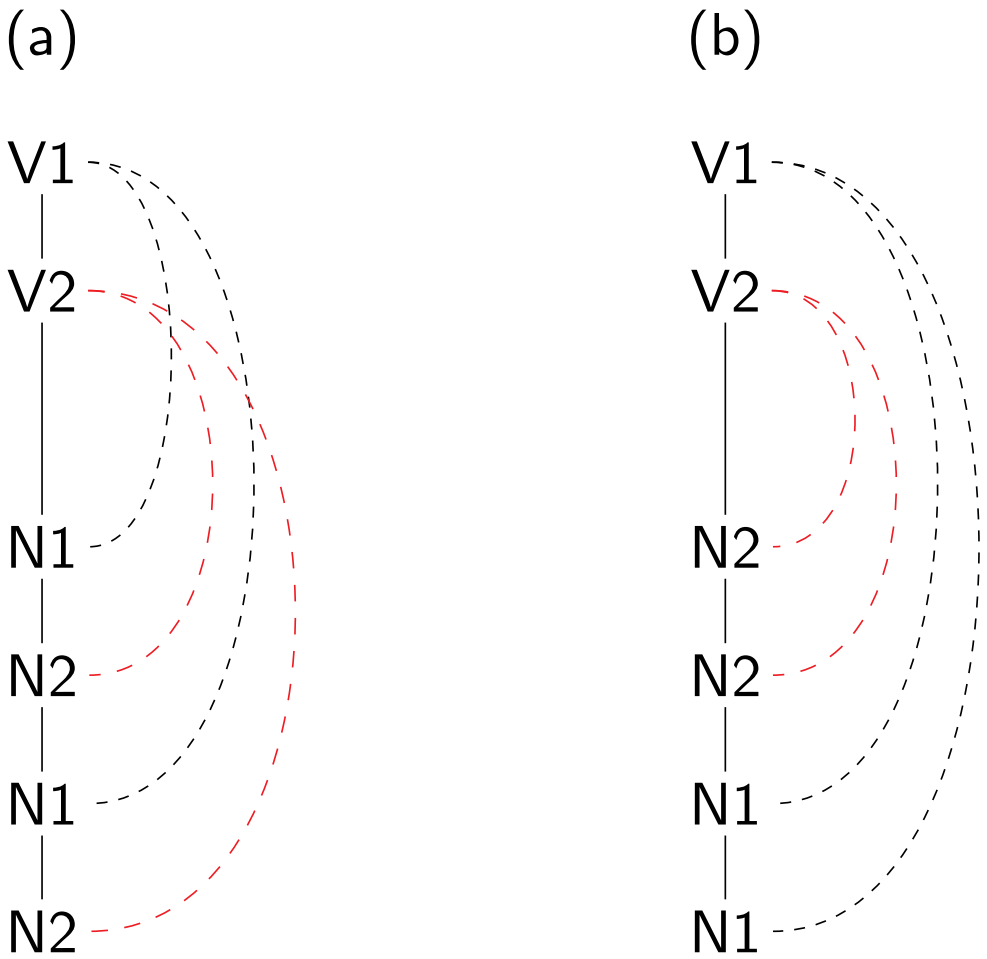
- The MIX language can be generated.
- The universal recognition problem is NP-complete.

k-TT-MCTAG:

- In each derivation step, the number of pending β trees is at most k.
- Mild context-sensitivity has been proven (Kallmeyer and Parmentier, 2008).

What k also means: maximal size of the complement set
What k does not mean: k-gap degree, arity from SN-MCTAG

k-TT-MCTAG: Example



$k \geq 4$

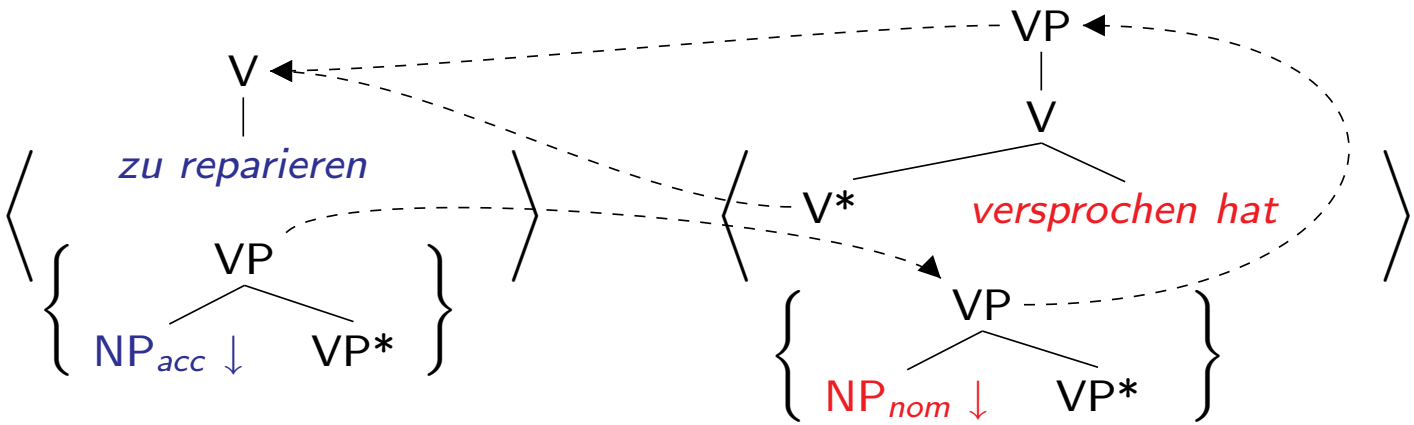
The linguistic application: Principles of complementation

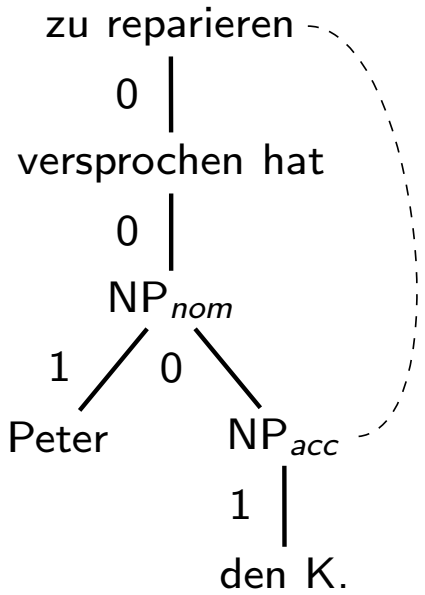
- A tuple = a subcategorization frame, i.e. a head and its complements (as substitution slots and footnodes)
- Substitution = strong islands
- no empty elements (traces, PRO)
- no base order of complements

$\Rightarrow$ **less elementary structures than in a German TAG**

The analyses: Coherent constructions

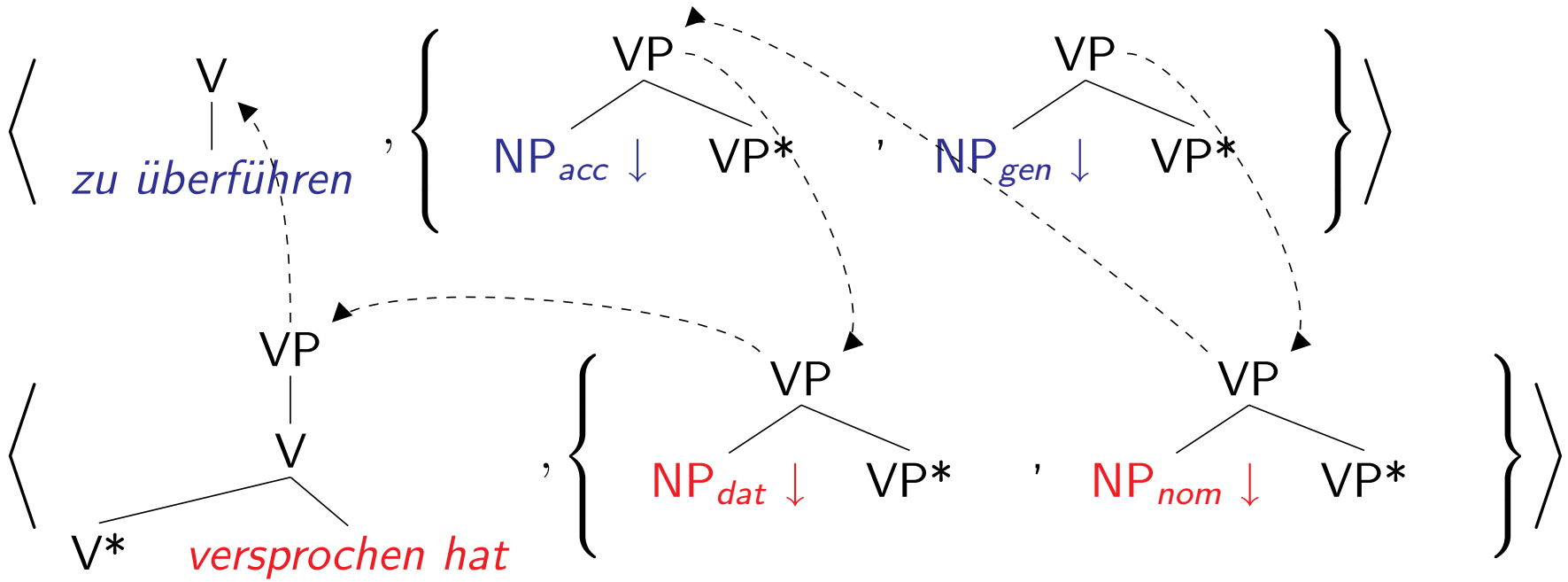
(4) *dass den Kühlschrank Peter zu reparieren versprochen hat*
('that Peter has promised to repair the fridge today')





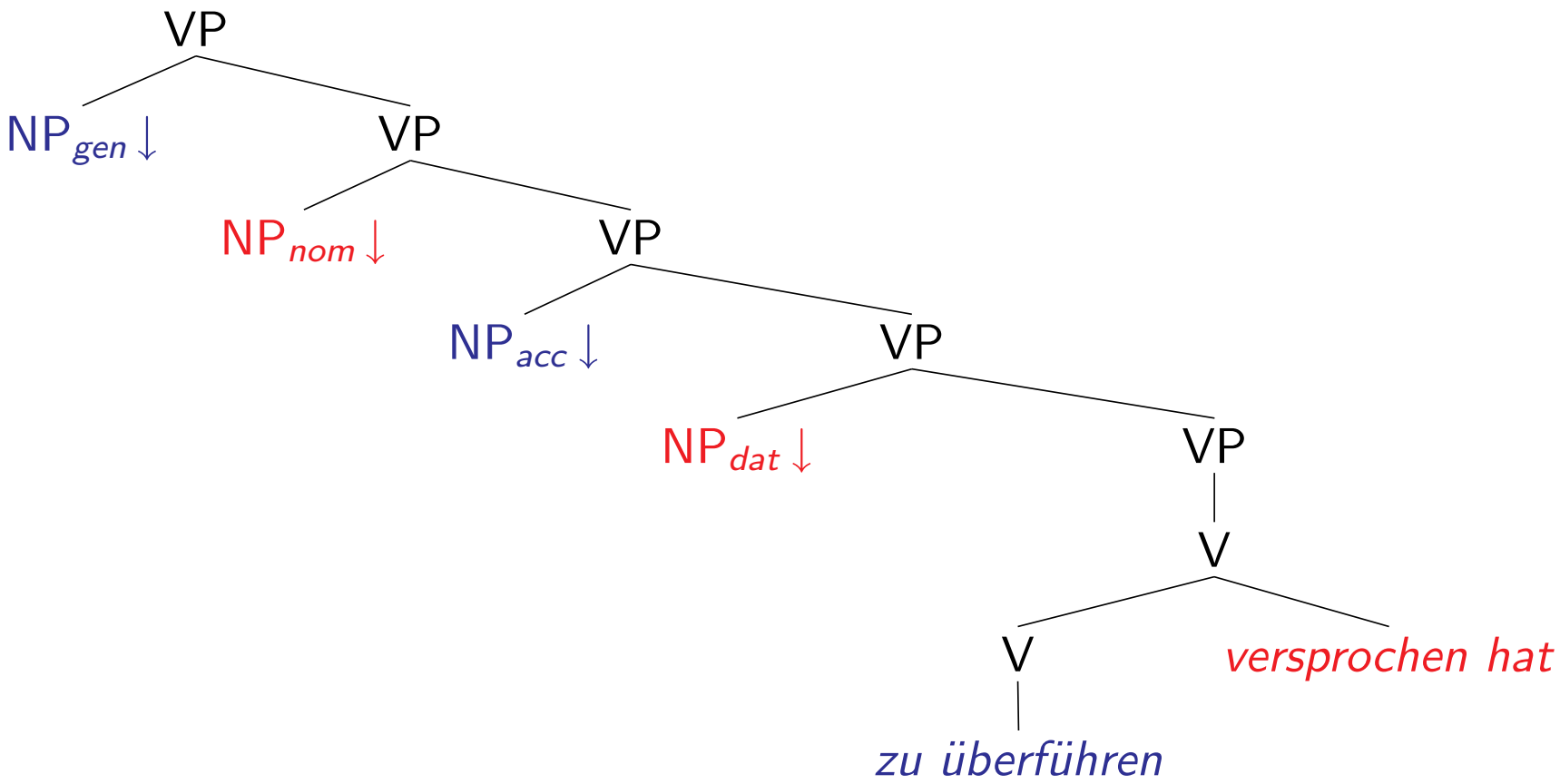
The analysis: Coherent constructions

(5) *dass des Verbrechens der Detektiv den Verdächtigen dem Klienten | zu überführen versprochen hat*
('that the detective has promised the client to indict the suspect of the crime')



The analysis: Coherent constructions

(5) *dass des Verbrechens der Detektiv den Verdächtigen dem Klienten | zu überführen versprochen hat*
('that the detective has promised the client to indict the suspect of the crime')



The analyses: Wh-"extraction"

(6) a. . . . , **wen/den** *heute Peter repariert*
which today Peter repairs
b. . . . , **wann** *ihn heute Peter repariert*
when him today Peter repairs

(7) *dass ihn heute Peter repariert*
that him today Peter repairs

Generalisation (for sentential complements)

- Verb-final sentences can be marked by its leftmost element.
- Verb-final sentences are derived using one and the same tree tuple for the main verb.

The analyses: Wh-"extraction"

(6) a. . . . , **wen/den** *heute Peter repariert*
which today Peter repairs
b. . . . , **wann** *ihn heute Peter repariert*
when him today Peter repairs

(7) *dass ihn heute Peter repariert*
that him today Peter repairs

Generalisation (for sentential complements)

- Verb-final sentences can be marked by its leftmost element.
- Verb-final sentences are derived using one and the same tree tuple for the main verb.

The analyses: Wh-"extraction"

The complex feature S-TYPE:

$$\begin{bmatrix} \text{S-TYPE} & \begin{bmatrix} \text{CONFIG} & (\text{V12}|\text{V3}) \\ \text{MARKED} & (+|\text{-}) \\ \text{MARKING} & (\text{dass}|\text{ob}|\text{wh}|\text{rel}|\ldots) \\ \text{COMP} & (+|\text{-}) \end{bmatrix} \end{bmatrix}$$

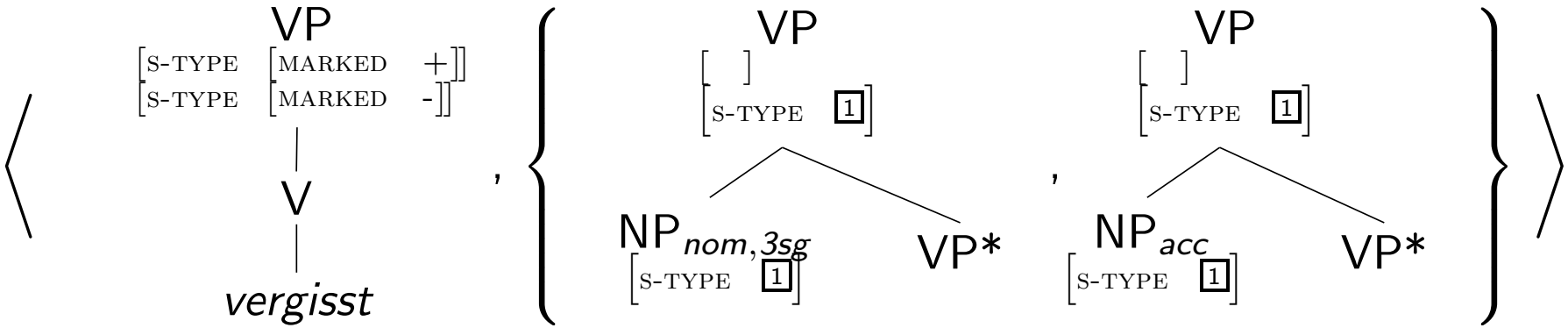
The basic tuple for verb-final sentences:

The analyses: Wh-"extraction"

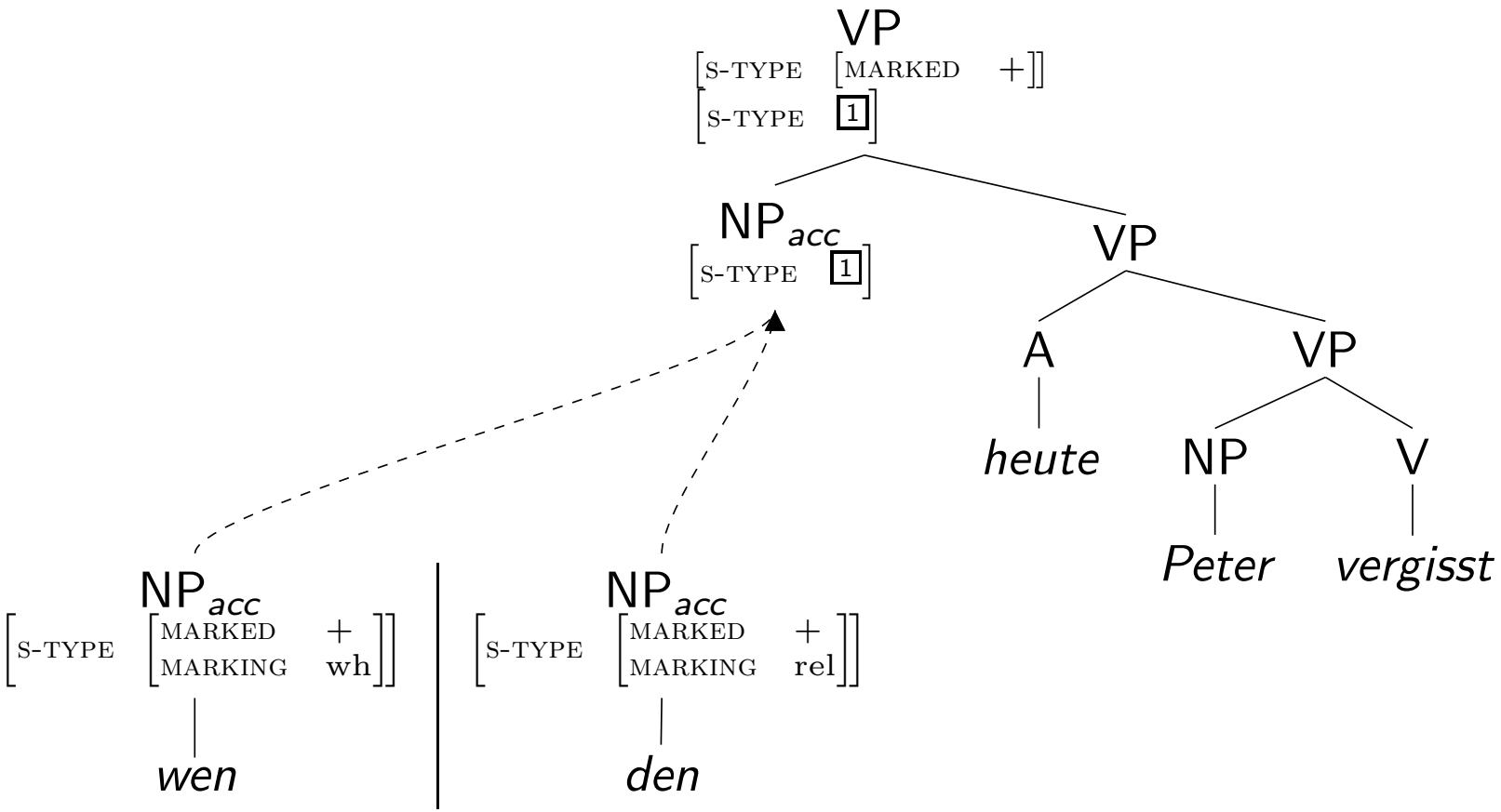
The complex feature S-TYPE:

$$\begin{bmatrix} \text{S-TYPE} & \begin{bmatrix} \text{CONFIG} & (\text{V12}|\text{V3}) \\ \text{MARKED} & (+|\text{-}) \\ \text{MARKING} & (\text{dass}|\text{ob}|\text{wh}|\text{rel}|\ldots) \\ \text{COMP} & (+|\text{-}) \end{bmatrix} \end{bmatrix}$$

The basic tuple for verb-final sentences:



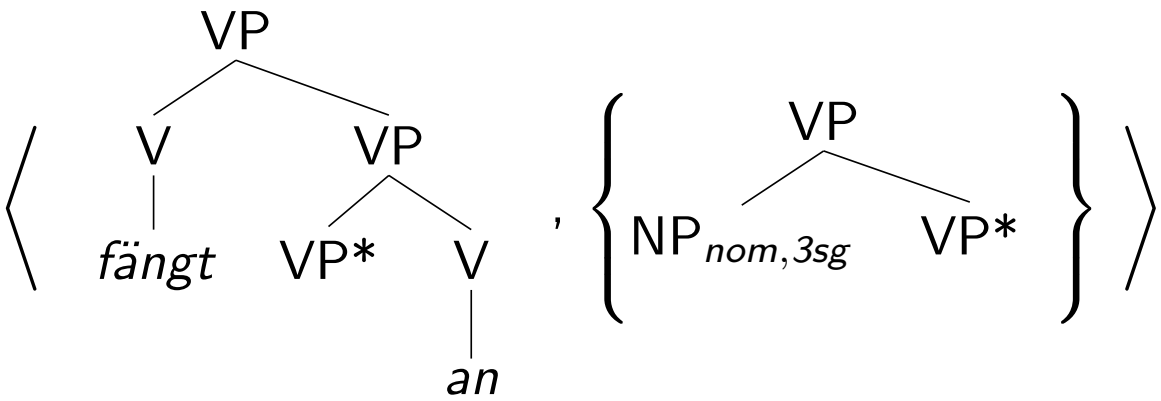
The analyses: Wh-"extraction"



The limits: Scrambling in the Mittelfeld

Scrambling in the V2-Mittelfeld: The governed verb and its object are not adjacent in the Mittelfeld!

(8) *Heute fängt ihn Peter zu reparieren an.*
today begins him Peter to repair PART
('Peter begins to repair him/it')



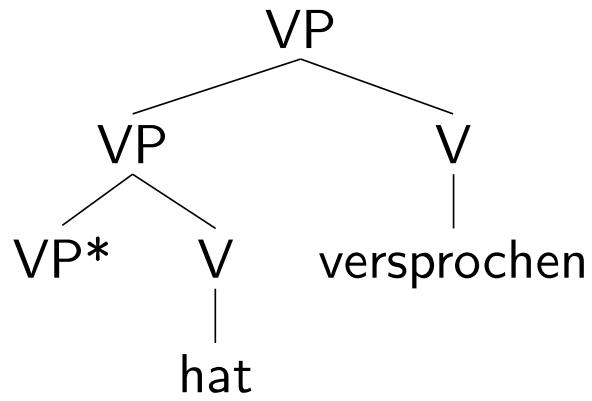
⇝ **"spine-sharing" ??**

The limits: Fronting

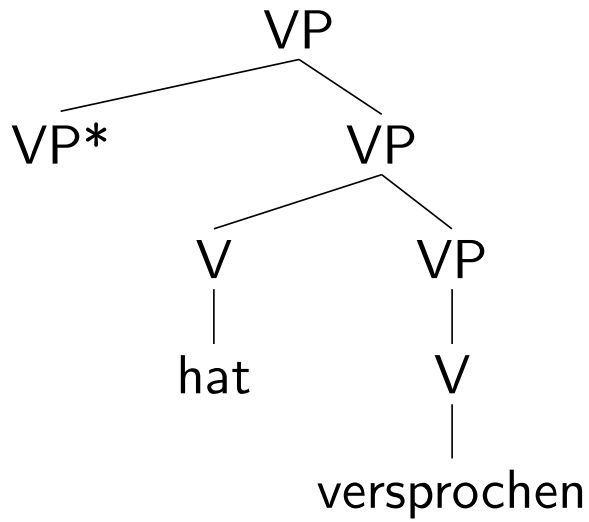
Partial fronting: The verb is to the left of its complement!

(9) *Zu reparieren hat Peter den Kühlschrank versprochen.*
to repair has Peter the fridge promised
('Peter has promised to repair the fridge')

δ_1:



δ_2:



$\rightsquigarrow$ **"tree-sharing" ??**

The limits: Extraposition and elliding coordination

Extraposed relative clauses:

(10) *[Die Feigen [von jenem* **Baum***]] sind reif,* **an dem** *der Tourist jeden Morgen vorbeikommt.*
('the figs of the tree are ripe, that the tourist comes by every day')

The antecedent can be arbitrarily deeply embedded in a noun phrase.

Elliding coordination:

(11) *Der Tourist pflückt eine Feige und*
dann ~~pflückt~~ ~~der Tourist~~ nochmal eine.
('the tourist picks a fig and then again another one')

Complements and even heads can be ellided in coordination structures.

The limits: Extraposition and elliding coordination

Extraposed relative clauses:

(10) *[Die Feigen [von jenem* **Baum***]] sind reif,* **an dem** *der Tourist jeden Morgen vorbeikommt.*
('the figs of the tree are ripe, that the tourist comes by every day')

The antecedent can be arbitrarily deeply embedded in a noun phrase.

Elliding coordination:

(11) *Der Tourist pflückt eine Feige und*
dann ~~pflückt~~ ~~der Tourist~~ nochmal eine.
('the tourist picks a fig and then again another one')

Complements and even heads can be ellided in coordination structures.

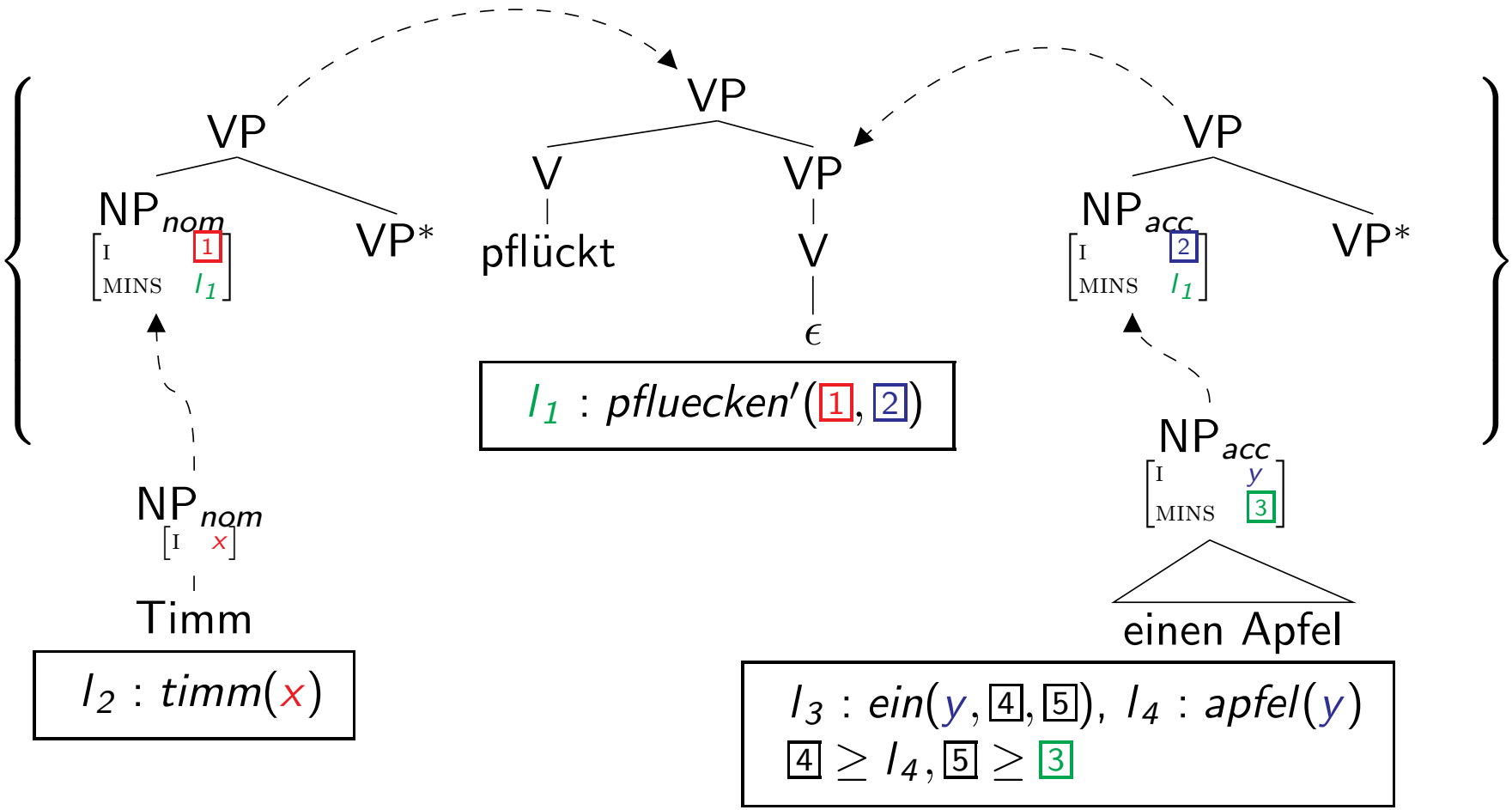
Adding semantics to GerTT

The trees in GerTT are equipped with semantic representations (Gardent & Kallmeyer 2003, Kallmeyer & Romero 2008):

- in the TT-MCTAG each elementary tree can be linked to a semantic representation containing meta-variables for values that will be obtained form other representations during semantic computation;
- identifications of these meta-variables with possible values is done via the feature unifications already implemented in TAG;
- in addition, scope constraints allow to generate underspecified representations.

Adding semantics to GerTT

(12) *Timm pflückt einen Apfel.*
('Timm$_{nom}$ picks an apple$_{acc}$')



The grammar implementation: Syntax

Implementation framework: **XMG** and **TuLiPA** (see next talk)

State of implementation:

- free word order phenomena:
  scrambling, coherent constructions, verbal clustering
- extraction phenomena:
  relative clauses, wh-questions, bridging constructions
- ca. 70 XMG-classes + 26 features (without semantics)

adjectives	3 + 4
adpositions	1 + 5
determiners	1 + 2
...	...
nouns	6 + 3
verbs	18 + **20**

subcat-frame related

The grammar implementation: Syntax

Implementation framework: **XMG** and **TuLiPA** (see next talk)

State of implementation:

- free word order phenomena:
  scrambling, coherent constructions, verbal clustering
- extraction phenomena:
  relative clauses, wh-questions, bridging constructions
- ca. 70 XMG-classes + 26 features (without semantics)

adjectives	3 + 4
adpositions	1 + 5
determiners	1 + 2
...	...
nouns	6 + 3
verbs	18 + **20**

subcat-frame related

The grammar implementation: Syntax

Implementation framework: **XMG** and **TuLiPA** (see next talk)

State of implementation:

- free word order phenomena:
  scrambling, coherent constructions, verbal clustering
- extraction phenomena:
  relative clauses, wh-questions, bridging constructions
- ca. 70 XMG-classes + 26 features (without semantics)

Currently, coverage testing is prepared based on the TSNLP test suite.

The grammar implementation: Semantics

Semantic coverage of the first release of GerTT, GerTT-8.05:

- verbal predicates
- intersective N-modifiers (adjectives and relative clauses)
- quantifiers
- control, raising
- intensional adverbs

But the currently implemented semantics is simplified:

- only predicate-argument relations and scope relations, no situation variables;
- quantifiers are too free in their scope behaviour, word order is not sufficiently taken into account;
- auxiliaries don't have a semantics.

The grammar implementation: Semantics

Semantic coverage of the first release of GerTT, GerTT-8.05:

- verbal predicates
- intersective N-modifiers (adjectives and relative clauses)
- quantifiers
- control, raising
- intensional adverbs

But the currently implemented semantics is simplified:

- only predicate-argument relations and scope relations, no situation variables;
- quantifiers are too free in their scope behaviour, word order is not sufficiently taken into account;
- auxiliaries don't have a semantics.

The lexicon: A 2-layered format

Morphological lexicon

maps an (inflected) token to some lemma form, while preserving morphological information in a feature structure.

vergisst	vergessen	[pos=v; num=sg; per=3;]

Lemma lexicon

maps a lemma onto tree tuple families, while also containing selectional restrictions (e.g., case assignment).

*ENTRY: vergessen
*CAT: v
*SEM: BinaryRel[pred=vergessen]
*ACC: 1
*FAM: Vnp2
*FILTERS: []
*EX:
*EQUATIONS:
NParg1 → cas = nom
NParg2 → cas = acc
*COANCHORS:

The lexicon: Sources for extension

Sources of an extended 2-layered lexicon:

- Morphologial features: NEGRA
- Lemmatization and POS-tags: TüPP-D/Z
- Subcategorization frames: Resources from Schulte im Walde (2002)

$\rightsquigarrow$ ca. 25 000 morph entries

On the agenda:

- Handling of lexical gaps
- Coverage testing with real-world data

The lexicon: Sources for extension

Sources of an extended 2-layered lexicon:

- Morphologial features: NEGRA
- Lemmatization and POS-tags: TüPP-D/Z
- Subcategorization frames: Resources from Schulte im Walde (2002)

$\rightsquigarrow$ ca. 25 000 morph entries

On the agenda:

- Handling of lexical gaps
- Coverage testing with real-world data

Summary and outlook

- GerTT is a large coverage TAG-based grammar for German that is under development
- goal: cover a substantial number of German syntactic constructions ($\rightarrow$ TSNLP)
- partly extended with semantics

- TAG not adequate for German, therefore TT-MCTAG is used
- lexical entries are tuples that factor complementation
- free word order is modelled via "delayed" adjunctions (tree-locality under node sharing)
- advantages of k-TT-MCTAG: mild context-sensitivity, locality (leads to natural definitions of islands)

Summary and outlook

- GerTT is a large coverage TAG-based grammar for German that is under development
- goal: cover a substantial number of German syntactic constructions ($\rightarrow$ TSNLP)
- partly extended with semantics

- TAG not adequate for German, therefore TT-MCTAG is used
- lexical entries are tuples that factor complementation
- free word order is modelled via "delayed" adjunctions (tree-locality under node sharing)
- advantages of k-TT-MCTAG: mild context-sensitivity, locality (leads to natural definitions of islands)