
Syntax-Based Models

Philipp Koehn

(Lecture given by Shuoyang Ding)

29 March 2016



Tree-Based Models



- Traditional statistical models operate on sequences of words
- Many translation problems can be best explained by pointing to syntax
 - reordering, e.g., verb movement in German–English translation
 - long distance agreement (e.g., subject-verb) in output

⇒ Translation models based on tree representation of language

- significant ongoing research
- state-of-the art for some language pairs

Tree-Based Translation Model



- Translation by parsing
 - synchronous grammar has to parse entire input sentence
 - output tree is generated at the same time
 - process is broken up into a number of rule applications
- Translation probability

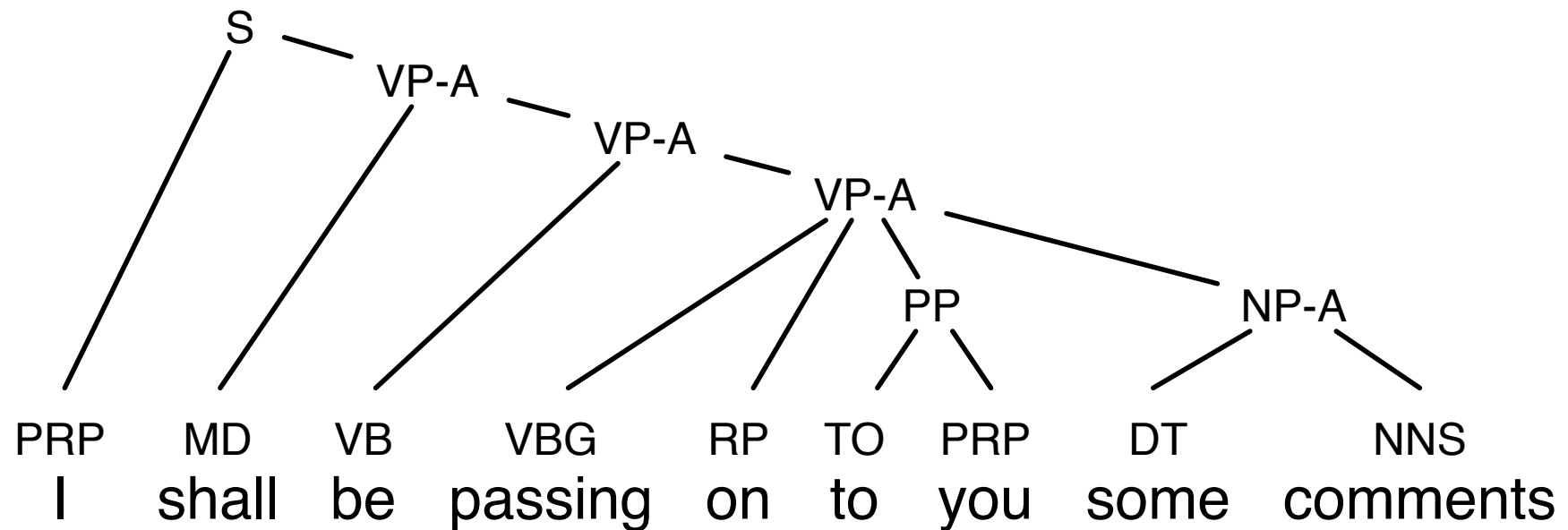
$$\text{SCORE}(\text{TREE}, E, F) = \prod_i \text{RULE}_i$$

- Many ways to assign probabilities to rules

Syntactic Rules

- Phrase structure
 - noun phrases: *the big man*, *a house*, ...
 - prepositional phrases: *at 5 o'clock*, *in Edinburgh*, ...
 - verb phrases: *going out of business*, *eat chicken*, ...
 - adjective phrases, ...
- Context-free Grammars (CFG)
 - non-terminal symbols: phrase structure labels, part-of-speech tags
 - terminal symbols: words
 - production rules: $NT \rightarrow [NT, T]^+$
example: $NP \rightarrow DET\ NN$

Phrase Structure Grammar



Phrase structure grammar tree for an English sentence
(as produced Collins' parser)

- English rule

$NP \rightarrow DET\ JJ\ NN$

- French rule

$NP \rightarrow DET\ NN\ JJ$

- Synchronous rule (indices indicate alignment):

$NP \rightarrow DET_1\ NN_2\ JJ_3 \mid DET_1\ JJ_3\ NN_2$

Synchronous Grammar Rules



7

- Nonterminal rules

$$\text{NP} \rightarrow \text{DET}_1 \text{NN}_2 \text{JJ}_3 \mid \text{DET}_1 \text{JJ}_3 \text{NN}_2$$

- Terminal rules

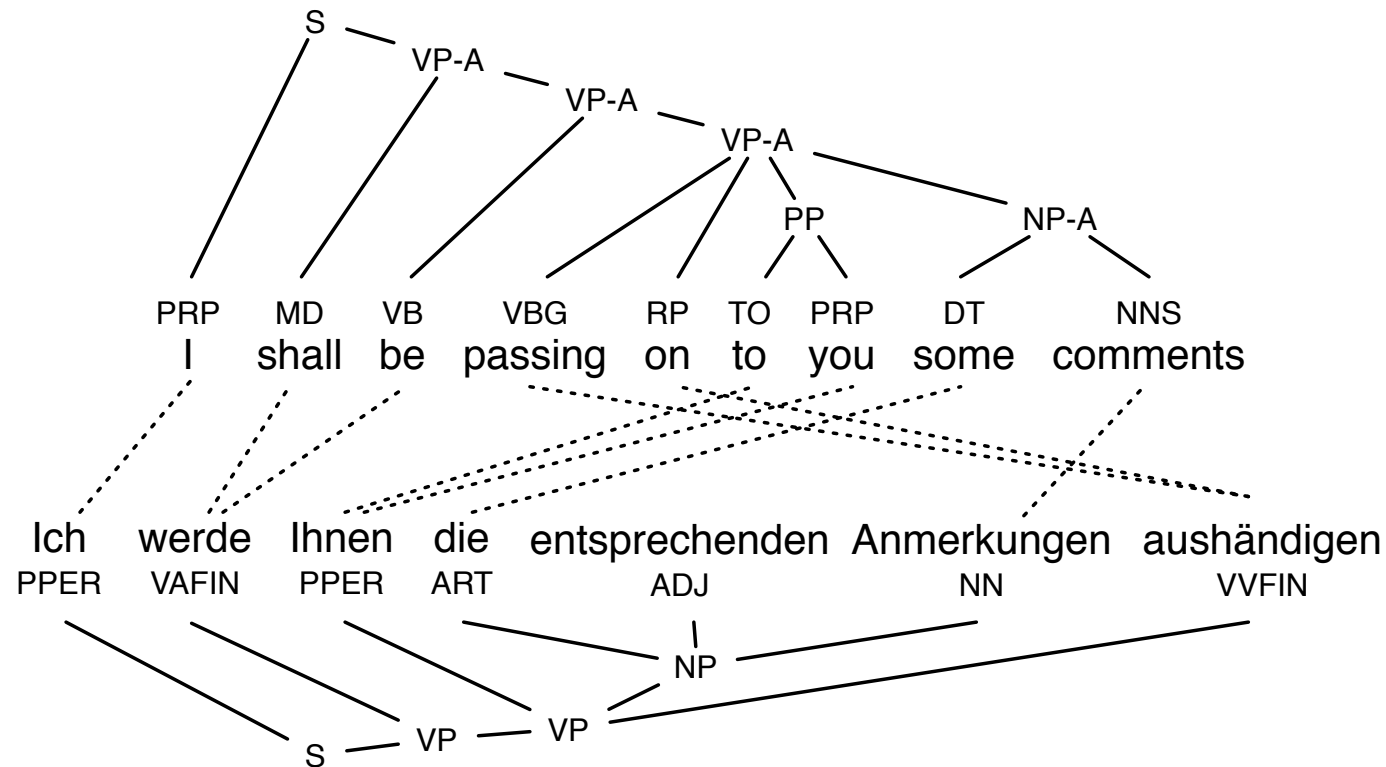
$$\text{N} \rightarrow \text{maison} \mid \text{house}$$
$$\text{NP} \rightarrow \text{la maison bleue} \mid \text{the blue house}$$

- Mixed rules

$$\text{NP} \rightarrow \text{la maison JJ}_1 \mid \text{the JJ}_1 \text{house}$$

Learning Syntactic Rules

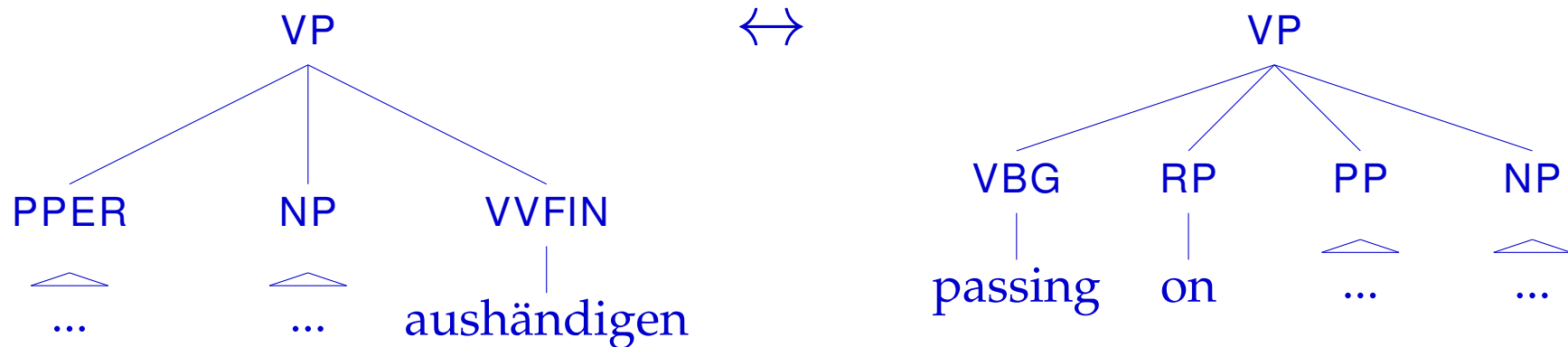
Aligned Tree Pair



Phrase structure grammar trees with word alignment
(German–English sentence pair.)

Reordering Rule

- Subtree alignment



- Synchronous grammar rule

$VP \rightarrow PPER_1 NP_2 \text{ aushändigen} \mid \text{passing on } PP_1 NP_2$

- Note:

- one word **aushändigen** mapped to two words **passing on** ok
- but: fully non-terminal rule not possible
(one-to-one mapping constraint for nonterminals)

Another Rule

- Subtree alignment



- Synchronous grammar rule (stripping out English internal structure)

$\text{PRO/PP} \rightarrow \text{Ihnen} \mid \text{to you}$

- Rule with internal structure

$\text{PRO/PP} \rightarrow \text{Ihnen} \mid \begin{array}{cc} \text{TO} & \text{PRP} \\ | & | \\ \text{to} & \text{you} \end{array}$

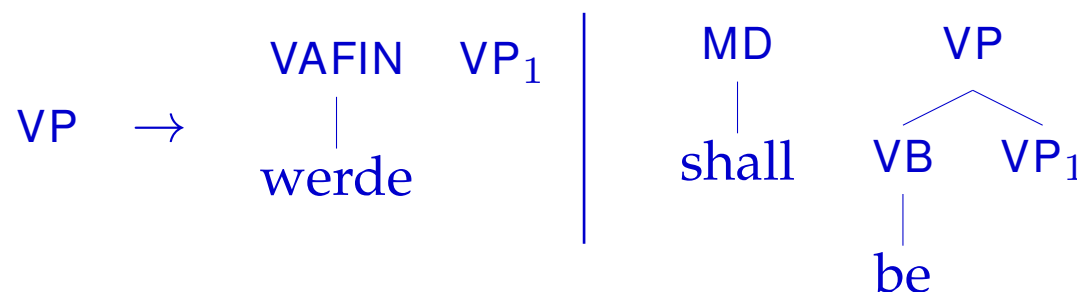
Another Rule

- Translation of German **werde** to English **shall be**



- Translation rule needs to include mapping of **VP**

\Rightarrow Complex rule



Internal Structure

- Stripping out internal structure

VP \rightarrow werde VP₁ | shall be VP₁

\Rightarrow synchronous context free grammar

- Maintaining internal structure

VP →

VP	→	VAFIN	VP ₁		MD	VP
						/ \
		werde			shall	VB VP ₁
						be

\Rightarrow synchronous tree substitution grammar

- Extracting rules from a word-aligned parallel corpus
- First: Hierarchical phrase-based model
 - only one non-terminal symbol **X**
 - no linguistic syntax, just a formally syntactic model
- Then: Synchronous phrase structure model
 - non-terminals for words and phrases: **NP, VP, PP, ADJ, ...**
 - corpus must also be parsed with syntactic parser

Extracting Phrase Translation Rules

15



	Ich	werde	Ihnen	die	entsprechenden	Anmerkungen	aushändigen
I							
shall							
be							
passing							
on							
to							
you							
some							
comments							

.....▶ shall be = werde

Extracting Phrase Translation Rules

16



	Ich	werde	Ihnen	die	entsprechenden	Anmerkungen	aushändigen
I							
shall							
be							
passing							
on							
to							
you							
some							
comments							

some comments =
die entsprechenden Anmerkungen

Extracting Phrase Translation Rules

17



	Ich	werde	Ihnen	die	entsprechenden	Anmerkungen	aushändigen
I							
shall							
be							
passing							
on							
to							
you							
some							
comments							

werde Ihnen die entsprechenden
Anmerkungen aushändigen
= shall be passing on to you
some comments

Extracting Hierarchical Phrase Translation Rules



	Ich	werde	Ihnen	die	entsprechenden	Anmerkungen	aushändigen
I							
shall							
be							
passing							
on							
to							
you							
some							
comments							

subtracting
subphrase

werde X aushändigen
= shall be passing on X

- Recall: consistent phrase pairs

(\bar{e}, \bar{f}) consistent with $A \Leftrightarrow$

$$\begin{aligned} & \forall e_i \in \bar{e} : (e_i, f_j) \in A \rightarrow f_j \in \bar{f} \\ \text{AND } & \forall f_j \in \bar{f} : (e_i, f_j) \in A \rightarrow e_i \in \bar{e} \\ \text{AND } & \exists e_i \in \bar{e}, f_j \in \bar{f} : (e_i, f_j) \in A \end{aligned}$$

- Let P be the set of all extracted phrase pairs (\bar{e}, \bar{f})

- Extend recursively:

if $(\bar{e}, \bar{f}) \in P$ AND $(\bar{e}_{\text{SUB}}, \bar{f}_{\text{SUB}}) \in P$

AND $\bar{e} = \bar{e}_{\text{PRE}} + \bar{e}_{\text{SUB}} + \bar{e}_{\text{POST}}$

AND $\bar{f} = \bar{f}_{\text{PRE}} + \bar{f}_{\text{SUB}} + \bar{f}_{\text{POST}}$

AND $\bar{e} \neq \bar{e}_{\text{SUB}}$ AND $\bar{f} \neq \bar{f}_{\text{SUB}}$

add $(e_{\text{PRE}} + \mathbf{X} + e_{\text{POST}}, f_{\text{PRE}} + \mathbf{X} + f_{\text{POST}})$ to P

(note: any of e_{PRE} , e_{POST} , f_{PRE} , or f_{POST} may be empty)

- Set of hierarchical phrase pairs is the closure under this extension mechanism

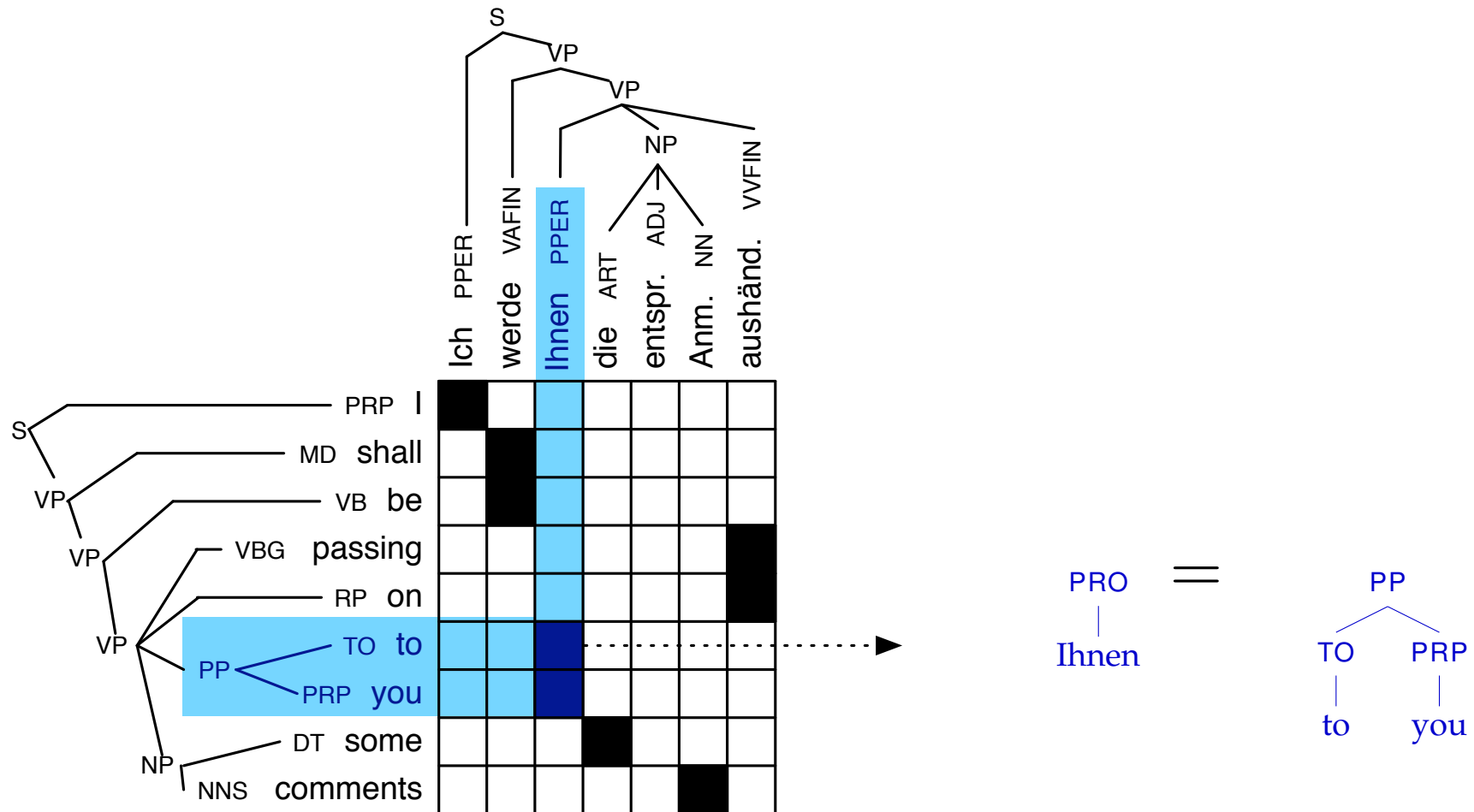


- Removal of multiple sub-phrases leads to rules with multiple non-terminals, such as:

$$Y \rightarrow X_1 X_2 \mid X_2 \text{ of } X_1$$

- Typical restrictions to limit complexity (Chiang, 2005)
 - at most 2 nonterminal symbols
 - no neighboring non-terminals on the source side
 - at least 1 but at most 5 words per language
 - span at most 15 words (counting gaps)

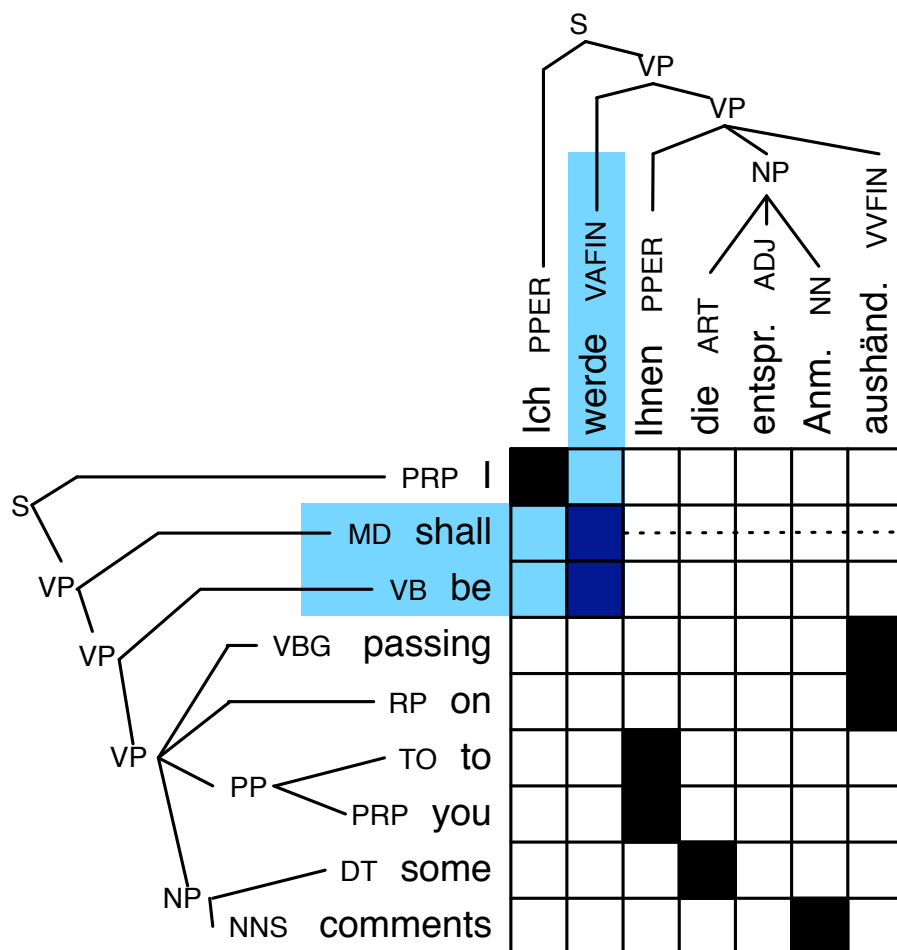
Learning Syntactic Translation Rules



Constraints on Syntactic Rules

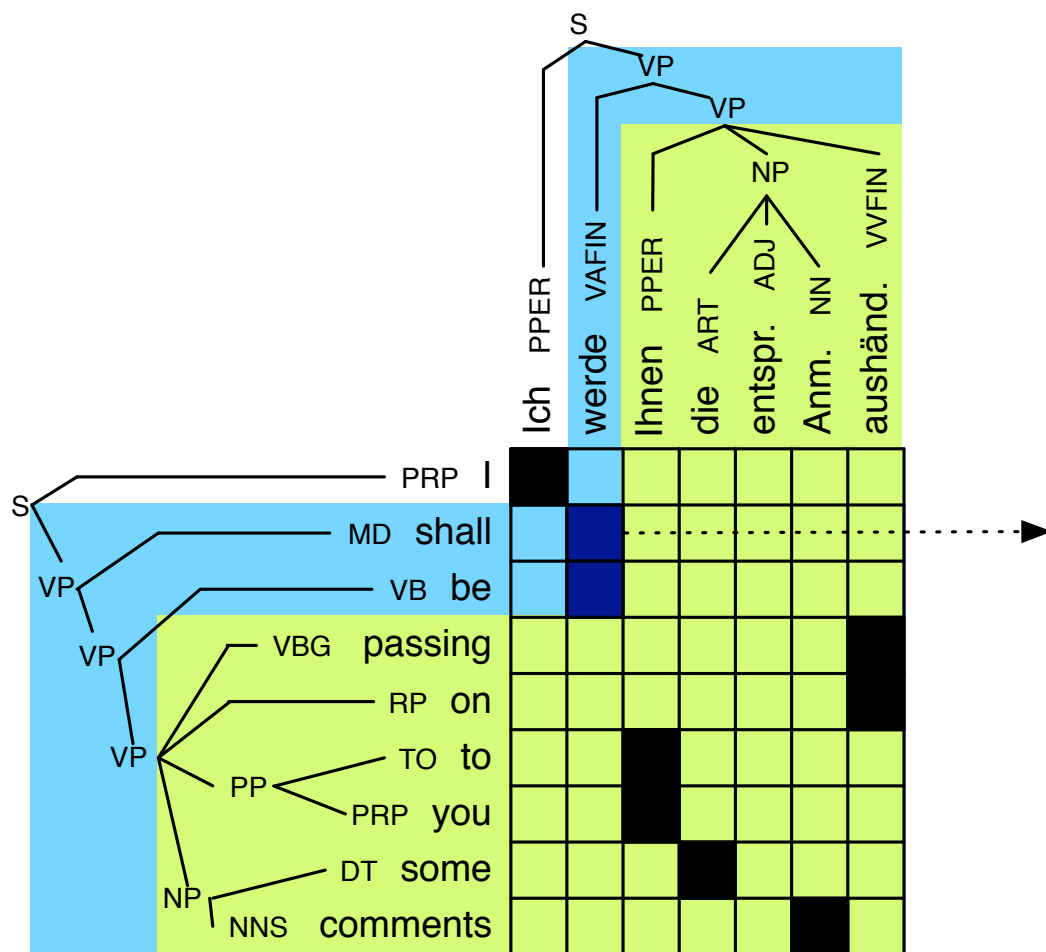
- Same word alignment constraints as hierarchical models
- Hierarchical: rule can cover any span
 \Leftrightarrow syntactic rules must cover constituents in the tree
- Hierarchical: gaps may cover any span
 \Leftrightarrow gaps must cover constituents in the tree
- Much less rules are extracted (all things being equal)

Impossible Rules

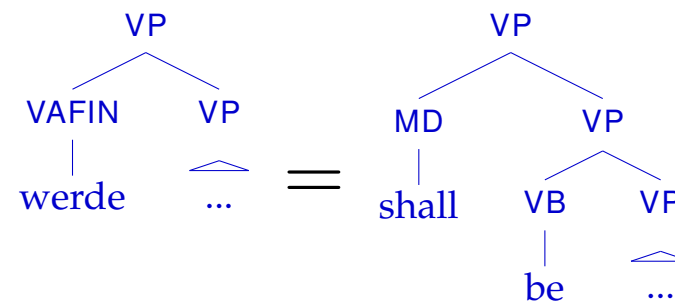


English span not a
constituent
no rule extracted

Rules with Context



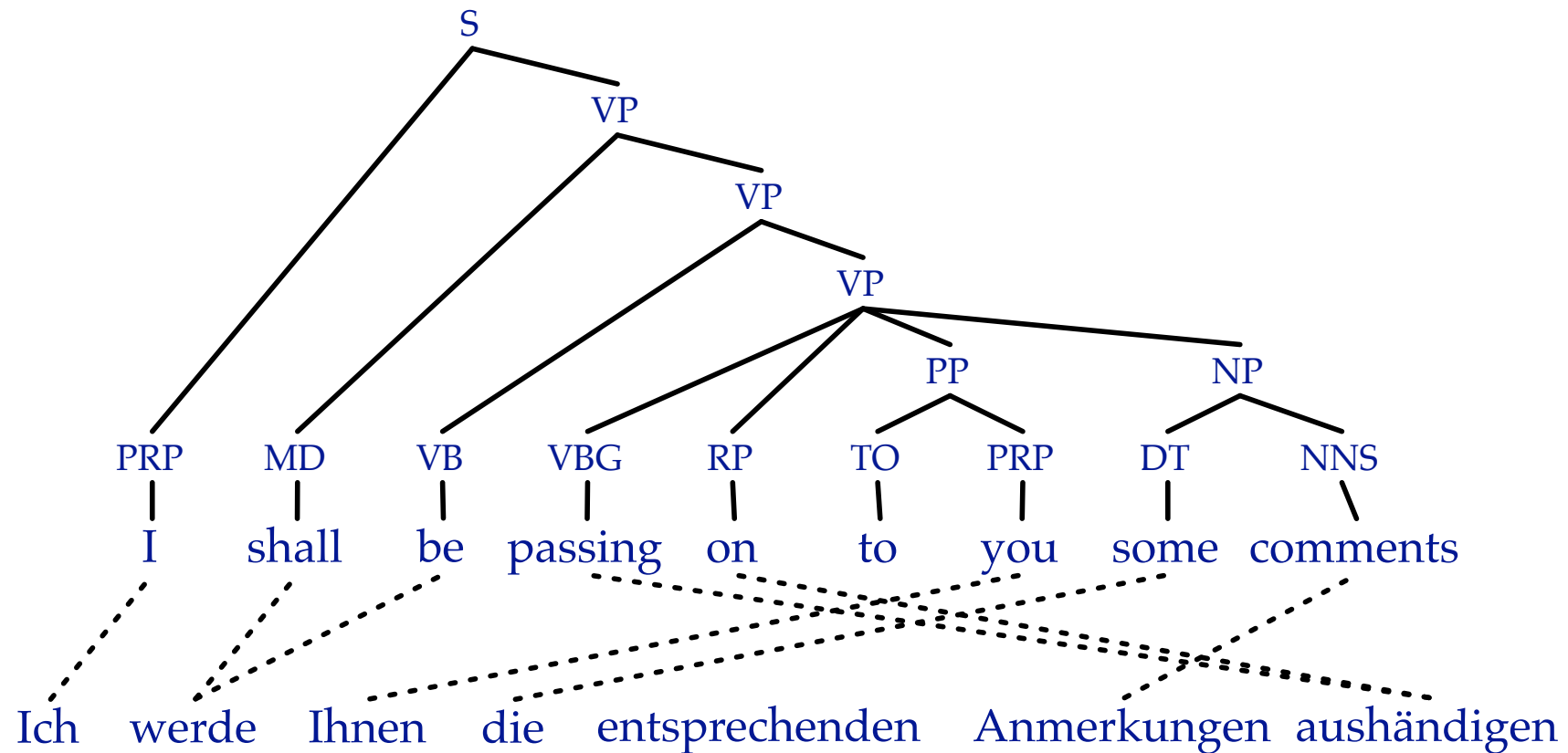
Rule with this phrase pair
requires syntactic context



Too Many... or Too Few?

- Need to limit which rules to extract
 - Option 1: similar restriction as for hierarchical model
(maximum span size, maximum number of terminals and non-terminals, etc.)
 - Option 2: only extract minimal rules ("GHKM" rules)
- Having trees on both sides could potentially fire the rule constraint too often
 - Only use syntactic tree on one side.

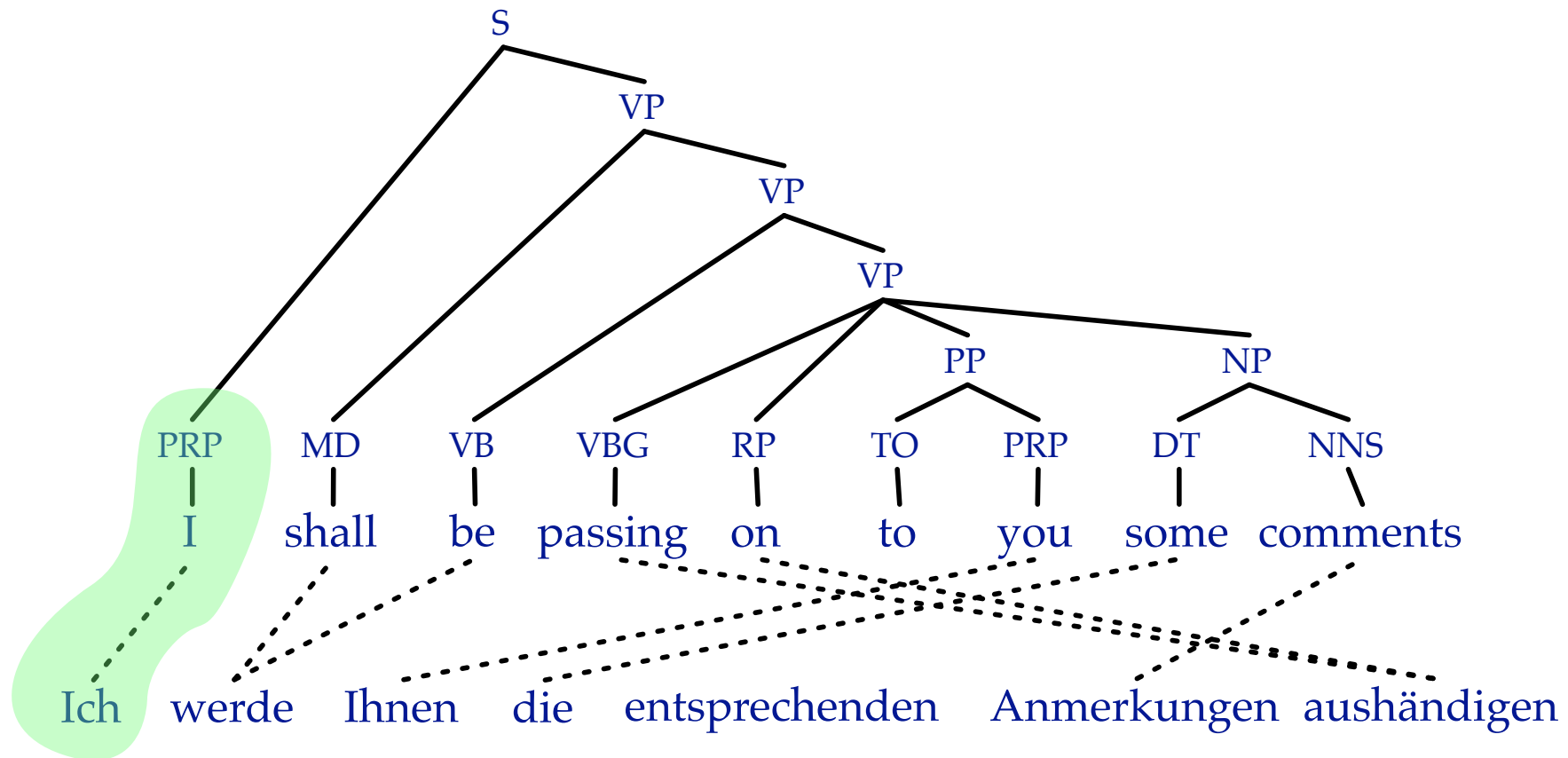
Minimal Rules



Extract: set of smallest rules required to explain the sentence pair

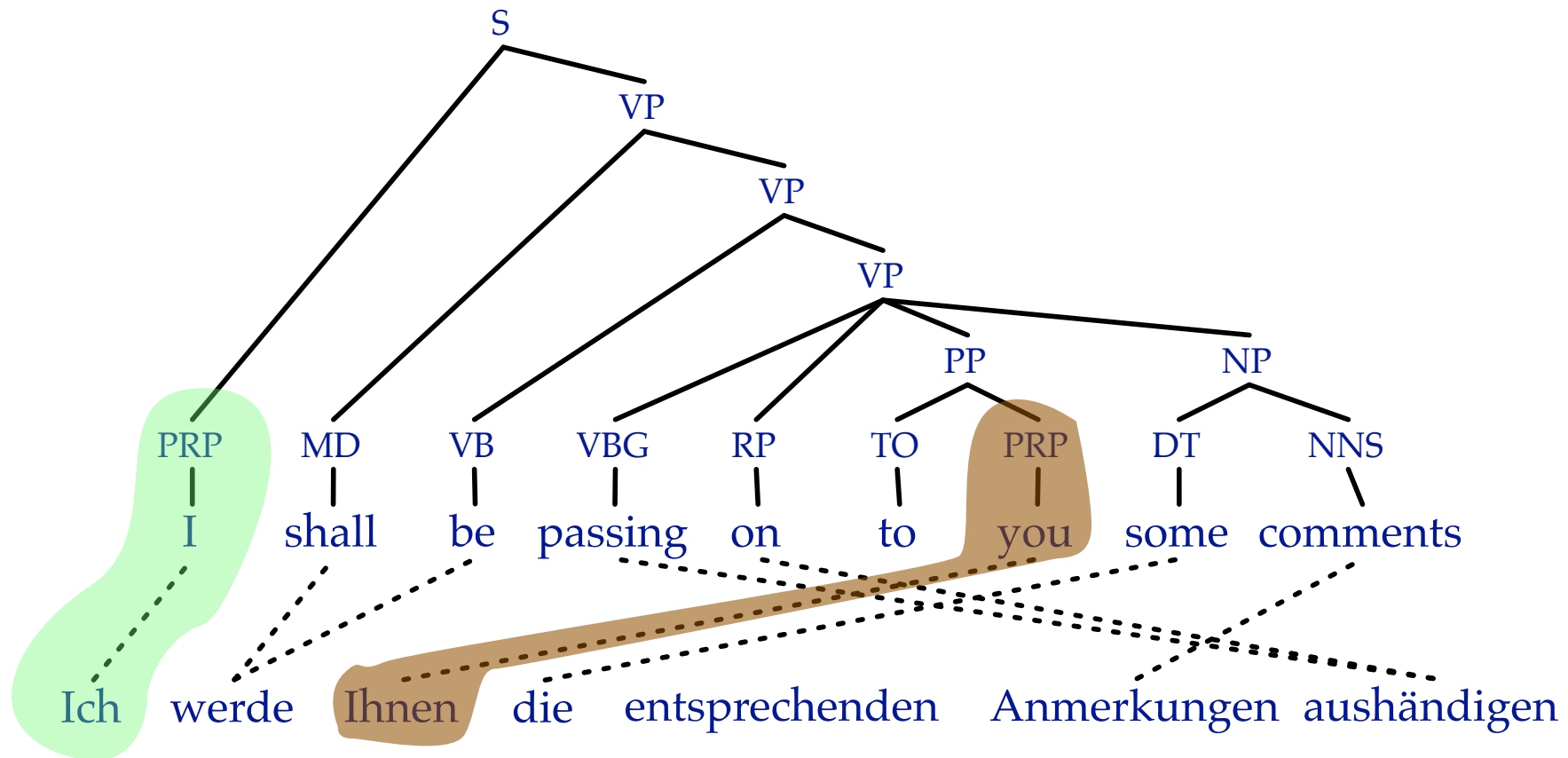
Lexical Rule

28



Extracted rule: $\text{PRP} \rightarrow \text{Ich} \mid \text{I}$

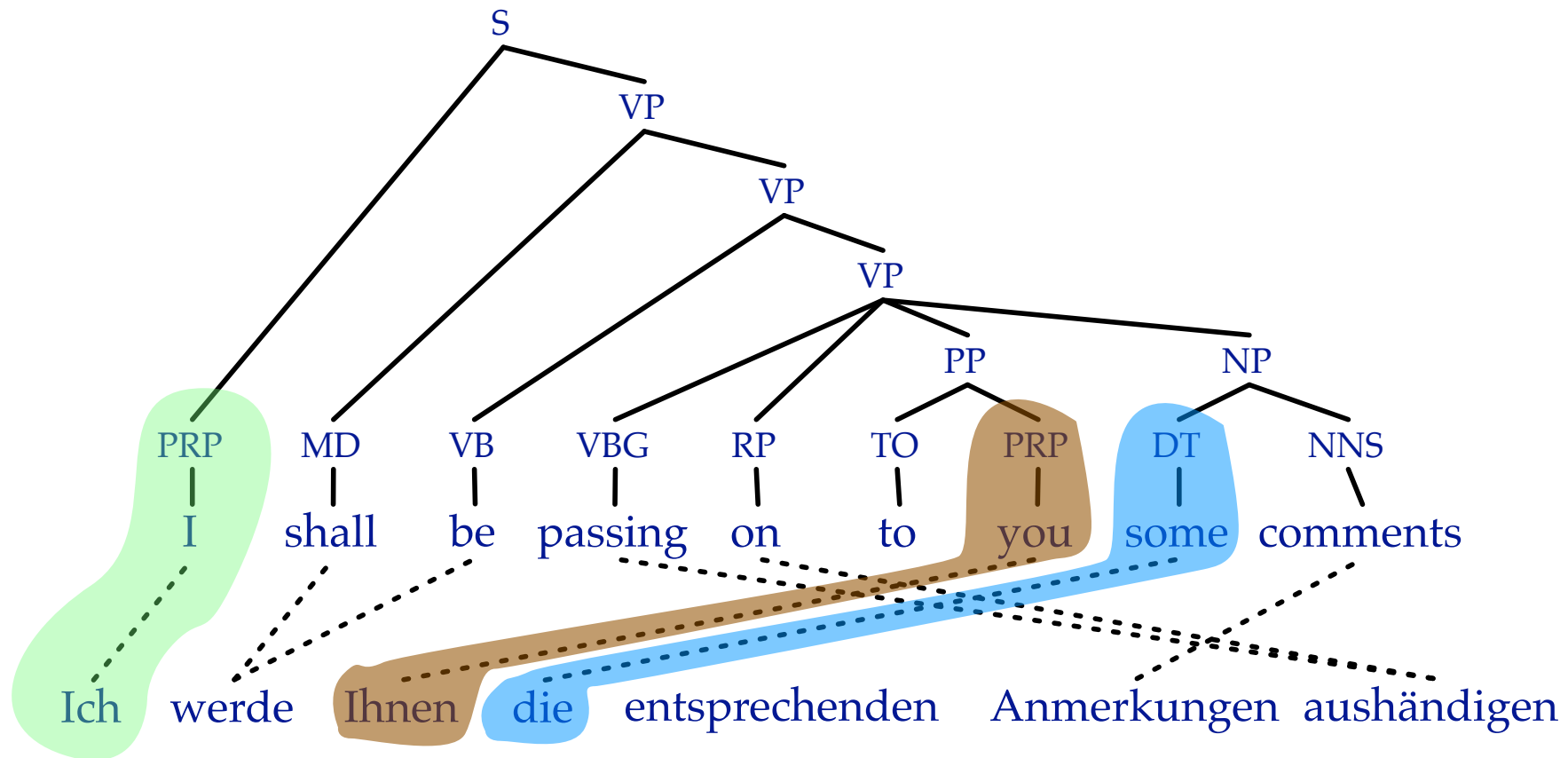
Lexical Rule



Extracted rule: $PRP \rightarrow \text{Ihnen} \mid \text{you}$

Lexical Rule

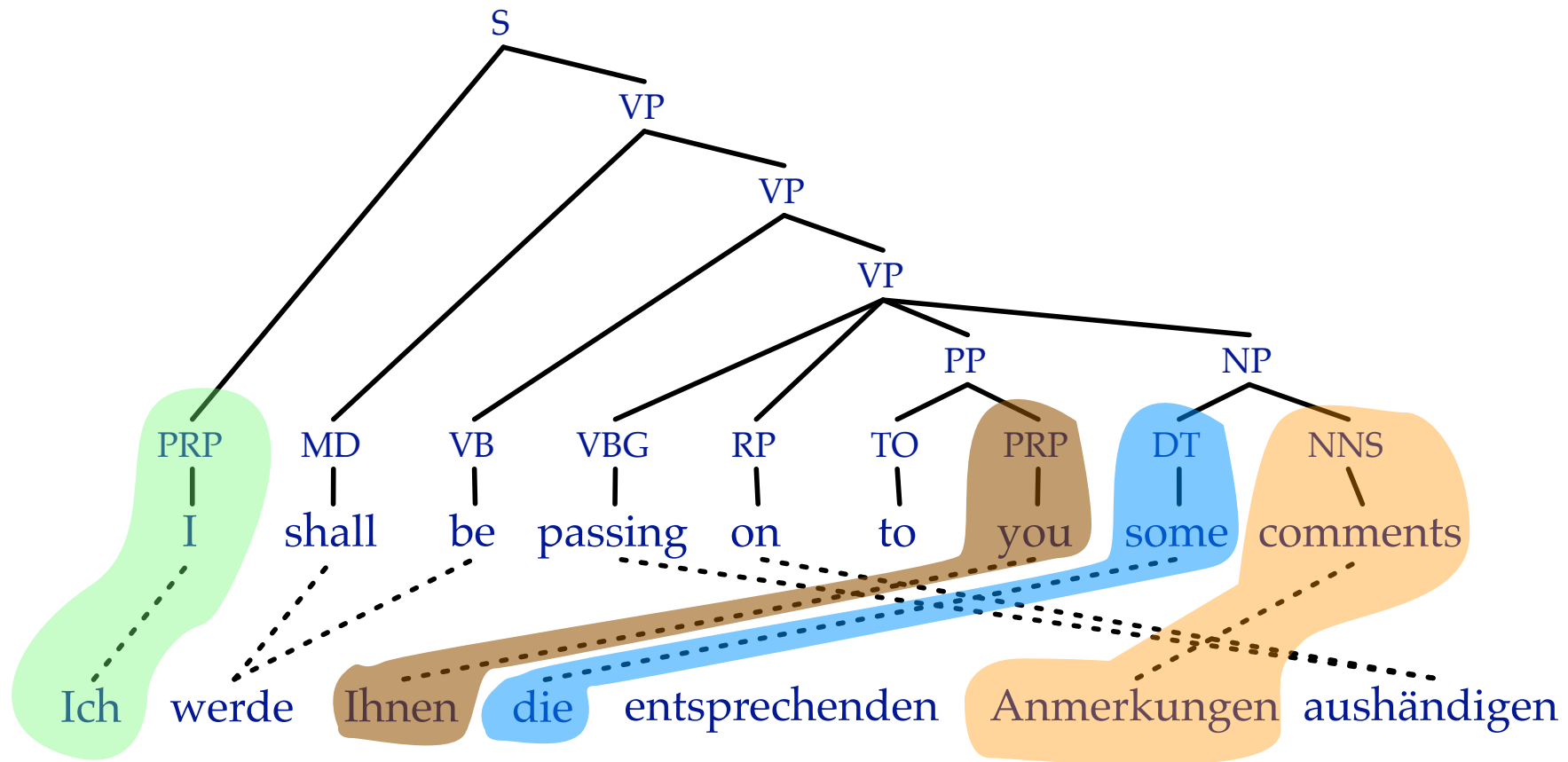
30



Extracted rule: DT → die | some

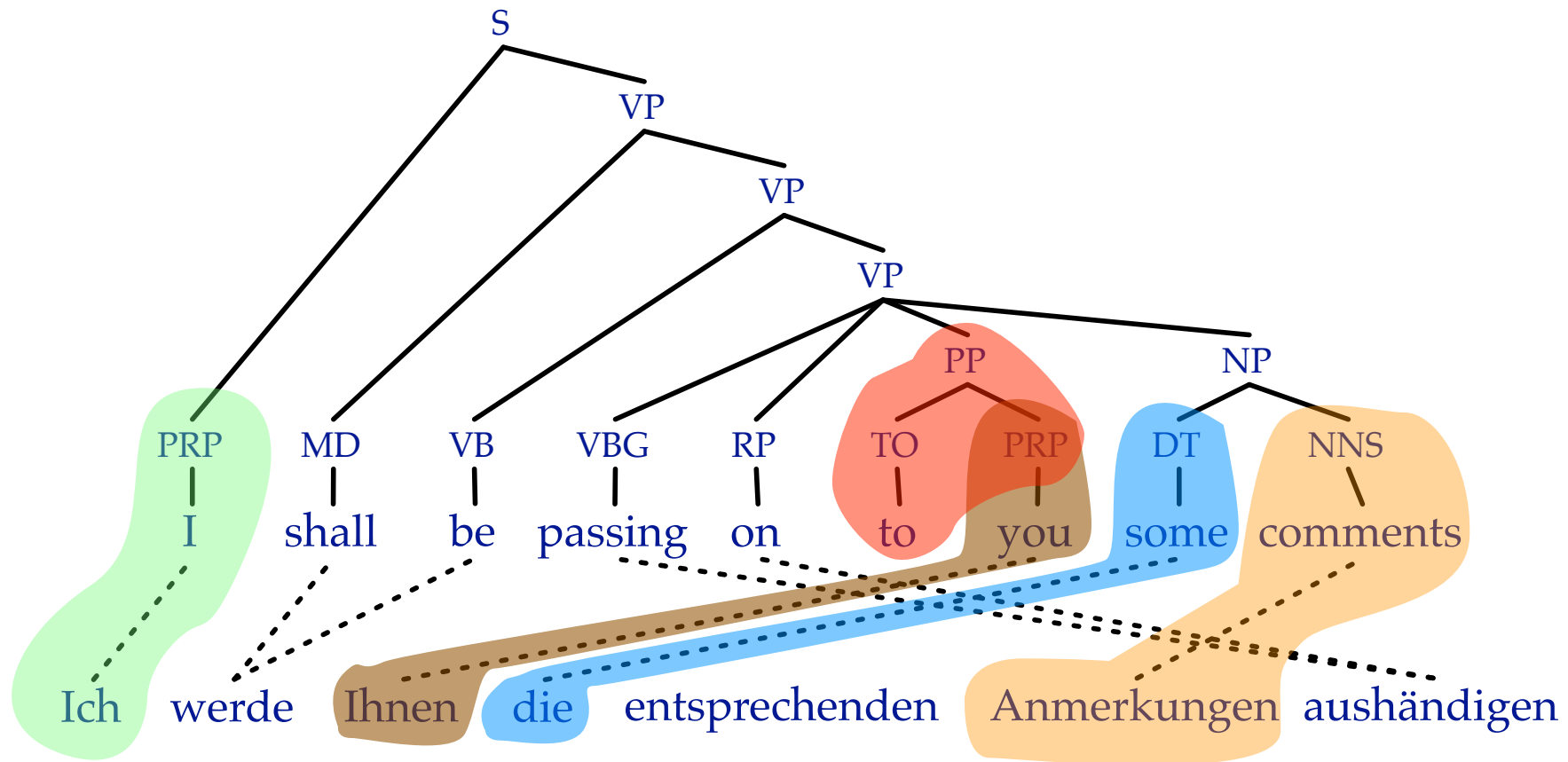
Lexical Rule

31



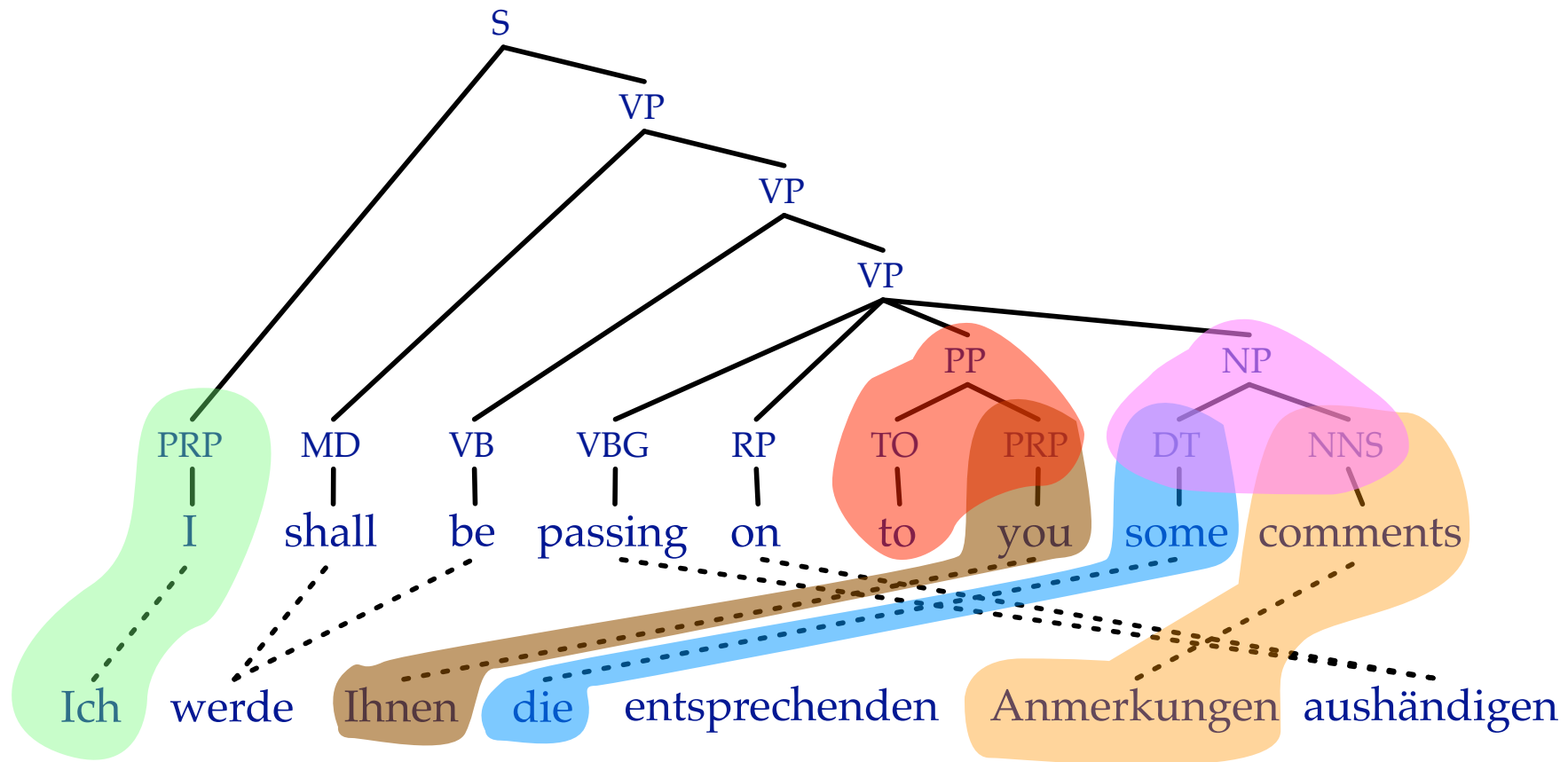
Extracted rule: $NNS \rightarrow \text{Anmerkungen} \mid \text{comments}$

Insertion Rule



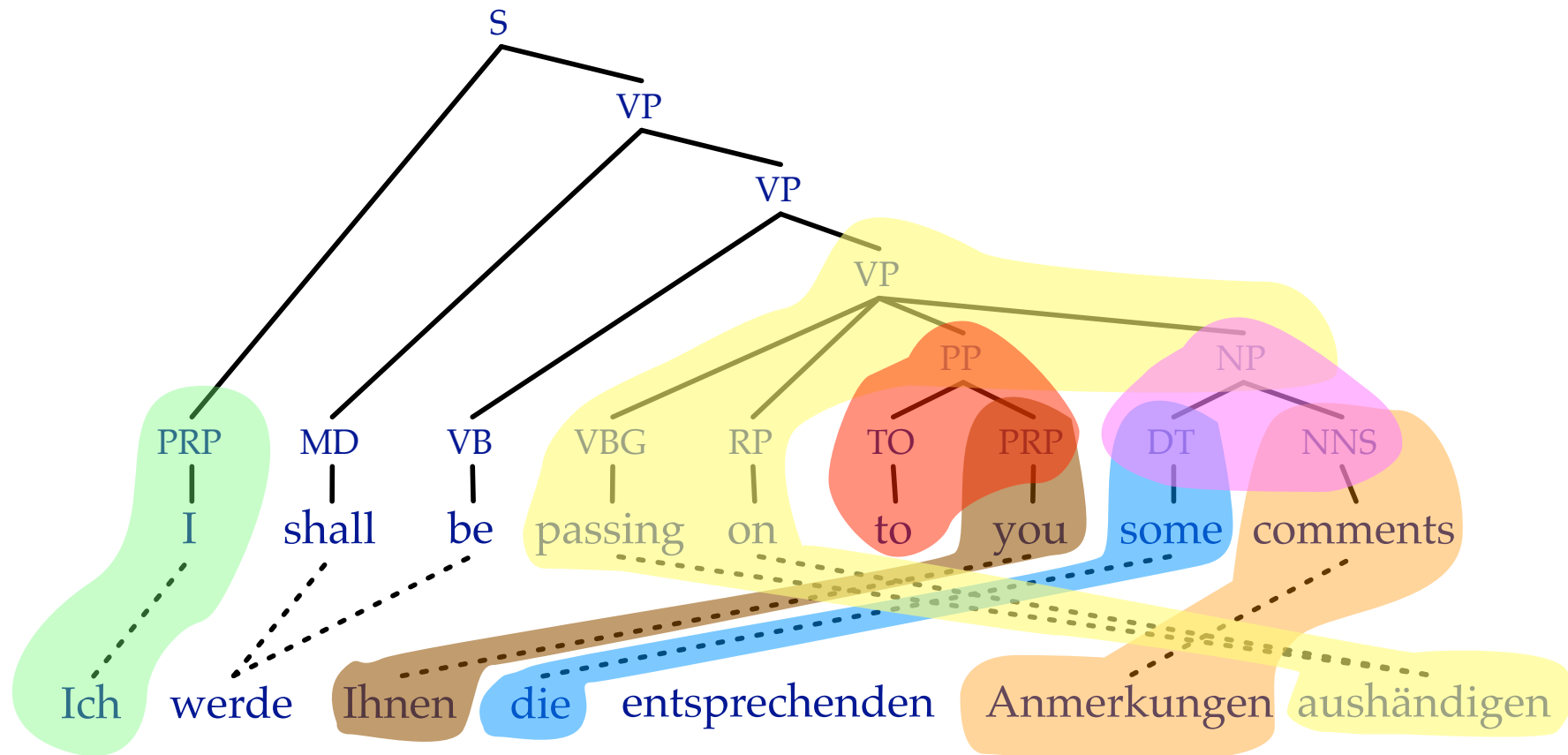
Extracted rule: $PP \rightarrow X \mid \text{to PRP}$

Non-Lexical Rule



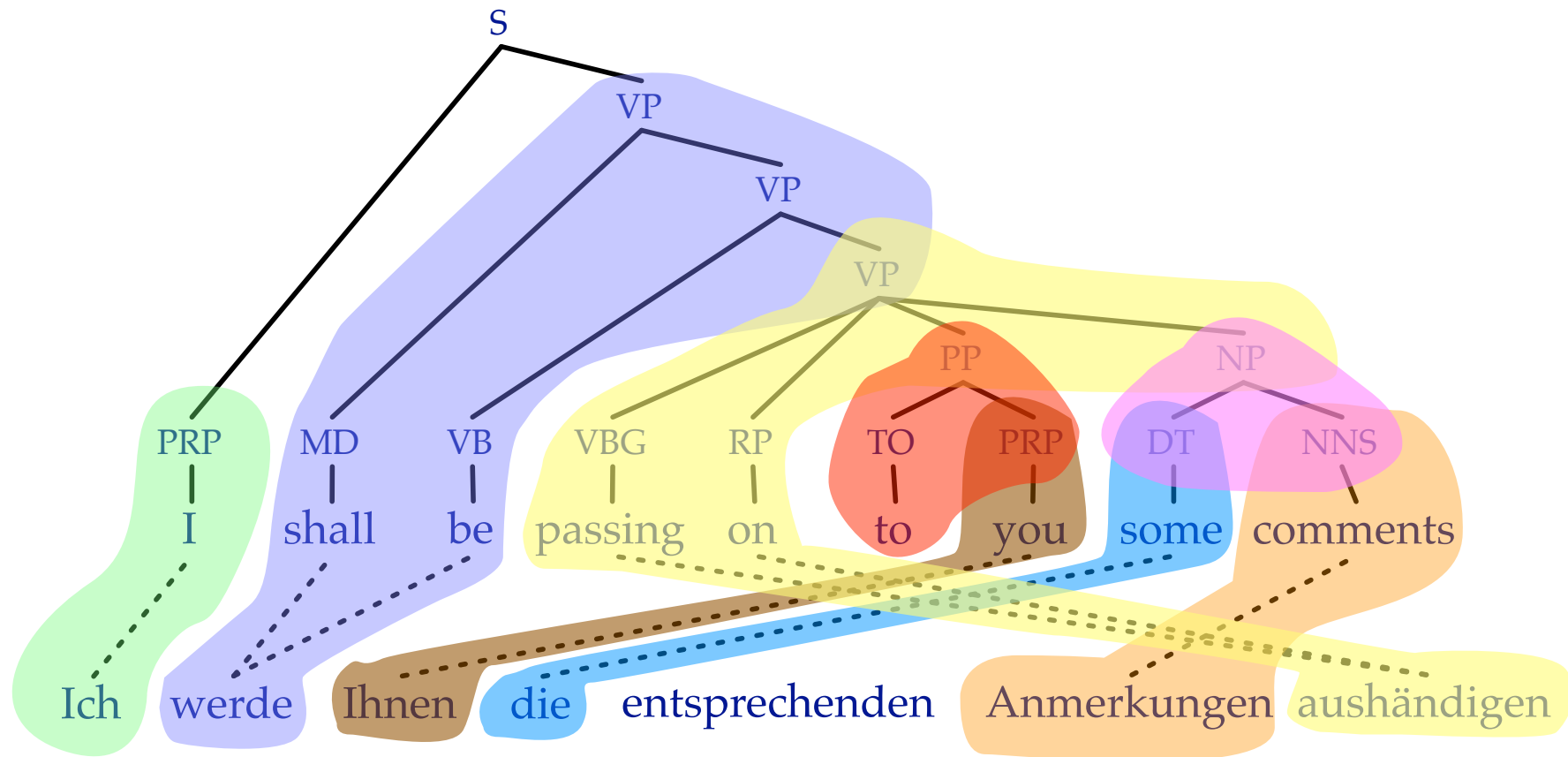
Extracted rule: $NP \rightarrow X_1 X_2 \mid DT_1 NNS_2$

Lexical Rule with Syntactic Context



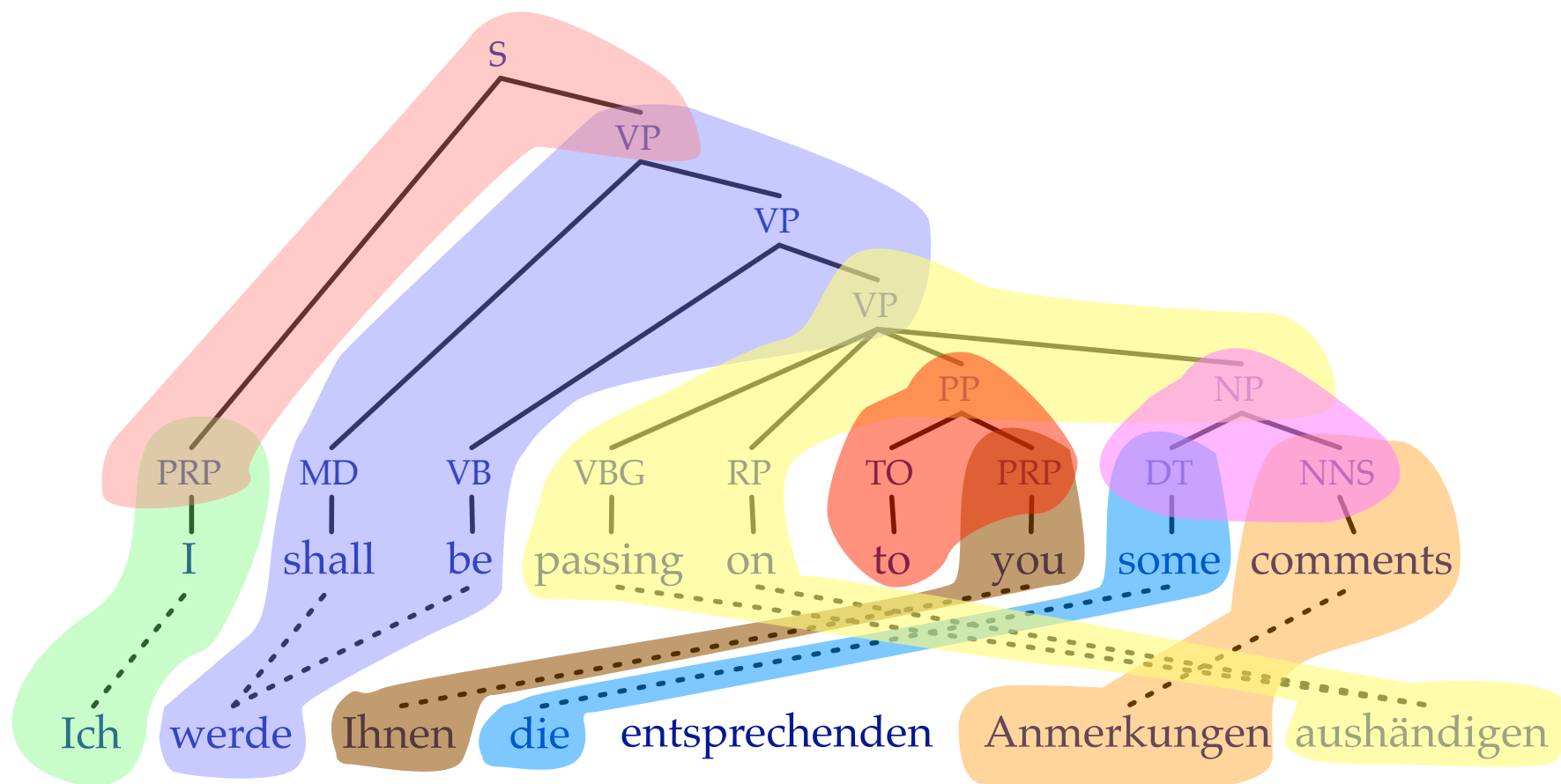
Extracted rule: $VP \rightarrow X_1 X_2 \text{ aushändigen} \mid \text{passing on } PP_1 NP_2$

Lexical Rule with Syntactic Context



Extracted rule: $VP \rightarrow \text{werde } X \mid \text{shall be } VP$ (ignoring internal structure)

Non-Lexical Rule



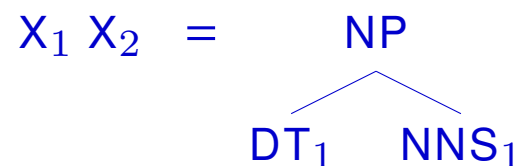
Extracted rule: $S \rightarrow X_1 X_2 \mid \text{PRP}_1 \text{VP}_2$
DONE — note: one rule per alignable constituent

Too Few Phrasal Rules?

- Lexical rules will be 1-to-1 mappings (unless word alignment requires otherwise)
- But: phrasal rules very beneficial in phrase-based models
- Solutions
 - combine rules that contain a maximum number of symbols (as in hierarchical models, recall: "Option 1")
 - compose minimal rules to cover a maximum number of non-leaf nodes

Composed Rules

- Current rules



die = DT
|
some

entsprechenden Anmerkungen = NNS
|
comments

- Composed rule

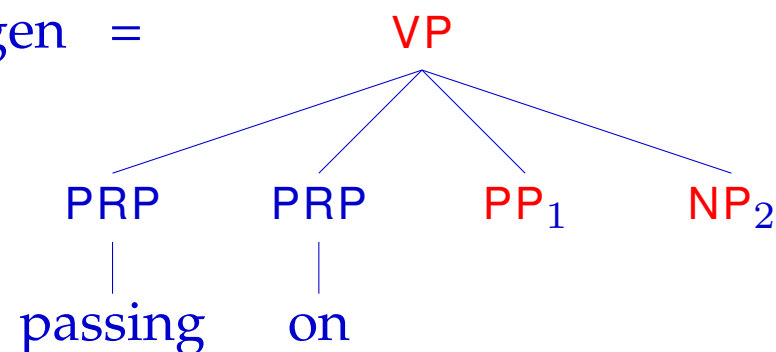


(1 non-leaf node: NP)

Composed Rules

- Minimal rule:

$X_1 X_2$ aushändigen =

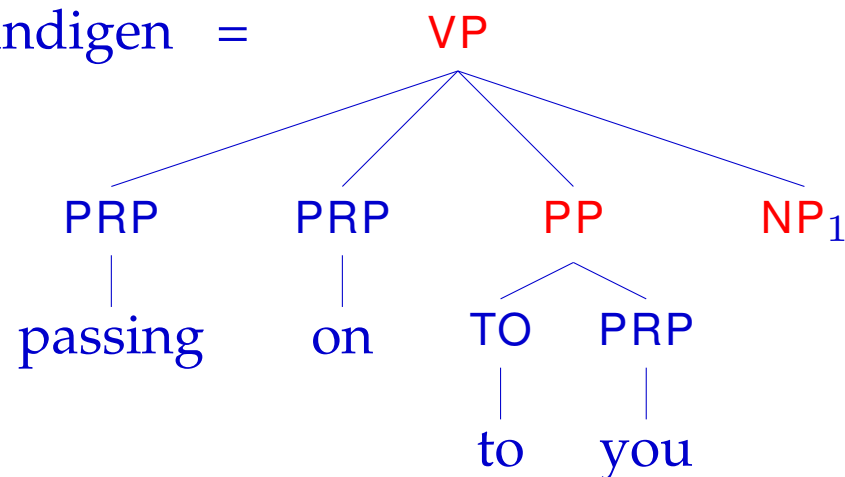


3 non-leaf nodes:

VP, PP, NP

- Composed rule:

Ihnen X₁ aushändigen =

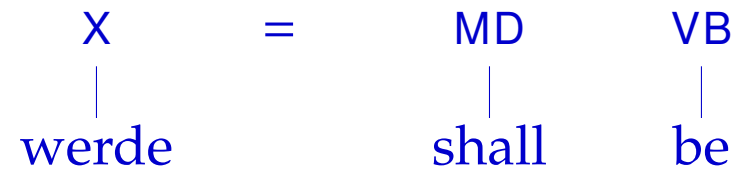


3 non-leaf nodes:

VP, PP and NP

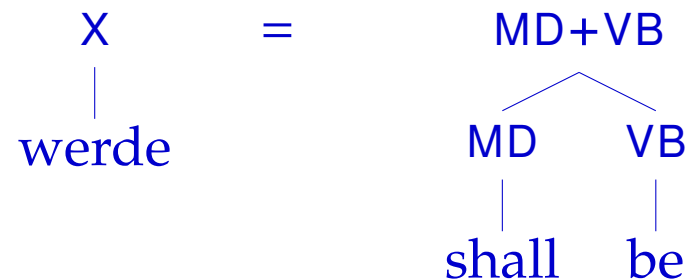
Relaxing Tree Constraints

- Impossible rule



- Create new non-terminal label: MD+VB

⇒ New rule



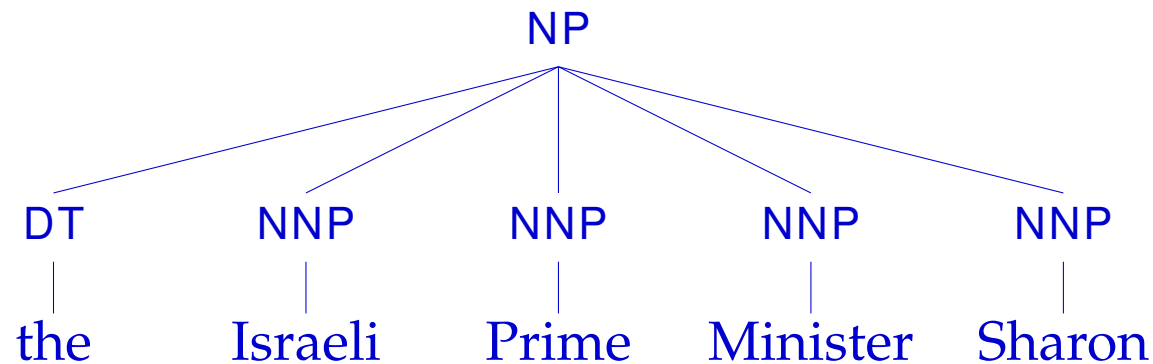
Zollmann Venugopal Relaxation

- If span consists of two constituents , join them: $X+Y$
- If span consists of three constituents, join them: $X+Y+Z$
- If span covers constituents with the same parent x and include
 - every but the first child Y , label as $X\backslash Y$
 - every but the last child Y , label as X/Y
- For all other cases, label as **FAIL**

⇒ More rules can be extracted, but number of non-terminals blows up

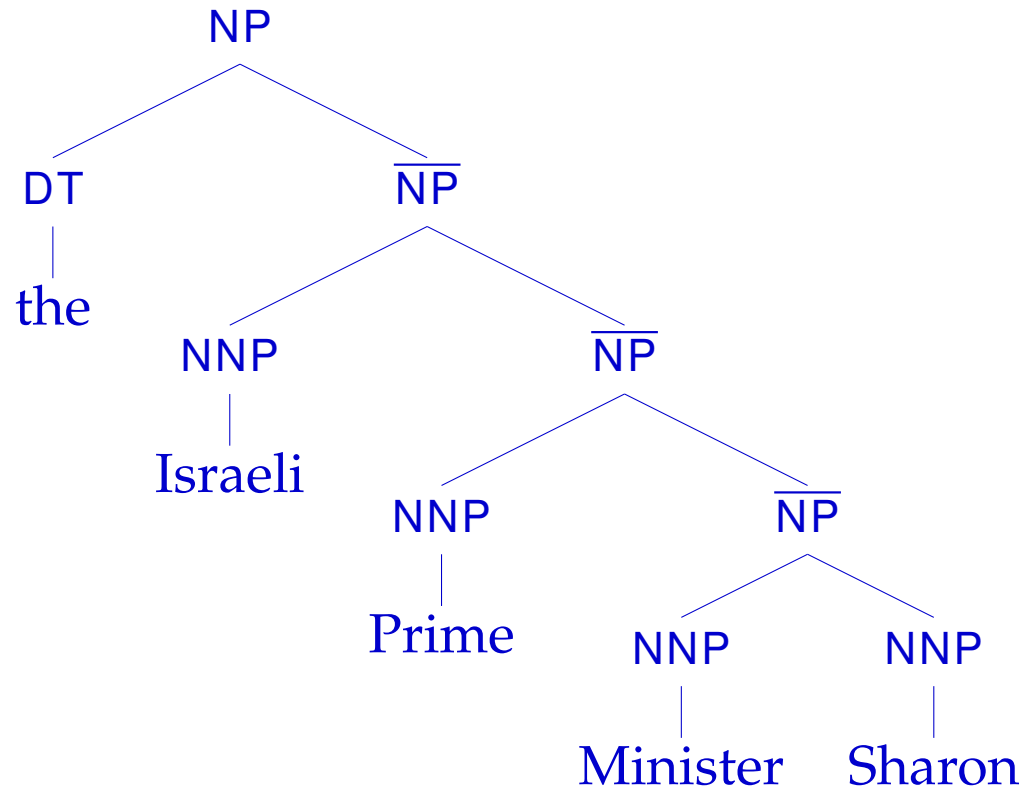
Special Problem: Flat Structures

- Flat structures severely limit rule extraction



- Can only extract rules for individual words or entire phrase

Relaxation by Tree Binarization



More rules can be extracted

Left-binarization or right-binarization?

Scoring Translation Rules

- Extract all rules from corpus
- Score based on counts
 - joint rule probability: $p(\text{LHS}, \text{RHS}_f, \text{RHS}_e)$
 - rule application probability: $p(\text{RHS}_f, \text{RHS}_e | \text{LHS})$
 - direct translation probability: $p(\text{RHS}_e | \text{RHS}_f, \text{LHS})$
 - noisy channel translation probability: $p(\text{RHS}_f | \text{RHS}_e, \text{LHS})$
 - lexical translation probability: $\prod_{e_i \in \text{RHS}_e} p(e_i | \text{RHS}_f, a)$