Syntax-Based Models

Philipp Koehn

7 November 2017

what is syntax?

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

Dependency Structure

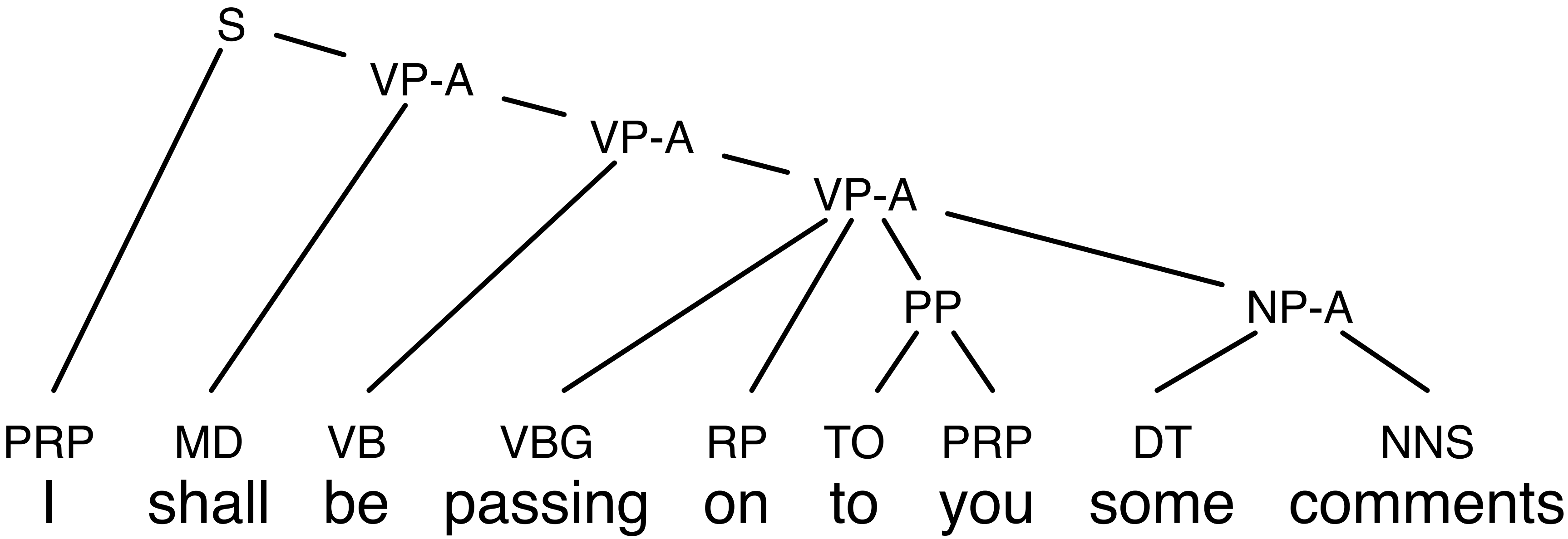
I	like	the	interesting	lecture
PRO	VB	DET	JJ	NN
↓		↓	↓	↓
like		lecture	lecture	like

- Center of a sentence is the verb
- Its dependents are its arguments (e.g., subject noun)
- These may have further dependents (adjective of noun)

Phrase Structure Grammar

- 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 → [NT,T]+
    example: NP → DET NN

Phrase Structure Grammar



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

syntactic transfer

Synchronous Phrase Structure Grammar

- English rule

$$\text{NP} \rightarrow \text{DET JJ NN}$$

- French rule

$$\text{NP} \rightarrow \text{DET NN JJ}$$

- Synchronous rule (indices indicate alignment):

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

Synchronous Grammar Rules

- Nonterminal rules

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

- Terminal rules

$$\mathsf{N} \rightarrow \text{maison} \mid \text{house}$$

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

- Mixed rules

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

Tree-Based Translation Model

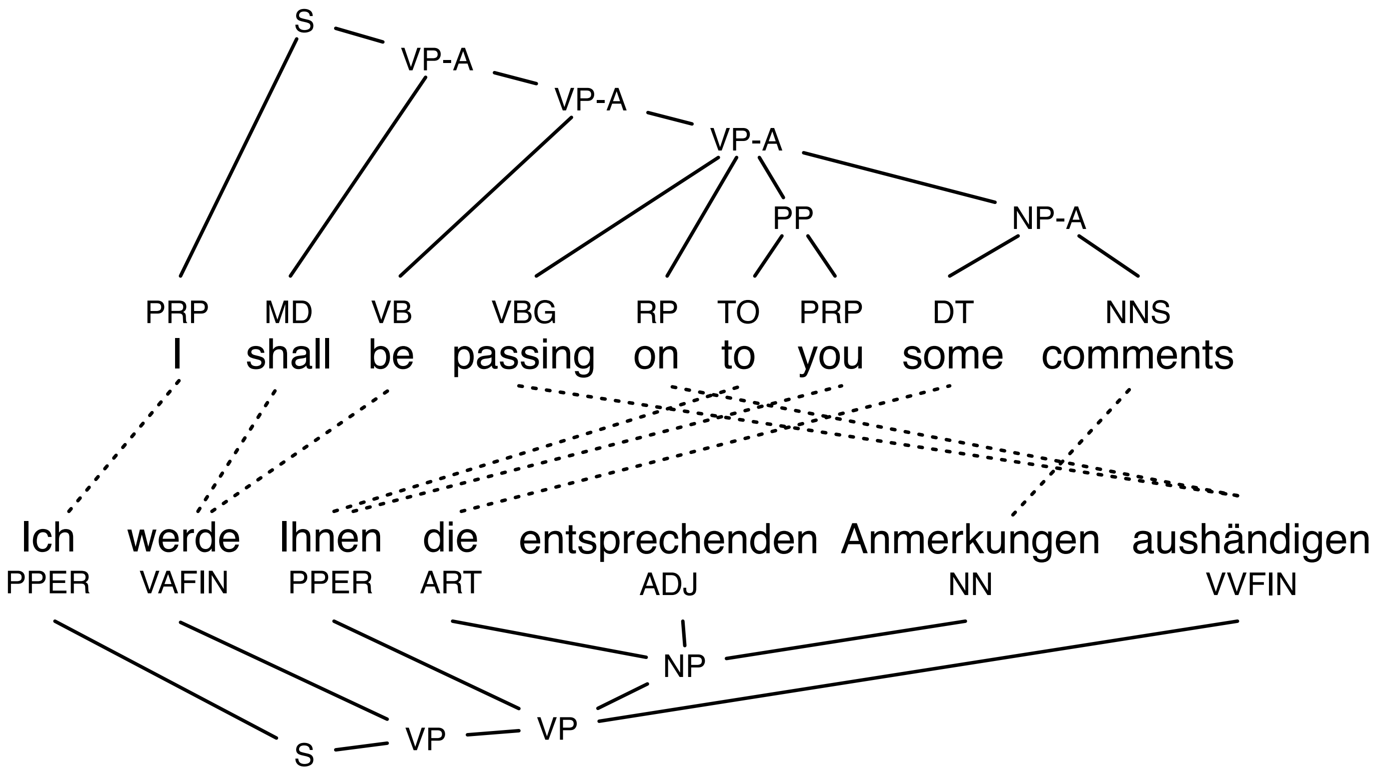
- Translation by parsing
  - synchronous grammar has to parse entire input sentence
  - output tree is generated at the same time
  - process is broken up into a number of rule applications

- Translation probability

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

- Many ways to assign probabilities to rules

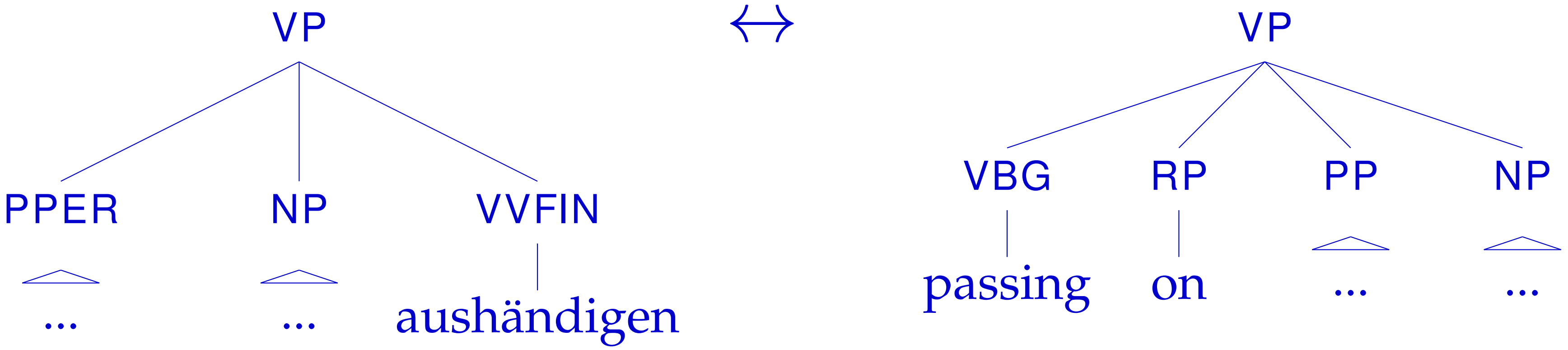
Aligned Tree Pair



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

Reordering Rule

- Subtree alignment



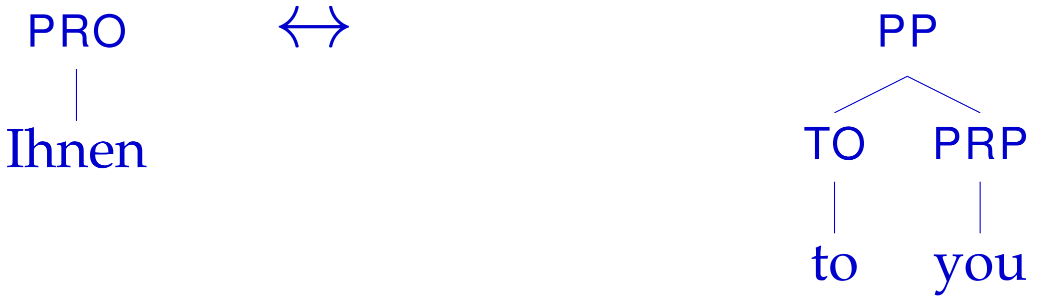
- Synchronous grammar rule

$$\text{VP} \rightarrow \text{PPER}_1 \; \text{NP}_2 \; \text{aushändigen} \mid \text{passing on} \; \text{PP}_1 \; \text{NP}_2$$

- 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

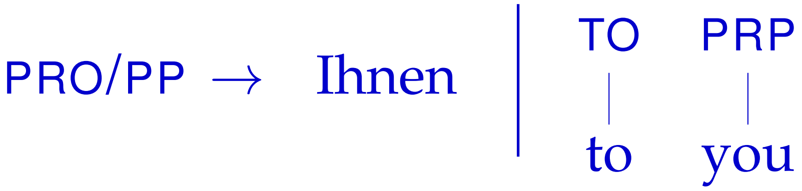
- Subtree alignment



- Synchronous grammar rule (stripping out English internal structure)

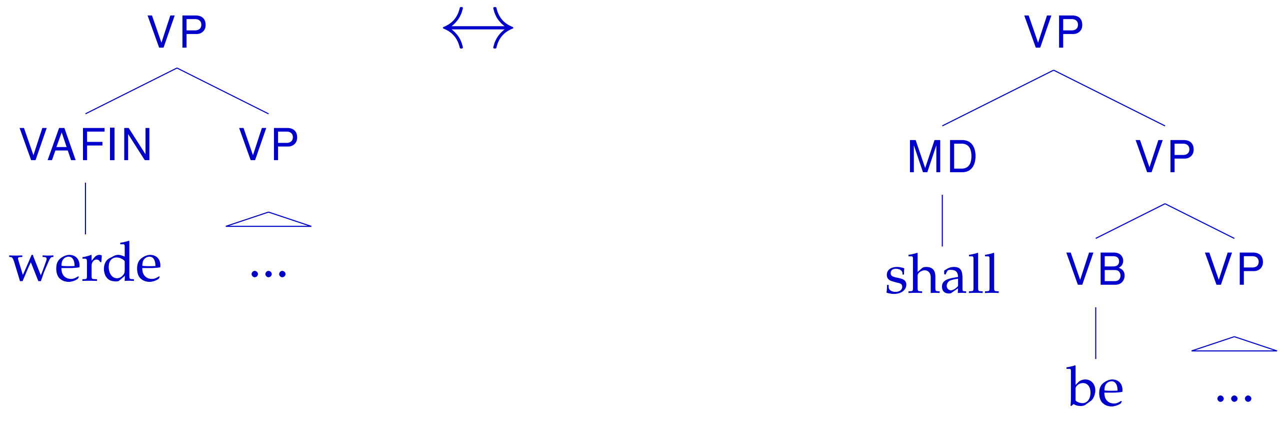
PRO/PP → Ihnen | to you

- Rule with internal structure



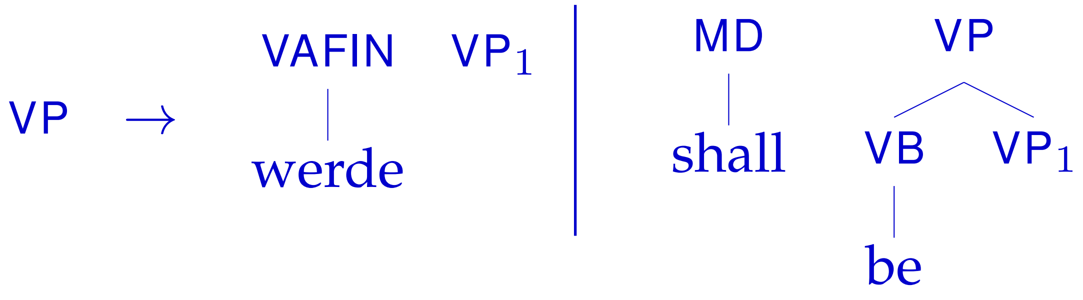
Another Rule

- Translation of German werde to English shall be



- Translation rule needs to include mapping of VP

⇒ Complex rule



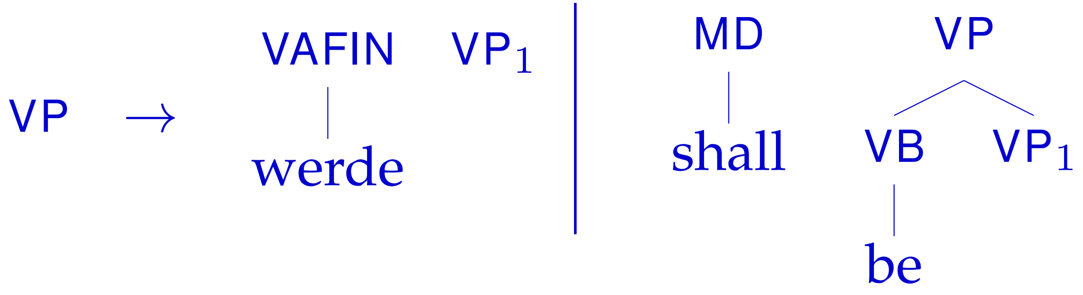
Internal Structure

- Stripping out internal structure

$$\mathsf{VP} \rightarrow \text{werde } \mathsf{VP}_1 \mid \text{shall be } \mathsf{VP}_1$$

⇒ synchronous context free grammar

- Maintaining internal structure



⇒ synchronous tree substitution grammar

learning

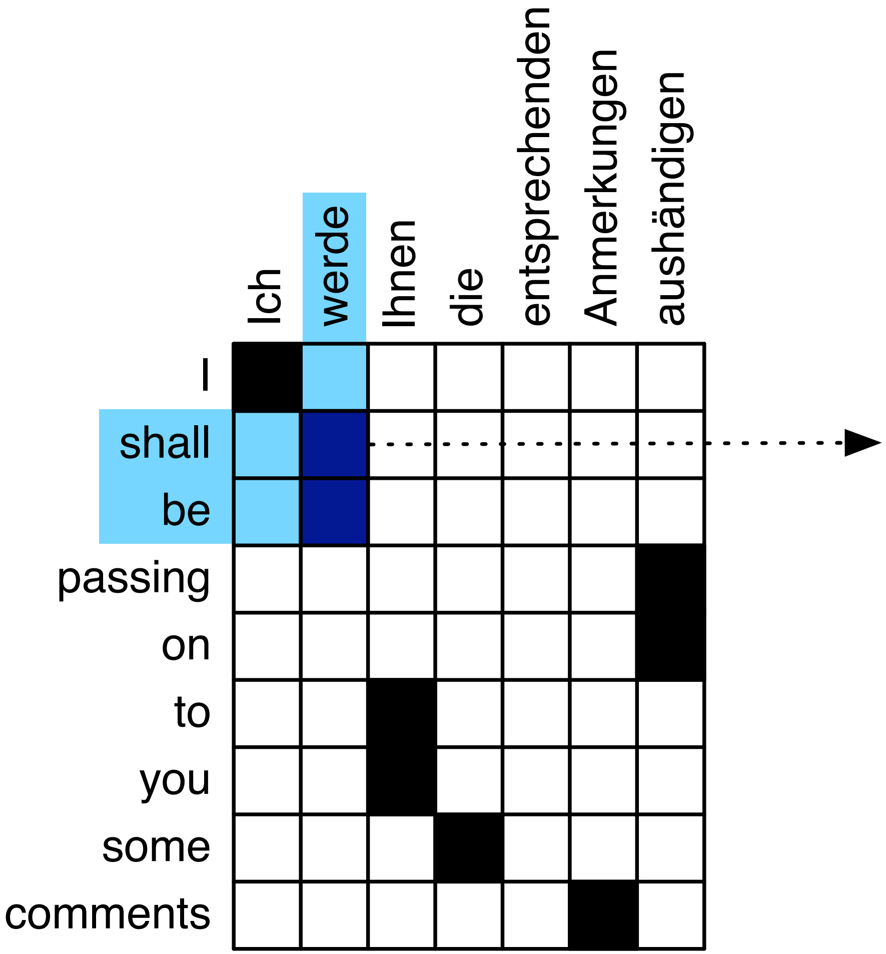
Learning Synchronous Grammars

- 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

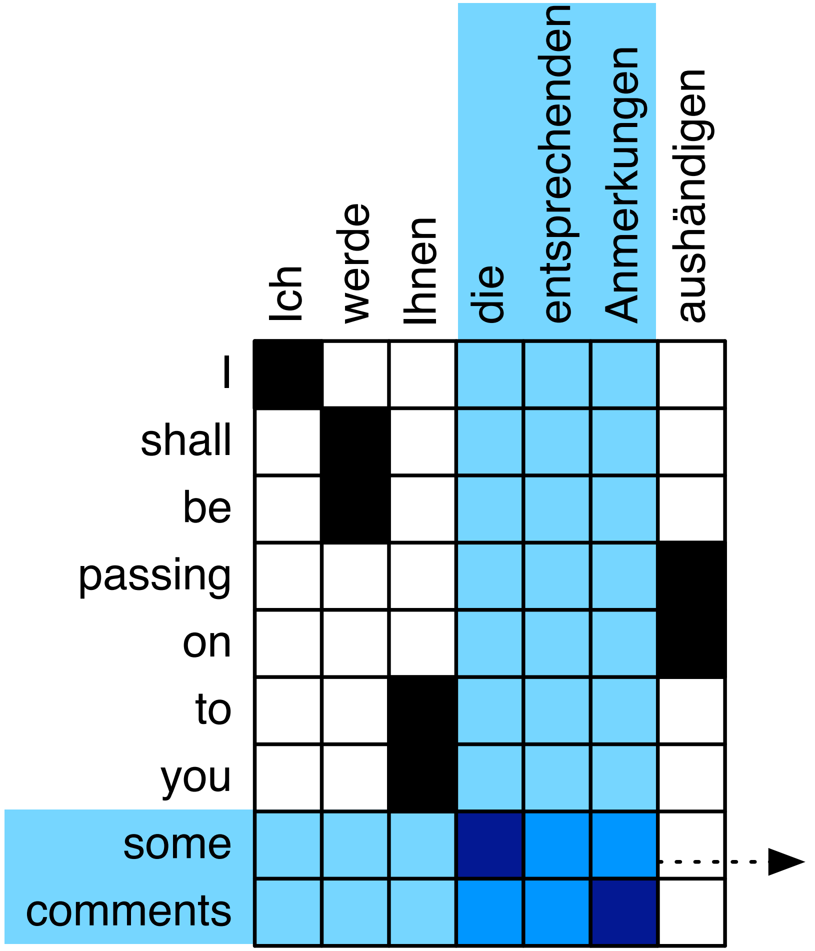
- Then: Synchronous phrase structure model
  - non-terminals for words and phrases: NP, VP, PP, ADJ, ...
  - corpus must also be parsed with syntactic parser

Extracting Phrase Translation Rules



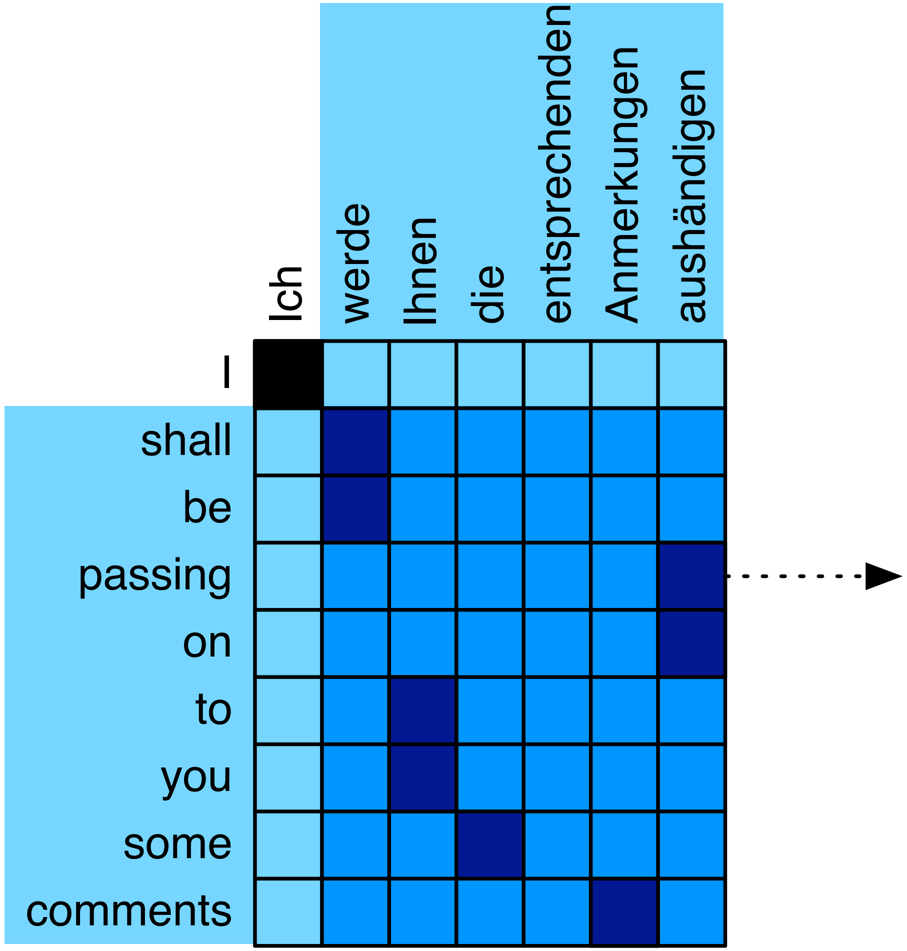
shall be = werde

Extracting Phrase Translation Rules



some comments =
die entsprechenden Anmerkungen

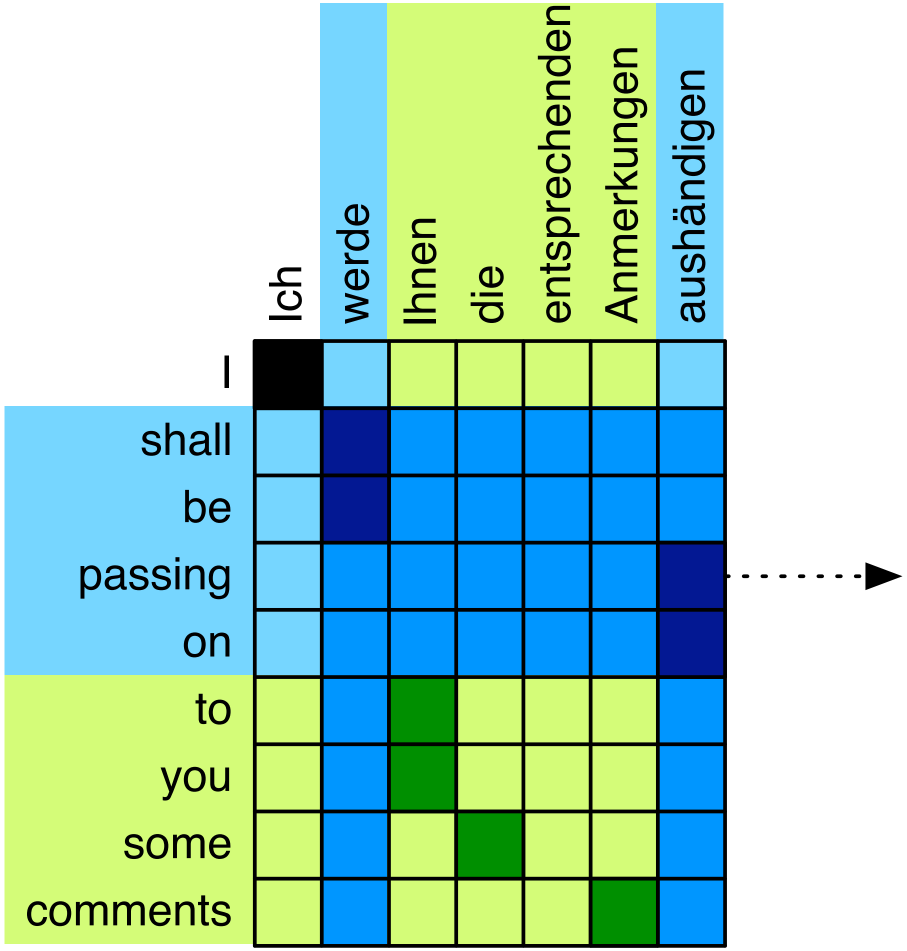
Extracting Phrase Translation Rules



	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



subtracting subphrase

werde X aushändigen
= shall be passing on X

Formal Definition

- Recall: consistent phrase pairs

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

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

Formal Definition

- Extend recursively:

$$
\begin{aligned}
&\text{if } (\bar{e}, \bar{f}) \in P \text{ AND } (\bar{e}_{\text{SUB}}, \bar{f}_{\text{SUB}}) \in P \\
&\quad \text{AND } \bar{e} = \bar{e}_{\text{PRE}} + \bar{e}_{\text{SUB}} + \bar{e}_{\text{POST}} \\
&\quad \text{AND } \bar{f} = \bar{f}_{\text{PRE}} + \bar{f}_{\text{SUB}} + \bar{f}_{\text{POST}} \\
&\quad \text{AND } \bar{e} \neq \bar{e}_{\text{SUB}} \text{ AND } \bar{f} \neq \bar{f}_{\text{SUB}} \\
&\text{add } (e_{\text{PRE}} + \text{X} + e_{\text{POST}}, f_{\text{PRE}} + \text{X} + f_{\text{POST}}) \text{ to } P
\end{aligned}
$$

(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

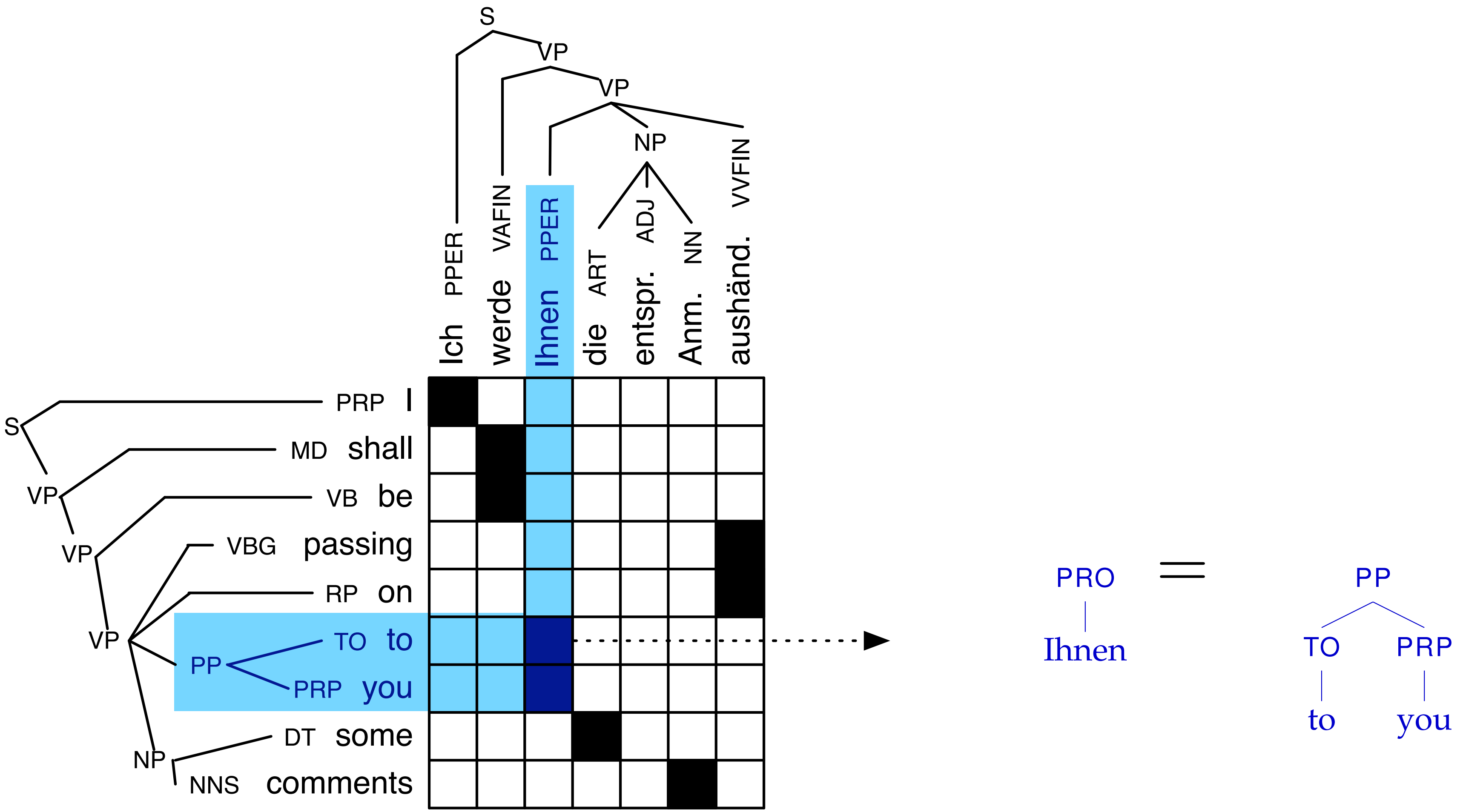
Comments

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

$$\mathsf{Y} \rightarrow \mathsf{X}_1\ \mathsf{X}_2 \ \mid\ \mathsf{X}_2\ \textit{of}\ \mathsf{X}_1$$

- 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

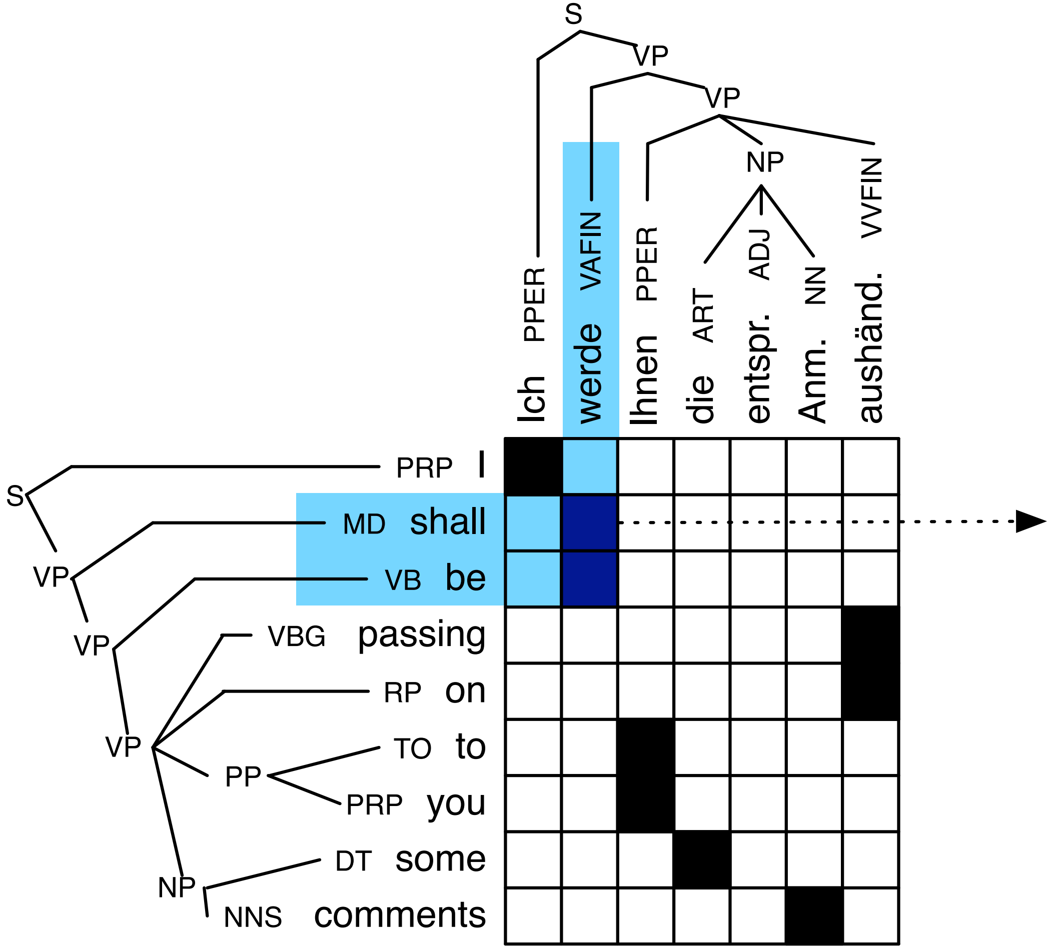
- Same word alignment constraints as hierarchical models

- Hierarchical: rule can cover any span
  ⇔ syntactic rules must cover constituents in the tree

- Hierarchical: gaps may cover any span
  ⇔ gaps must cover constituents in the tree

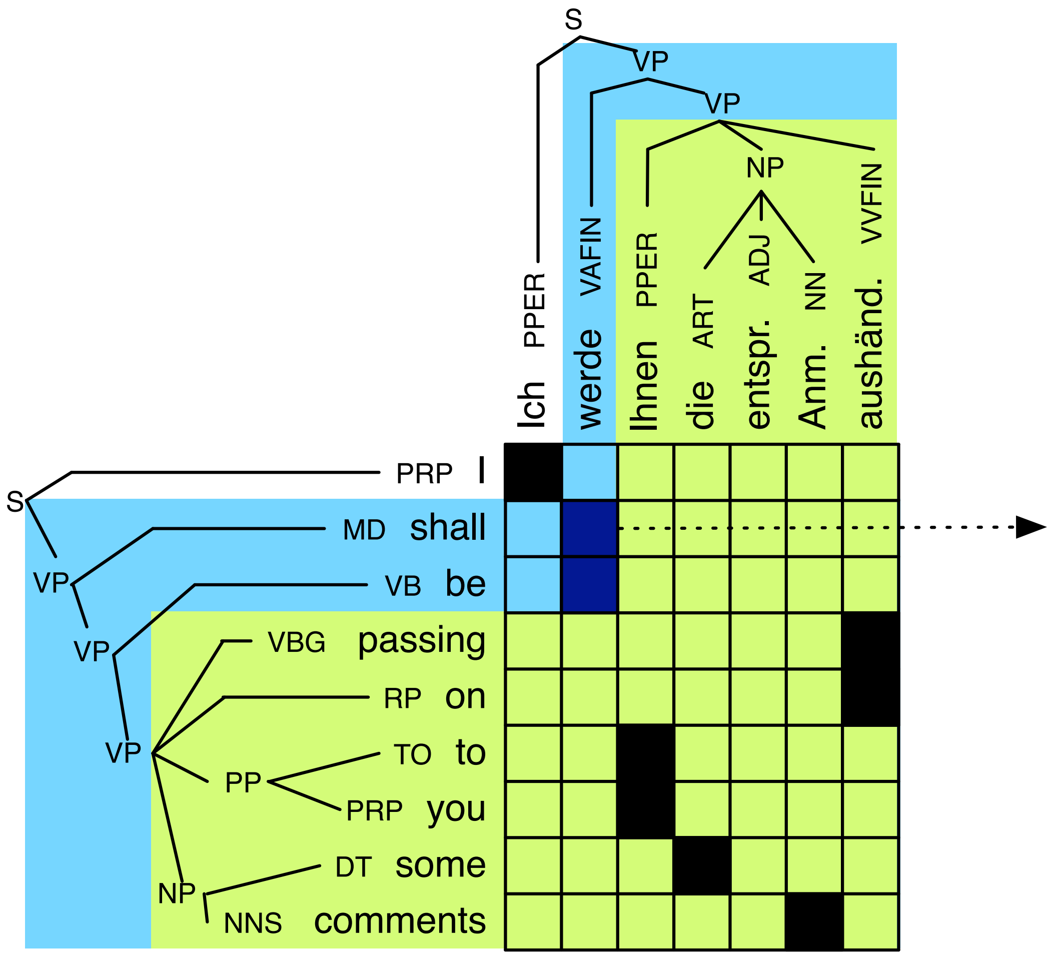
- Much less rules are extracted (all things being equal)

Impossible Rules

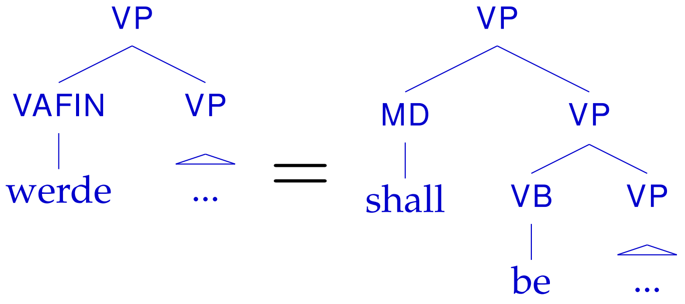


English span not a constituent
no rule extracted

Rules with Context



Rule with this phrase pair requires syntactic context



Too Many Rules Extractable

- Huge number of rules can be extracted
  (every alignable node may or may not be part of a rule → exponential number of rules)

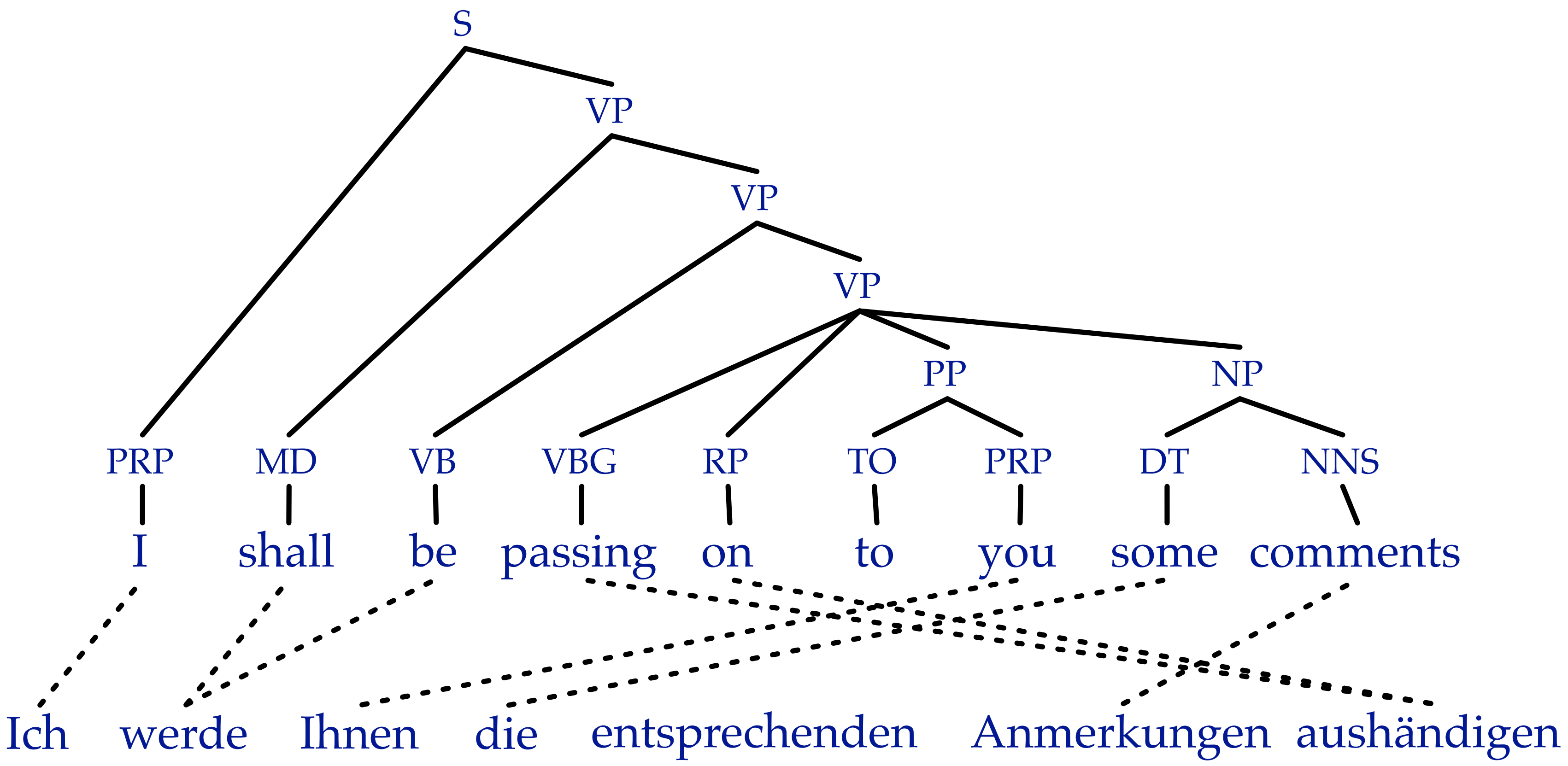
- Need to limit which rules to extract

- 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)

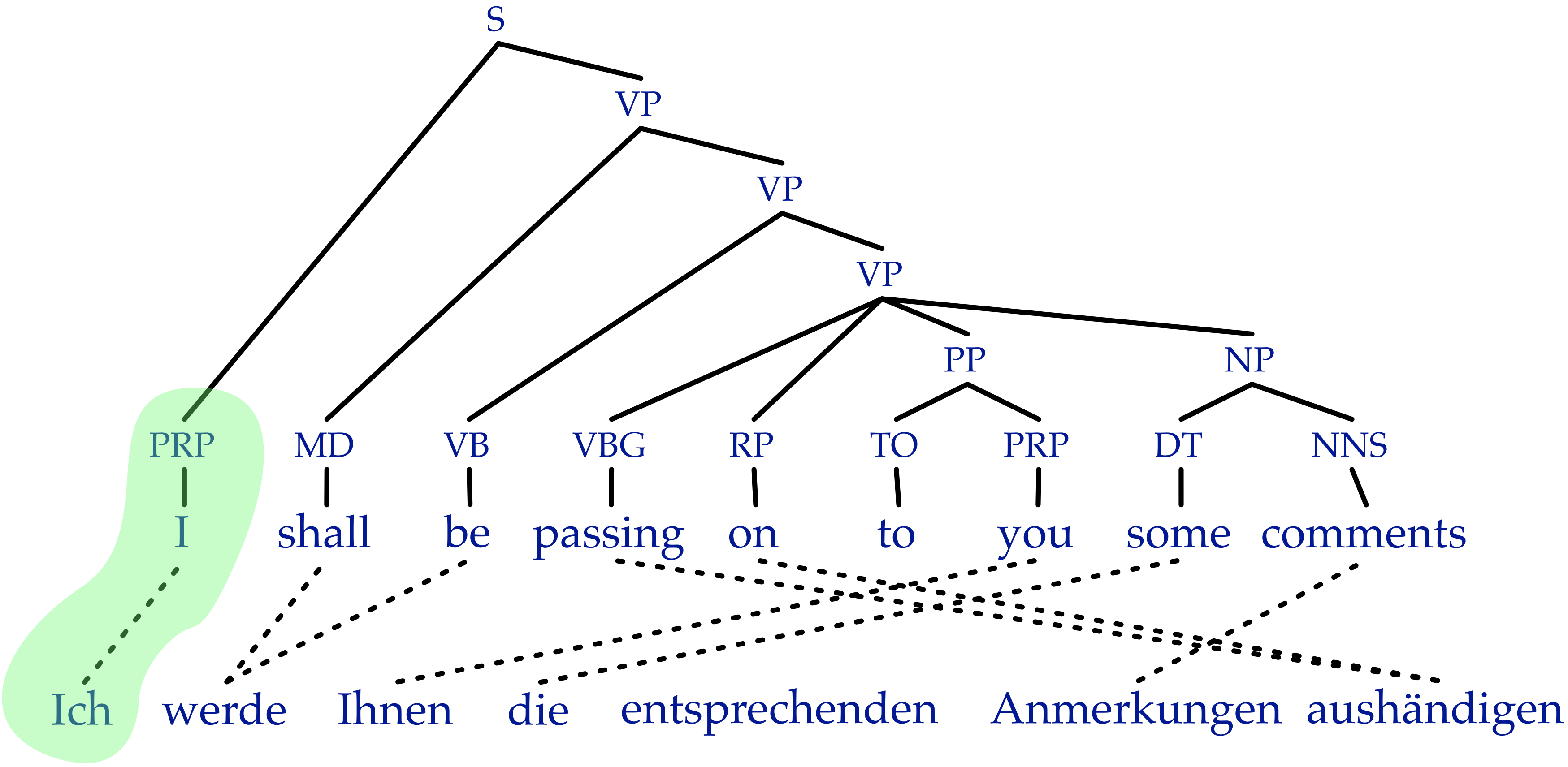
refinements

Minimal Rules



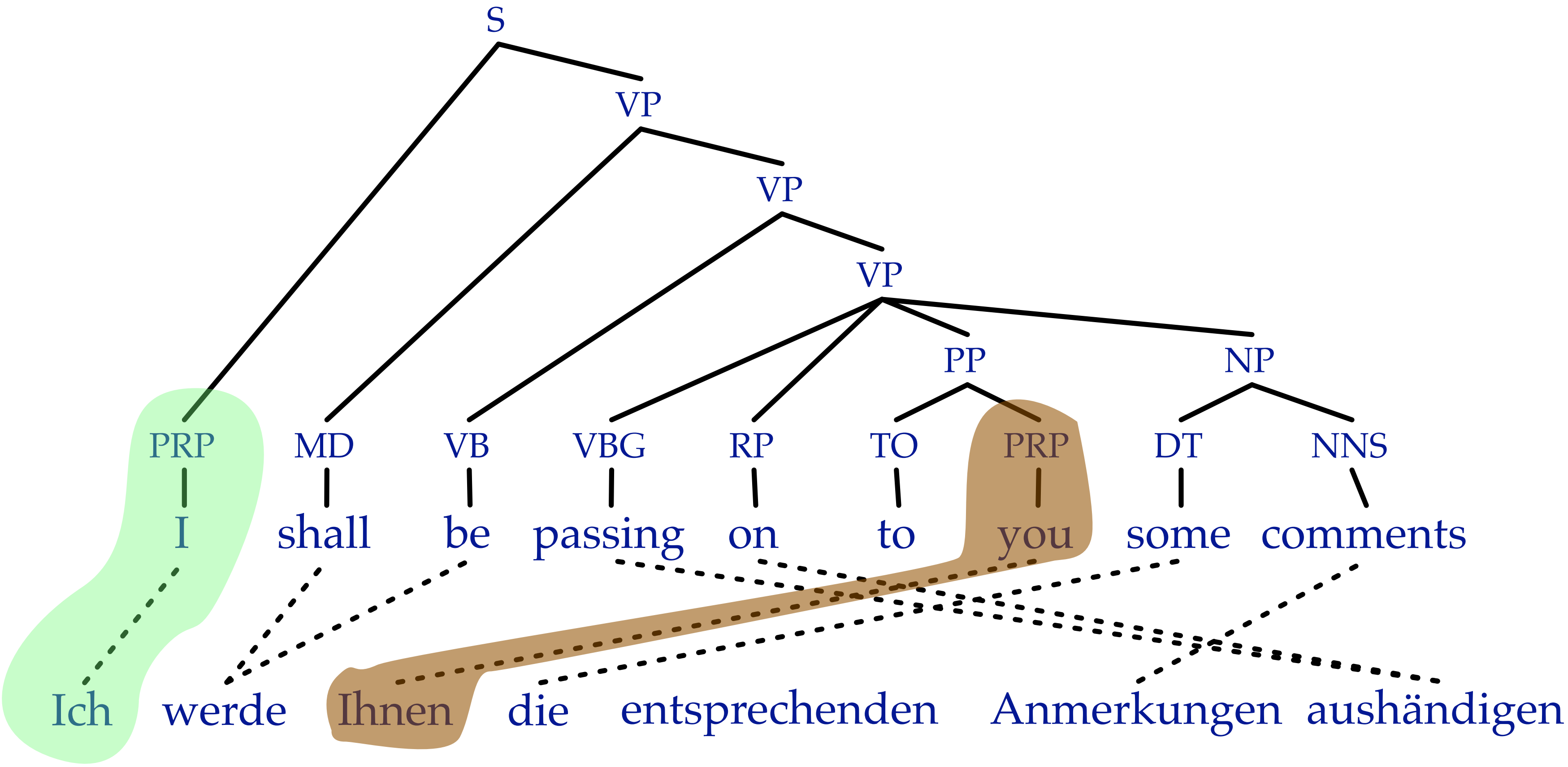
Extract: set of smallest rules required to explain the sentence pair

Lexical Rule



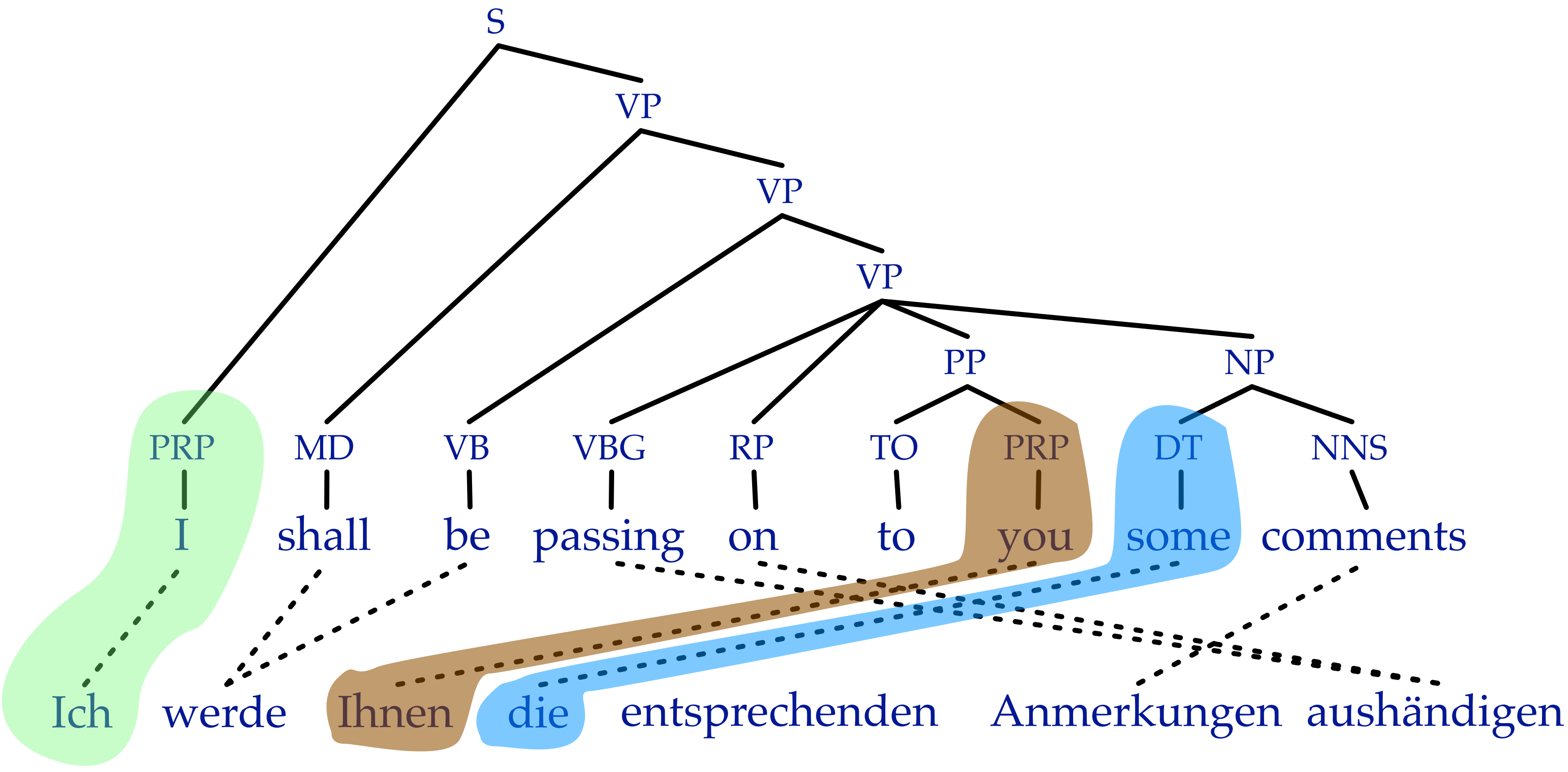
Extracted rule: PRP → Ich | I

Lexical Rule



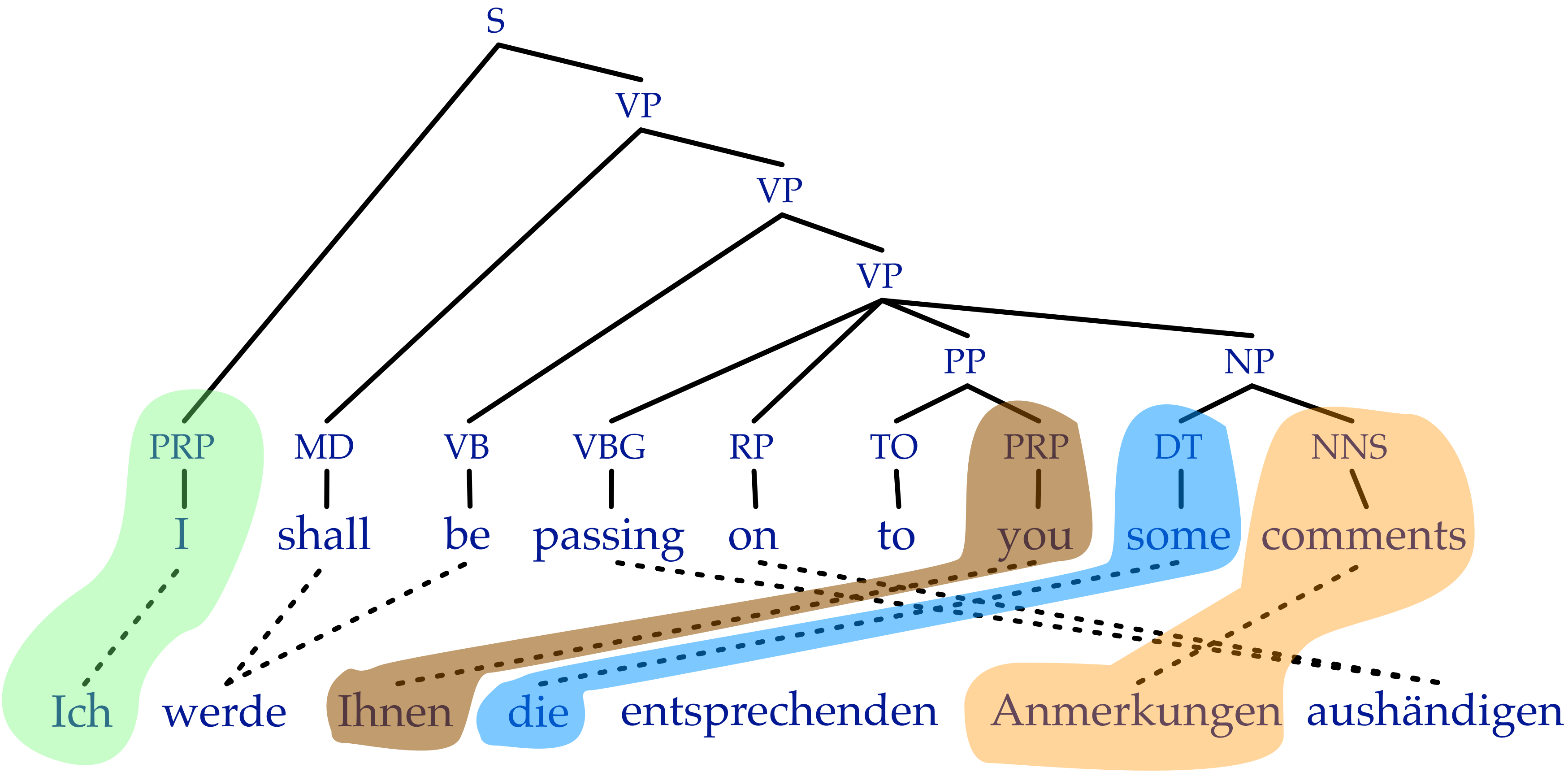
Extracted rule: PRP → Ihnen | you

Lexical Rule



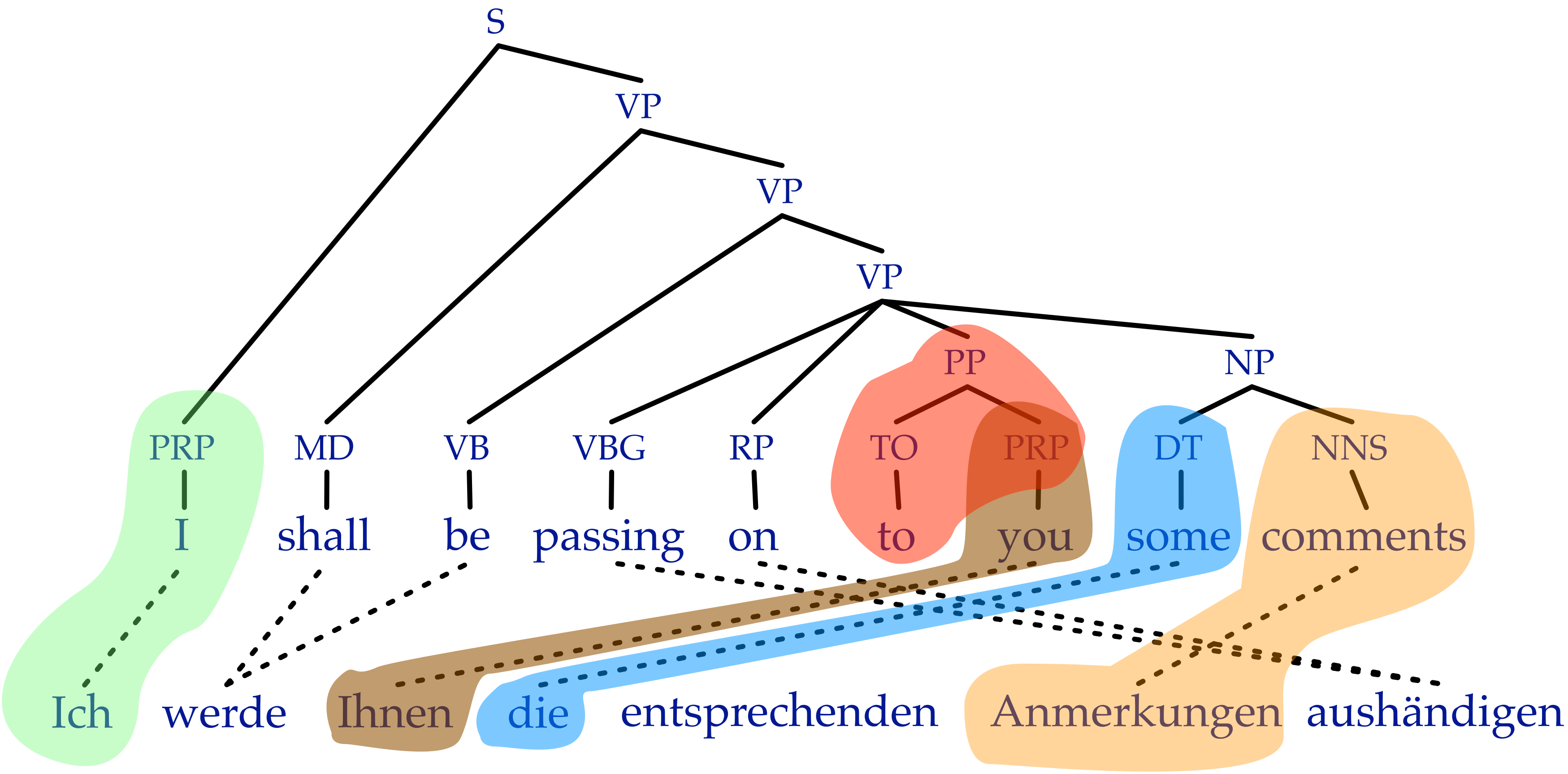
Extracted rule: DT → die | some

Lexical Rule



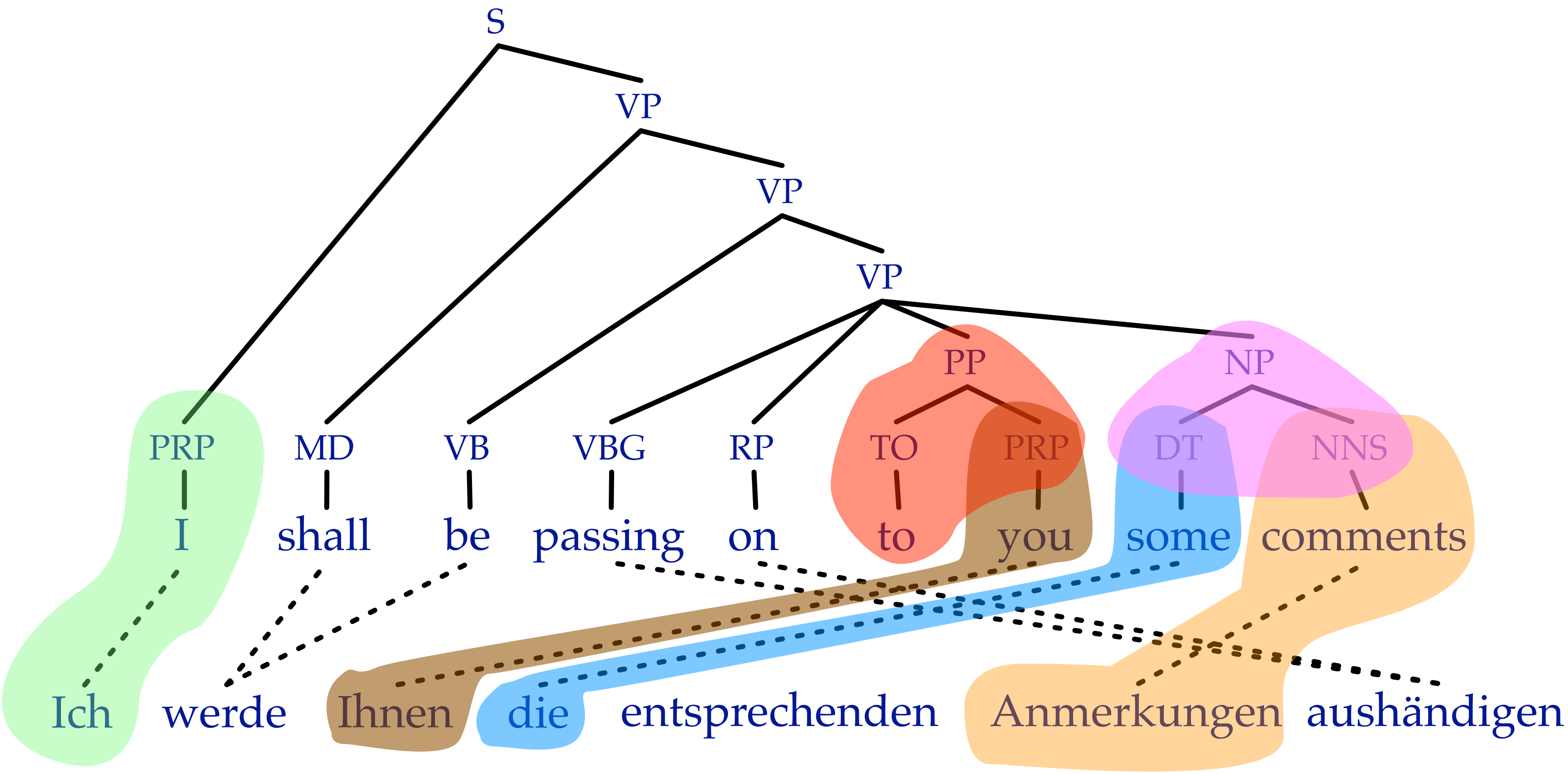
Extracted rule: NNS → Anmerkungen | comments

Insertion Rule



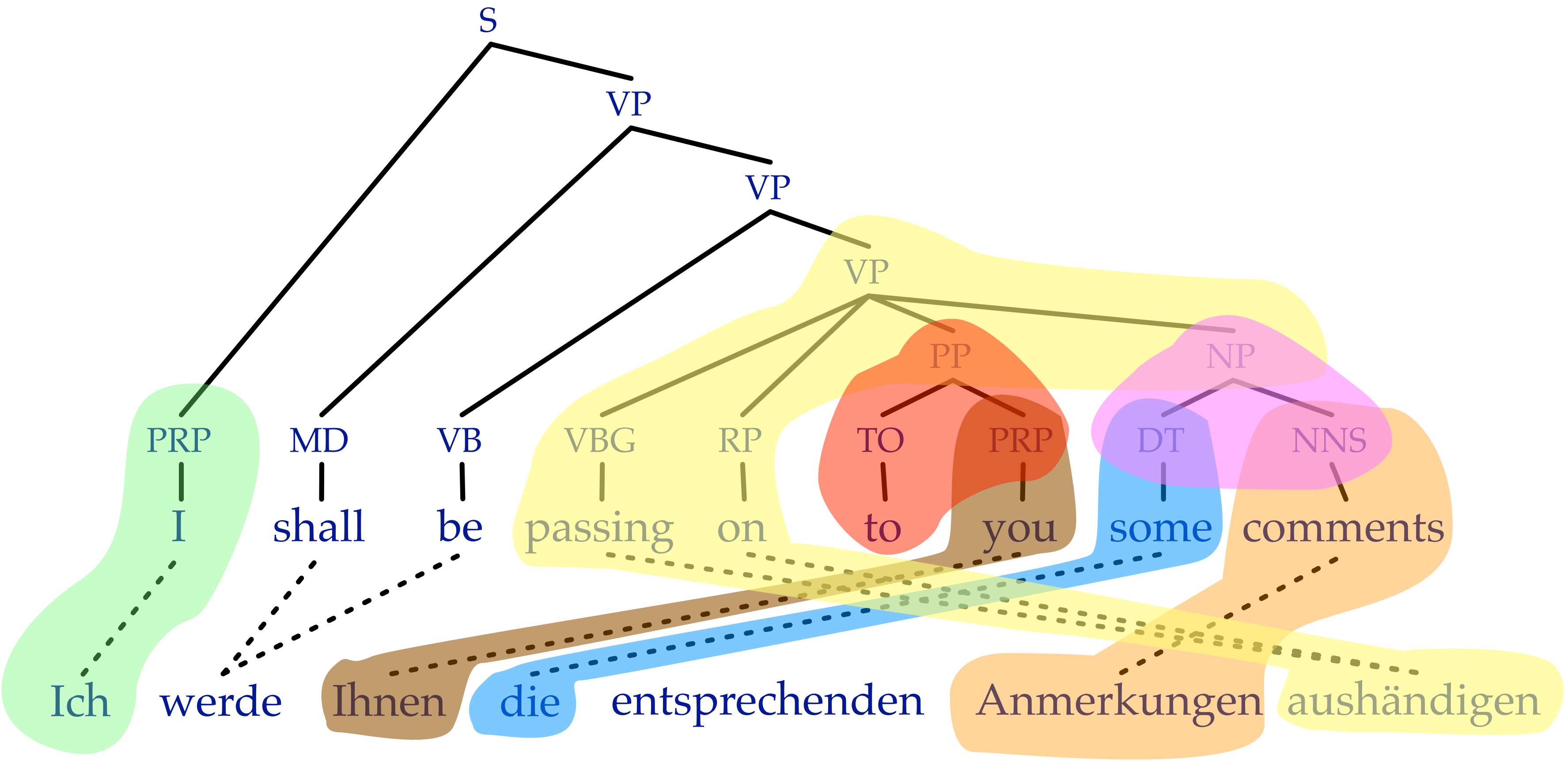
Extracted rule: PP → X | to PRP

Non-Lexical Rule



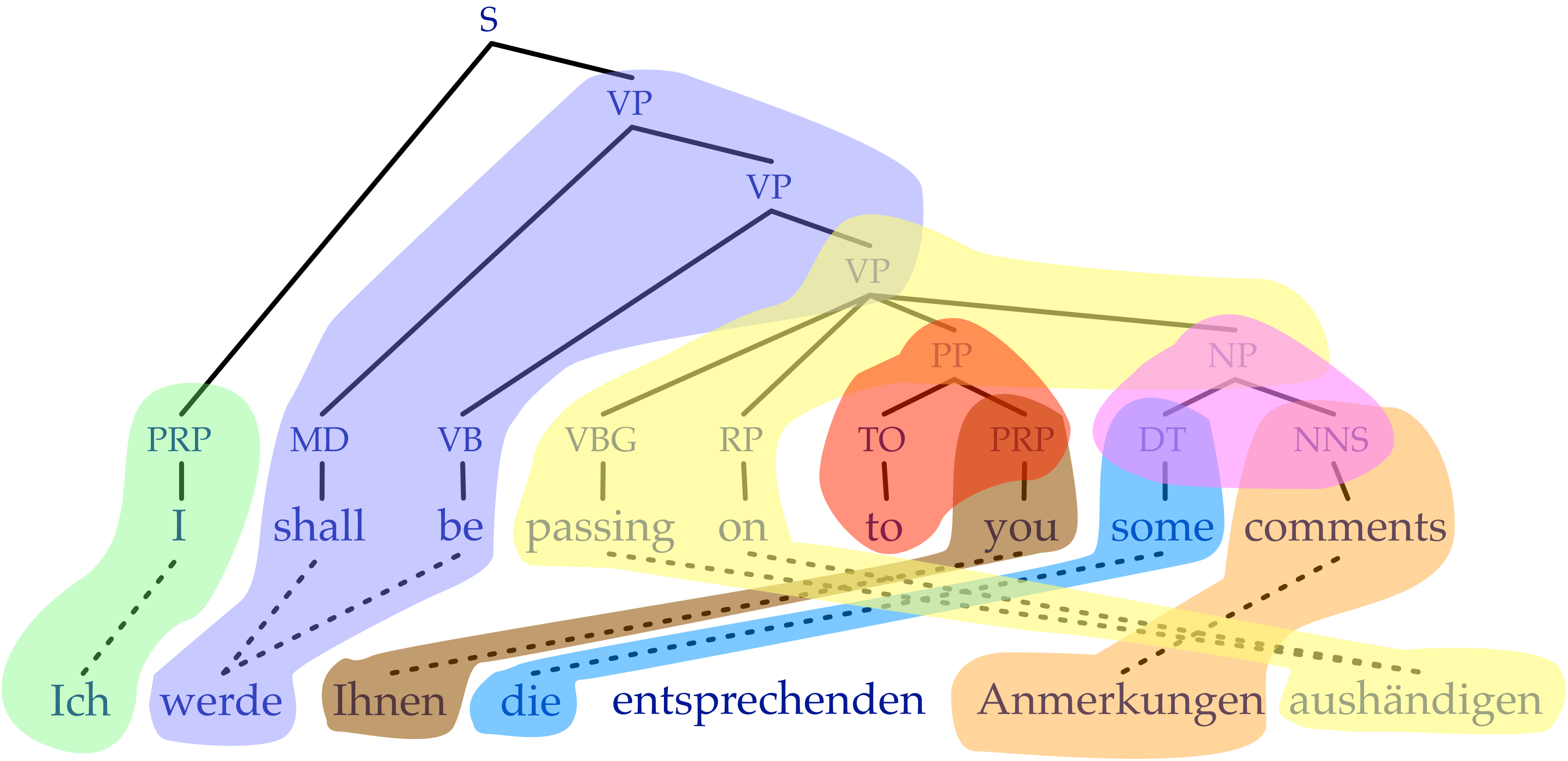
Extracted rule: $\text{NP} \rightarrow \text{X}_1\ \text{X}_2 \mid \text{DT}_1\ \text{NNS}_2$

Lexical Rule with Syntactic Context



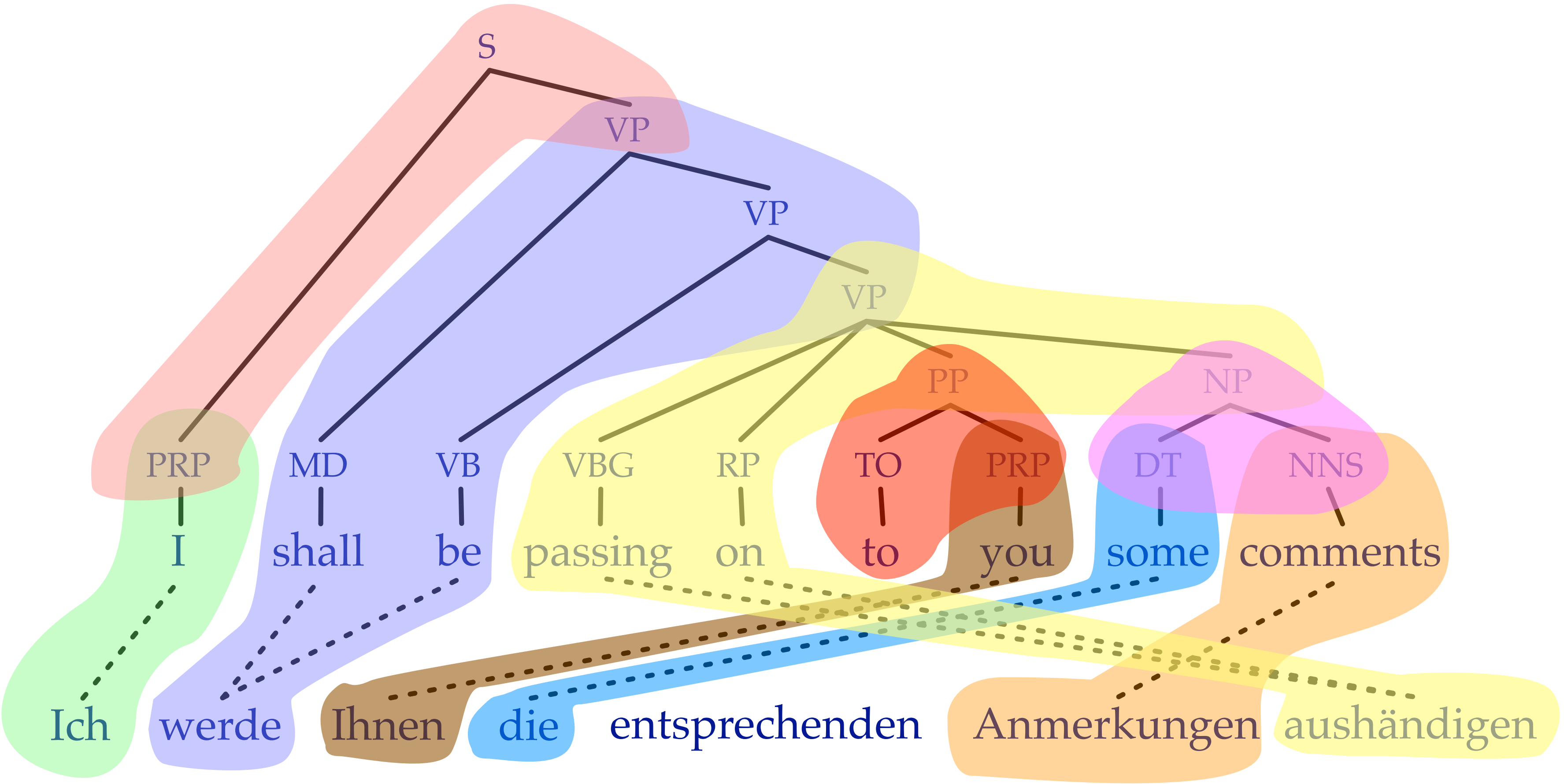
Extracted rule: $\texttt{VP} \rightarrow \texttt{X}_1\ \texttt{X}_2$ aushändigen | passing on $\texttt{PP}_1\ \texttt{NP}_2$

Lexical Rule with Syntactic Context



Extracted rule: VP → werde X | shall be VP (ignoring internal structure)

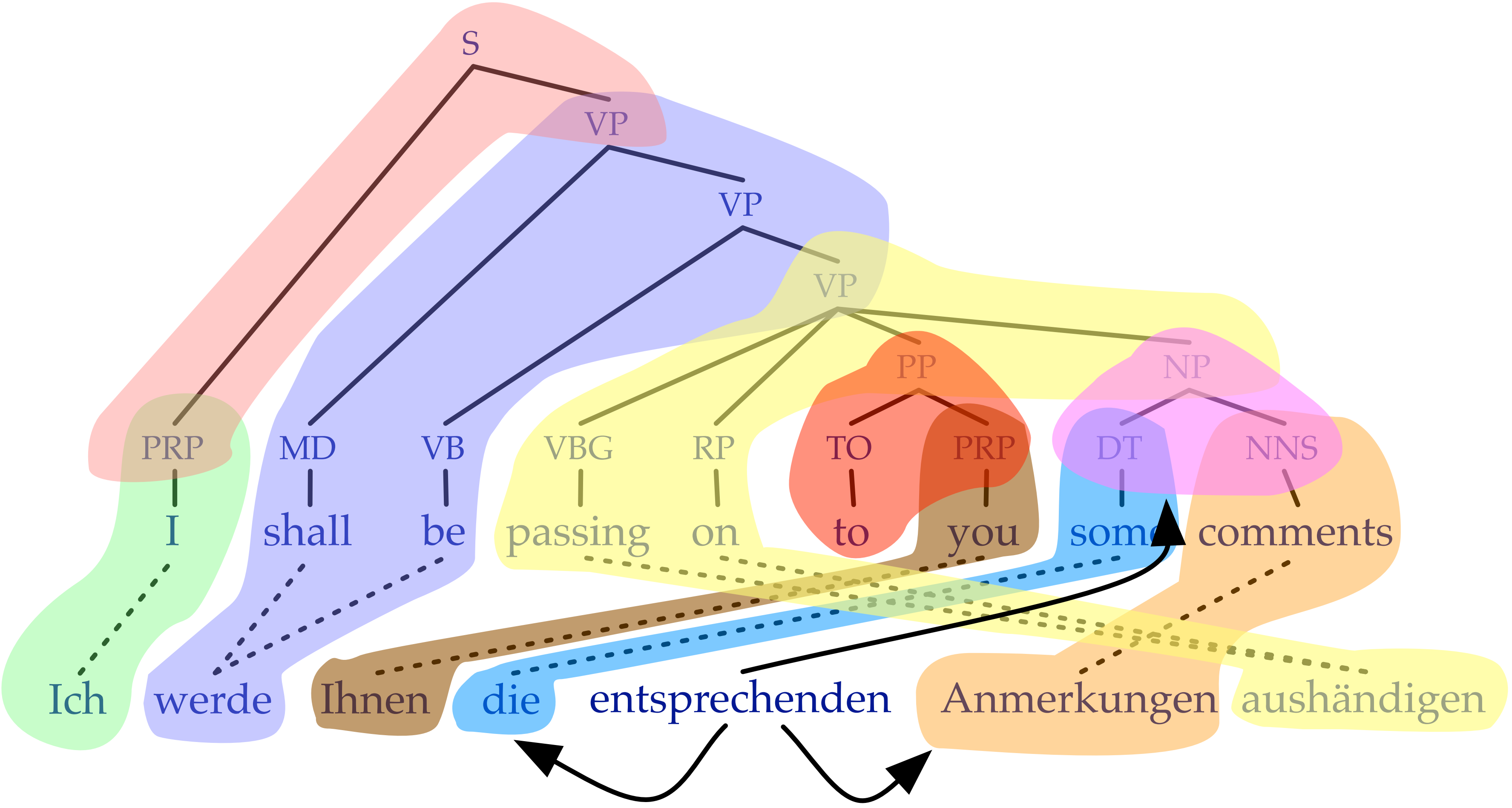
Non-Lexical Rule



Extracted rule: S → X_1 X_2 | PRP_1 VP_2

DONE — note: one rule per alignable constituent

Unaligned Source Words



Attach to neighboring words or higher nodes → additional rules

Too Few Phrasal Rules?

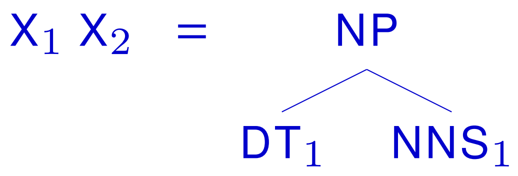
- Lexical rules will be 1-to-1 mappings (unless word alignment requires otherwise)

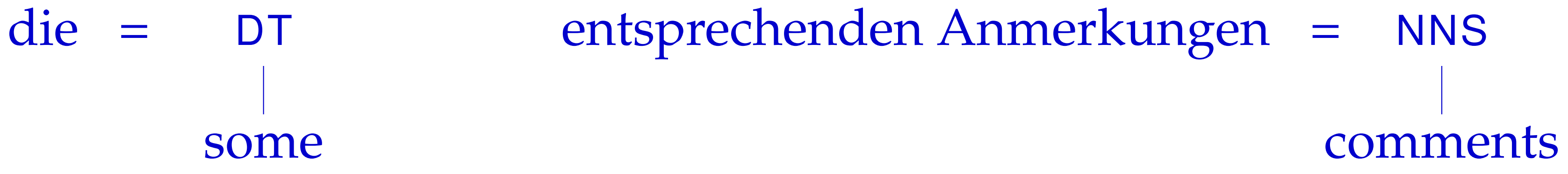
- But: phrasal rules very beneficial in phrase-based models

- 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

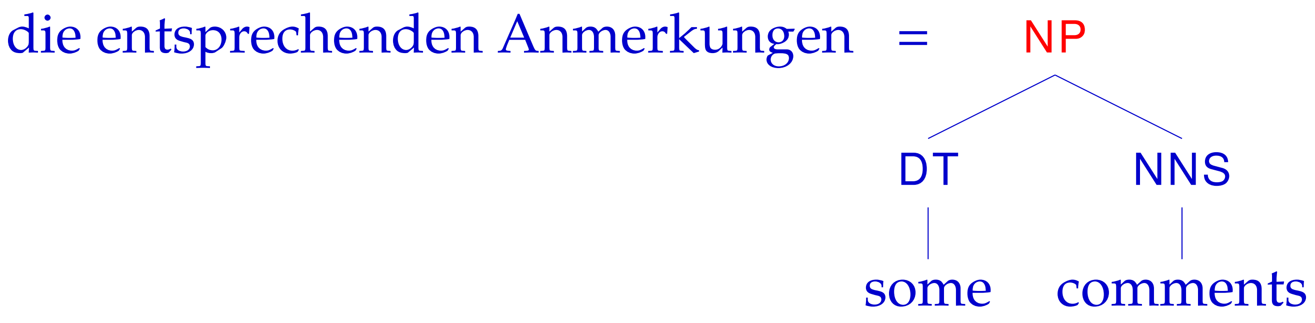
Composed Rules

- Current rules





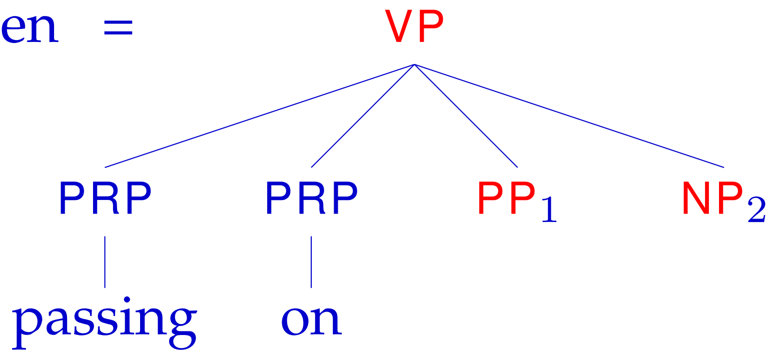
- Composed rule



(1 non-leaf node: NP)

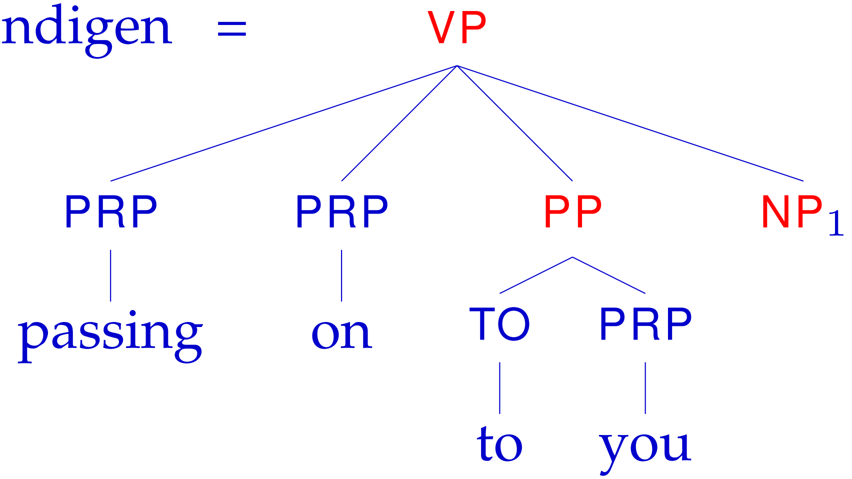
Composed Rules

- Minimal rule: X_1 X_2 aushändigen =



3 non-leaf nodes:
VP, PP, NP

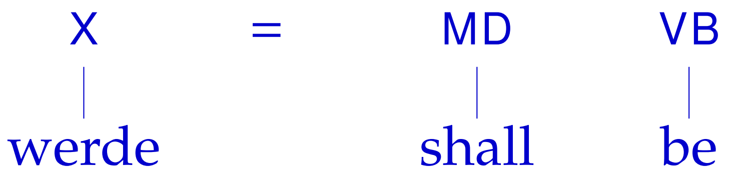
- Composed rule: Ihnen X_1 aushändigen =



3 non-leaf nodes:
VP, PP and NP

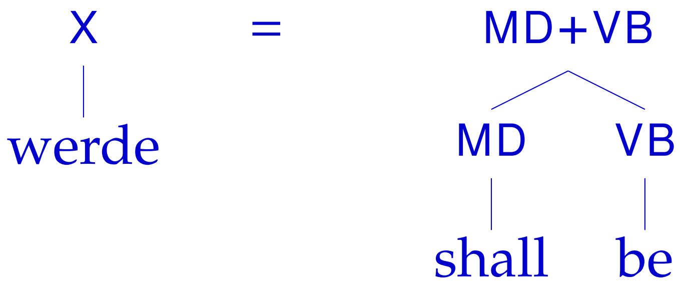
Relaxing Tree Constraints

- Impossible rule



- Create new non-terminal label: MD+VB

⇒ New rule



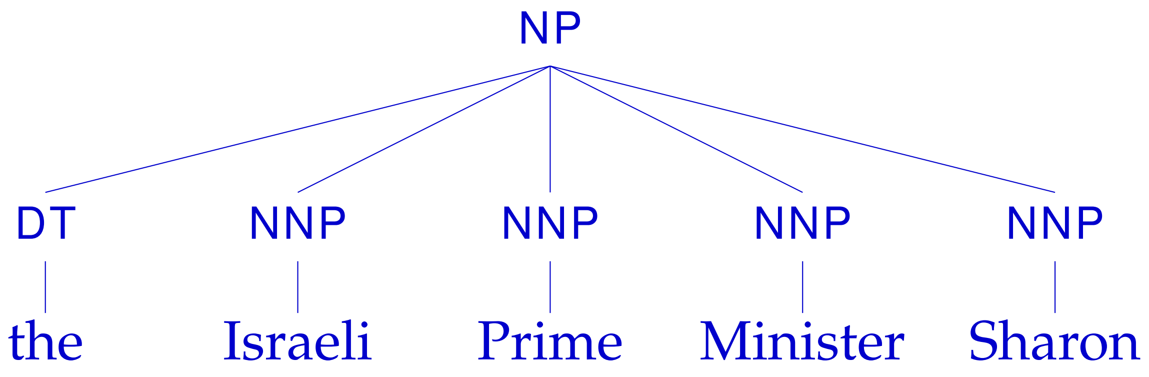
Zollmann Venugopal Relaxation

- If span consists of two constituents , join them: X+Y
- If span conststs 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\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

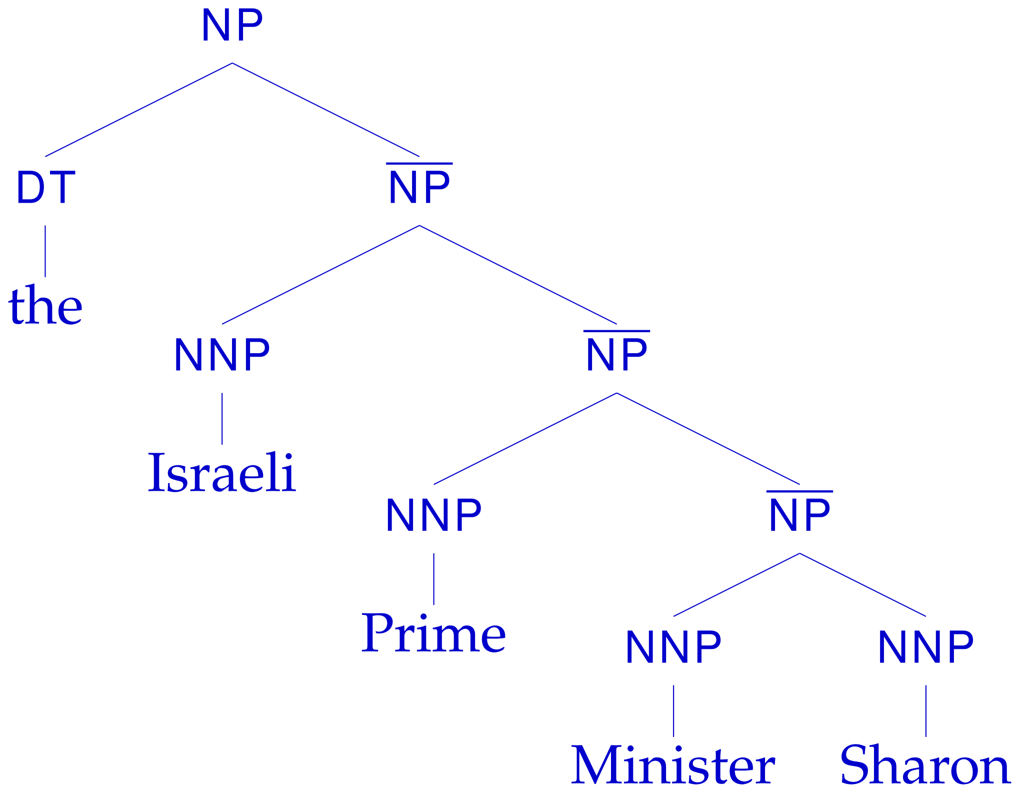
Special Problem: Flat Structures

- 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: $p(\textsf{LHS}, \textsf{RHS}_f, \textsf{RHS}_e)$
  - rule application probability: $p(\textsf{RHS}_f, \textsf{RHS}_e | \textsf{LHS})$
  - direct translation probability: $p(\textsf{RHS}_e | \textsf{RHS}_f, \textsf{LHS})$
  - noisy channel translation probability: $p(\textsf{RHS}_f | \textsf{RHS}_e, \textsf{LHS})$
  - lexical translation probability: $\prod_{e_i \in \textsf{RHS}_e} p(e_i | \textsf{RHS}_f, a)$

next lecture: decoding