
Machine Translation

Philipp Koehn

2 May 2017



Machine Translation: Chinese



因为一项2011年赤字削减协议，如果无法与共和党达成折中方案，奥巴马总统可能在年底面临联邦预算自动减少1000多亿美元的局面。在外交政策辩论中，奥巴马说，他的军事预算不会“减少”而将“维持”。

A 2011 deficit reduction agreement, if a compromise can not be reached with the Republican Party, President Obama may face at the end of the federal budget situation automatically reduced by more than 1000 billion dollars. In the foreign policy debate, Obama said, his military budget will not "reduce" and "maintain".

Machine Translation: French



Obama et Romney prévoient de mener campagne dans les «swing states» à un rythme effréné pour les quatre derniers jours avant l'élection. L'Ohio se présente comme l'Etat le plus disputé du pays.

Obama and Romney plan to campaign in the "swing states" at a breakneck pace for the last four days before the election. The Ohio State presents itself as the most played country.

No Single Right Answer



这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.

Israel is in charge of the security at this airport.

The security work for this airport is the responsibility of the Israel government.

Israeli side was in charge of the security of this airport.

Israel is responsible for the airport's security.

Israel is responsible for safety work at this airport.

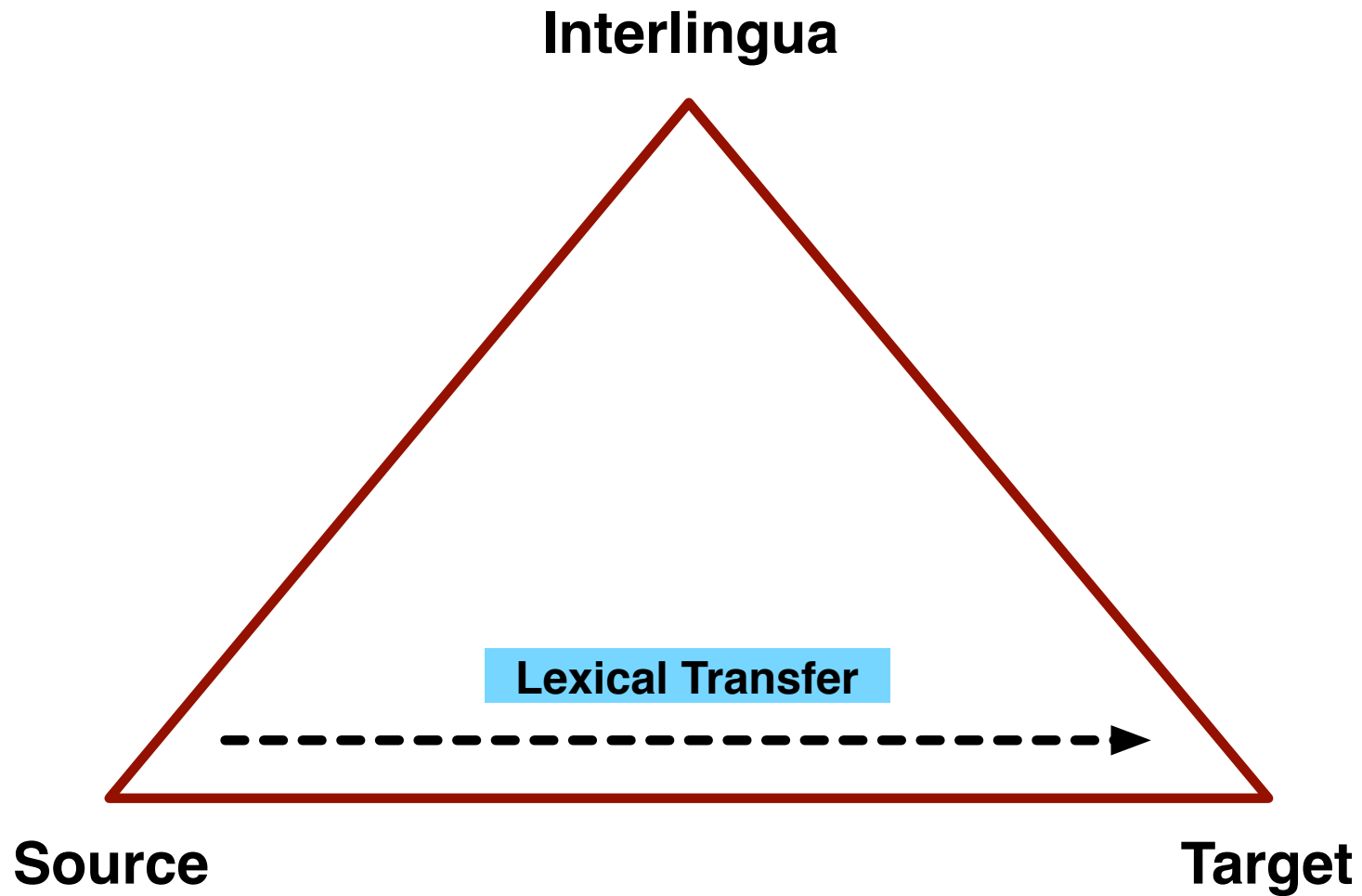
Israel presides over the security of the airport.

Israel took charge of the airport security.

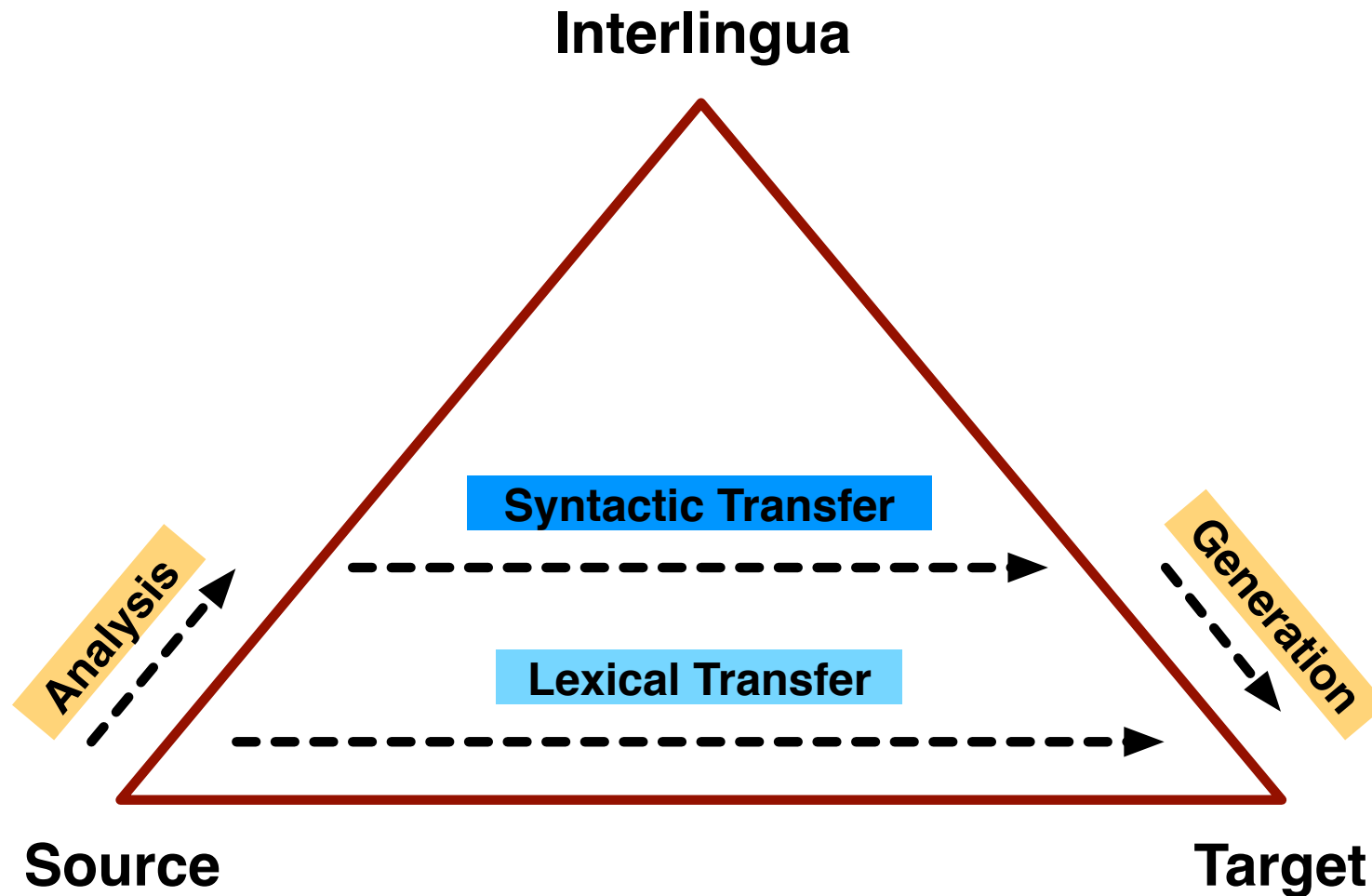
The safety of this airport is taken charge of by Israel.

This airport's security is the responsibility of the Israeli security officials.

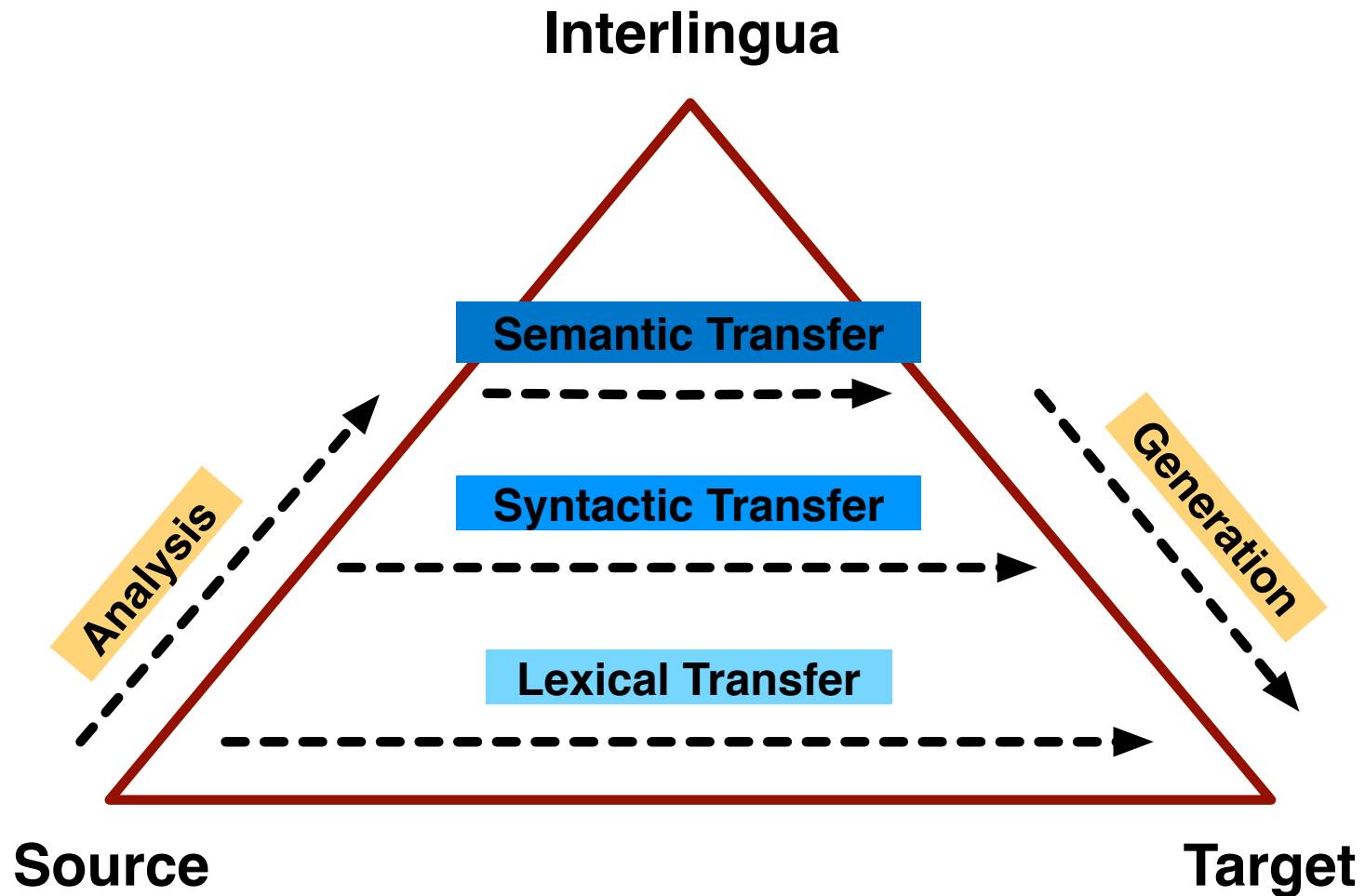
A Clear Plan



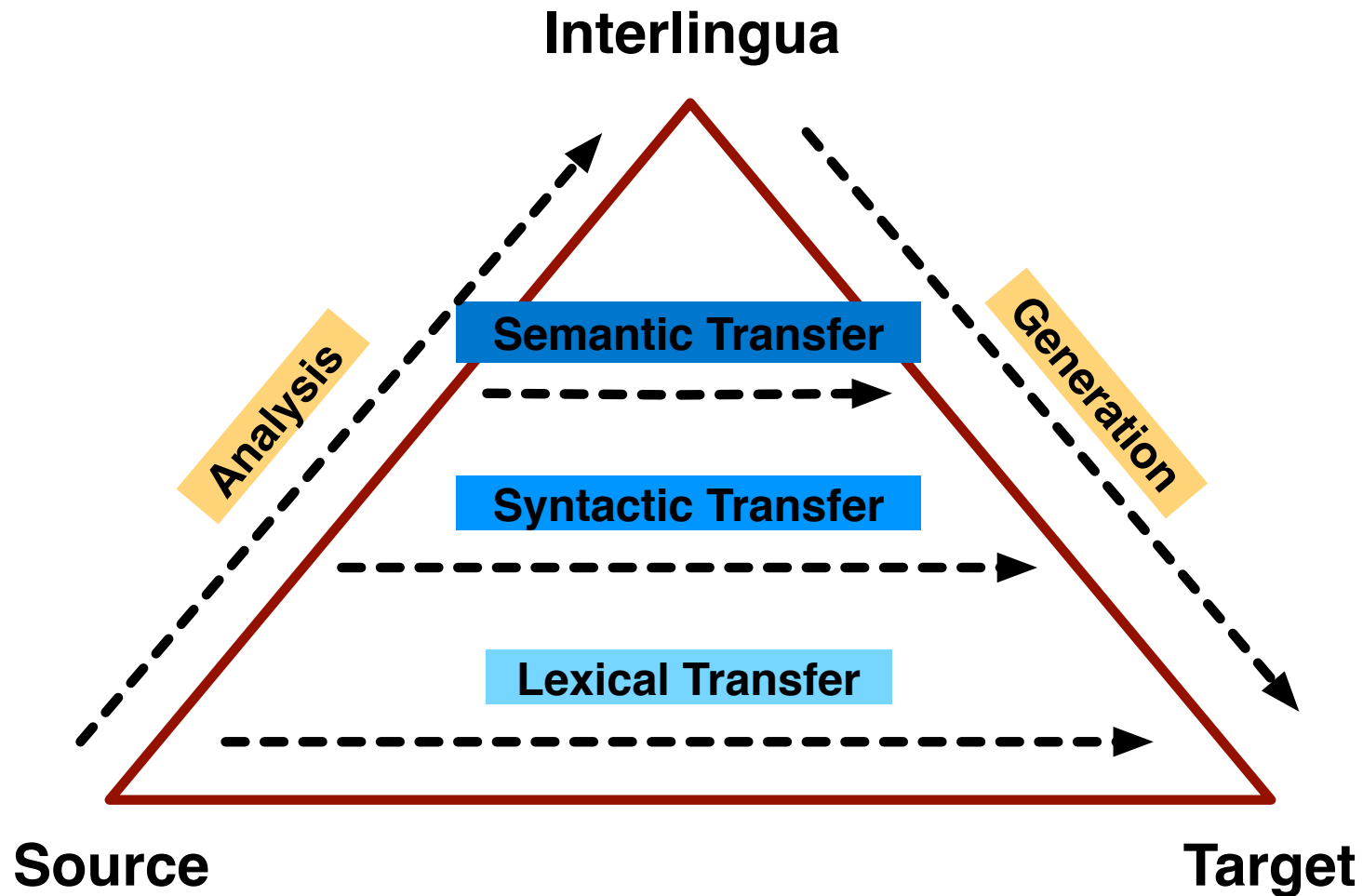
A Clear Plan



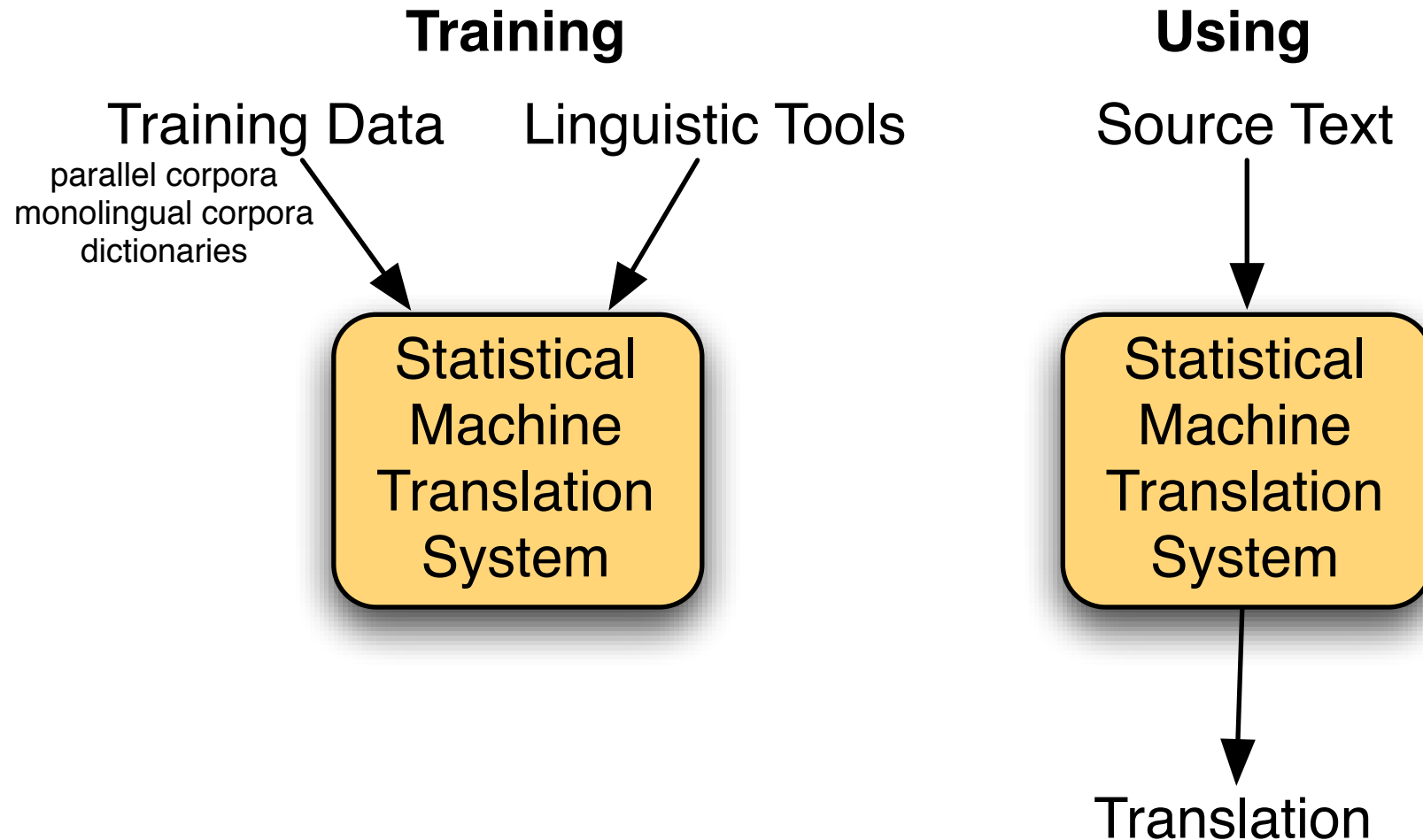
A Clear Plan



A Clear Plan



Learning from Data



why is that a good plan?

Word Translation Problems

- Words are ambiguous

He deposited money in a **bank** account
with a high **interest** rate.

Sitting on the **bank** of the Mississippi,
a passing ship piqued his **interest**.

- How do we find the right meaning, and thus translation?
- Context should be helpful

Syntactic Translation Problems

- Languages have different sentence structure

das	behaupten	sie	wenigstens
this	claim	they	at least
the		she	

- Convert from object-verb-subject (OVS) to subject-verb-object (SVO)
- Ambiguities can be resolved through syntactic analysis
 - the meaning **the** of **das** not possible (not a noun phrase)
 - the meaning **she** of **sie** not possible (subject-verb agreement)

- Pronominal anaphora

I saw the movie and **it** is good.

- How to translate **it** into German (or French)?
 - **it** refers to **movie**
 - **movie** translates to **Film**
 - **Film** has masculine gender
 - ergo: **it** must be translated into masculine pronoun **er**
- We are not handling this very well [Le Nagard and Koehn, 2010]

Semantic Translation Problems



- Coreference

Whenever I visit my uncle and his daughters,
I can't decide who is my favorite **cousin**.

- How to translate **cousin** into German? Male or female?
- Complex inference required

Semantic Translation Problems

- Discourse

Since you brought it up, I do not agree with you.

Since you brought it up, we have been working on it.

- How to translated *since*? Temporal or conditional?
- Analysis of discourse structure — a hard problem

Learning from Data



- What is the best translation?

Sicherheit → security

Sicherheit → safety

Sicherheit → certainty

Learning from Data

- What is the best translation?

Sicherheit → security 14,516

Sicherheit → safety 10,015

Sicherheit → certainty 334

- Counts in European Parliament corpus

Learning from Data



- What is the best translation?

Sicherheit → security 14,516

Sicherheit → safety 10,015

Sicherheit → certainty 334

- Phrasal rules

Sicherheitspolitik → security policy 1580

Sicherheitspolitik → safety policy 13

Sicherheitspolitik → certainty policy 0

Lebensmittelsicherheit → food security 51

Lebensmittelsicherheit → food safety 1084

Lebensmittelsicherheit → food certainty 0

Rechtssicherheit → legal security 156

Rechtssicherheit → legal safety 5

Rechtssicherheit → legal certainty 723

Learning from Data

- What is most fluent?

a problem for translation

a problem of translation

a problem in translation

Learning from Data



- What is most fluent?

a problem for translation 13,000

a problem of translation 61,600

a problem in translation 81,700

- Hits on Google

Learning from Data



- What is most fluent?

a problem for translation 13,000

a problem of translation 61,600

a problem in translation 81,700

a translation problem 235,000

Learning from Data

- What is most fluent?

police disrupted the demonstration

police broke up the demonstration

police dispersed the demonstration

police ended the demonstration

police dissolved the demonstration

police stopped the demonstration

police suppressed the demonstration

police shut down the demonstration

Learning from Data

- What is most fluent?

police disrupted the demonstration 2,140

police broke up the demonstration 66,600

police dispersed the demonstration 25,800

police ended the demonstration 762

police dissolved the demonstration 2,030

police stopped the demonstration 722,000

police suppressed the demonstration 1,400

police shut down the demonstration 2,040

word alignment

- How to translate a word → look up in dictionary
 - Haus** — house, building, home, household, shell.
- Multiple translations
 - some more frequent than others
 - for instance: **house**, and **building** most common
 - special cases: **Haus** of a **snail** is its **shell**
- Note: In all lectures, we translate from a foreign language into English

Collect Statistics

Look at a parallel corpus (German text along with English translation)

Translation of <i>Haus</i>	Count
house	8,000
building	1,600
home	200
household	150
shell	50

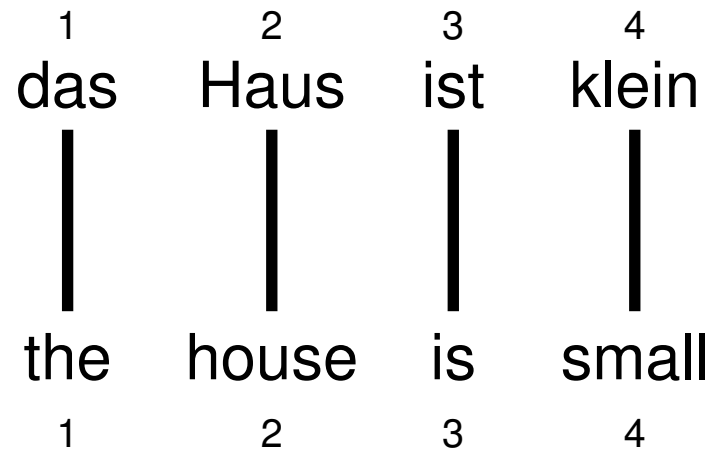
Estimate Translation Probabilities

Maximum likelihood estimation

$$p_f(e) = \begin{cases} 0.8 & \text{if } e = \text{house,} \\ 0.16 & \text{if } e = \text{building,} \\ 0.02 & \text{if } e = \text{home,} \\ 0.015 & \text{if } e = \text{household,} \\ 0.005 & \text{if } e = \text{shell.} \end{cases}$$

Alignment

- In a parallel text (or when we translate), we align words in one language with the words in the other



- Word positions are numbered 1–4

Alignment Function

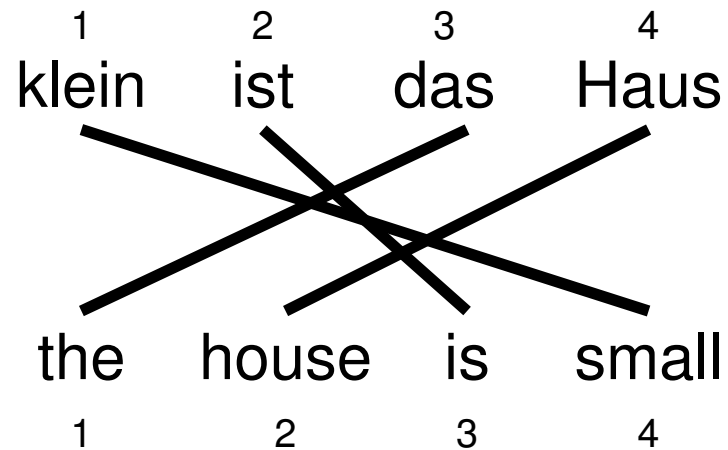
- Formalizing alignment with an alignment function
- Mapping an English target word at position i to a German source word at position j with a function $a : i \rightarrow j$

- Example

$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\}$$

Reordering

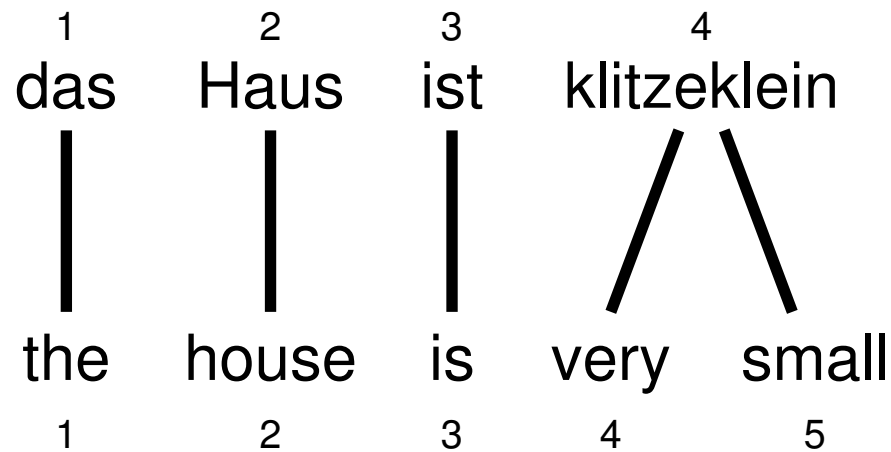
Words may be reordered during translation



$$a : \{1 \rightarrow 3, 2 \rightarrow 4, 3 \rightarrow 2, 4 \rightarrow 1\}$$

One-to-Many Translation

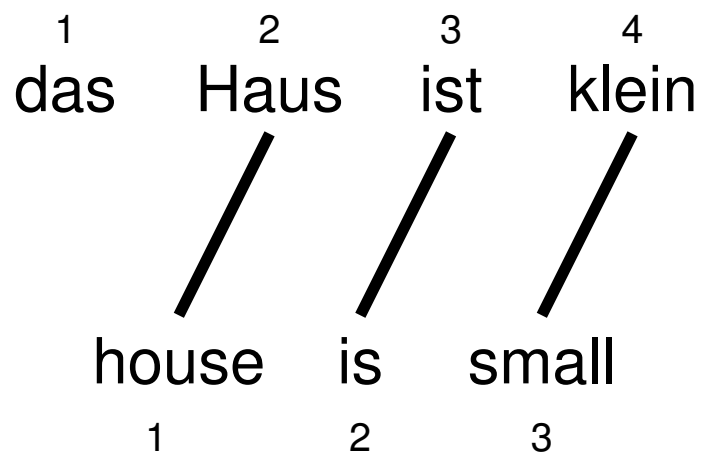
A source word may translate into multiple target words



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4\}$$

Dropping Words

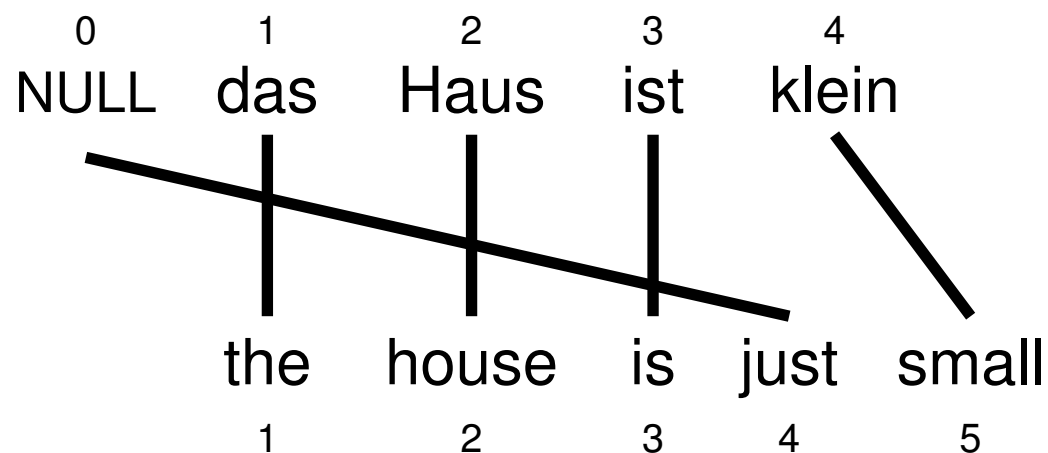
Words may be dropped when translated
(German article **das** is dropped)



$$a : \{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4\}$$

Inserting Words

- Words may be added during translation
 - The English **just** does not have an equivalent in German
 - We still need to map it to something: special NULL token



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 0, 5 \rightarrow 4\}$$

IBM Model 1

- Generative model: break up translation process into smaller steps
 - IBM Model 1 only uses lexical translation
- Translation probability
 - for a foreign sentence $\mathbf{f} = (f_1, \dots, f_{l_f})$ of length l_f
 - to an English sentence $\mathbf{e} = (e_1, \dots, e_{l_e})$ of length l_e
 - with an alignment of each English word e_j to a foreign word f_i according to the alignment function $a : j \rightarrow i$

$$p(\mathbf{e}, a | \mathbf{f}) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j | f_{a(j)})$$

- parameter ϵ is a normalization constant

Example

das

e	$t(e f)$
the	0.7
that	0.15
which	0.075
who	0.05
this	0.025

Haus

e	$t(e f)$
house	0.8
building	0.16
home	0.02
household	0.015
shell	0.005

ist

e	$t(e f)$
is	0.8
's	0.16
exists	0.02
has	0.015
are	0.005

klein

e	$t(e f)$
small	0.4
little	0.4
short	0.1
minor	0.06
petty	0.04

$$\begin{aligned} p(e, a|f) &= \frac{\epsilon}{4^3} \times t(\text{the}|\text{das}) \times t(\text{house}|\text{Haus}) \times t(\text{is}|\text{ist}) \times t(\text{small}|\text{klein}) \\ &= \frac{\epsilon}{4^3} \times 0.7 \times 0.8 \times 0.8 \times 0.4 \\ &= 0.0028\epsilon \end{aligned}$$

em algorithm

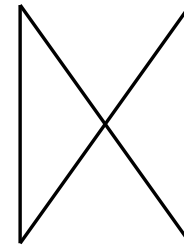
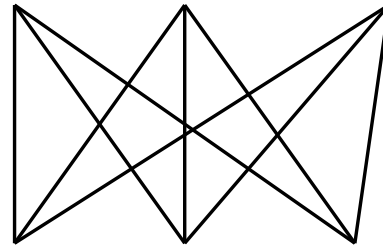
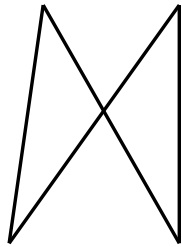
- We would like to estimate the lexical translation probabilities $t(e|f)$ from a parallel corpus
- ... but we do not have the alignments
- Chicken and egg problem
 - if we had the *alignments*,
→ we could estimate the *parameters* of our generative model
 - if we had the *parameters*,
→ we could estimate the *alignments*

EM Algorithm

- Incomplete data
 - if we had *complete data*, would could estimate *model*
 - if we had *model*, we could fill in the *gaps in the data*
- Expectation Maximization (EM) in a nutshell
 1. initialize model parameters (e.g. uniform)
 2. assign probabilities to the missing data
 3. estimate model parameters from completed data
 4. iterate steps 2–3 until convergence

EM Algorithm

... la maison ... la maison blue ... la fleur ...

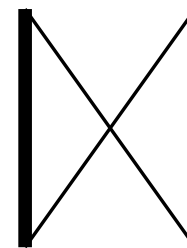
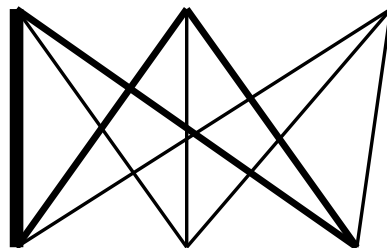
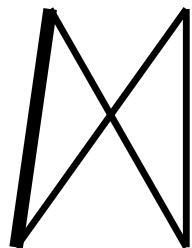


... the house ... the blue house ... the flower ...

- Initial step: all alignments equally likely
- Model learns that, e.g., **la** is often aligned with **the**

EM Algorithm

... la maison ... la maison blue ... la fleur ...

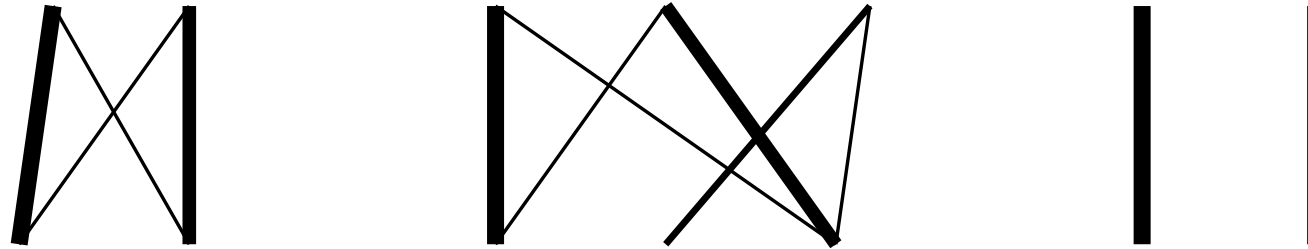


... the house ... the blue house ... the flower ...

- After one iteration
- Alignments, e.g., between **la** and **the** are more likely

EM Algorithm

... la maison ... la maison bleu ... la fleur ...



... the house ... the blue house ... the flower ...

- After another iteration
- It becomes apparent that alignments, e.g., between **fleur** and **flower** are more likely (pigeon hole principle)

EM Algorithm

... la maison ... la maison bleu ... la fleur ...
/ | | X | |
... the house ... the blue house ... the flower ...

- Convergence
- Inherent hidden structure revealed by EM

EM Algorithm

... la maison ... la maison bleu ... la fleur ...
/ | | X | |
... the house ... the blue house ... the flower ...



$p(\text{la}|\text{the}) = 0.453$
 $p(\text{le}|\text{the}) = 0.334$
 $p(\text{maison}|\text{house}) = 0.876$
 $p(\text{bleu}|\text{blue}) = 0.563$
...

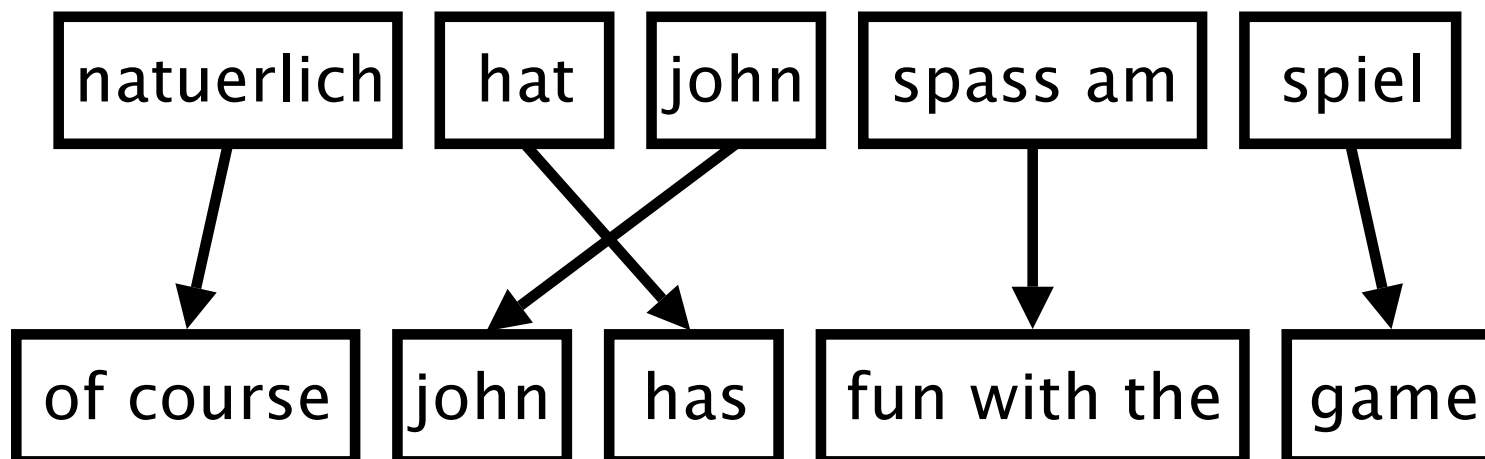
- Parameter estimation from the aligned corpus

IBM Model 1 and EM

- EM Algorithm consists of two steps
- Expectation-Step: Apply model to the data
 - parts of the model are hidden (here: alignments)
 - using the model, assign probabilities to possible values
- Maximization-Step: Estimate model from data
 - take assign values as fact
 - collect counts (weighted by probabilities)
 - estimate model from counts
- Iterate these steps until convergence

phrase-based models

Phrase-Based Model



- Foreign input is segmented in phrases
- Each phrase is translated into English
- Phrases are reordered

Phrase Translation Table

- Main knowledge source: table with phrase translations and their probabilities
- Example: phrase translations for *natuerlich*

Translation	Probability $\phi(\bar{e} f)$
of course	0.5
naturally	0.3
of course ,	0.15
, of course ,	0.05

Real Example

- Phrase translations for **den Vorschlag** learned from the Europarl corpus:

English	$\phi(\bar{e} f)$	English	$\phi(\bar{e} f)$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159

- lexical variation (**proposal** vs **suggestions**)
- morphological variation (**proposal** vs **proposals**)
- included function words (**the**, **a**, ...)
- noise (**it**)

decoding

Decoding

- We have a mathematical model for translation

$$p(\mathbf{e}|\mathbf{f})$$

- Task of decoding: find the translation \mathbf{e}_{best} with highest probability

$$\mathbf{e}_{\text{best}} = \operatorname{argmax}_{\mathbf{e}} p(\mathbf{e}|\mathbf{f})$$

- Two types of error
 - the most probable translation is bad → fix the model
 - search does not find the most probably translation → fix the search
- Decoding is evaluated by search error, not quality of translations (although these are often correlated)

Translation Process



- Task: translate this sentence from German into English

er **geht** **ja** **nicht** **nach** **hause**

Translation Process

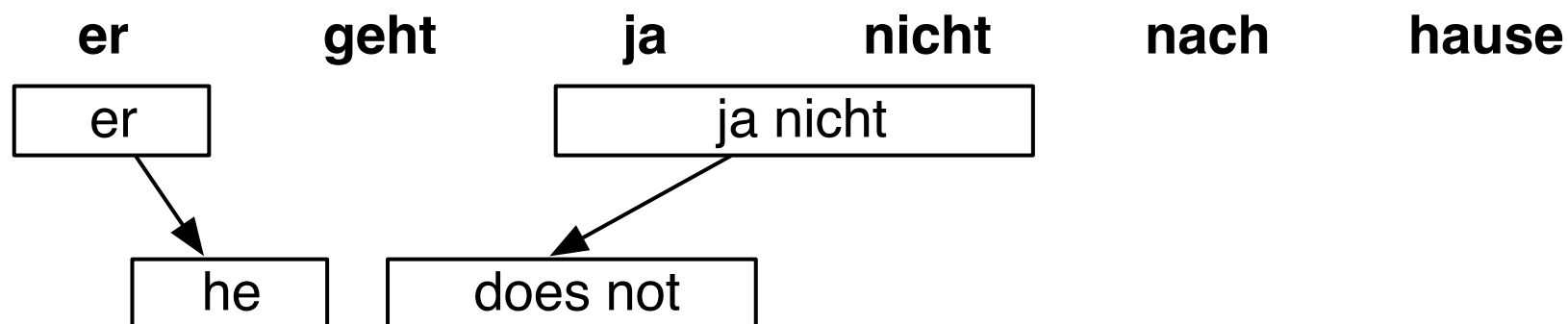
- Task: translate this sentence from German into English



- Pick phrase in input, translate

Translation Process

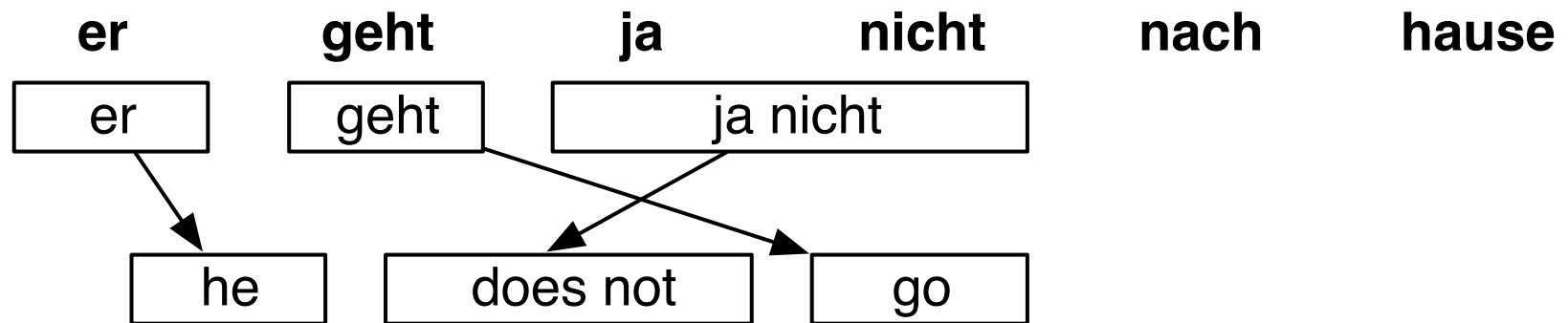
- Task: translate this sentence from German into English



- Pick phrase in input, translate
 - it is allowed to pick words out of sequence reordering
 - phrases may have multiple words: many-to-many translation

Translation Process

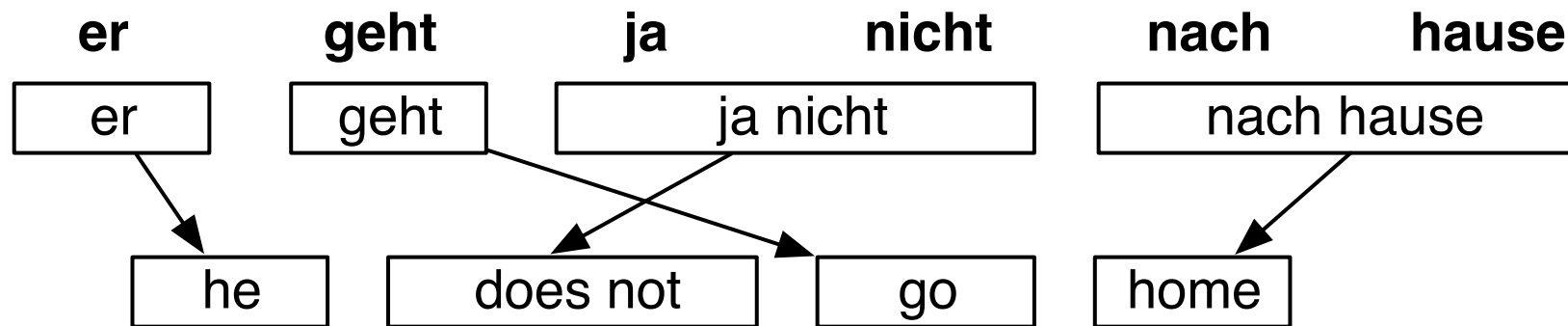
- Task: translate this sentence from German into English



- Pick phrase in input, translate

Translation Process

- Task: translate this sentence from German into English



- Pick phrase in input, translate

Computing Translation Probability

- Probabilistic model for phrase-based translation:

$$e_{\text{best}} = \operatorname{argmax}_e \prod_{i=1}^I \phi(\bar{f}_i | \bar{e}_i) d(\text{start}_i - \text{end}_{i-1} - 1) p_{\text{LM}}(\mathbf{e})$$

- Score is computed incrementally for each partial hypothesis
- Components

Phrase translation Picking phrase \bar{f}_i to be translated as a phrase \bar{e}_i

→ look up score $\phi(\bar{f}_i | \bar{e}_i)$ from phrase translation table

Reordering Previous phrase ended in end_{i-1} , current phrase starts at start_i

→ compute $d(\text{start}_i - \text{end}_{i-1} - 1)$

Language model For n -gram model, need to keep track of last $n - 1$ words

→ compute score $p_{\text{LM}}(w_i | w_{i-(n-1)}, \dots, w_{i-1})$ for added words w_i

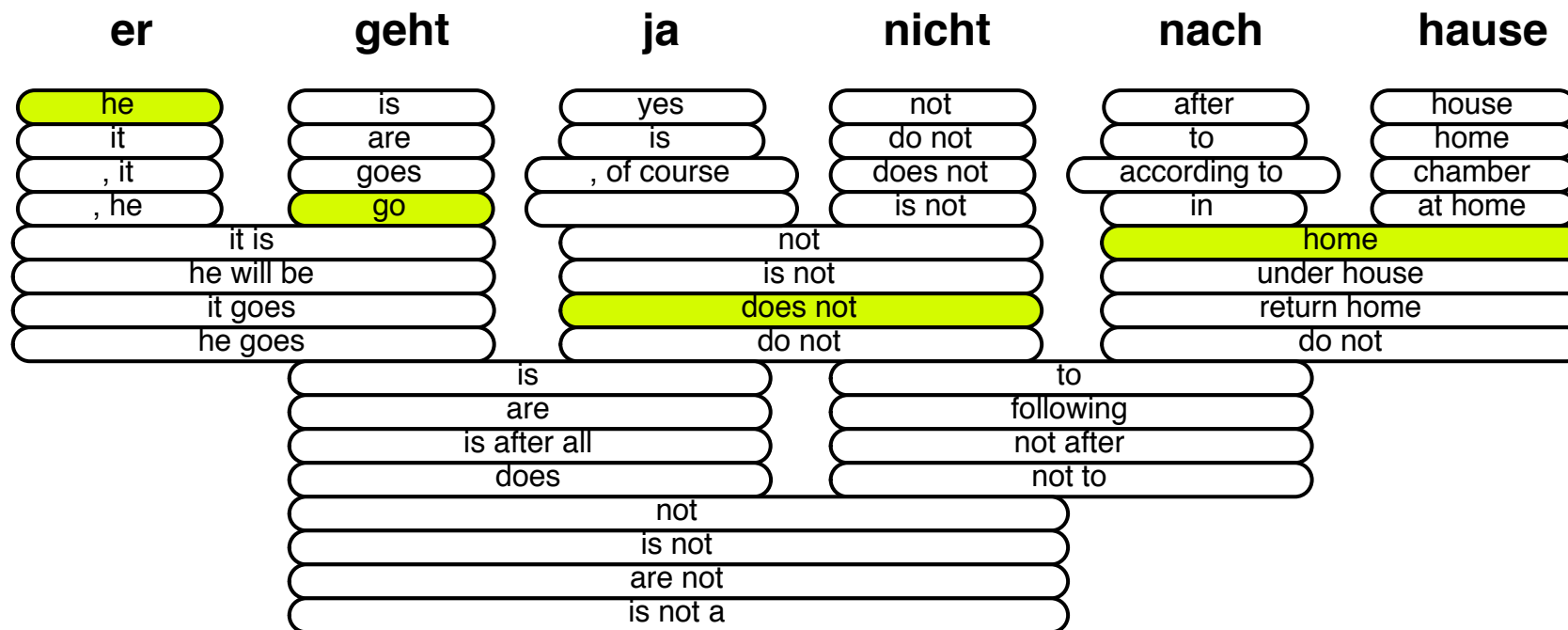
decoding process

Translation Options

er	geht	ja	nicht	nach	hause
he	is	yes	not	after	house
it	are	is	do not	to	home
, it	goes	, of course	does not	according to	chamber
, he	go	,	is not	in	at home
it is		not		home	
he will be		is not		under house	
it goes		does not		return home	
he goes		do not		do not	
		is		to	
		are		following	
		is after all		not after	
		does		not to	
		not			
		is not			
		are not			
		is not a			

- Many translation options to choose from
 - in Europarl phrase table: 2727 matching phrase pairs for this sentence
 - by pruning to the top 20 per phrase, 202 translation options remain

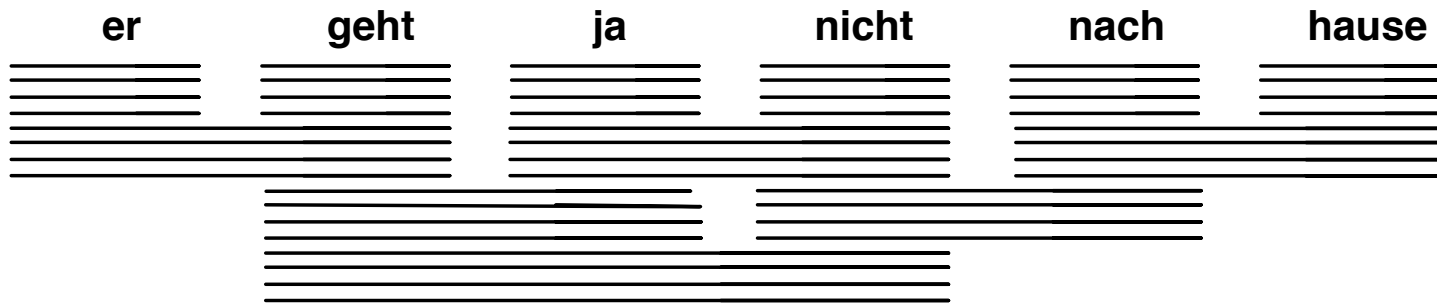
Translation Options



- The machine translation decoder does not know the right answer
 - picking the right translation options
 - arranging them in the right order
- Search problem solved by heuristic beam search

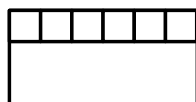
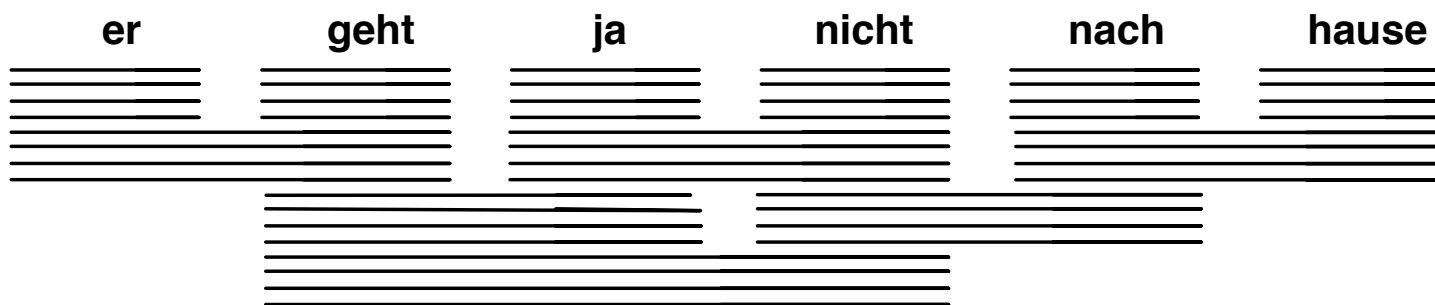
Decoding: Precompute Translation Options

59



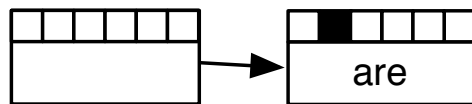
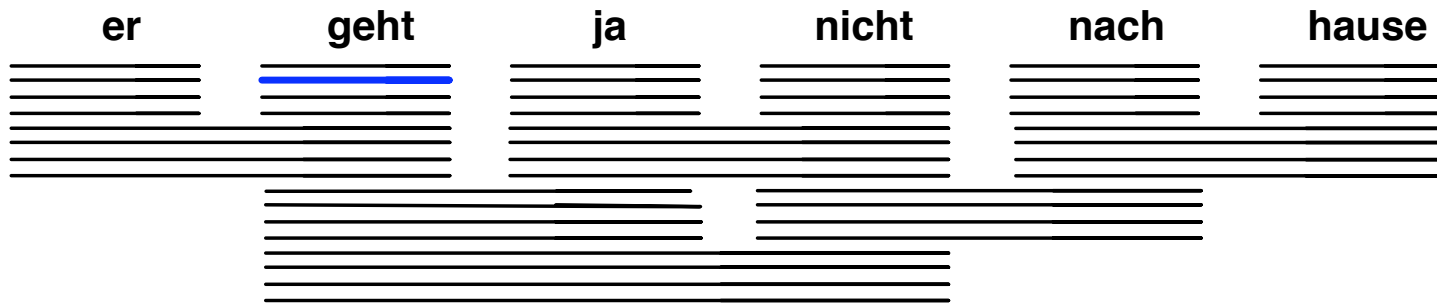
consult phrase translation table for all input phrases

Decoding: Start with Initial Hypothesis



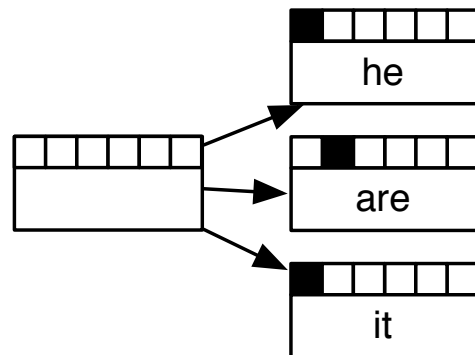
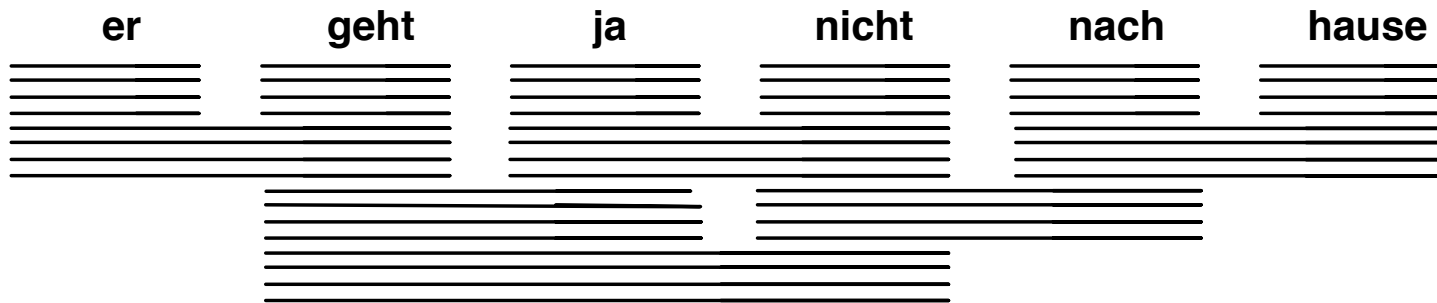
initial hypothesis: no input words covered, no output produced

Decoding: Hypothesis Expansion



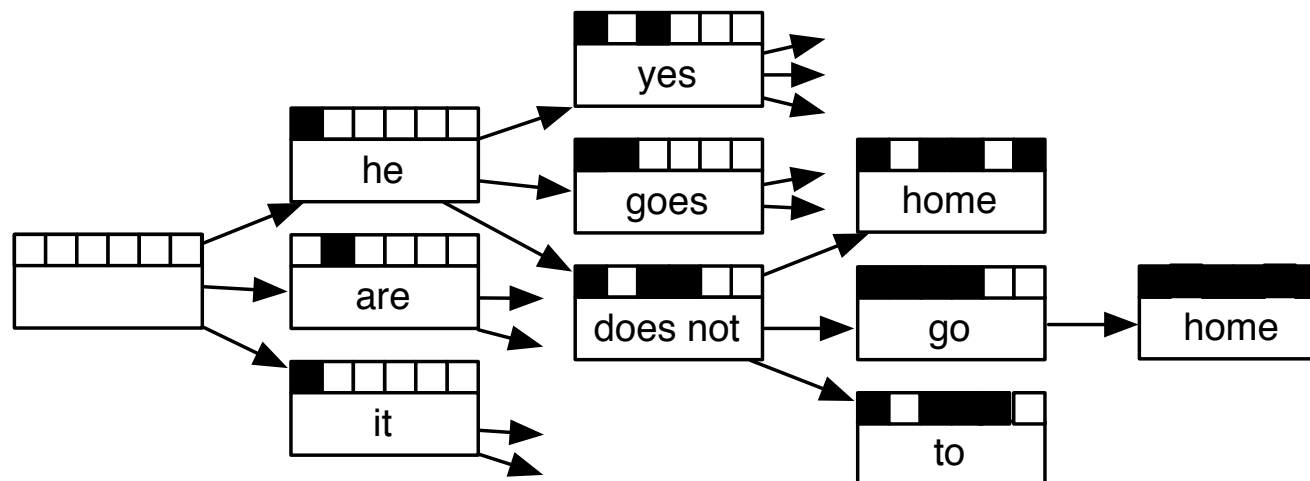
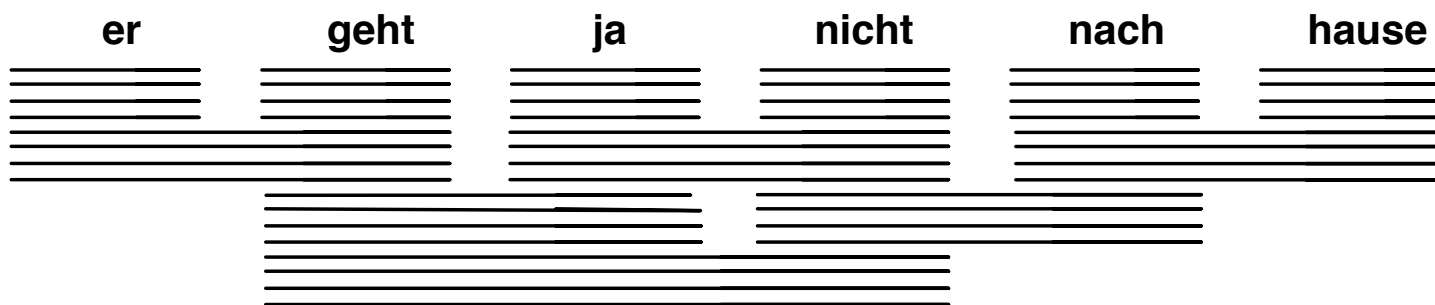
pick any translation option, create new hypothesis

Decoding: Hypothesis Expansion



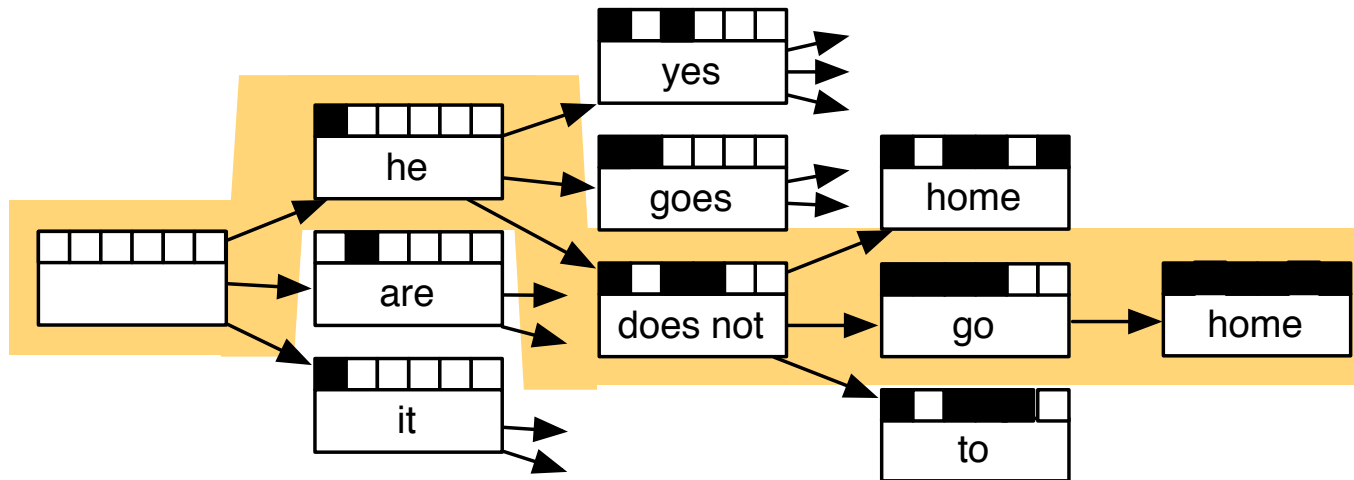
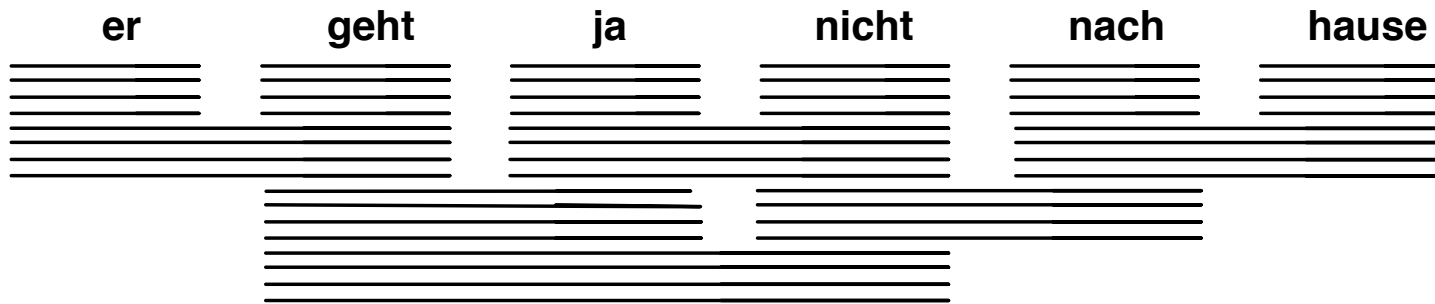
create hypotheses for all other translation options

Decoding: Hypothesis Expansion



also create hypotheses from created partial hypothesis

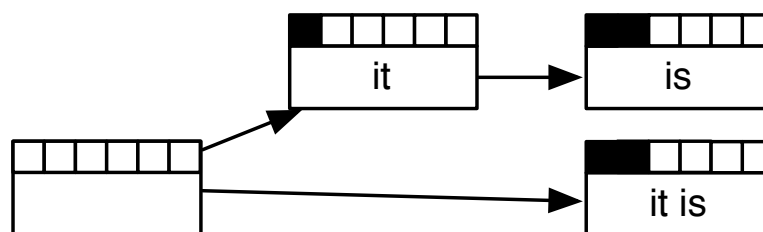
Decoding: Find Best Path



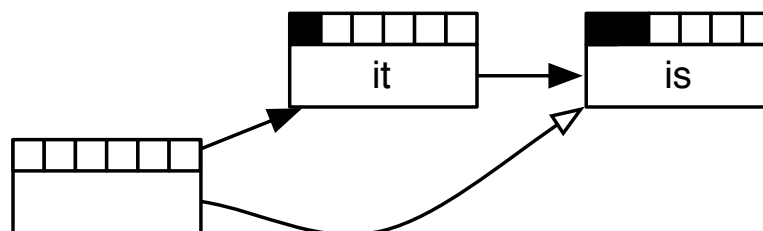
backtrack from highest scoring complete hypothesis

Recombination

- Two hypothesis paths lead to two matching hypotheses
 - same number of foreign words translated
 - same English words in the output
 - different scores

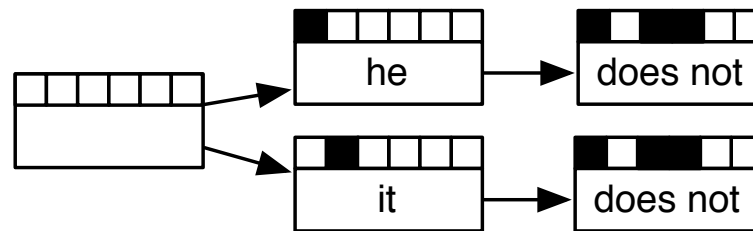


- Worse hypothesis is dropped

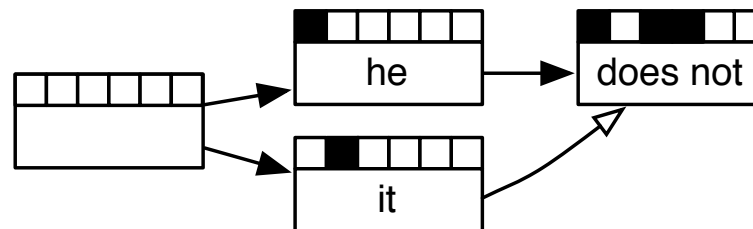


Recombination

- Two hypothesis paths lead to hypotheses indistinguishable in subsequent search
 - same number of foreign words translated
 - same last two English words in output (assuming trigram language model)
 - same last foreign word translated
 - different scores



- Worse hypothesis is dropped

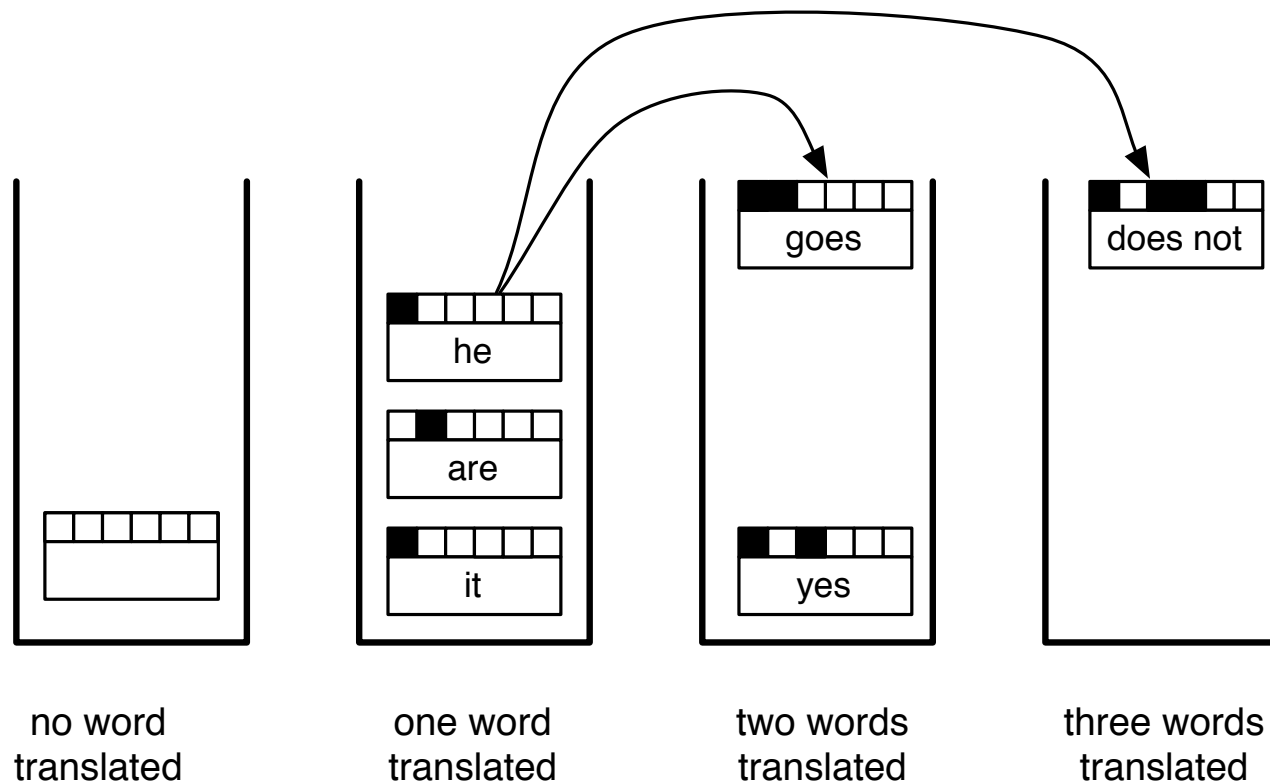


Pruning



- Recombination reduces search space, but not enough (we still have a NP complete problem on our hands)
- Pruning: remove bad hypotheses early
 - put comparable hypothesis into stacks (hypotheses that have translated same number of input words)
 - limit number of hypotheses in each stack

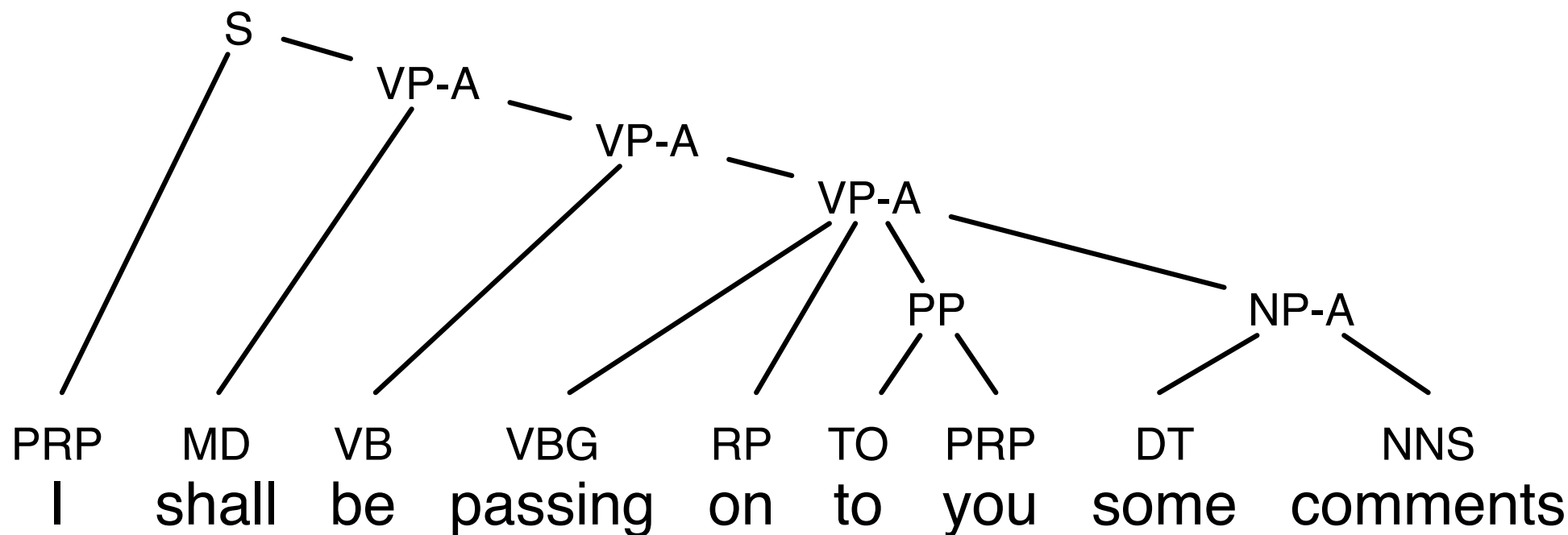
Stacks



- Hypothesis expansion in a stack decoder
 - translation option is applied to hypothesis
 - new hypothesis is dropped into a stack further down

syntax-based models

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

- Nonterminal rules

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

- Terminal rules

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

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

- Mixed rules

$NP \rightarrow \text{la maison } JJ_1 \mid \text{the } JJ_1 \text{ house}$

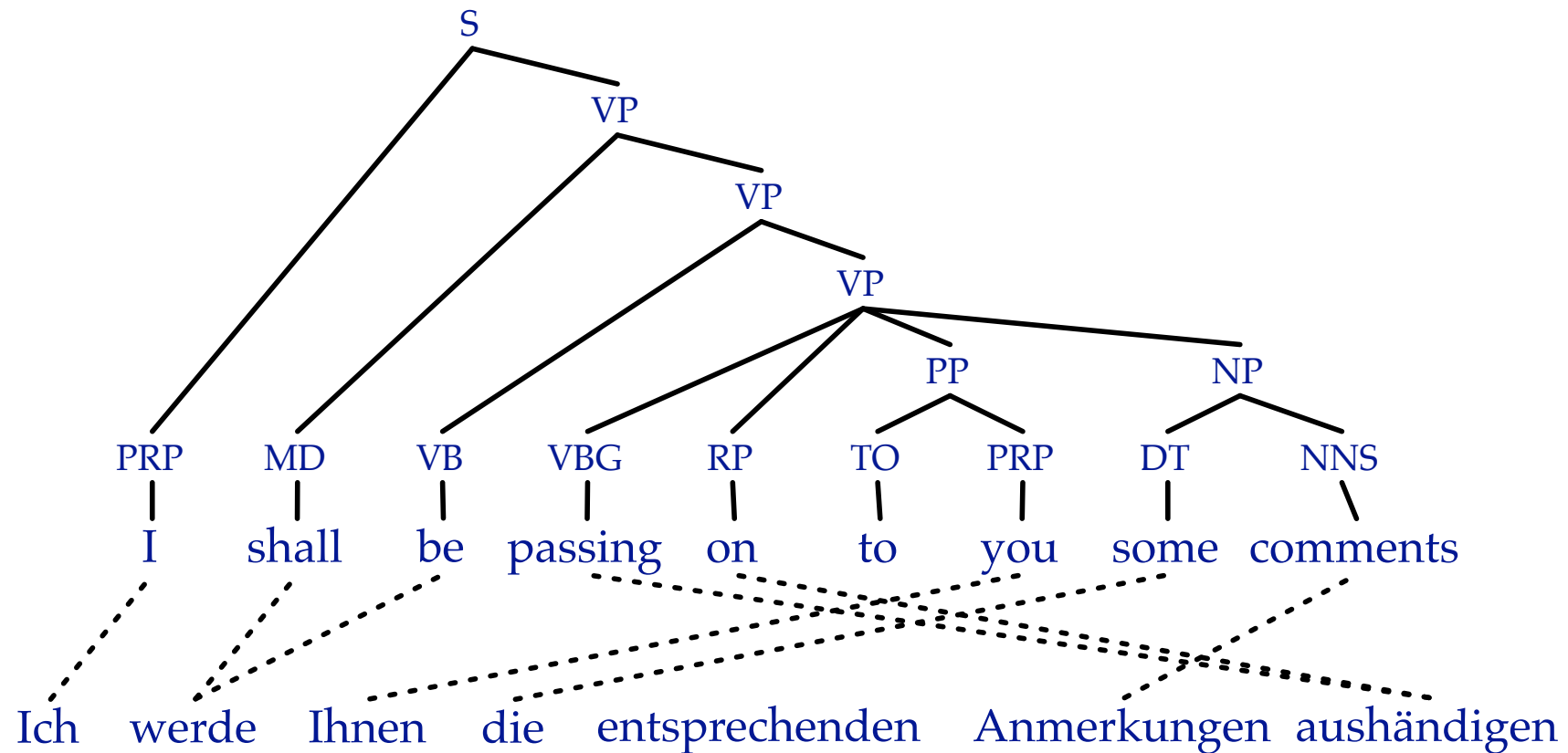
Tree-Based Translation Model

- 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

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

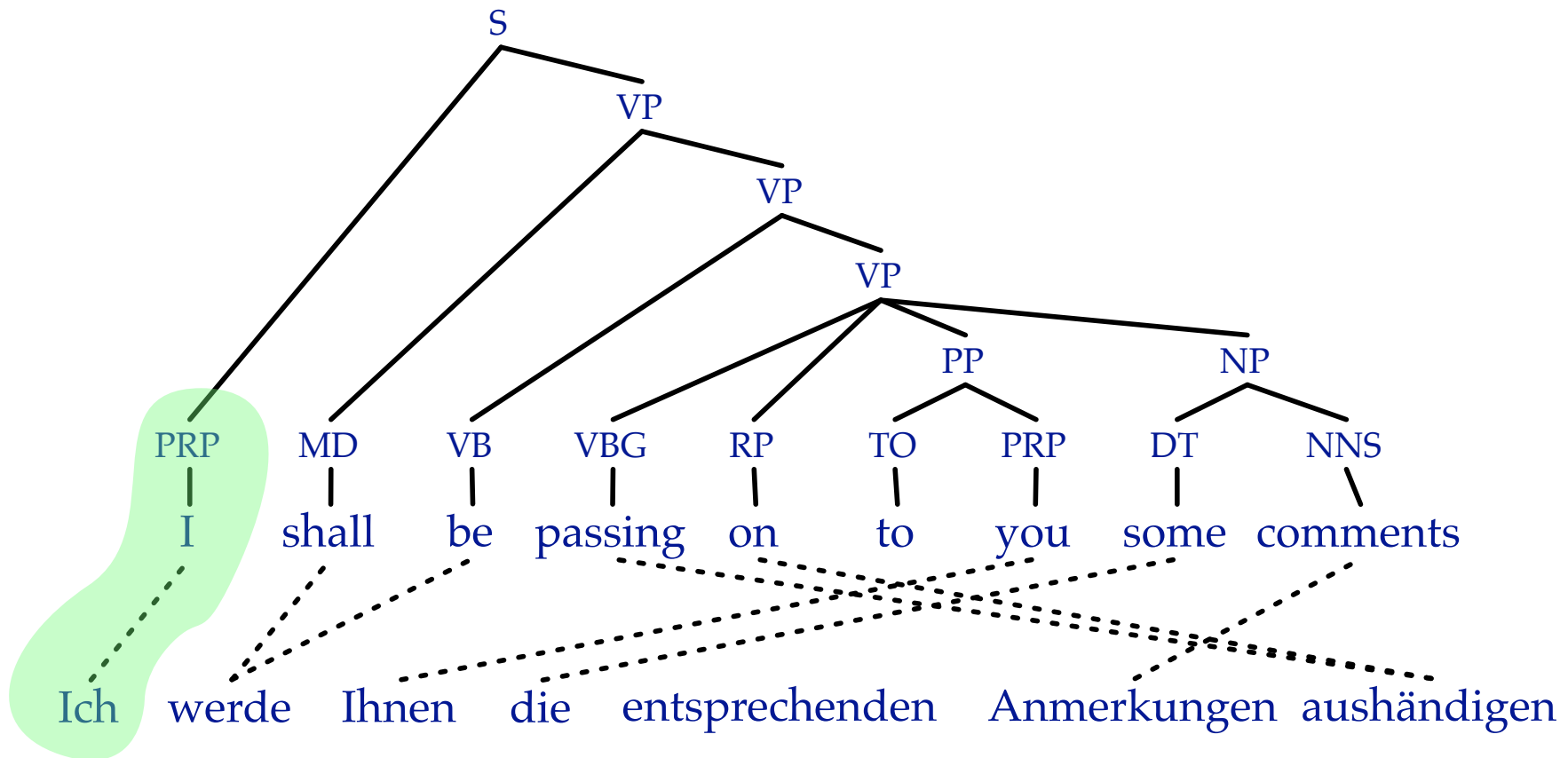
- Many ways to assign probabilities to rules

Minimal Rules

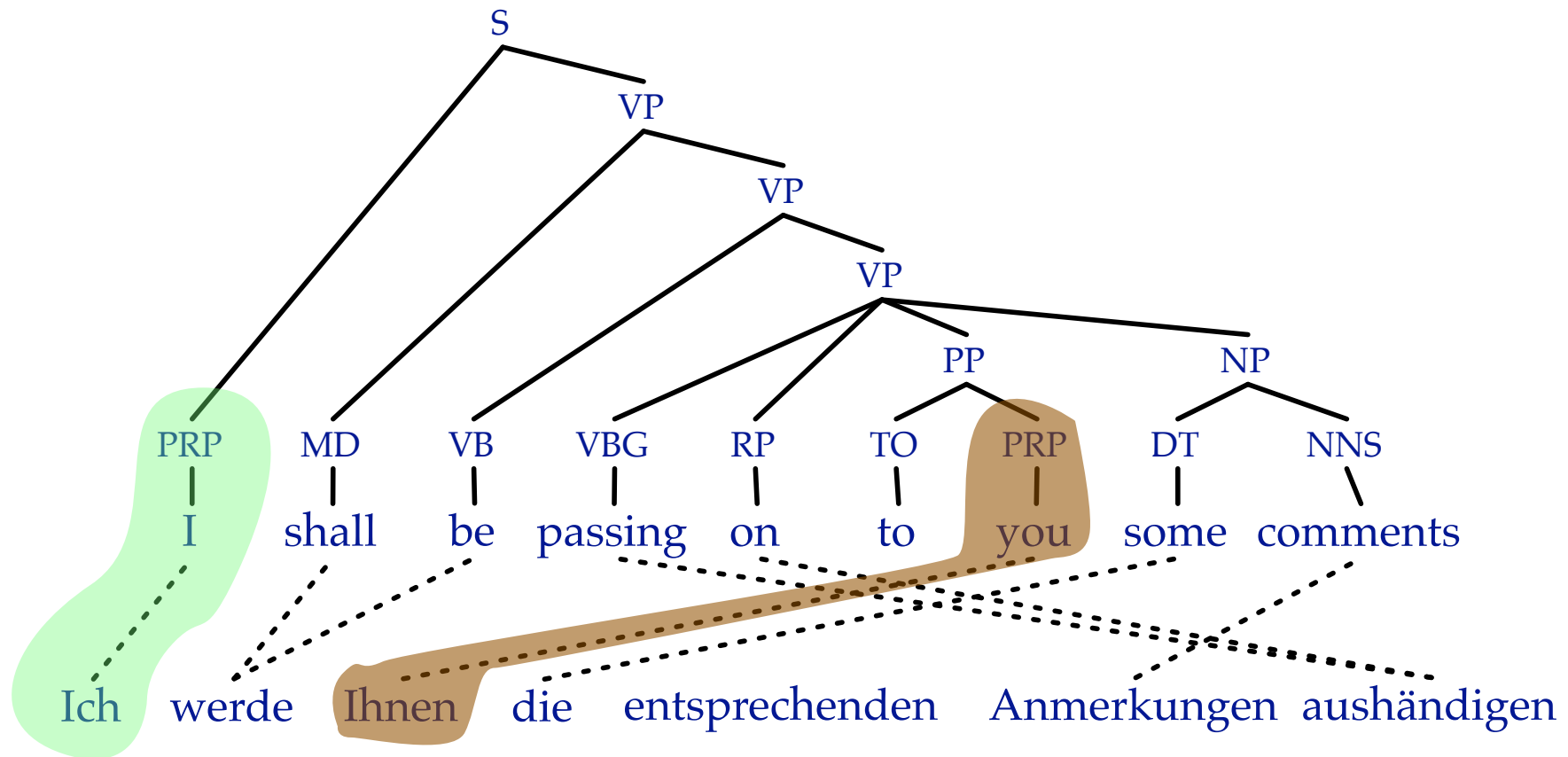


Extract: set of smallest rules required to explain the sentence pair

Lexical Rule

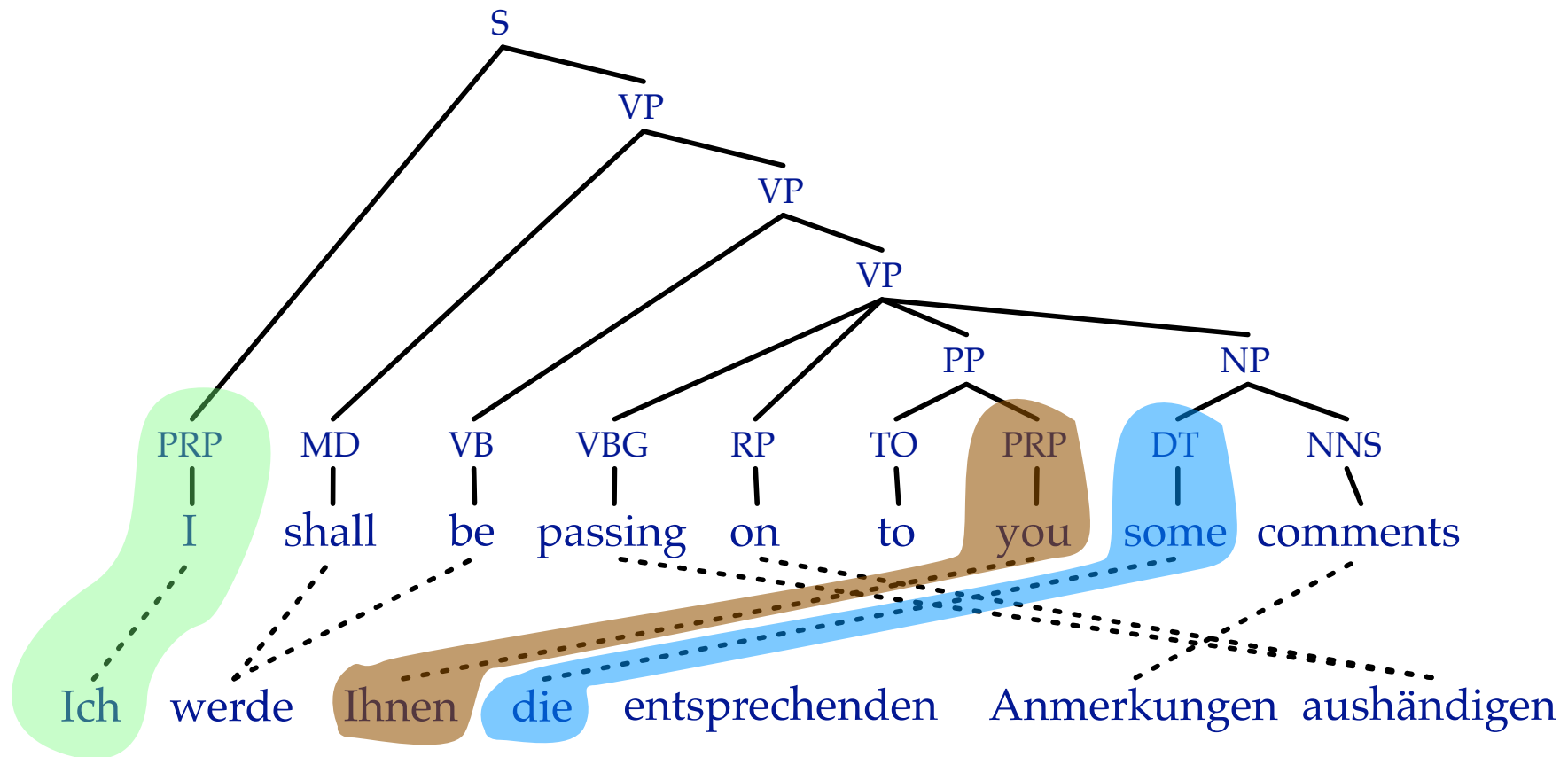


Lexical Rule



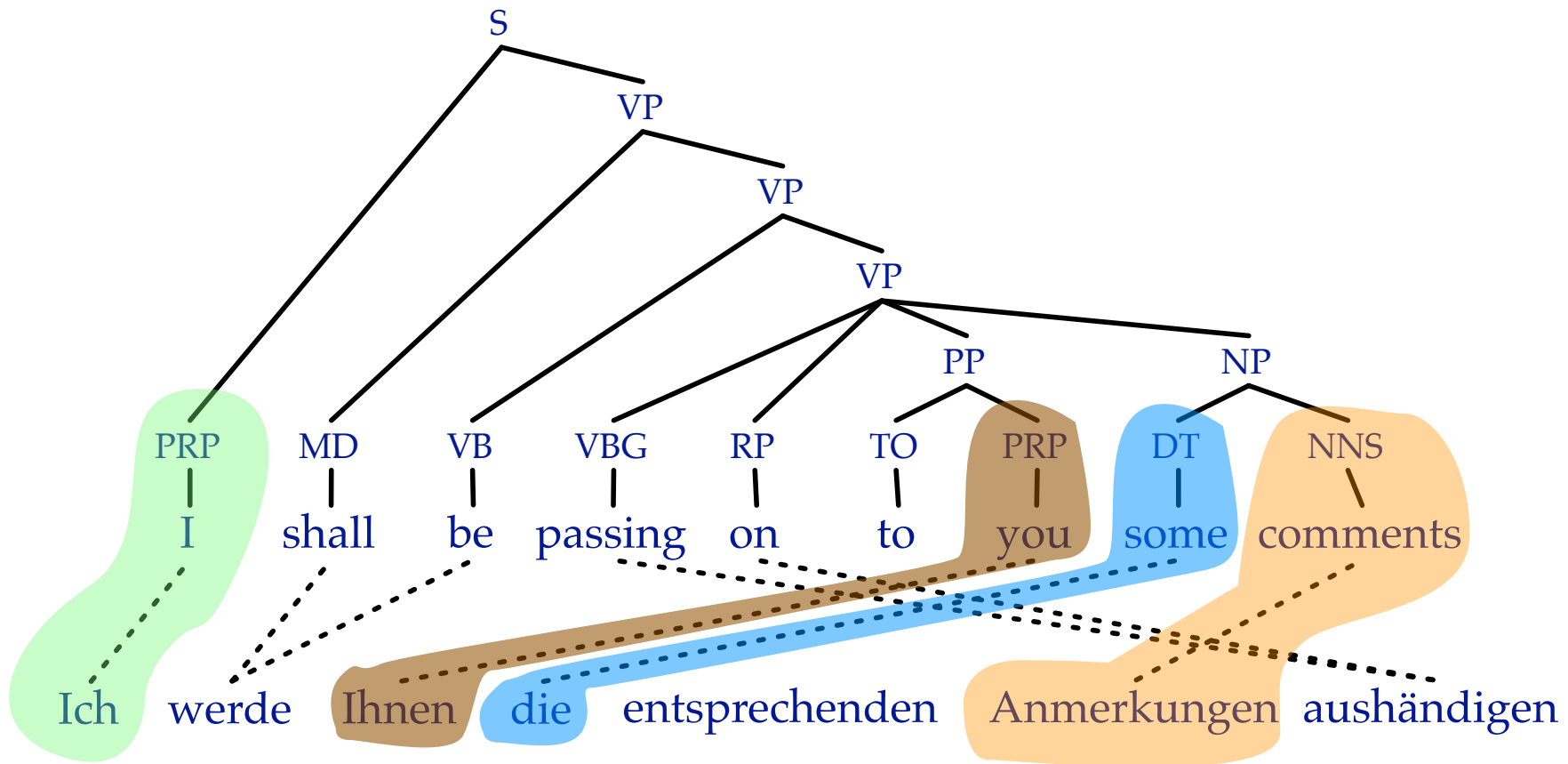
Extracted rule: PRP → Ihnen | you

Lexical Rule



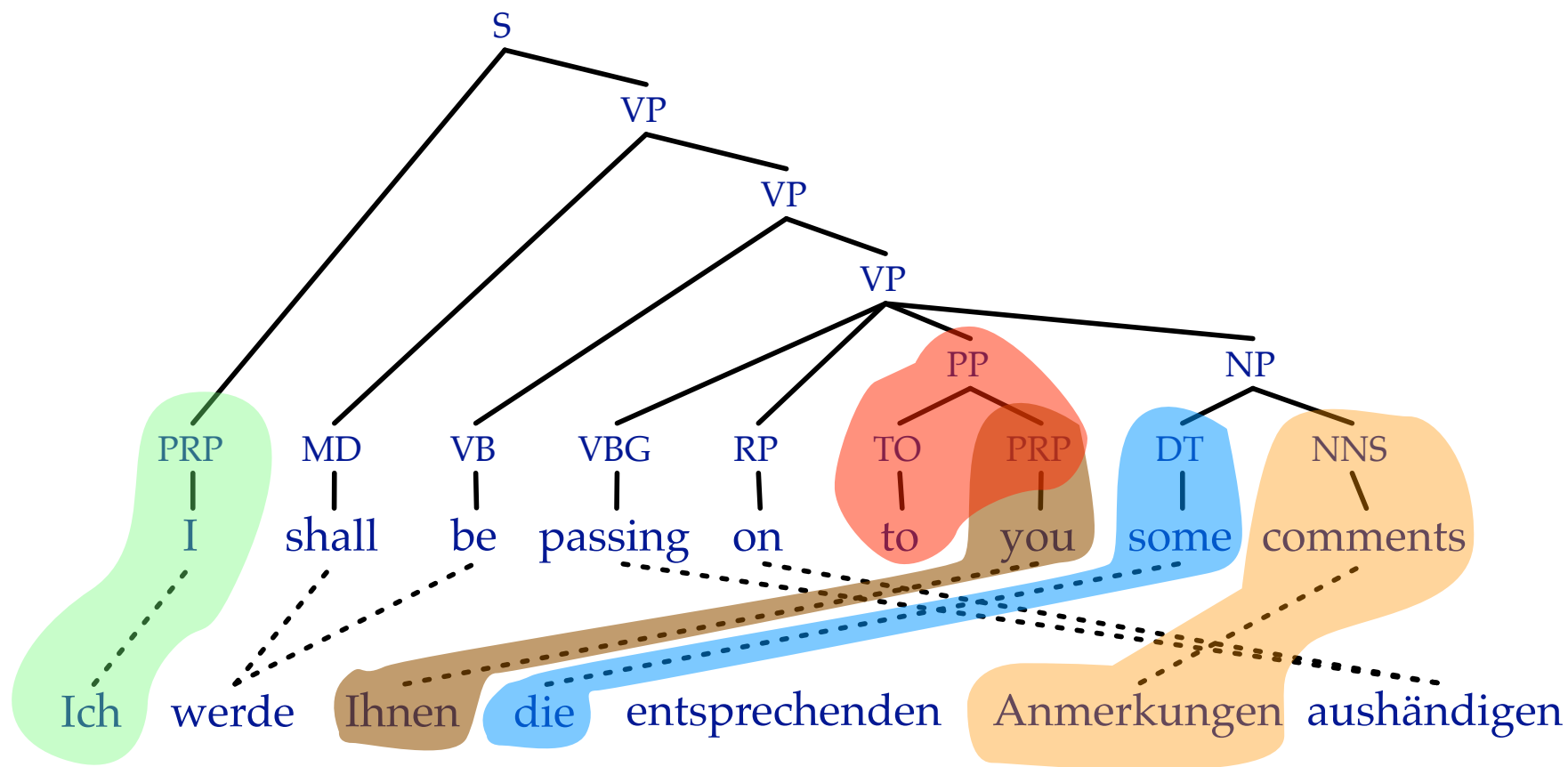
Extracted rule: DT → die | some

Lexical Rule



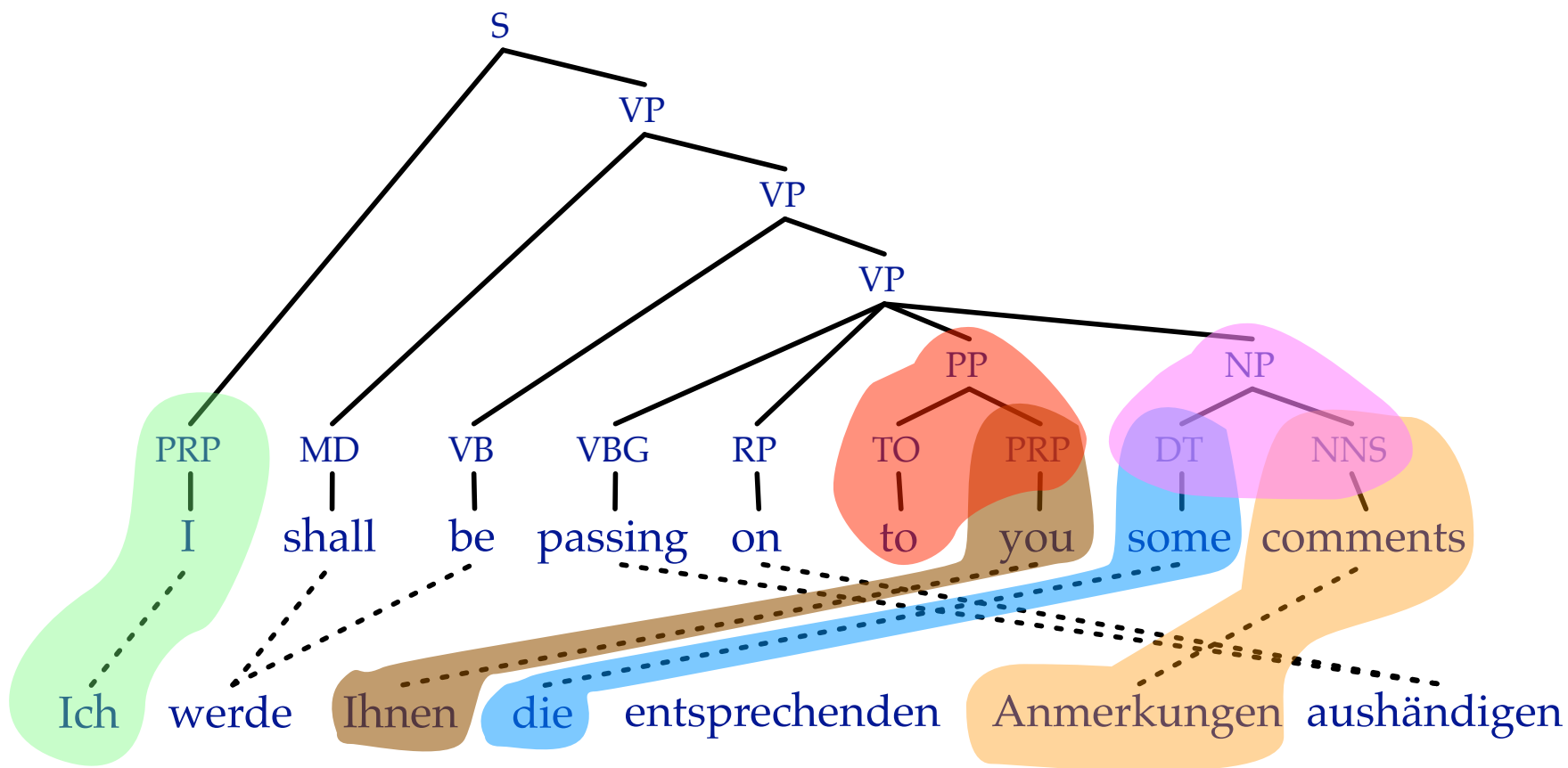
Extracted rule: **NNS** → **Anmerkungen** | **comments**

Insertion Rule



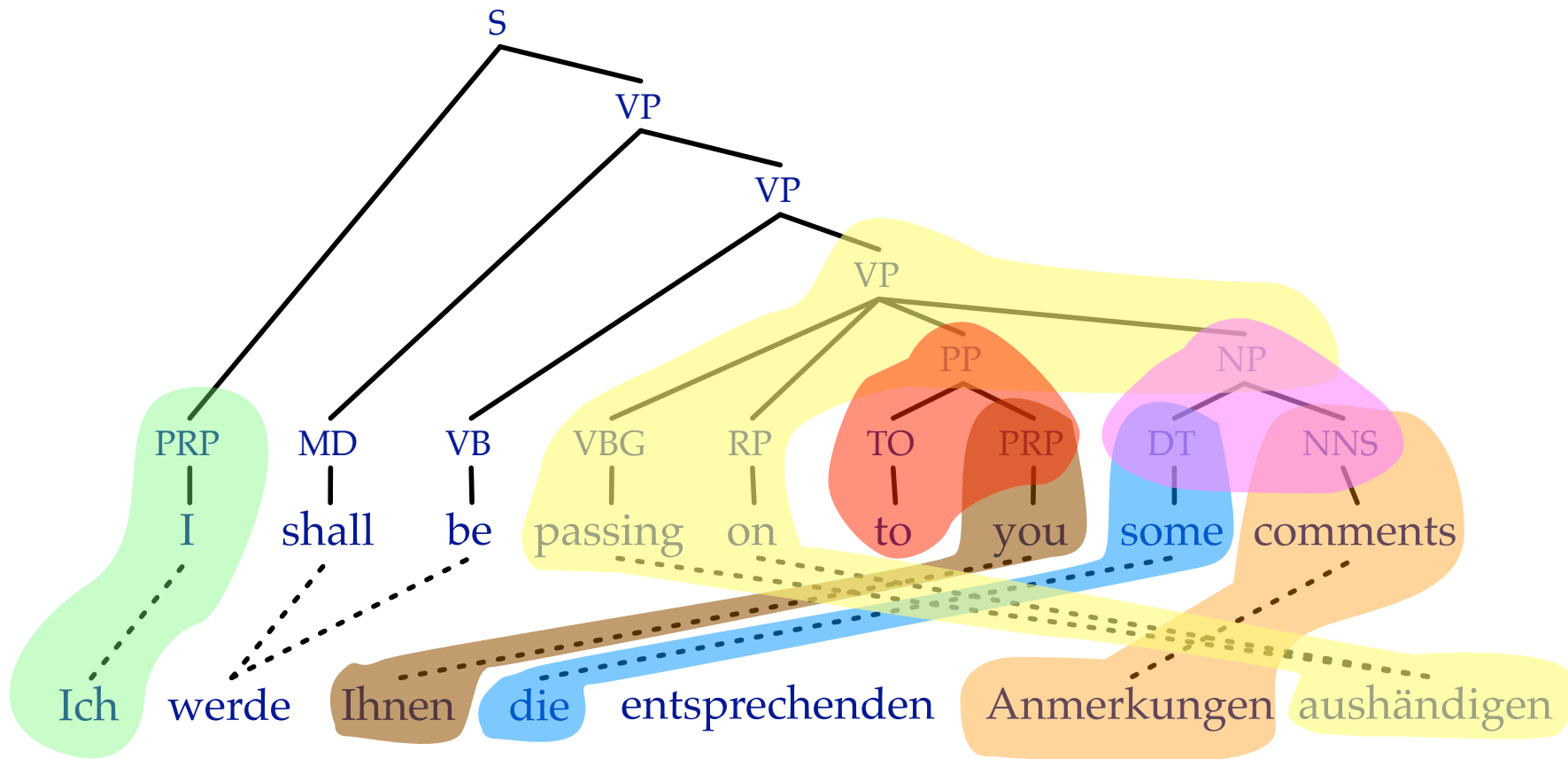
Extracted rule: $PP \rightarrow X \mid \text{to PRP}$

Non-Lexical Rule



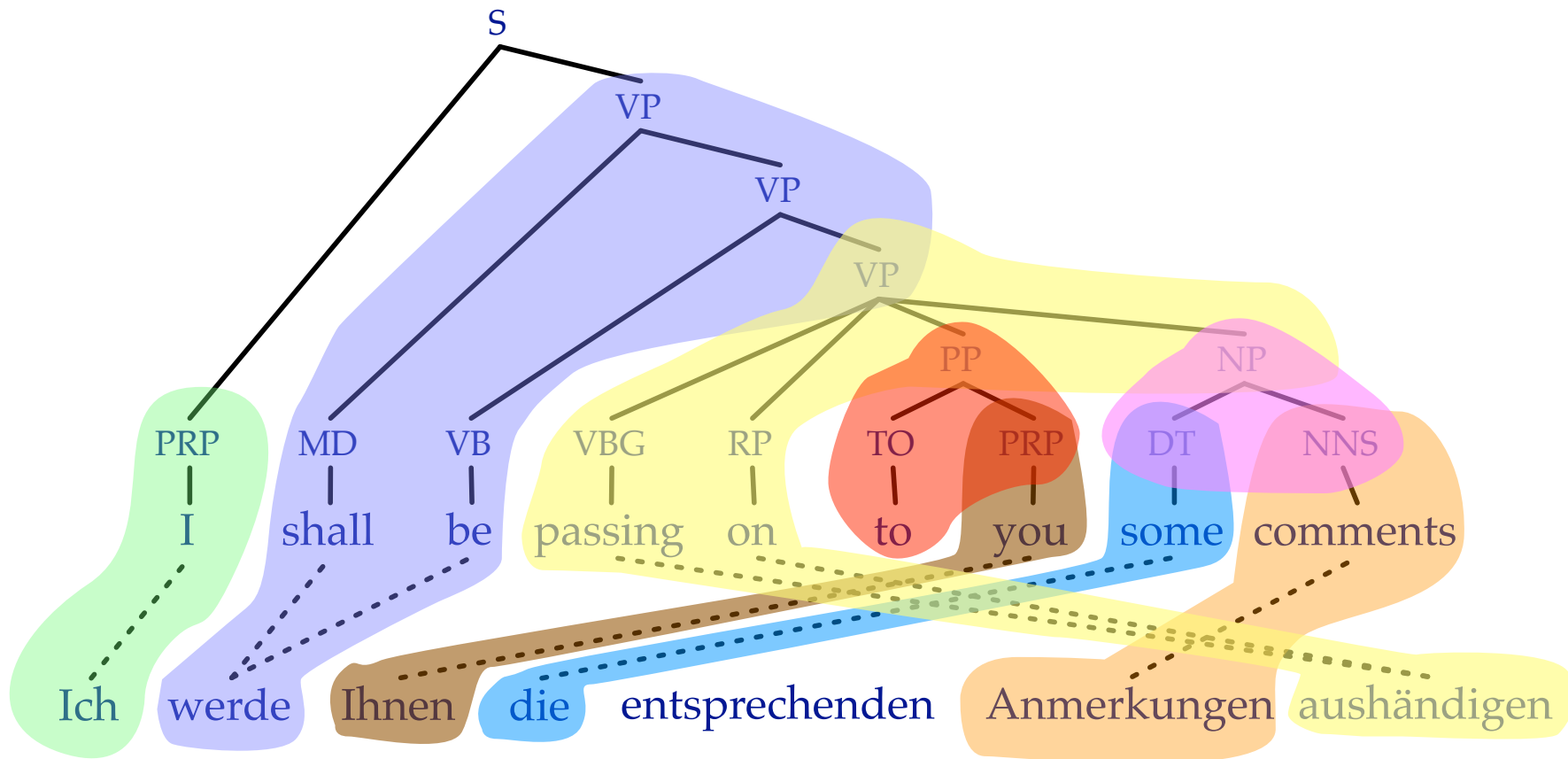
Extracted rule: $NP \rightarrow X_1 X_2 \mid DT_1 NNS_2$

Lexical Rule with Syntactic Context



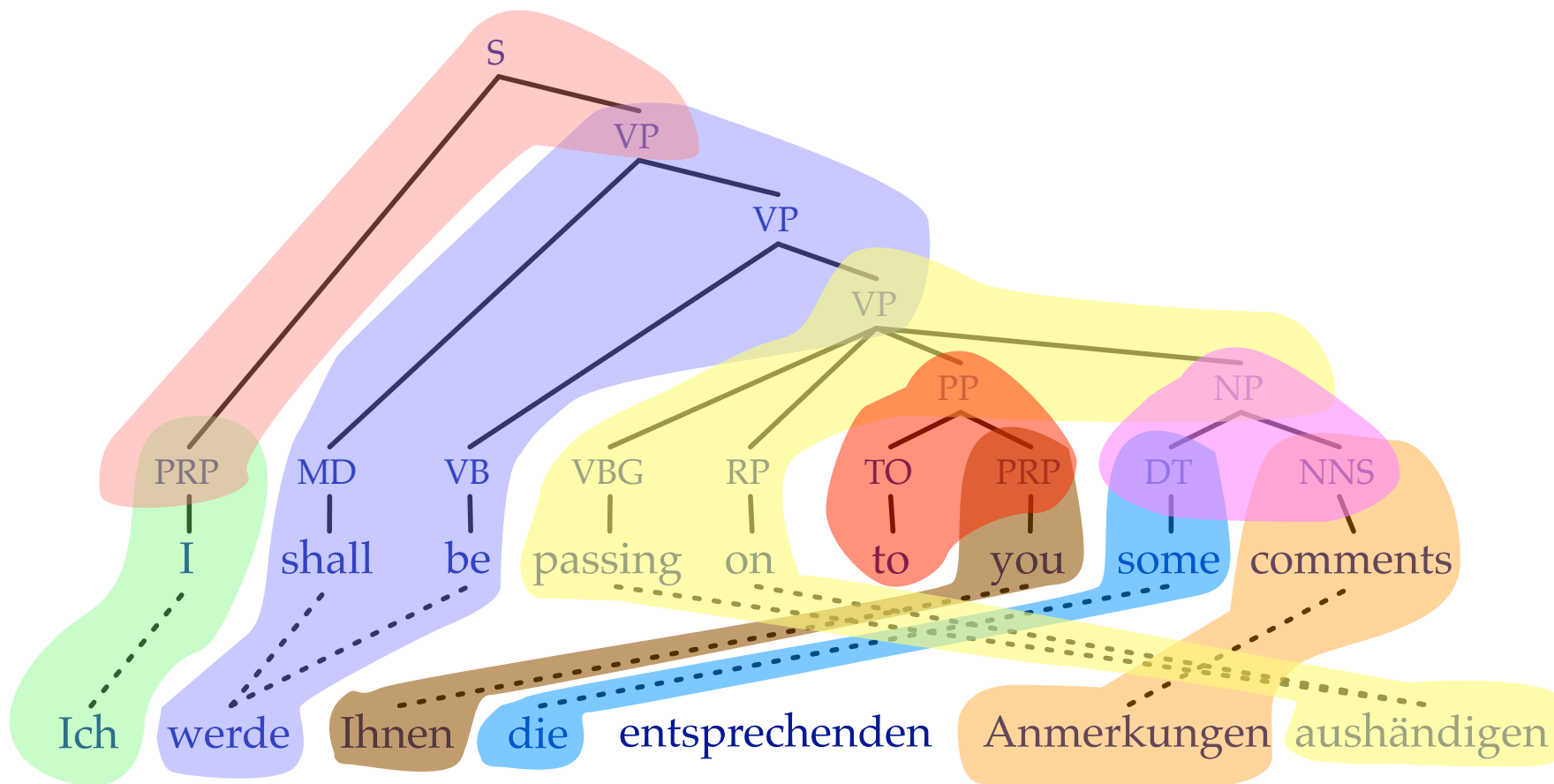
Extracted rule: $VP \rightarrow X_1 X_2 \text{ aushändigen} \mid \text{passing on } PP_1 NP_2$

Lexical Rule with Syntactic Context



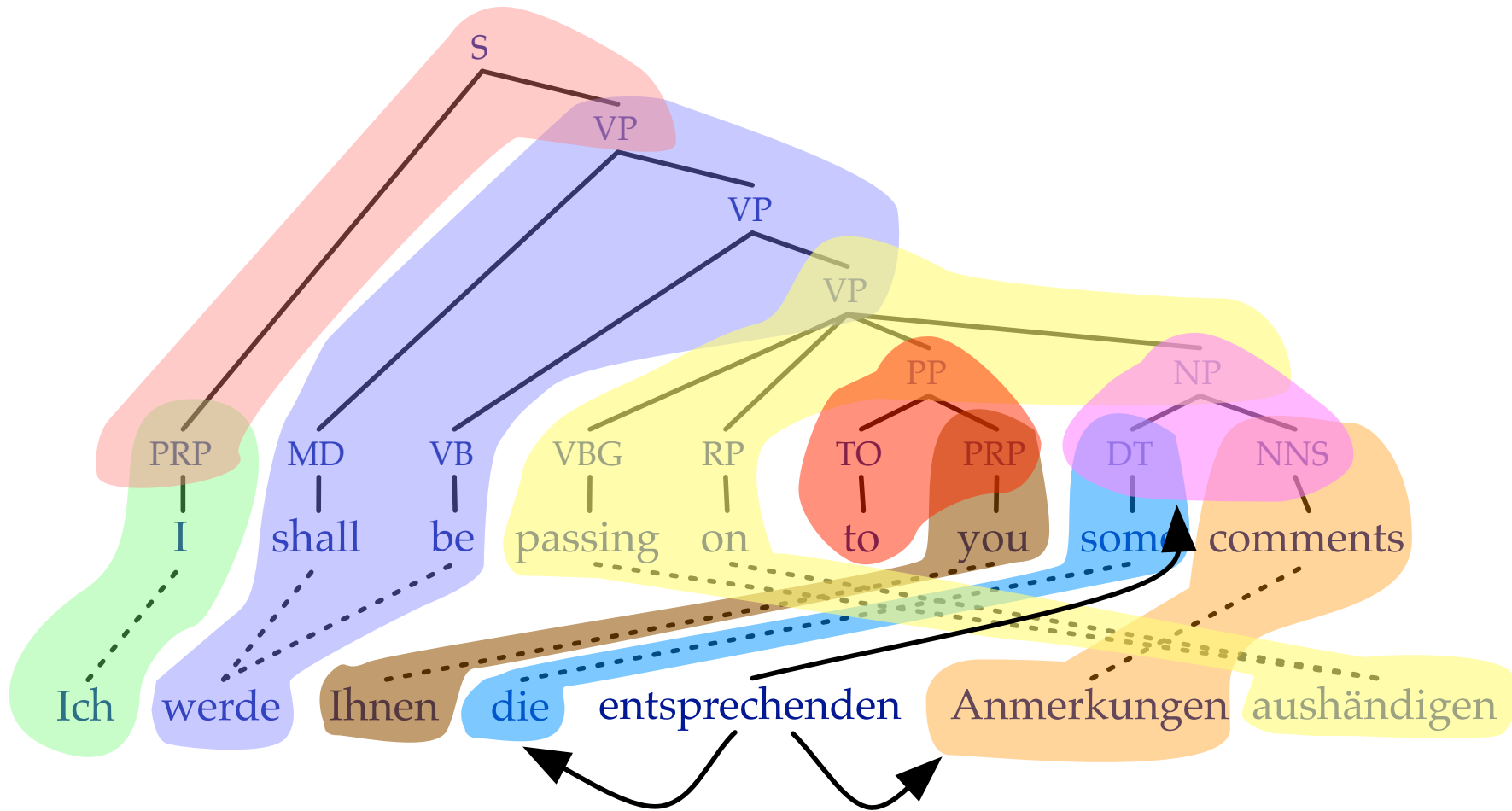
Extracted rule: $VP \rightarrow \text{werde } X \mid \text{shall be } VP$ (ignoring internal structure)

Non-Lexical Rule



Extracted rule: $S \rightarrow X_1 X_2 \mid PRP_1 VP_2$
 DONE — note: one rule per alignable constituent

Unaligned Source Words

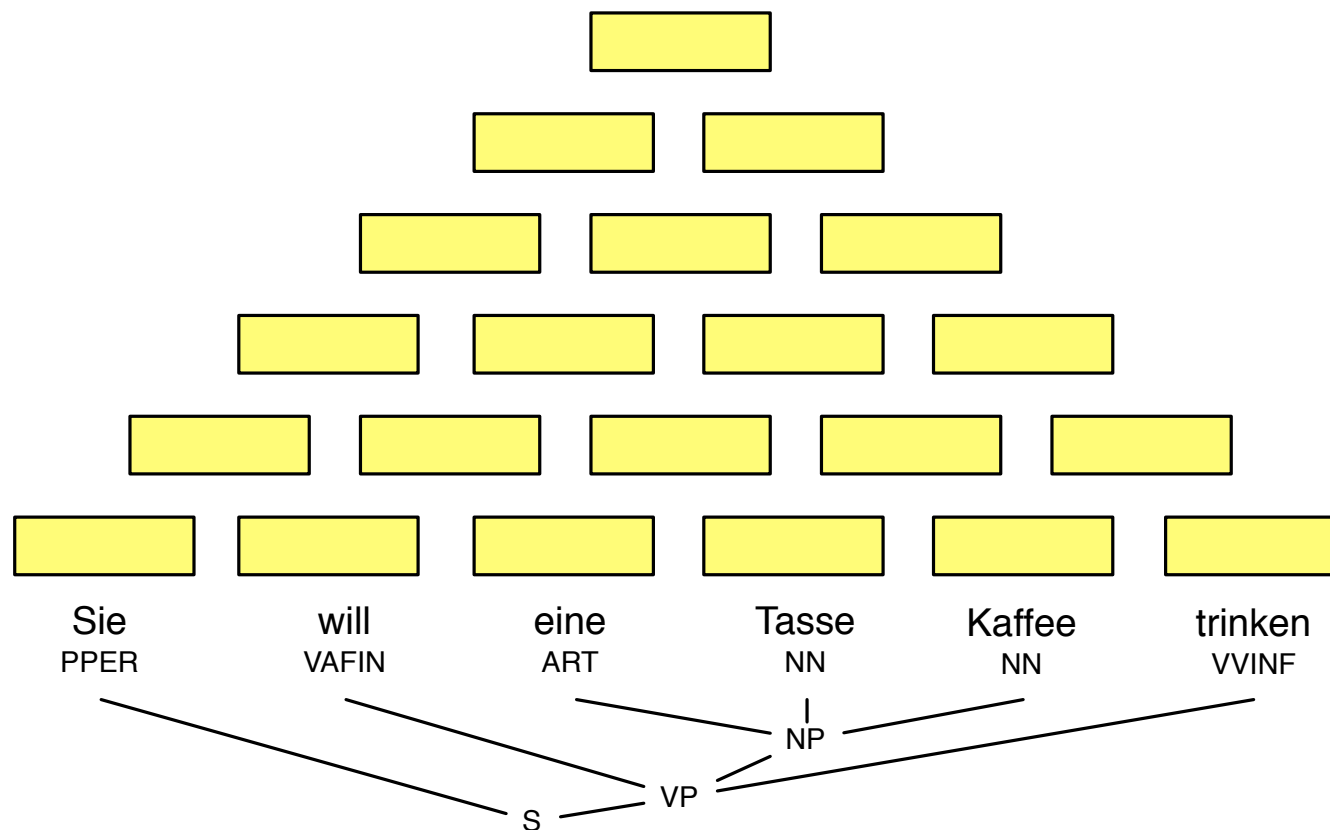


Attach to neighboring words or higher nodes → additional rules

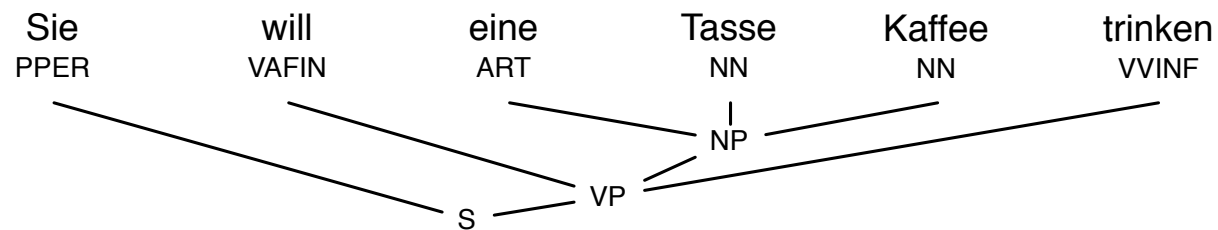
Syntactic Decoding

Inspired by monolingual syntactic chart parsing:

During decoding of the source sentence,
a chart with translations for the $O(n^2)$ spans has to be filled

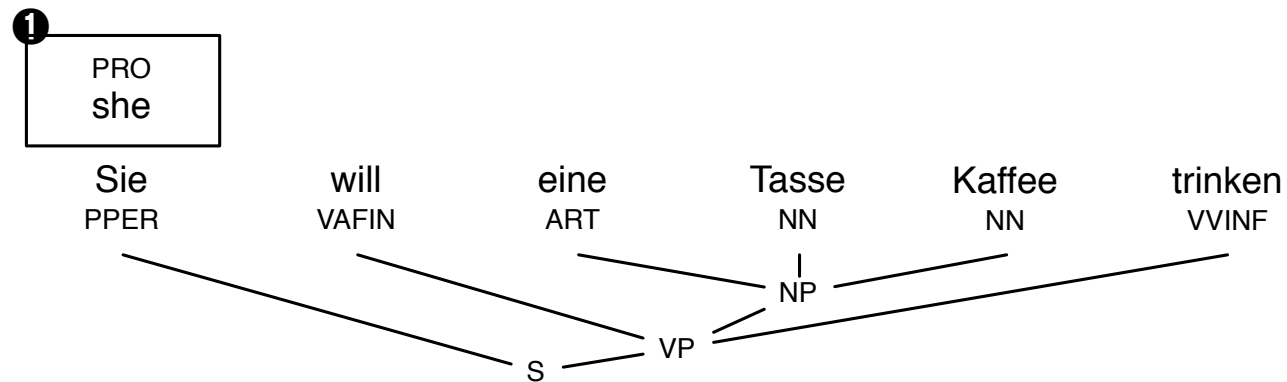


Syntax Decoding



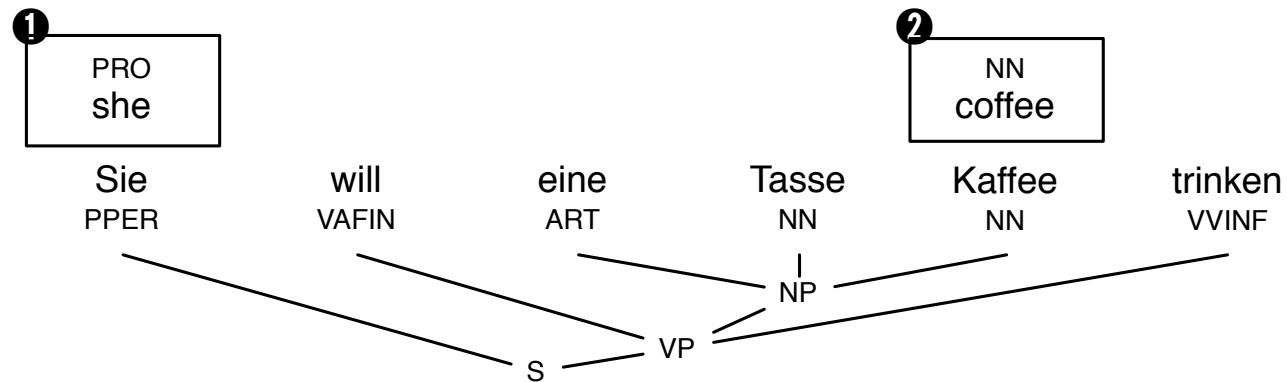
German input sentence with tree

Syntax Decoding



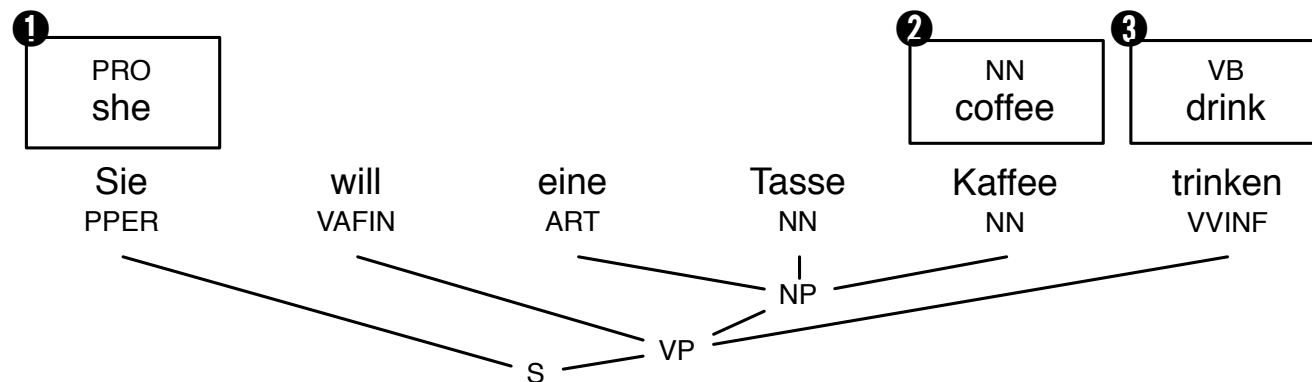
Purely lexical rule: filling a span with a translation (a constituent in the chart)

Syntax Decoding



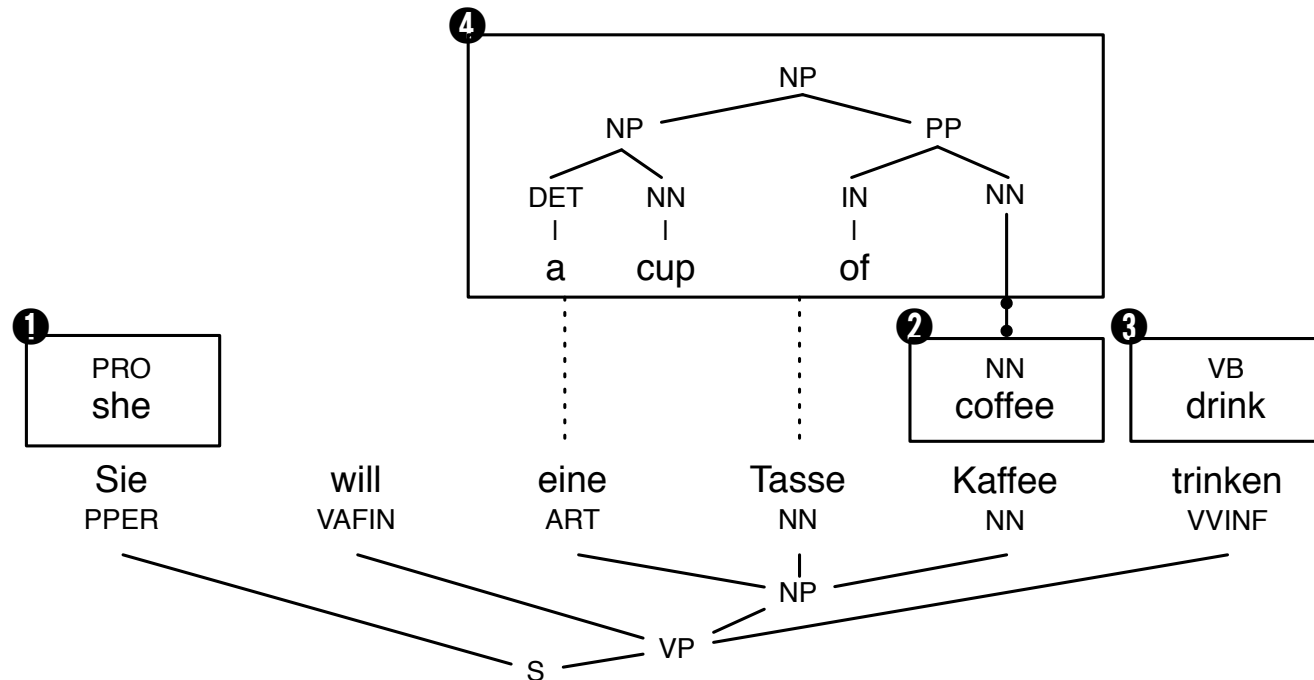
Purely lexical rule: filling a span with a translation (a constituent in the chart)

Syntax Decoding



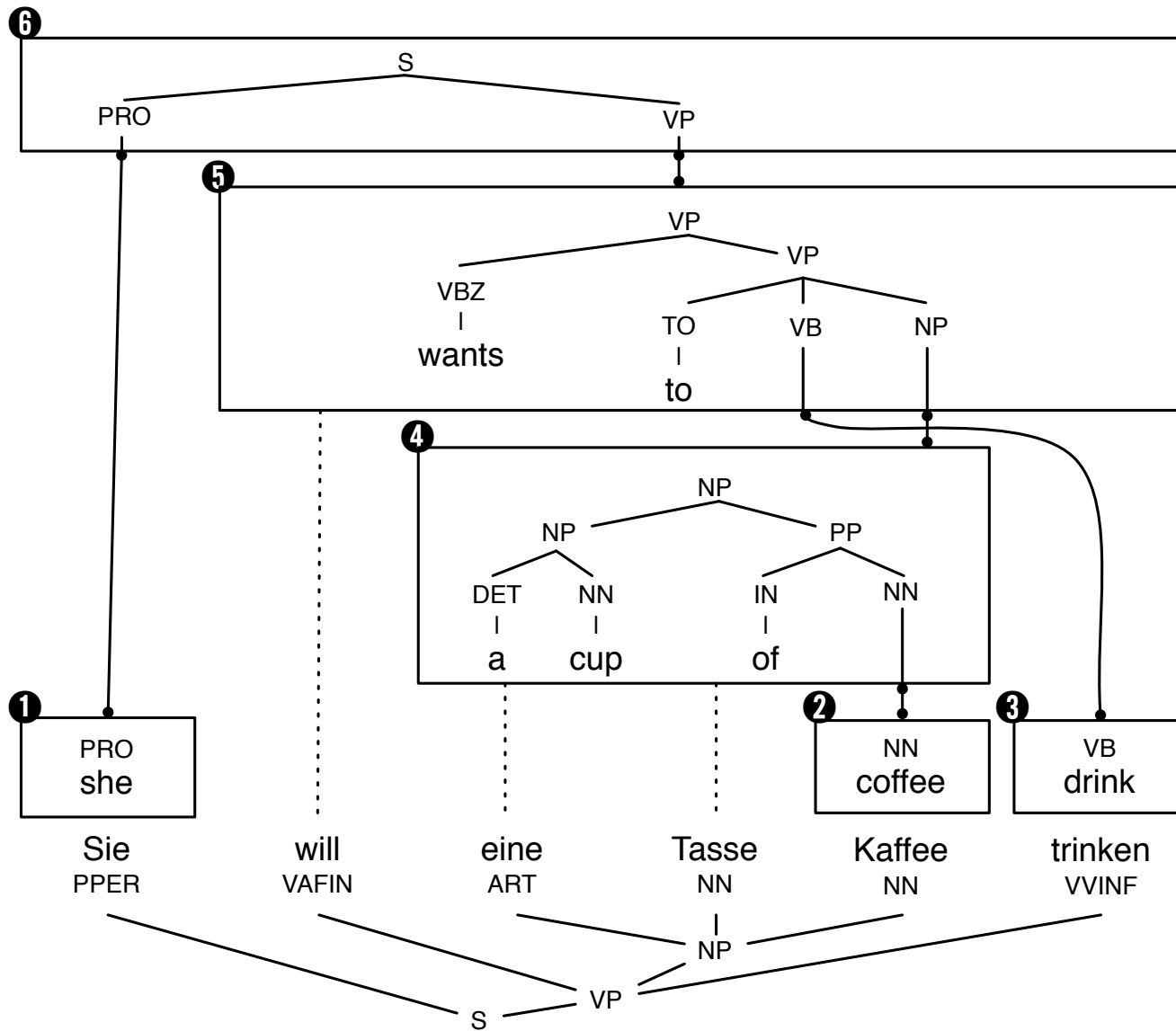
Purely lexical rule: filling a span with a translation (a constituent in the chart)

Syntax Decoding



Complex rule: matching underlying constituent spans, and covering words

Syntax Decoding



there is more...

Major Challenges

- Linguistically motivated models
- Machine learning
(esp. neural network models)
- Beyond sentence level
(pronouns, discourse relationships, inference)
- Evaluation

questions?