COS 484

Natural Language Processing

L8: Constituency Parsing

Spring 2023

Feature templates and sparse features

Tags	DT	NN	VB	DT	NN
Words	The	old	man	the	boat
	w_{i-1}	w_i	w_{i+1}	w_{i+2}	w_{i+3}

Feature template = abstract specification of features

Feature = what we actually use in classifier (can be very sparse!)

Which of these feature templates would help most to tag 'old' correctly?

A) $\langle t_i, t_{i-1}, w_i, w_{i-1}, w_{i+1} \rangle$

B) $\langle t_i, t_{i-1}, w_i, w_{i-1} \rangle$

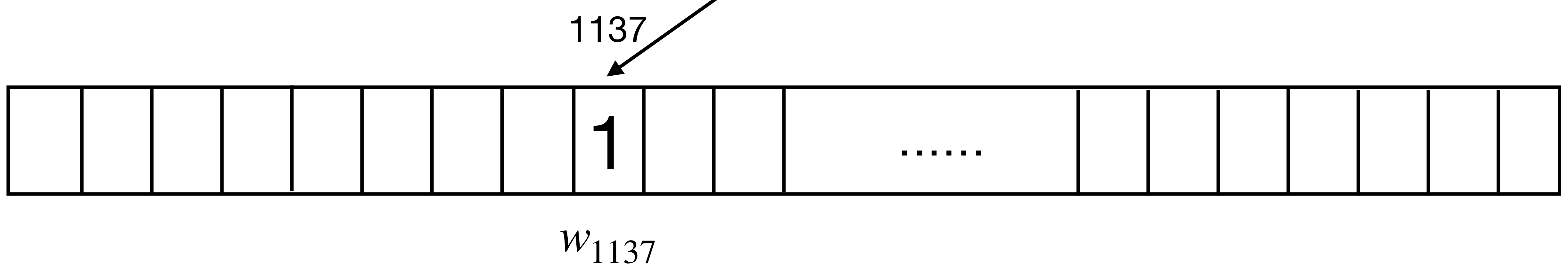
C) $\langle t_i, w_i, w_{i-1}, w_{i+1} \rangle$

D) $\langle t_i, w_i, w_{i-1}, w_{i+1}, w_{i+2} \rangle$

Feature template = $\langle t_i, t_{i-1}, w_i, w_{i-1} \rangle$

Feature =

$$\mathbb{I}(t_i = \mathrm{NN} \wedge t_{i-1} = \mathrm{DT} \wedge w_i = \mathrm{old} \wedge w_{i-1} = The)$$



We can only count the # of features that appear in the training set (sometimes keeping them when frequency $\geq$ threshold)

Feature templates and sparse features

- Bigram MEMM: $P(s_i = s \mid s_{i-1}, O) = \dfrac{\exp(\boxed{\mathbf{w}} \cdot \mathbf{f}(\boxed{s_i = s}, s_{i-1}, O, i))}{\sum_{s'=1}^{K} \exp(\mathbf{w} \cdot \mathbf{f}(s_i = s', s_{i-1}, O, i))}$

- Multinominal logistic regression: $P(y = c \,|\, x) = \dfrac{e^{\boxed{\mathbf{w}_c} \cdot \mathbf{x} + b_c}}{\sum_{j=1}^{m} e^{\mathbf{w}_j \cdot \mathbf{x} + b_j}}$

Feature 8 = $\text{bigram(American breakfast)}$

Weight vector for class 'positive': $w_{\text{pos},8}$

Equivalent as: **Feature 137** = $\text{bigram(American breakfast)} \wedge y = \text{positive}$

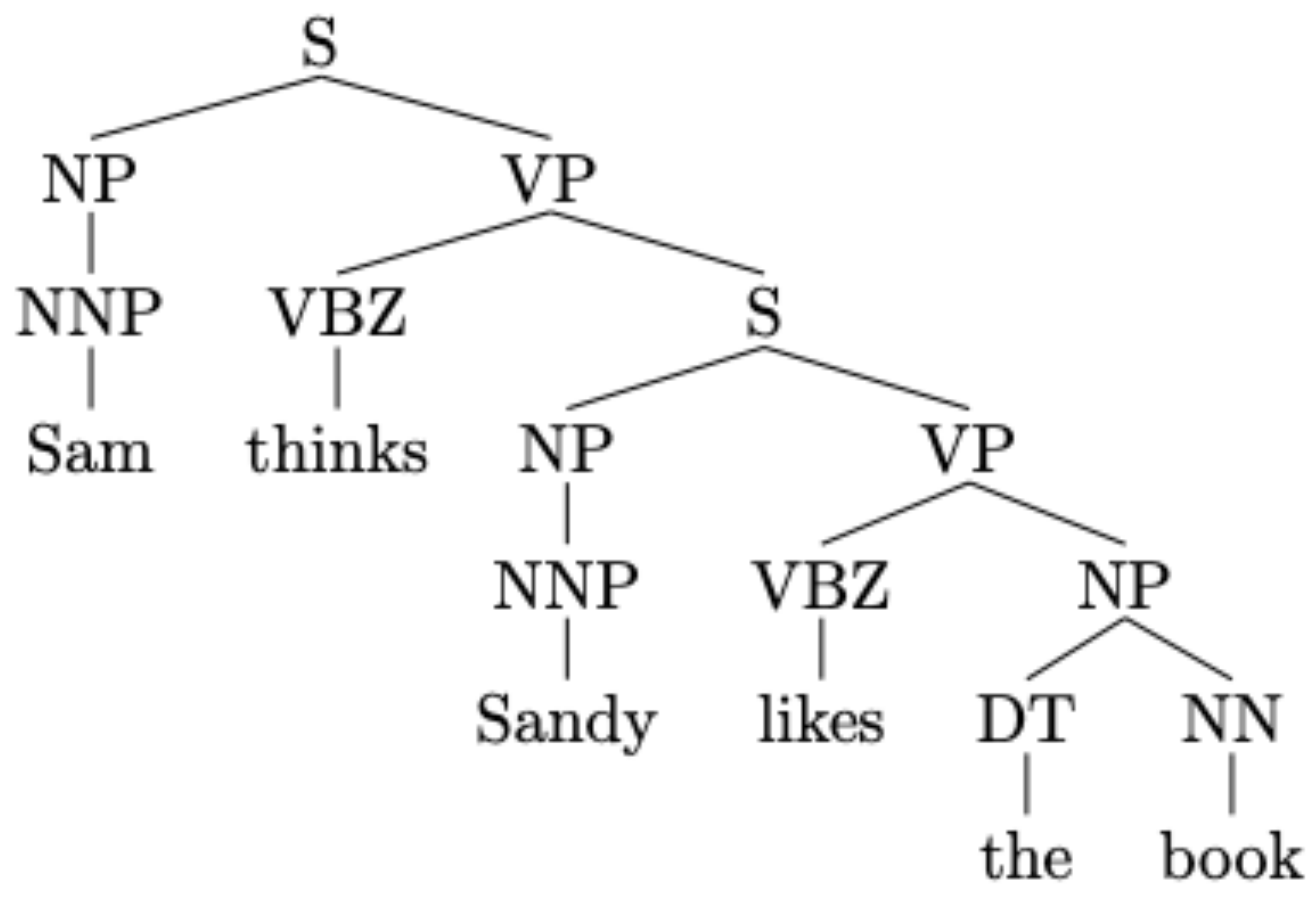
Weight vector: w_{137}

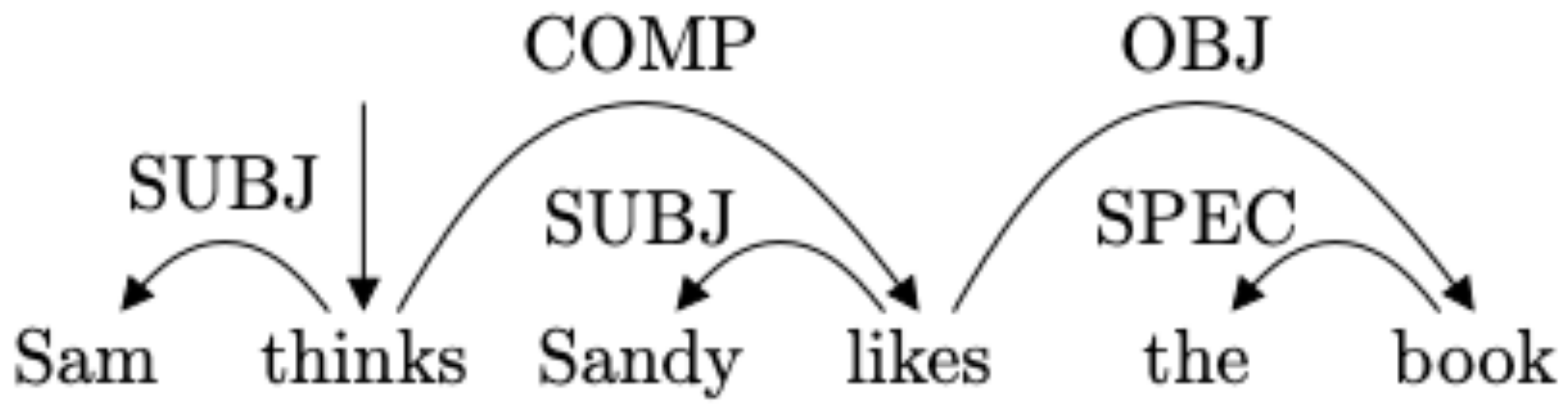
Syntactic structure: constituency and dependency

Theme: How do we represent the structure of **sentences** using (syntax) **trees**?

Two views of linguistic structure

- **Constituency** (today)
  - = phrase structure grammar
  - based on context-free grammars (CFGs)
- **Dependency** (next class)





Constituency structure

- **Phrase structure** organizes words into **nested constituents**
- Starting units: words are given a category: part-of-speech tags

  the, cuddly, cat, by, the, door

  DT, JJ, NN, IN, DT, NN

- Words combine into phrases with categories

  the cuddly cat, by the door

  NP→DT JJ NN PP → IN DT NN

- Phrases can combine into bigger phrases recursively

  the cuddly cat by the door

  NP → NP PP

NP: noun phrase, PP: prepositional phrase

Syntactic parsing

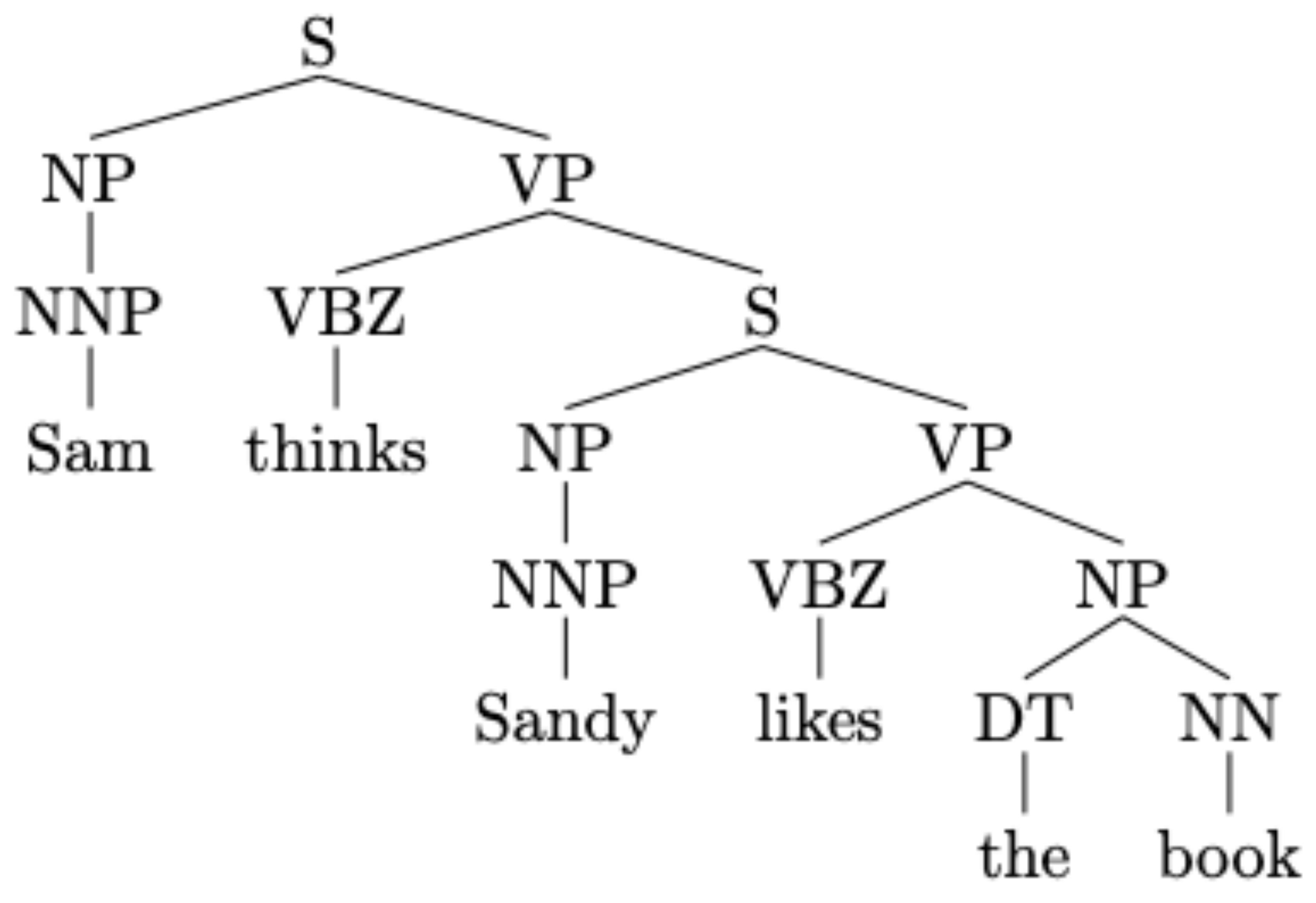
Syntactic parsing is the task of recognizing a sentence and assigning a structure to it.

Constituency parsing is the task of recognizing a sentence and assigning a **constituency** structure to it.

Input

Sam thinks Sandy likes the book

Output



Syntactic parsing: applications

- Grammar checking
  - If a sentence can't be parsed, it may have grammatical errors (or at least hard to read)
- Used as intermediate representations for downstream tasks
  - Machine translation (syntax-based statistical MT)
  - Information extraction
  - Question answering

Syntactic parsing: applications

Used as intermediate representation for downstream applications

English word order: subject — verb — object

Japanese word order: subject — object — verb

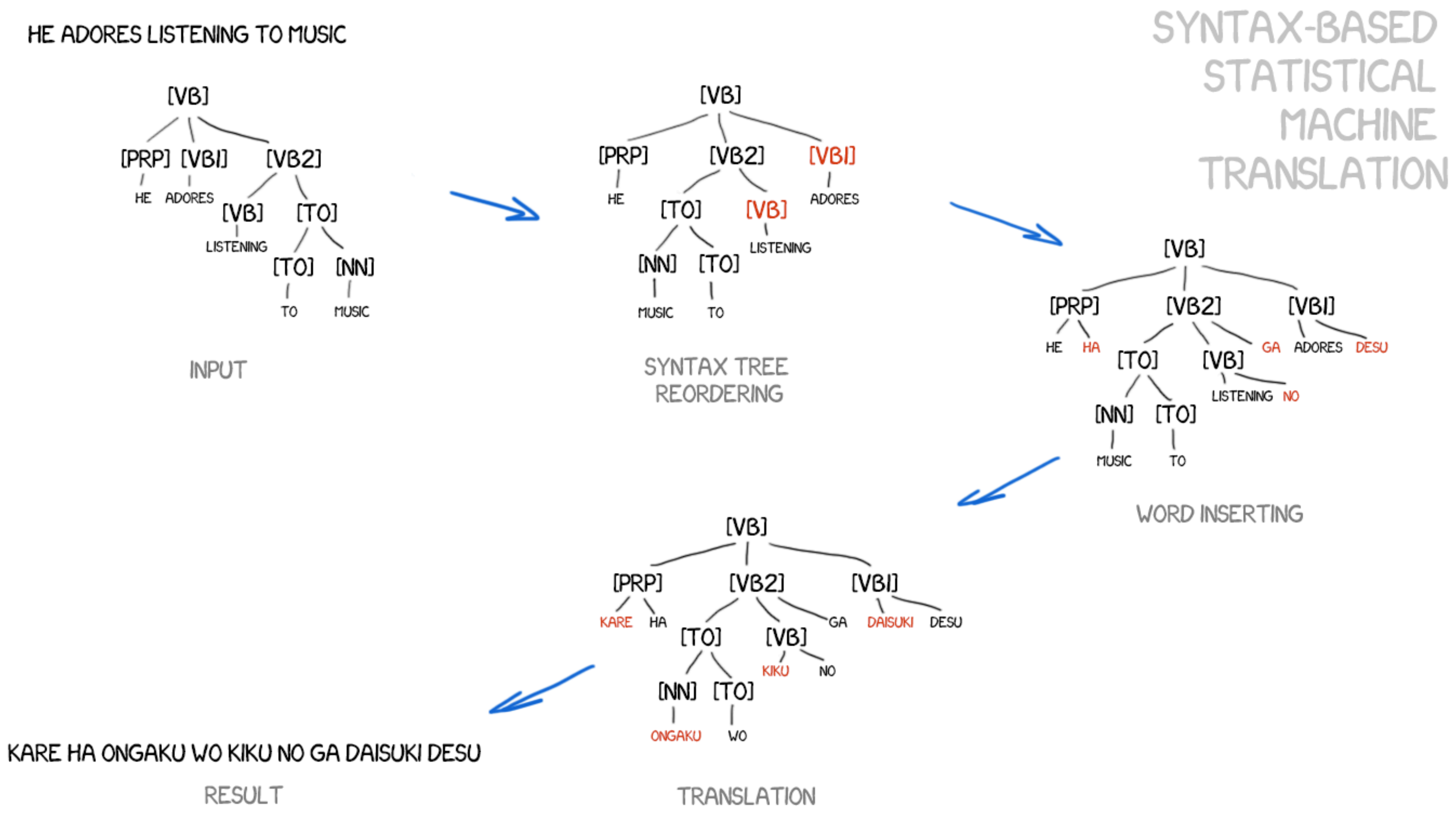
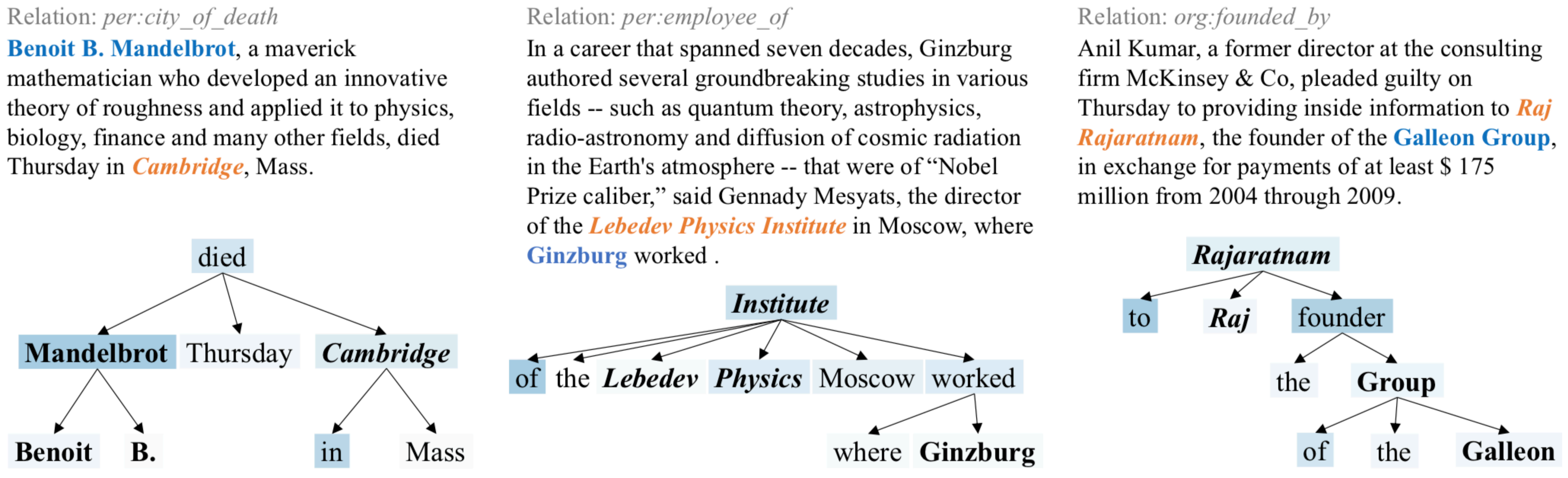


Image credit: http://vas3k.com/blog/machine_translation/

Syntactic parsing: applications

Used as intermediate representation for downstream applications

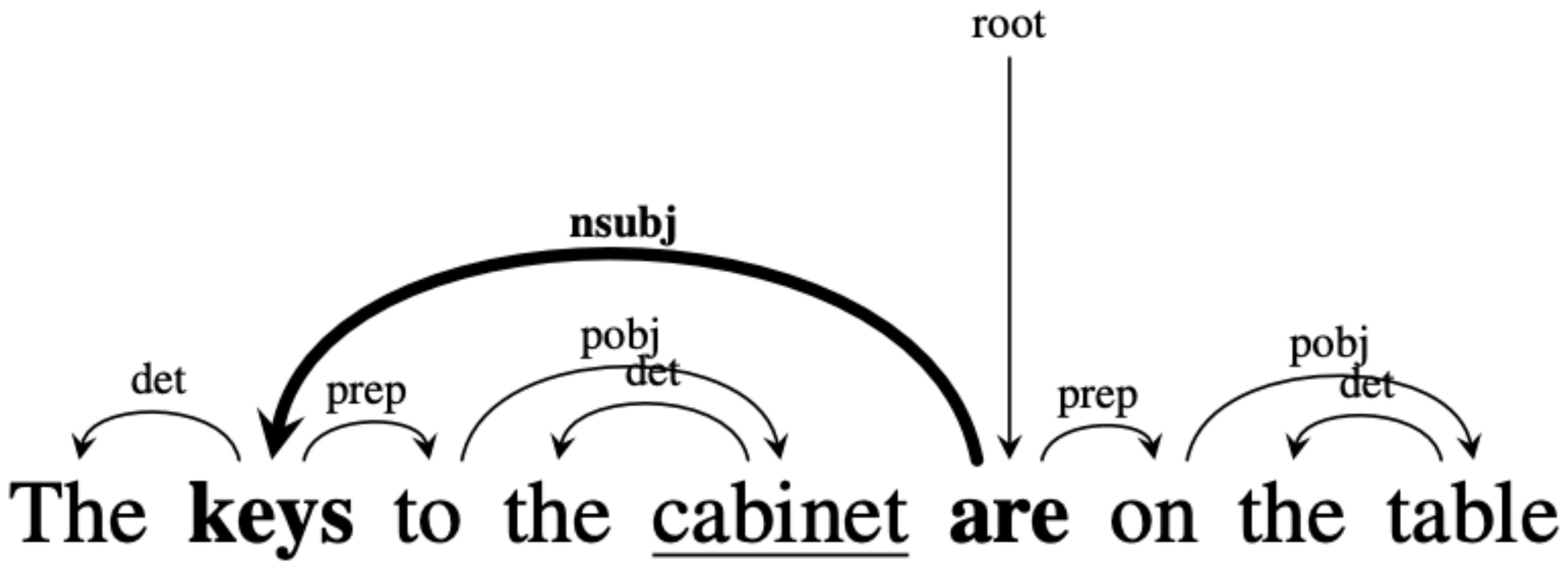


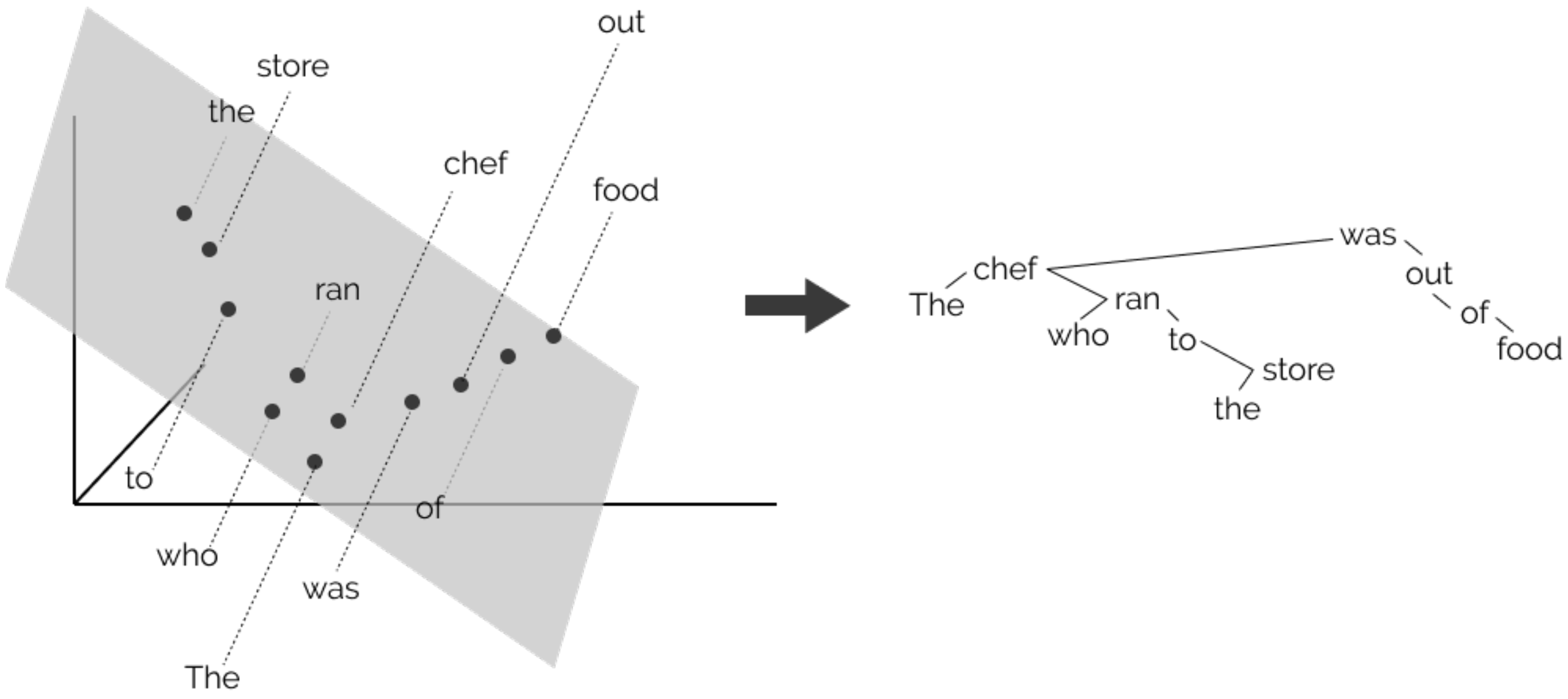
(Note: these are dependency parses)

Image credit: (Zhang et al, 2018)

Tree structures in the deep learning era

The keys to the cabinet <u>is/are</u> on the table.





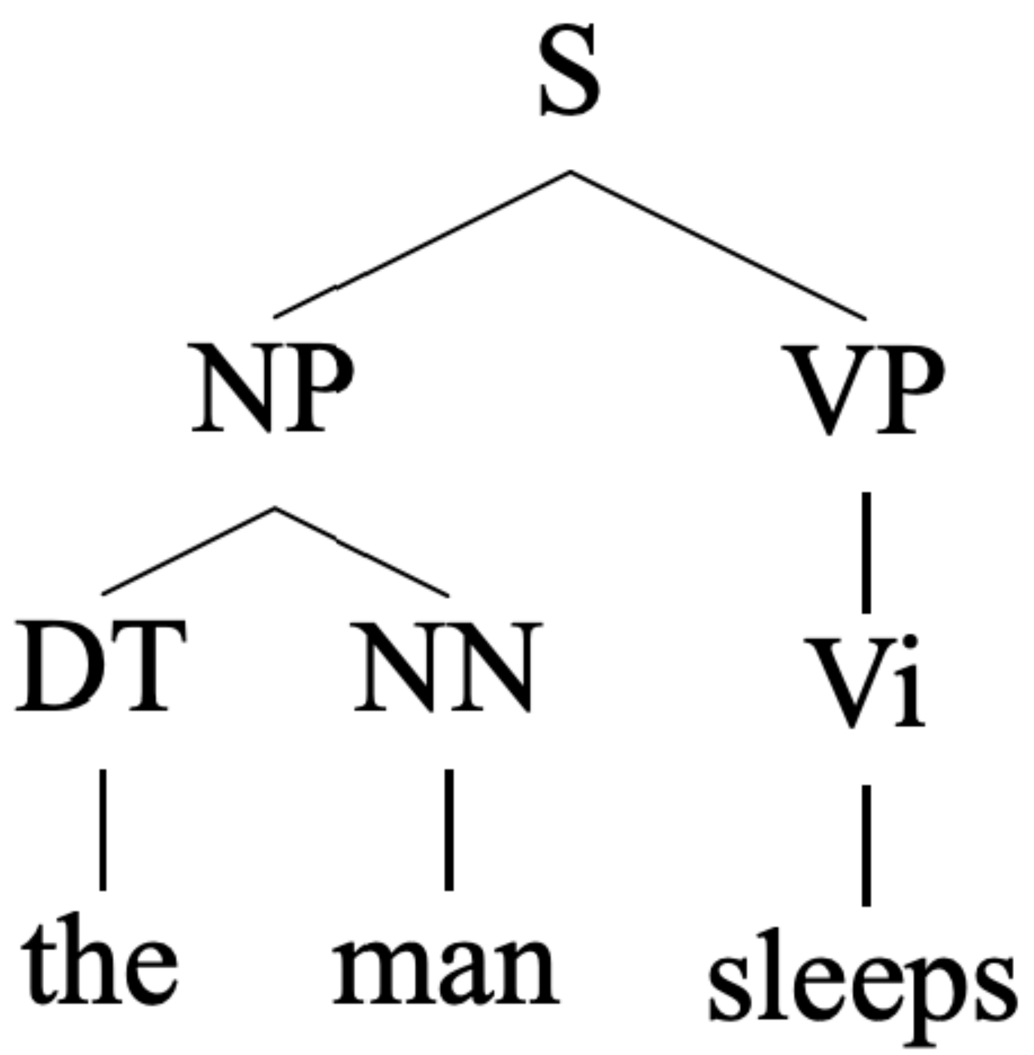
(Linzen et al., 2016): Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies

(Hewitt and Manning, 2019): A Structural Probe for Finding Syntax in Word Representations

Context-free grammars (CFGs)

- The most widely used formal system for modeling constituency structure in English and other natural languages
- A context free grammar $G = (N, \Sigma, R, S)$ where
  - N is a set of non-terminal symbols
    - Phrasal categories: S, NP, VP, …
    - Parts-of-speech (pre-terminals): DT, NN, Vi, …
  - Σ is a set of terminal symbols: the, man, sleeps, ..
  - R is a set of rules of the form $X \to Y_1 Y_2 \ldots Y_n$ for $n \geq 1$, $X \in N, Y_i \in (N \cup \Sigma)$
    - Examples: S → NP VP, NP → DT NN, NN → man
  - $S \in N$ is a distinguished start symbol

Not always the sentence non-terminal S



S:sentence, VP:verb phrase, NP: noun phrase, DT:determiner, NN: noun, Vi: intransitive verb…

A context-free grammar for English

$$N = \{\underline{\text{S, NP, VP, PP,}}\ \underline{\text{DT, Vi, Vt, NN, IN}}\}$$
$$S = \text{S}$$
$$\Sigma = \{\text{sleeps, saw, man, woman, telescope, the, with, in}\}$$

$R =$

S	$\rightarrow$	NP	VP
VP	$\rightarrow$	Vi	
VP	$\rightarrow$	Vt	NP
VP	$\rightarrow$	VP	PP
NP	$\rightarrow$	DT	NN
NP	$\rightarrow$	NP	PP
PP	$\rightarrow$	IN	NP

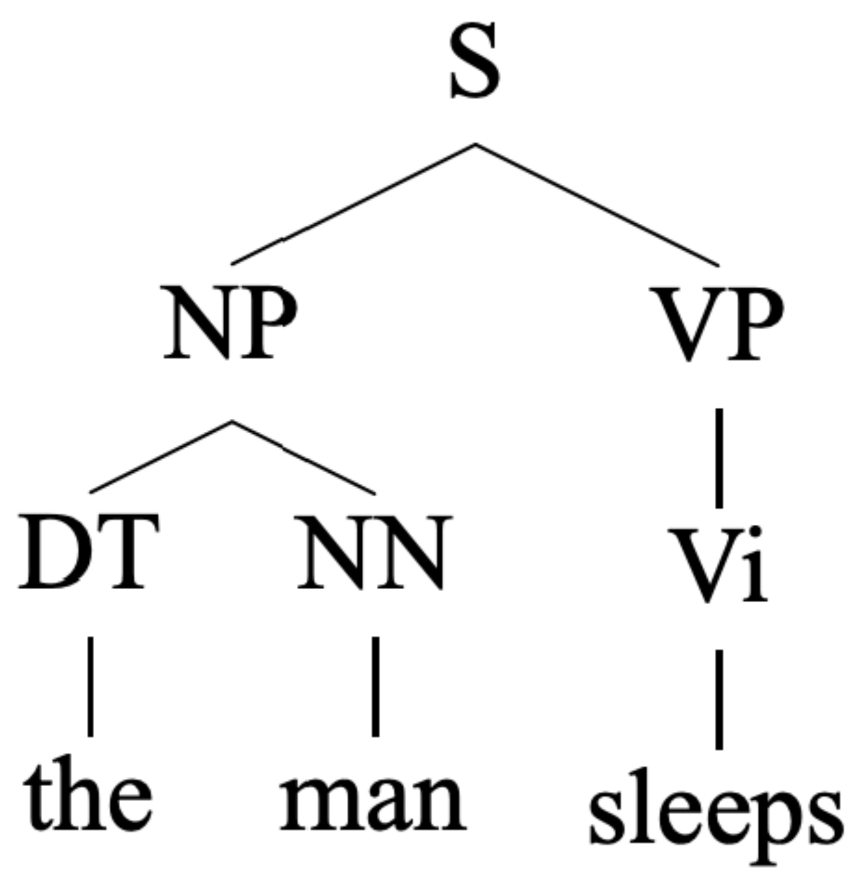
Grammar

Vi	$\rightarrow$	sleeps
Vt	$\rightarrow$	saw
NN	$\rightarrow$	man
NN	$\rightarrow$	woman
NN	$\rightarrow$	telescope
NN	$\rightarrow$	dog
DT	$\rightarrow$	the
IN	$\rightarrow$	with
IN	$\rightarrow$	in

Lexicon

S:sentence, VP:verb phrase, NP: noun phrase, PP:prepositional phrase,
DT:determiner, Vi:intransitive verb, Vt:transitive verb, NN: noun, IN:preposition

(Left-most) Derivations

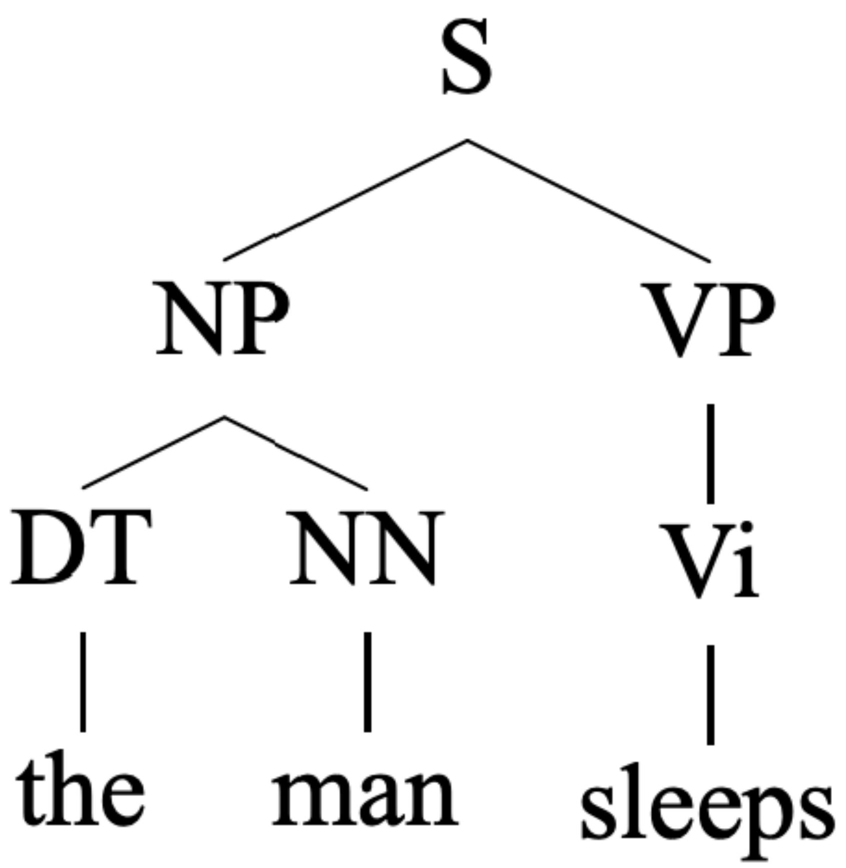


- A string “the man sleeps” can be **derived** from S
- **Derivation** = the sequence of rule expansions

- Given a CFG G, a left-most derivation is a sequence of strings $s_1, s_2, \ldots, s_n$, where
  - $s_1 = S$
  - $s_n \in \Sigma^*$: all possible strings made up of words from Σ
  - Each s_i for $i = 2, \ldots, n$ is derived from s_{i-1} by picking the left-most non-terminal X in s_{i-1} and replacing it by some β where $X \to \beta \in R$
- s_n: yield of the derivation

(Left-most) Derivations

- $s_1 = \text{S}$
- $s_2 = \text{NP VP}$
- $s_3 = \text{DT NN VP}$
- $s_4 = \text{the NN VP}$
- $s_5 = \text{the man VP}$
- $s_6 = \text{the man Vi}$
- $s_7 = \text{the man sleeps}$



A derivation can be represented as a parse tree!

$R =$

S	$\rightarrow$	NP	VP
VP	$\rightarrow$	Vi	
VP	$\rightarrow$	Vt	NP
VP	$\rightarrow$	VP	PP
NP	$\rightarrow$	DT	NN
NP	$\rightarrow$	NP	PP
PP	$\rightarrow$	IN	NP

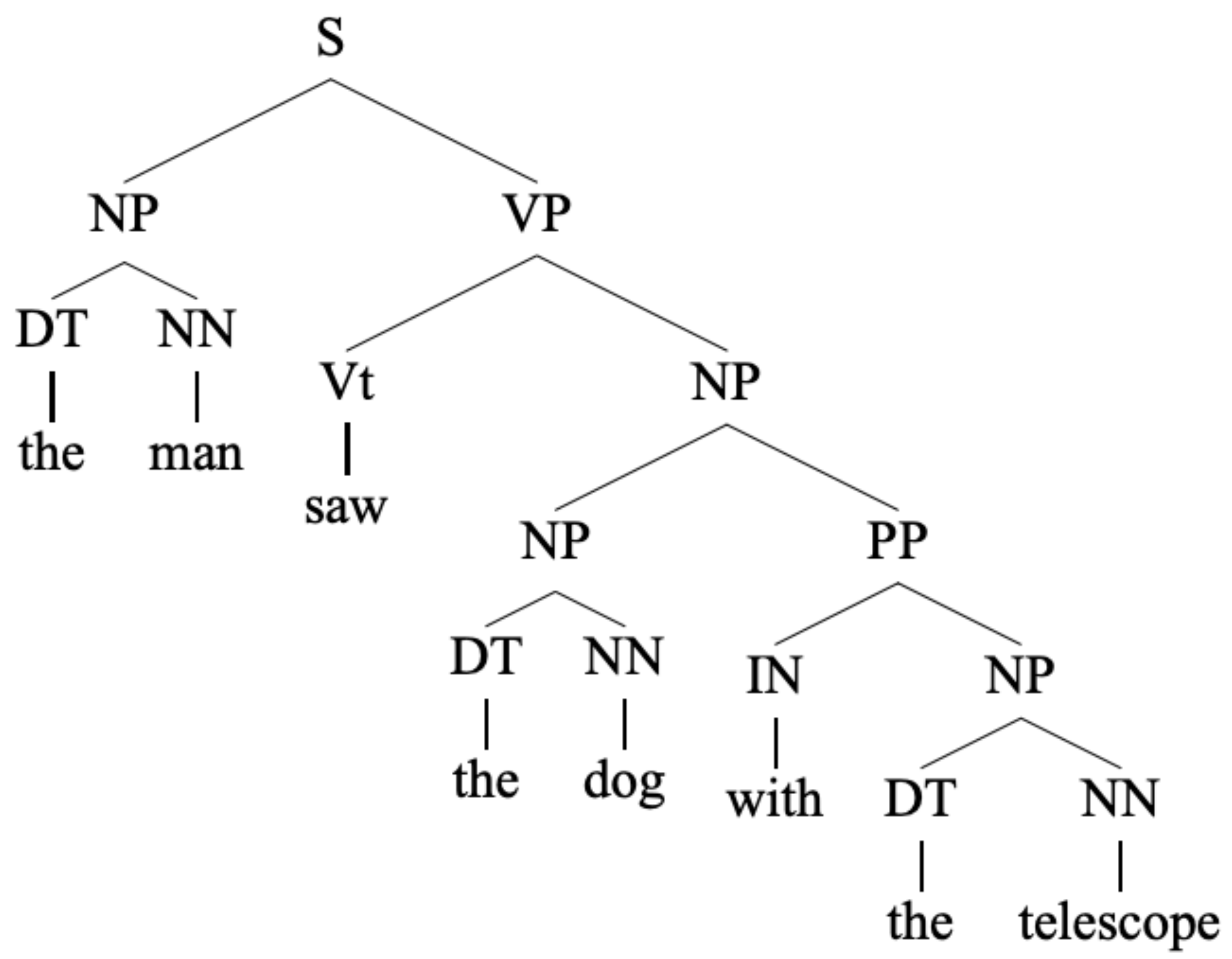
Vi	$\rightarrow$	sleeps
Vt	$\rightarrow$	saw
NN	$\rightarrow$	man
NN	$\rightarrow$	woman
NN	$\rightarrow$	telescope
NN	$\rightarrow$	dog
DT	$\rightarrow$	the
IN	$\rightarrow$	with
IN	$\rightarrow$	in

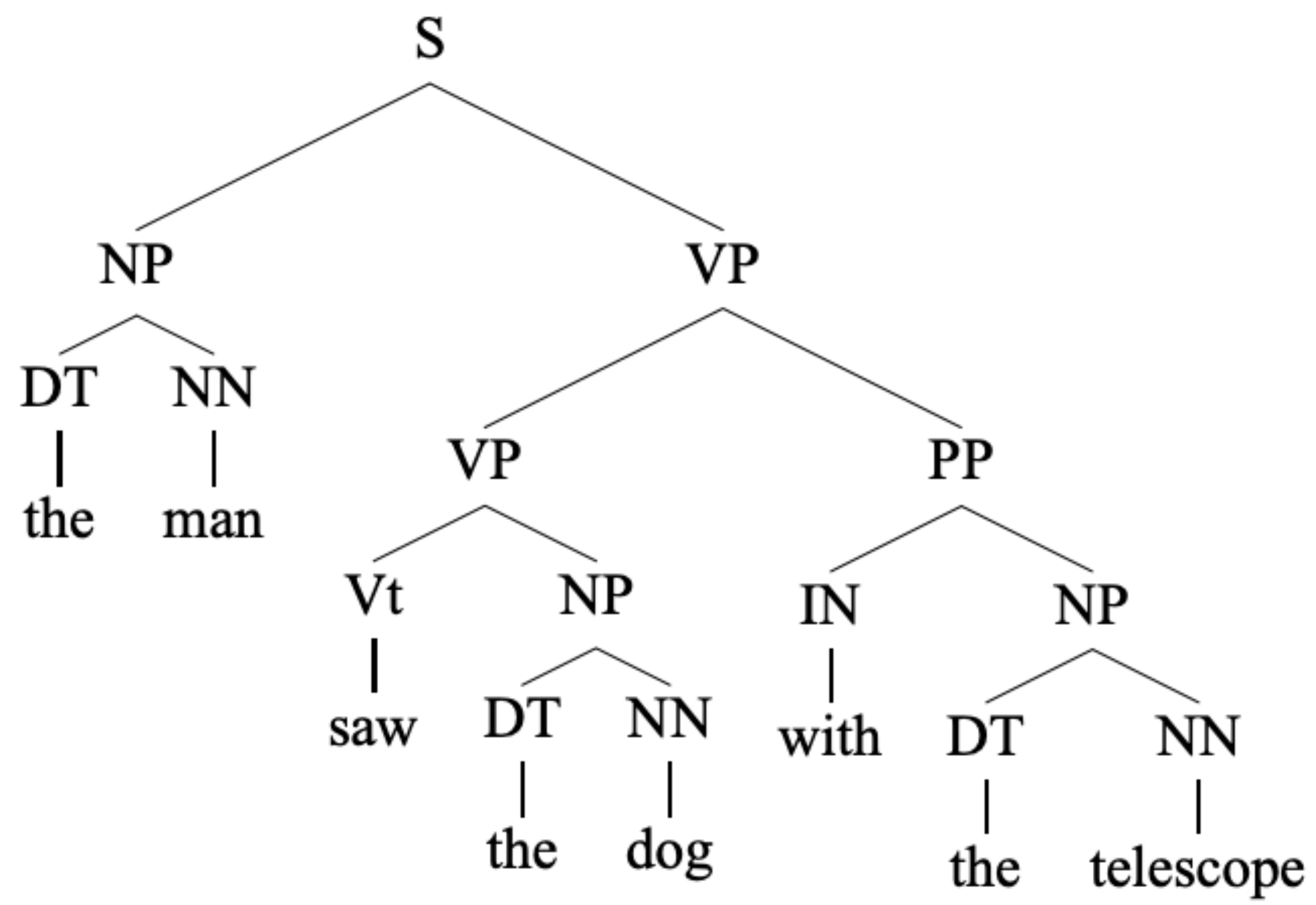
The set of possible derivations may be finite or infinite

Ambiguity

Some **sentences/phrases** may have more than one derivation (i.e. more than one parse tree!).

Attachment ambiguity (e.g., PP attachment)





Which one is the correct parse?

(a) Left (b) Right (c) both correct (d) both incorrect

The answer is (b).

Ambiguity

Some **sentences/phrases** may have more than one derivation (i.e. more than one parse tree!).

Coordination ambiguity

old men and women

old [men and women]   [old men] and women

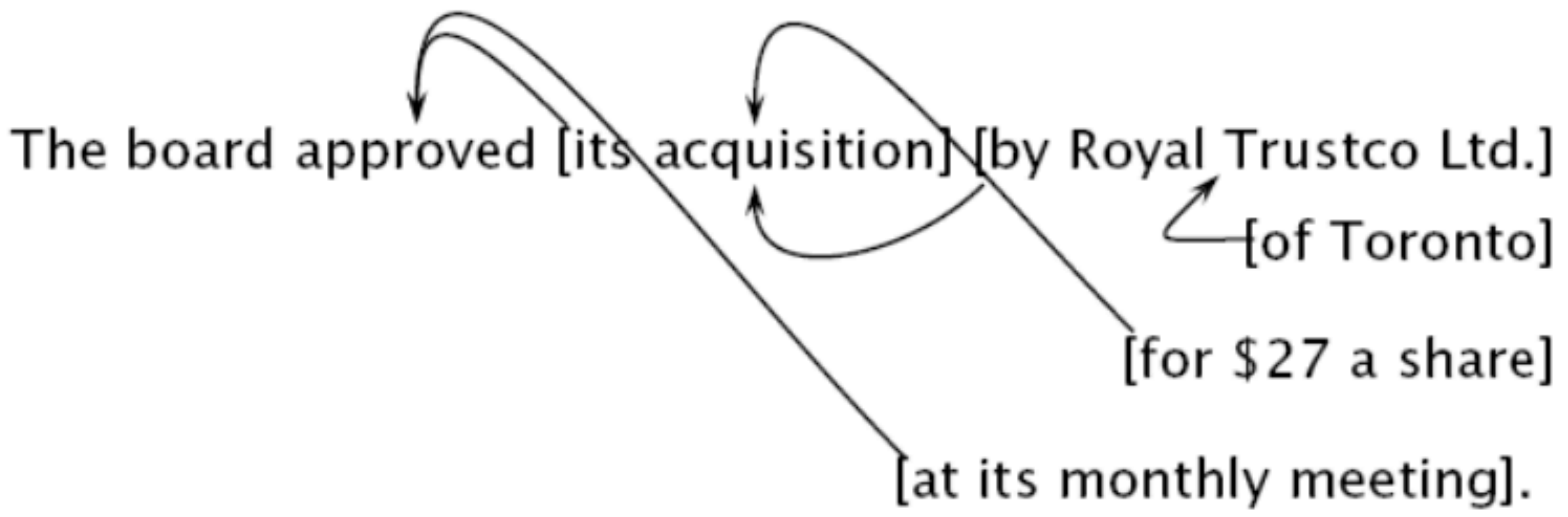
President Kennedy today pushed aside other White House business to devote all his time and attention to working on the Berlin crisis address he will deliver tomorrow night to the American people over nationwide television and radio.

Q: What ambiguities are there in this sentence?

Sentences can have a large number of parses

- In fact, sentences can have a very large number of possible parses

The board approved [its acquisition] [by Royal Trustco Ltd.] [of Toronto] [for $27 a share] [at its monthly meeting]



((ab)c)d (a(bc))d (ab)(cd) a((bc)d) a(b(cd))

Catalan number: $C_n = \frac{1}{n+1}\binom{2n}{n}$

1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786,…

- There is no way to choose the right parse!
- Constructing a grammar is difficult— a less constrained grammar can parse more sentences but result in more parses for even simple sentences

Probabilistic context-free grammars (PCFGs)

A probabilistic context-free grammar (PCFG) consists of:

- A context-free grammar: $G = (N, \Sigma, R, S)$
- For each rule $\alpha \to \beta \in R$, there is a parameter (probability) $q(\alpha \to \beta) \geq 0$. For any $X \in N$,

$$\sum_{\alpha \to \beta : \alpha = X} q(\alpha \to \beta) = 1$$

$R, q =$

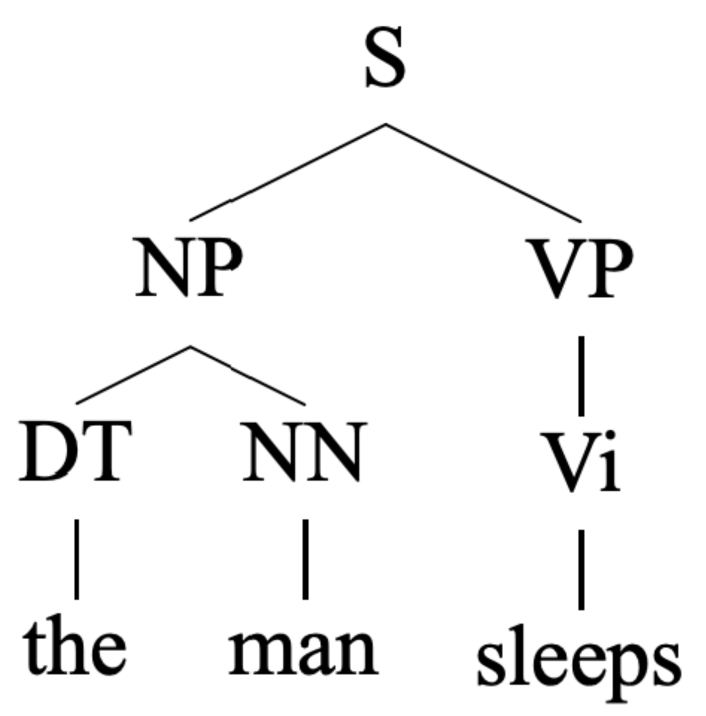
S	$\rightarrow$	NP	VP	1.0
VP	$\rightarrow$	Vi		0.3
VP	$\rightarrow$	Vt	NP	0.5
VP	$\rightarrow$	VP	PP	0.2
NP	$\rightarrow$	DT	NN	0.8
NP	$\rightarrow$	NP	PP	0.2
PP	$\rightarrow$	IN	NP	1.0

Vi	$\rightarrow$	sleeps	1.0
Vt	$\rightarrow$	saw	1.0
NN	$\rightarrow$	man	0.1
NN	$\rightarrow$	woman	0.1
NN	$\rightarrow$	telescope	0.3
NN	$\rightarrow$	dog	0.5
DT	$\rightarrow$	the	1.0
IN	$\rightarrow$	with	0.6
IN	$\rightarrow$	in	0.4

Probabilistic context-free grammars (PCFGs)

For any derivation (parse tree) containing rules:
$\alpha_1 \rightarrow \beta_1, \alpha_2 \rightarrow \beta_2, \ldots, \alpha_l \rightarrow \beta_l$, the probability of the parse is:

$$\prod_{i=1}^{l} q(\alpha_i \rightarrow \beta_i)$$



$R, q =$

S	$\rightarrow$	NP	VP	1.0
VP	$\rightarrow$	Vi		0.3
VP	$\rightarrow$	Vt	NP	0.5
VP	$\rightarrow$	VP	PP	0.2
NP	$\rightarrow$	DT	NN	0.8
NP	$\rightarrow$	NP	PP	0.2
PP	$\rightarrow$	IN	NP	1.0

Vi	$\rightarrow$	sleeps	1.0
Vt	$\rightarrow$	saw	1.0
NN	$\rightarrow$	man	0.1
NN	$\rightarrow$	woman	0.1
NN	$\rightarrow$	telescope	0.3
NN	$\rightarrow$	dog	0.5
DT	$\rightarrow$	the	1.0
IN	$\rightarrow$	with	0.6
IN	$\rightarrow$	in	0.4

$$P(t) = q(\text{S} \rightarrow \text{NP VP}) \times q(\text{NP} \rightarrow \text{DT NN}) \times q(\text{DT} \rightarrow \text{the})$$
$$\times\, q(\text{NN} \rightarrow \text{man}) \times q(\text{VP} \rightarrow \text{Vi}) \times q(\text{Vi} \rightarrow \text{sleeps})$$
$$= 1.0 \times 0.8 \times 1.0 \times 0.1 \times 0.3 \times 1.0 = 0.024$$

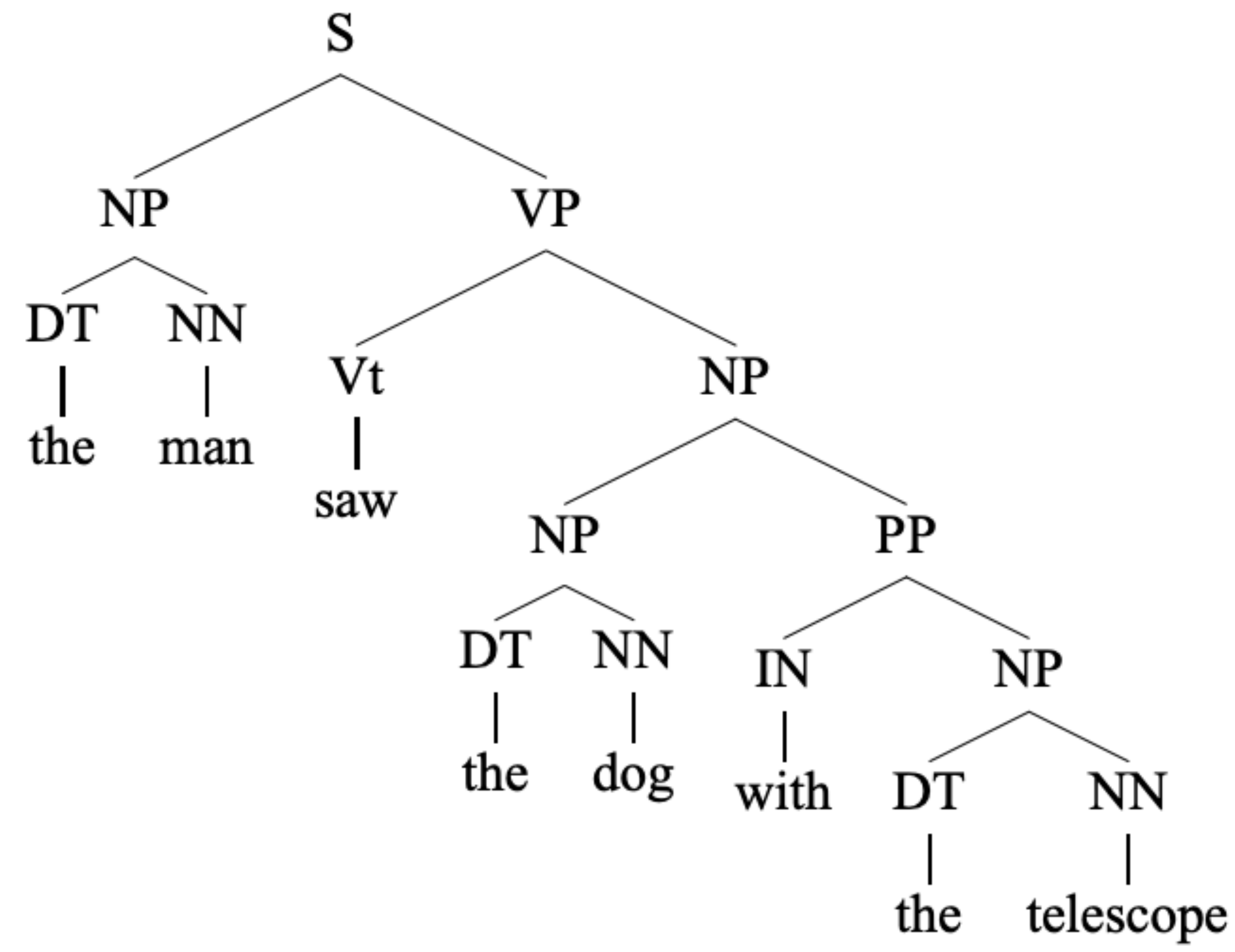
Q: Why do we want $\sum_{\alpha \rightarrow \beta : \alpha = X} q(\alpha \rightarrow \beta) = 1$?

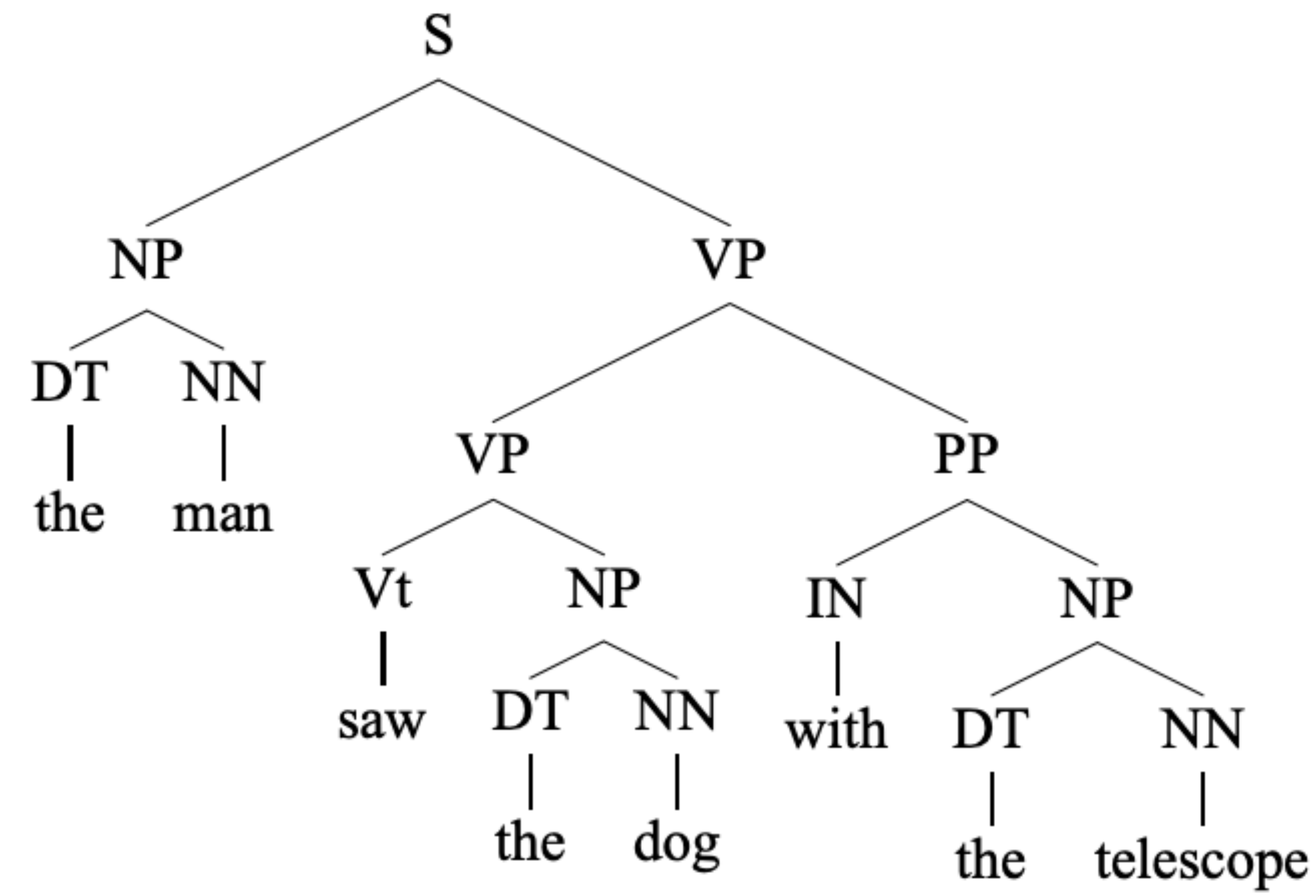
Which parse tree has a higher probability?

$R, q =$

S	$\rightarrow$	NP	VP	1.0
VP	$\rightarrow$	Vi		0.3
VP	$\rightarrow$	Vt	NP	0.5
VP	$\rightarrow$	VP	PP	0.2
NP	$\rightarrow$	DT	NN	0.8
NP	$\rightarrow$	NP	PP	0.2
PP	$\rightarrow$	IN	NP	1.0

Vi	$\rightarrow$	sleeps	1.0
Vt	$\rightarrow$	saw	1.0
NN	$\rightarrow$	man	0.1
NN	$\rightarrow$	woman	0.1
NN	$\rightarrow$	telescope	0.3
NN	$\rightarrow$	dog	0.5
DT	$\rightarrow$	the	1.0
IN	$\rightarrow$	with	0.6
IN	$\rightarrow$	in	0.4





$$q(\text{VP} \rightarrow \text{Vt NP}) \times q(\text{NP} \rightarrow \text{NP PP}) = 0.5 \times 0.2 = 0.1$$

$$q(\text{VP} \rightarrow \text{VP PP}) \times q(\text{VP} \rightarrow \text{Vt NP}) = 0.2 \times 0.5 = 0.1$$

This PCFG can't identify the correct parse tree!

Learning from data: treebanks

Adding probabilities to the rules: probabilistic CFGs

Treebanks: a collection of sentences paired with their annotated parse trees

```
((S
   (NP-SBJ (DT That)
     (JJ cold) (, ,)
     (JJ empty) (NN sky) )
   (VP (VBD was)
     (ADJP-PRD (JJ full)
       (PP (IN of)
         (NP (NN fire)
           (CC and)
           (NN light) ))))
   (. .) ))
```

(a)

```
((S
   (NP-SBJ The/DT flight/NN )
   (VP should/MD
     (VP arrive/VB
       (PP-TMP at/IN
         (NP eleven/CD a.m/RB ))
       (NP-TMP tomorrow/NN )))))
```

(b)

The Penn Treebank Project (Marcus et al, 1993)

Penn Treebank

Standard setup

- 40,000 sentences for training
- 1,700 for development
- 2,400 for testing

Phrasal categories	
ADJP	Adjective phrase
ADVP	Adverb phrase
NP	Noun phrase
PP	Prepositional phrase
S	Simple declarative clause
SBAR	Subordinate clause
SBARQ	Direct question introduced by *wh*-element
SINV	Declarative sentence with subject-aux inversion
SQ	Yes/no questions and subconstituent of SBARQ excluding *wh*-element
VP	Verb phrase
WHADVP	Wh-adverb phrase
WHNP	Wh-noun phrase
WHPP	Wh-prepositional phrase
X	Constituent of unknown or uncertain category
*	"Understood" subject of infinitive or imperative
0	Zero variant of *that* in subordinate clauses
T	Trace of wh-Constituent

Penn Treebank

Part-of-speech tagset

CC	Coordinating conj.	TO	infinitival *to*
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential there	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund/present pple
IN	Preposition	VBN	Verb, past participle
JJ	Adjective	VBP	Verb, non-3rd ps. sg. present
JJR	Adjective, comparative	VBZ	Verb, 3rd ps. sg. present
JJS	Adjective, superlative	WDT	Wh-determiner
LS	List item marker	WP	Wh-pronoun
MD	Modal	WP$	Possessive *wh*-pronoun
NN	Noun, singular or mass	WRB	Wh-adverb
NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	$	Dollar sign
NNPS	Proper noun, plural	.	Sentence-final punctuation
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Colon, semi-colon
PRP	Personal pronoun	(	Left bracket character
PP$	Possessive pronoun	)	Right bracket character
RB	Adverb	"	Straight double quote
RBR	Adverb, comparative	‘	Left open single quote
RBS	Adverb, superlative	“	Left open double quote
RP	Particle	’	Right close single quote
SYM	Symbol	”	Right close double quote

Treebanks

Which of the following statements is incorrect?

(a) A treebank can provide us frequencies and distributional information

(b) A treebank provides us a way to evaluate systems

(c) The treebank data can be biased to the selection of sentences/documents

(d) It is easy to scale up a treebank to multiple domains and languages

The answer is (d).

Deriving a PCFG from a treebank

- Training data: a set of parse trees $t_1, t_2, \ldots, t_m$
- A PCFG (N, Σ, S, R, q):
  - N is the set of all non-terminals seen in the trees
  - Σ is the set of all words seen in the trees
  - S is taken to be S.
  - R is taken to be the set of all rules $\alpha \rightarrow \beta$ seen in the trees

Deriving a PCFG from a treebank

```
((S
   (NP-SBJ (DT That)
     (JJ cold) (, ,)
     (JJ empty) (NN sky) )
   (VP (VBD was)
     (ADJP-PRD (JJ full)
       (PP (IN of)
         (NP (NN fire)
           (CC and)
           (NN light) ))))
   (. .) ))
```

(a)

```
((S
   (NP-SBJ The/DT flight/NN )
   (VP should/MD
     (VP arrive/VB
       (PP-TMP at/IN
         (NP eleven/CD a.m/RB ))
       (NP-TMP tomorrow/NN )))))
```

(b)

```
( (S (`` ``)
    (S-TPC-2
      (NP-SBJ-1 (PRP We) )
      (VP (MD would)
        (VP (VB have)
          (S
            (NP-SBJ (-NONE- *-1) )
            (VP (TO to)
              (VP (VB wait)
                (SBAR-TMP (IN until)
                  (S
                    (NP-SBJ (PRP we) )
                    (VP (VBP have)
                      (VP (VBN collected)
                        (PP-CLR (IN on)
                          (NP (DT those)(NNS assets)))))))))))))
    (, ,) ('' '')
    (NP-SBJ (PRP he) )
    (VP (VBD said)
      (S (-NONE- *T*-2) ))
    (. .) ))
```

Deriving a PCFG from a treebank

Grammar	Lexicon
$S \rightarrow NP\ VP\ .$	$PRP \rightarrow we \mid he$
$S \rightarrow NP\ VP$	$DT \rightarrow the \mid that \mid those$
$S \rightarrow$ " S " , $NP\ VP$.	$JJ \rightarrow cold \mid empty \mid full$
$S \rightarrow$ -NONE-	$NN \rightarrow sky \mid fire \mid light \mid flight \mid tomorrow$
$NP \rightarrow DT\ NN$	$NNS \rightarrow assets$
$NP \rightarrow DT\ NNS$	$CC \rightarrow and$
$NP \rightarrow NN\ CC\ NN$	$IN \rightarrow of \mid at \mid until \mid on$
$NP \rightarrow CD\ RB$	$CD \rightarrow eleven$
$NP \rightarrow DT\ JJ\ ,\ JJ\ NN$	$RB \rightarrow a.m.$
$NP \rightarrow PRP$	$VB \rightarrow arrive \mid have \mid wait$
$NP \rightarrow$ -NONE-	$VBD \rightarrow was \mid said$
$VP \rightarrow MD\ VP$	$VBP \rightarrow have$
$VP \rightarrow VBD\ ADJP$	$VBN \rightarrow collected$
$VP \rightarrow VBD\ S$	$MD \rightarrow should \mid would$
$VP \rightarrow VBN\ PP$	$TO \rightarrow to$
$VP \rightarrow VB\ S$	
$VP \rightarrow VB\ SBAR$	
$VP \rightarrow VBP\ VP$	
$VP \rightarrow VBN\ PP$	
$VP \rightarrow TO\ VP$	
$SBAR \rightarrow IN\ S$	
$ADJP \rightarrow JJ\ PP$	
$PP \rightarrow IN\ NP$	

A sample of the CFG grammar rules and lexical entries that would be extracted from the three treebank sentences

Deriving a PCFG from a treebank

- Training data: a set of parse trees $t_1, t_2, \ldots, t_m$
- A PCFG (N, Σ, S, R, q):
  - N is the set of all non-terminals seen in the trees
  - Σ is the set of all words seen in the trees
  - S is taken to be S.
  - R is taken to be the set of all rules $\alpha \to \beta$ seen in the trees

The maximum-likelihood parameter (MLE) estimates are:

$$q_{ML}(\alpha \to \beta) = \frac{\text{Count}(\alpha \to \beta)}{\text{Count}(\alpha)}$$

If we have seen the rule VP $\to$ Vt NP 105 times, and the the non-terminal VP 1000 times, $q(\text{VP} \to \text{Vt NP}) = 0.105$

Parsing with PCFGs

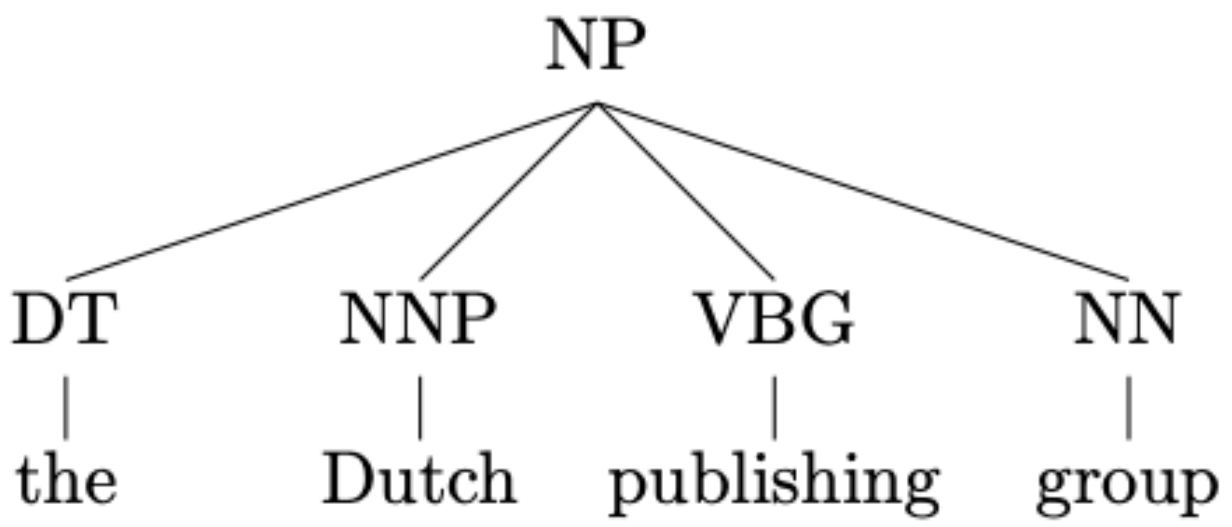
- Given a sentence s and a PCFG, how to find the highest scoring parse tree for s?

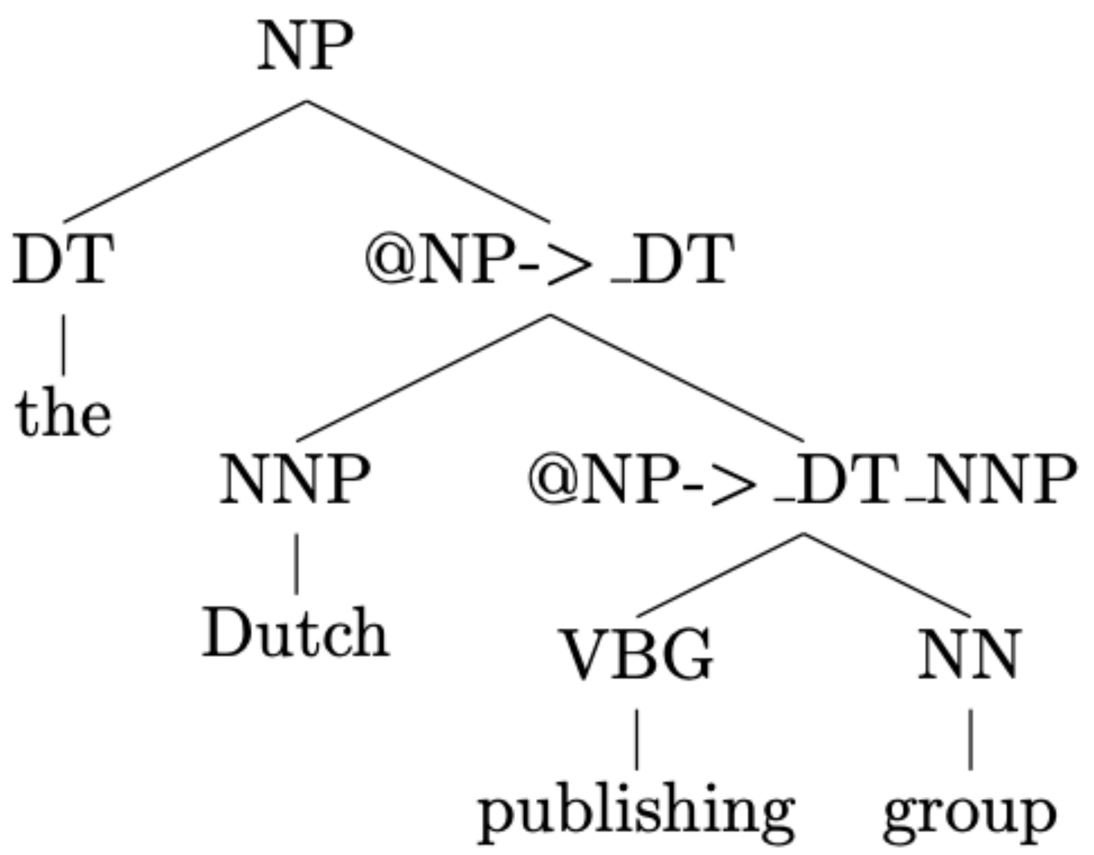
$$argmax_{t \in \mathcal{T}(s)} P(t)$$

- **The CKY algorithm**: applies to a restricted type of PCFG— a PCFG in Chomsky normal form (CNF)
  - CKY = the Cocke-Kasami-Younger algorithm
- **Chomsky Normal Form (CNF)**: all the rules take one of the two following forms:
  - $X \rightarrow Y_1 Y_2$ where $X \in N, Y_1 \in N, Y_2 \in N$
  - $X \rightarrow Y$ where $X \in N, Y \in \Sigma$
- It is possible to convert any PCFG into an equivalent grammar in CNF!
  - However, the trees will look different; It is possible to do “reverse transformation”

Converting PCFGs into a CNF grammar

- n-ary rules ($n > 2$): NP $\rightarrow$ DT NNP VBG NN



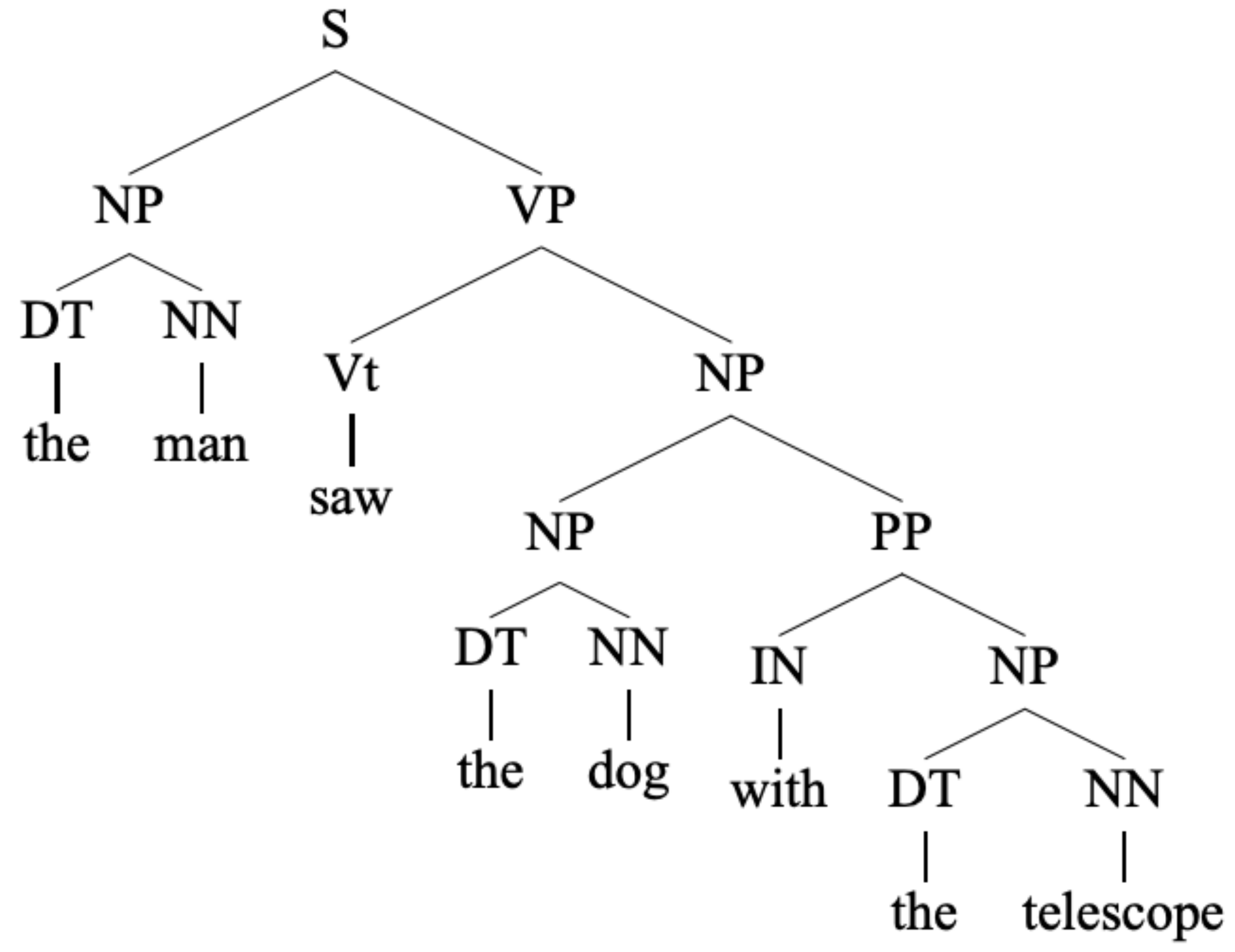


- Unary rules: VP $\rightarrow$ Vi, Vi $\rightarrow$ sleeps
  - Eliminate all the unary rules recursively by adding VP $\rightarrow$ sleeps

The CKY algorithm

- Dynamic programming
- Given a sentence $x_1, x_2, \ldots, x_n$, denote $\pi(i, j, X)$ as the highest score for any parse tree that *dominates* words $x_i, \ldots, x_j$ and has non-terminal $X \in N$ as its root.
- Output: $\pi(1, n, S)$
- Initially, for $i = 1, 2, \ldots, n$,

$$\pi(i, i, X) = \begin{cases} q(X \to x_i) & \text{if } X \to x_i \in R \\ 0 & \text{otherwise} \end{cases}$$

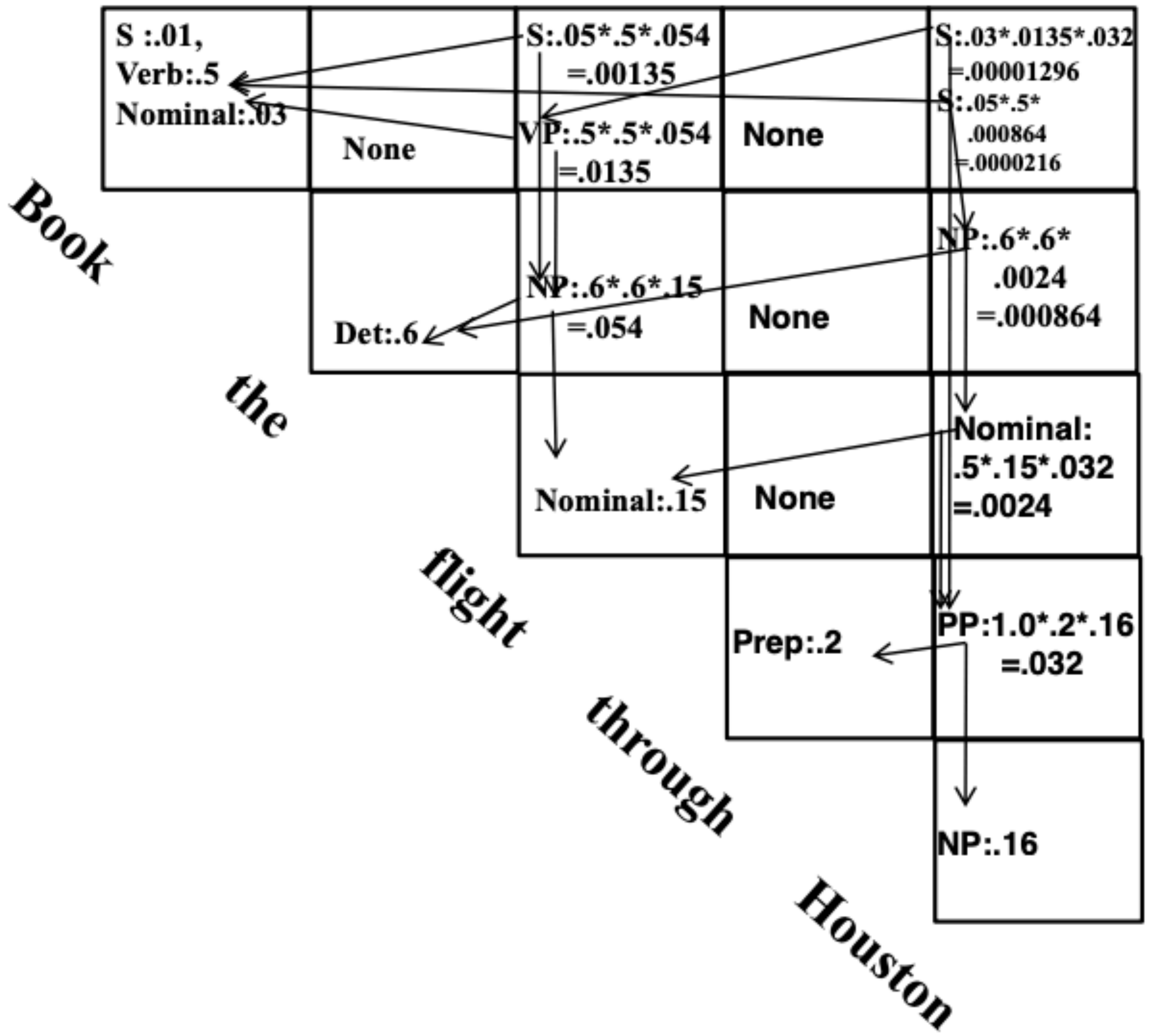


The CKY algorithm

- For all (i, j) such that $1 \leq i < j \leq n$ for all $X \in N$,

$$\pi(i,j,X) = \max_{X \to YZ \in R, i \leq k < j} q(X \to YZ) \times \pi(i,k,Y) \times \pi(k+1,j,Z)$$

Also stores backpointers which allow us to recover the parse tree



The CKY algorithm

- For all (i, j) such that $1 \leq i < j \leq n$ for all $X \in N$,

$$\pi(i, j, X) = \max_{X \to YZ \in R, i \leq k < j} q(X \to YZ) \times \pi(i, k, Y) \times \pi(k+1, j, Z)$$

What is the time complexity of the CKY algorithm?

a) $O(n^2 |R|)$

b) $O(n^3 |R|)$

c) $O(n^2 |N|^3)$

d) $O(n^3 |N|^3)$

N: set of non-terminal symbols

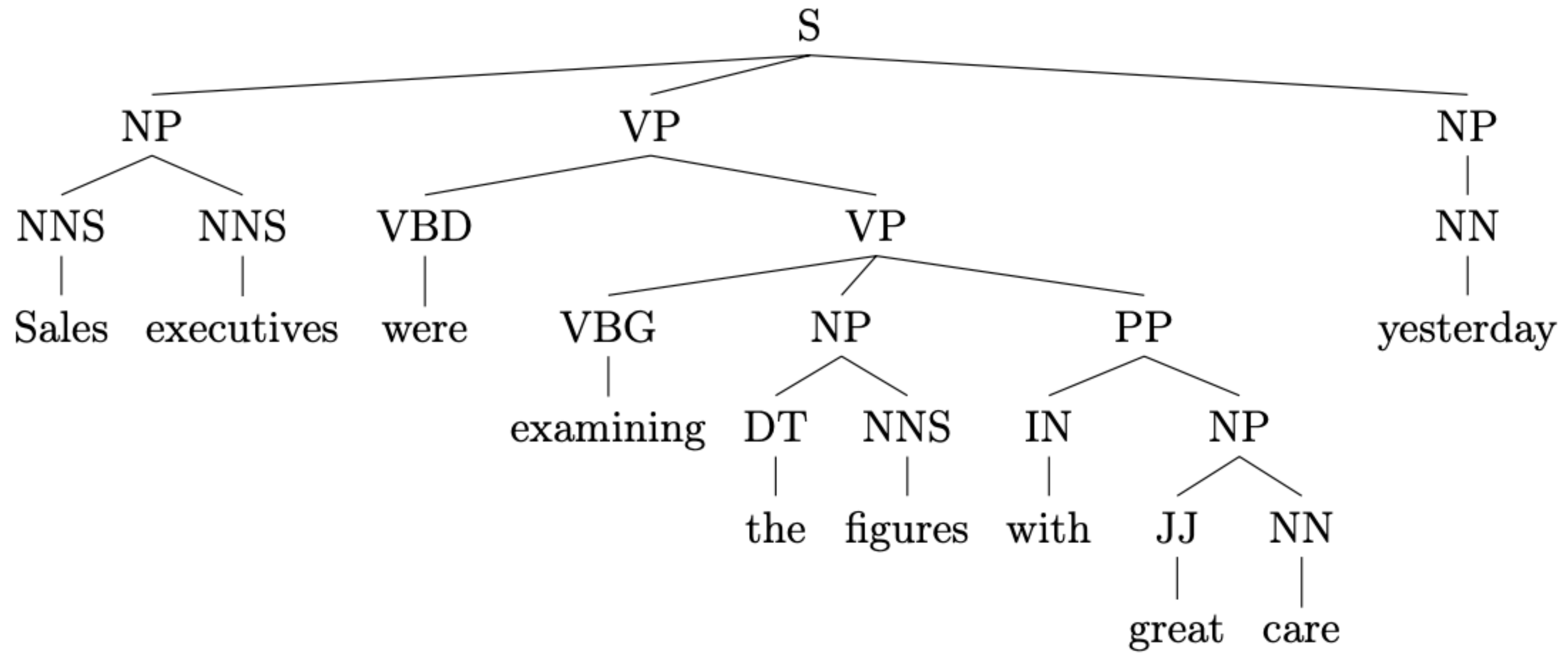
R: set of derivation rules

n: sentence length

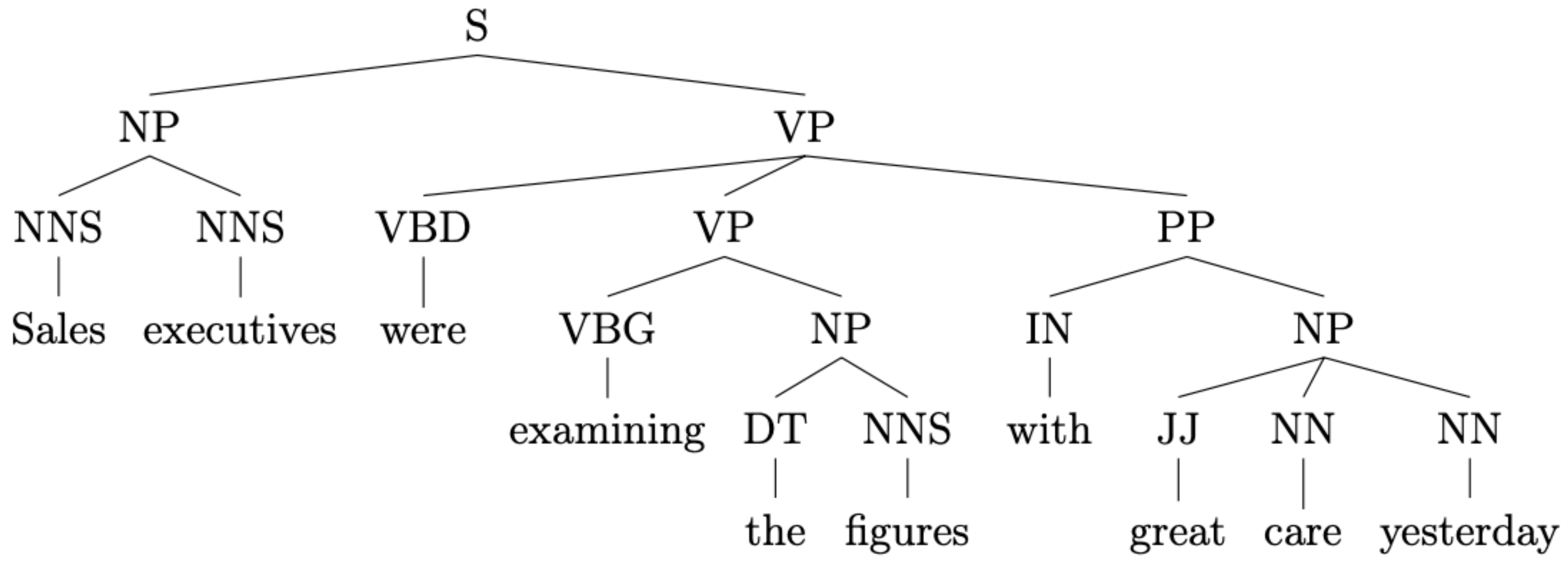
The answer is (b).

Evaluating constituency parsing

Gold: **(1, 10, S), (1, 2, NP)**, (3, 9, VP), (4, 9, VP), **(5, 6, NP)**, (7, 9, PP), (8, 9, NP), (10, 10, NP)



Predicted: **(1, 10, S), (1, 2, NP)**, (3, 10, VP), (4, 6, VP), **(5, 6, NP)**, (7, 10, PP), (8, 10, NP)



Evaluating constituency parsing

- **Labeled** recall: (# correct constituents in candidate) / (# constituents in gold tree)
- **Labeled** precision: (# correct constituents in candidate) / (# constituents in candidate)
- F1 is the harmonic mean of precision and recall = (2 * precision * recall) / (precision + recall)
- Part-of-speech tagging accuracy is evaluated separately
- A constituent is correct if there is a constituent in the gold tree with the same starting point, ending point, and **non-terminal symbol**.

Precision and Recall

Gold: **(1, 10, S), (1, 2, NP)**, (3, 9, VP), (4, 9, VP), **(5, 6, NP)**, (7, 9, PP), (8, 9, NP), (10, 10, NP)

Predicted: **(1, 10, S)**, **(1, 2, NP)**, (3, 10, VP), (4, 6, VP), **(5, 6, NP)**, (7, 10, PP), (8, 10, NP)

What are the **labeled** precision (P) / recall (R) in the above example?

(a) P = 3/8, R = 3/7

(b) P = 3/7, R = 3/8

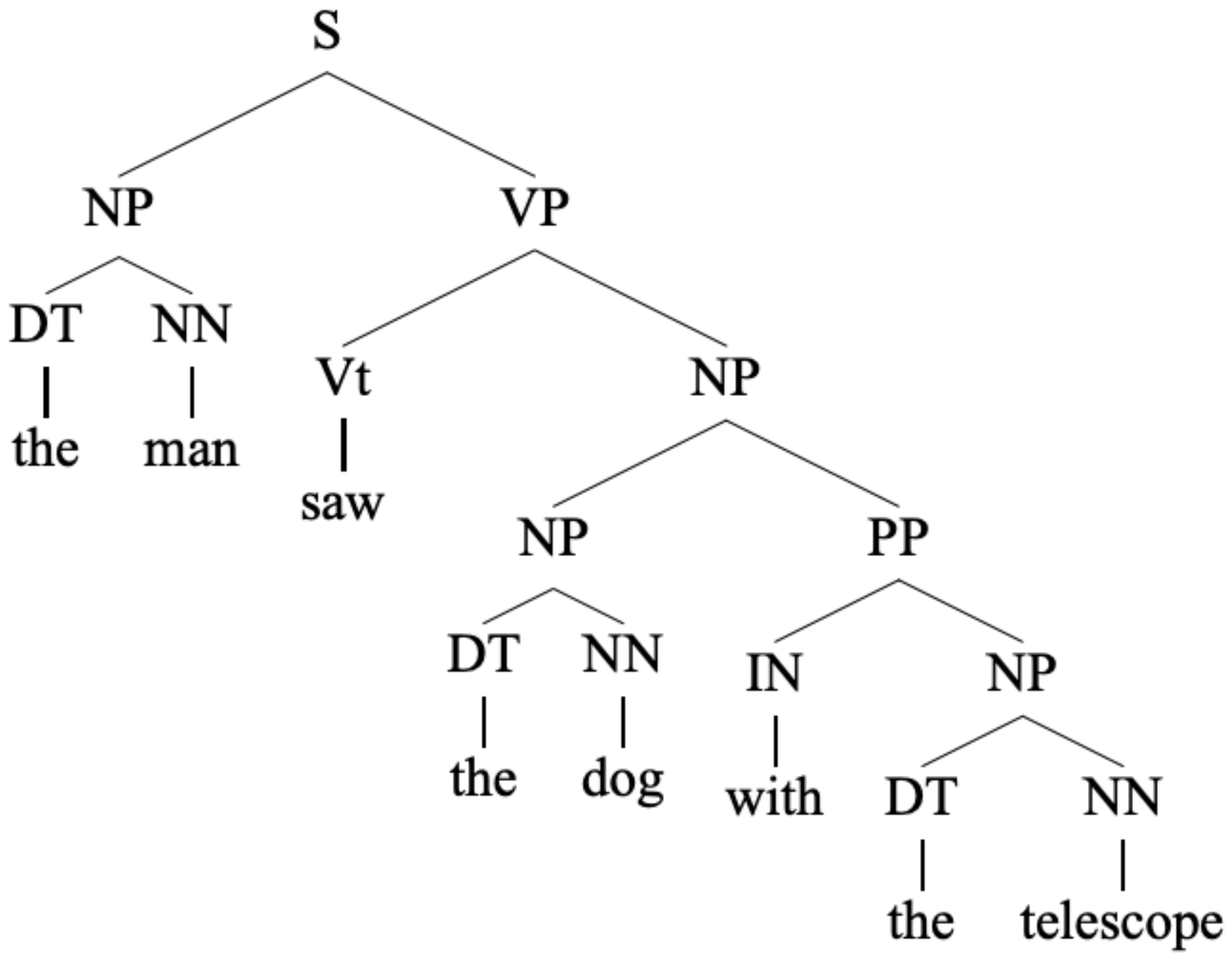
(c) P = 1/2, R = 1/2

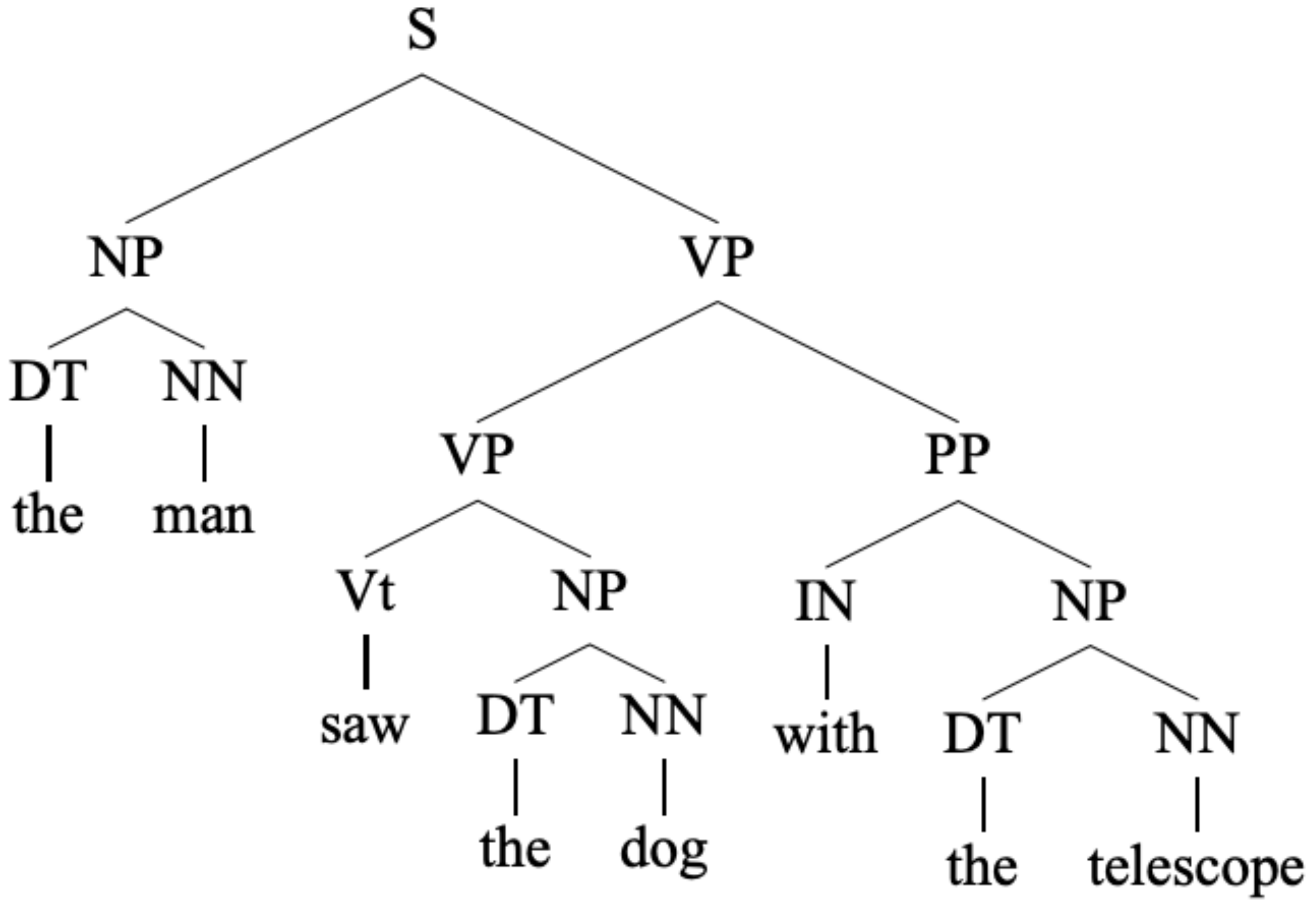
(d) P = 1, R = 1

The answer is (b). F1 = 40%, tagging accuracy = 100%

Weaknesses of PCFGs

Lack of sensitivity to lexical information (words)





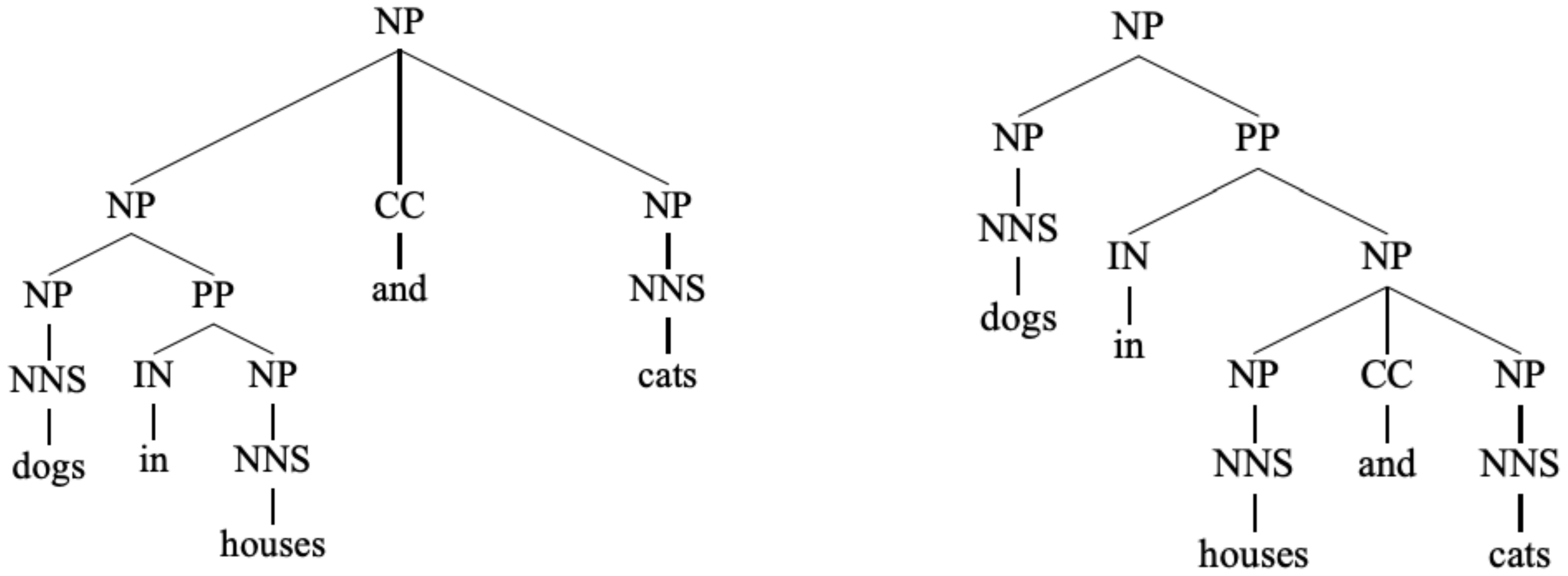
The only difference between these two parses:

$q(\text{VP} \rightarrow \text{VP PP})$ vs $q(\text{NP} \rightarrow \text{NP PP})$

… without looking at the words!

Weaknesses of PCFGs

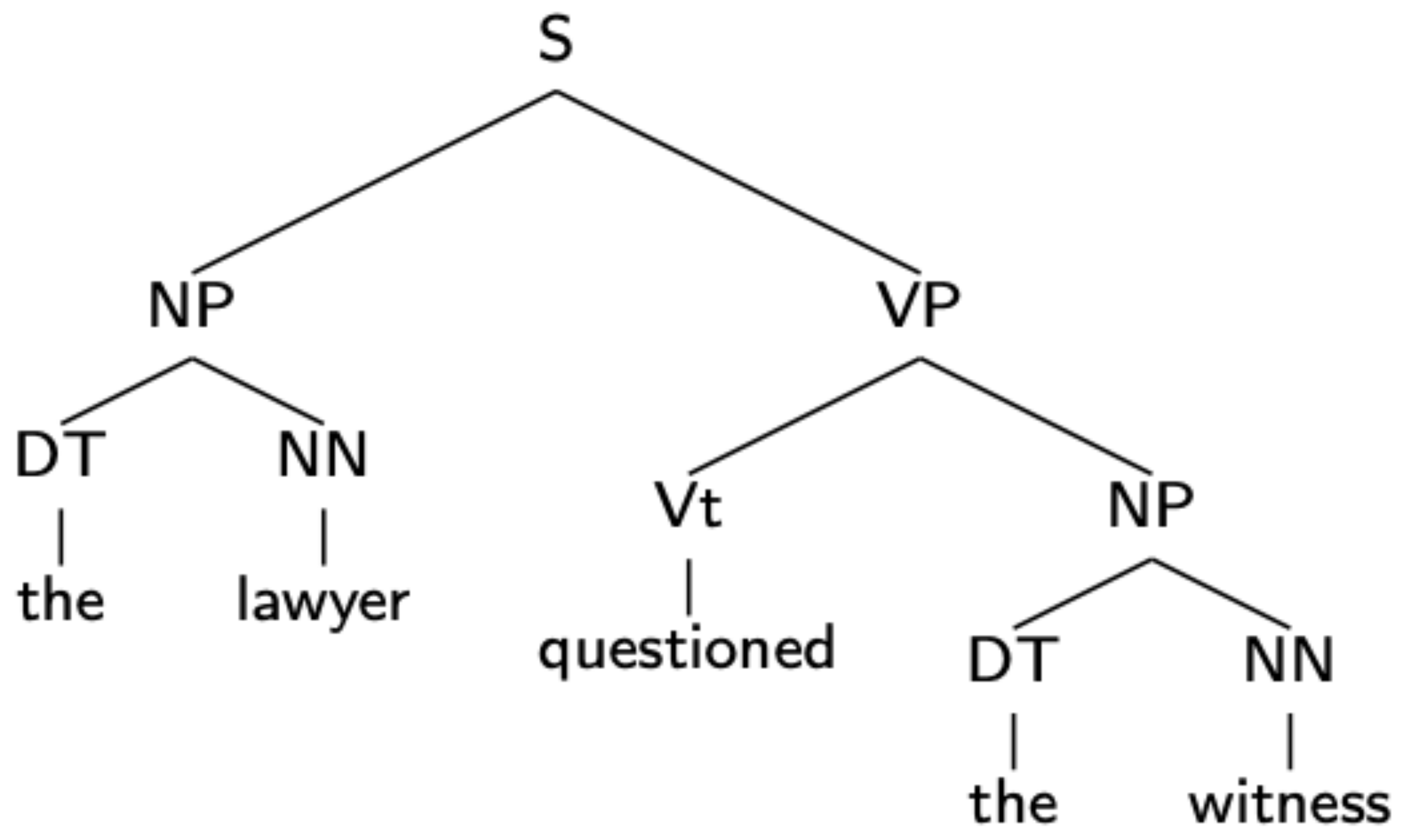
Lack of sensitivity to lexical information (words)

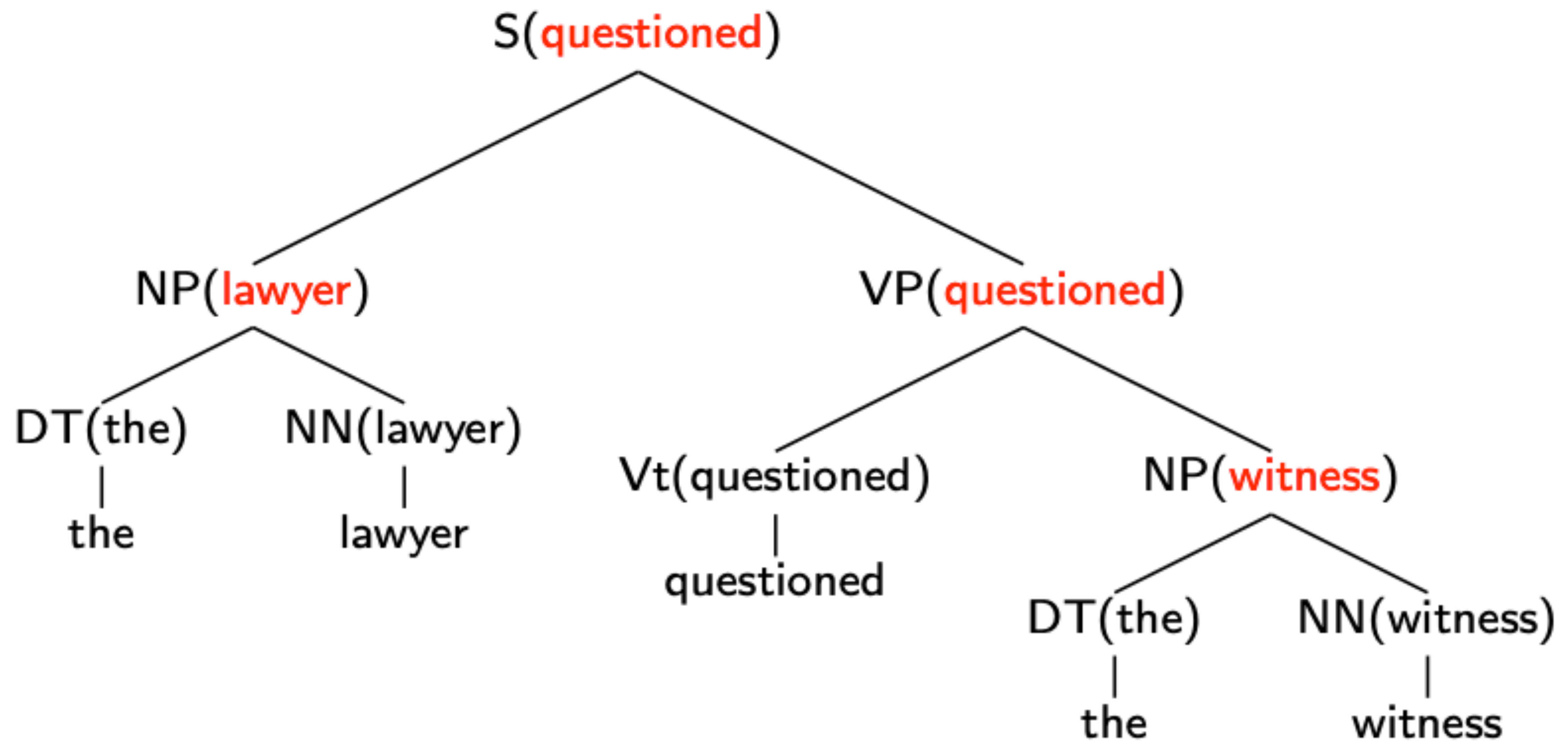


Exactly the same set of context-free rules!

Lexicalized PCFGs

- Key idea: add **headwords** to trees





- Each context-free rule has one special child that is the head of the rule (a core idea in syntax)

S	$\Rightarrow$	NP	VP		(VP is the head)
VP	$\Rightarrow$	Vt	NP		(Vt is the head)
NP	$\Rightarrow$	DT	NN	NN	(NN is the head)

Lexicalized PCFGs

The heads are decided by rules:

If the rule contains NN, NNS, or NNP:
Choose the rightmost NN, NNS, or NNP

Else If the rule contains an NP: Choose the leftmost NP

Else If the rule contains a JJ: Choose the rightmost JJ

Else If the rule contains a CD: Choose the rightmost CD

Else Choose the rightmost child

If the rule contains Vi or Vt: Choose the leftmost Vi or Vt

Else If the rule contains a VP: Choose the leftmost VP

Else Choose the leftmost child

Lexicalized PCFGs

$$
\begin{array}{llll}
\text{S(saw)} & \rightarrow_2 & \text{NP(man)} & \text{VP(saw)} \\
\text{VP(saw)} & \rightarrow_1 & \text{Vt(saw)} & \text{NP(dog)} \\
\text{NP(man)} & \rightarrow_2 & \text{DT(the)} & \text{NN(man)} \\
\text{NP(dog)} & \rightarrow_2 & \text{DT(the)} & \text{NN(dog)} \\
\text{Vt(saw)} & \rightarrow & \text{saw} & \\
\text{DT(the)} & \rightarrow & \text{the} & \\
\text{NN(man)} & \rightarrow & \text{man} & \\
\text{NN(dog)} & \rightarrow & \text{dog} &
\end{array}
$$

- Further reading: *Michael Collins. 2003. Head-Driven Statistical Models for Natural Language Parsing.*
- Results for a PCFG: 70.6% recall, 74.8% precision
- Results for a lexicalized PCFG: 88.1% recall, 88.3% precision