

2018/11/12 17:00 修正

計算言語学

構造分類 (形式文法)

東京大学生産技術研究所

吉永 直樹

site: <http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/class/cl/>

言語をカテゴリ(クラス)の構造に分類する

品詞タグ付け (part-of-speech tagging)

Lucy	in	the	sky	with	diamonds
NNP	IN	DT	NN	IN	NNS

単語に対する品詞の系列

固有表現認識 (named entity recognition)

Ringo	Star	has joined	the	Beatles
PERSON	Non-NE			ORGANIZATION

固有表現のチャンクとその分類

単語分割 (word segmentation)

中	国	人	参	政	权
0	1	1	0	0	1

単語境界の有無の系列

依存構造解析 (dependency parsing)

Ringo Star has joined the Beatles

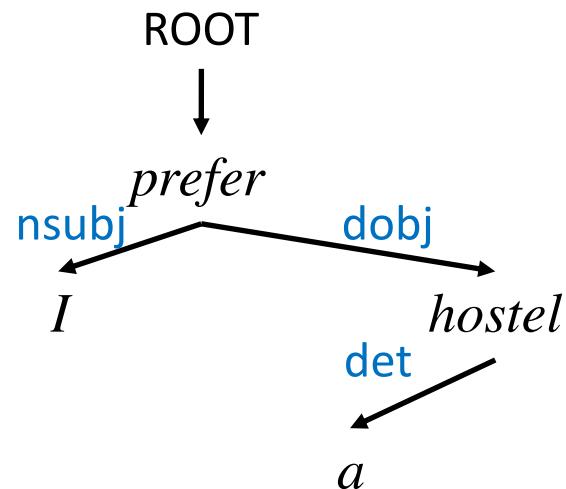
The diagram shows six words: Ringo, Star, has, joined, the, and Beatles. Blue curved arrows point from each word except 'joined' to the word 'joined'. This indicates that 'joined' is the root node or head of the sentence, and all other words depend on it.

依存構造木 or 結合操作の系列

個々のラベルの間に依存関係がある

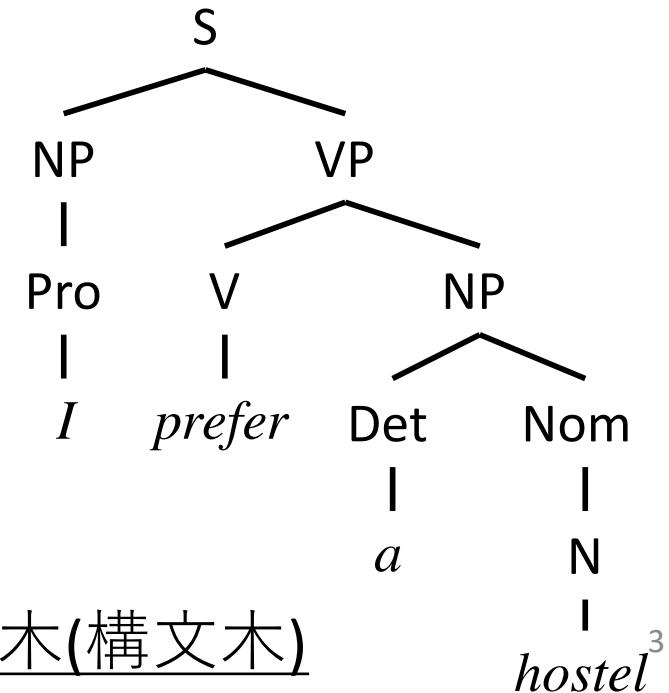
構文構造を記述する二つのアプローチ

- 依存文法 [Teschnière 1959]
 - 単語間の統語的依存関係を解析
 - 日本語やチェコ語など、語順が自由な言語で発達
 - 統計的手法



11/13/18

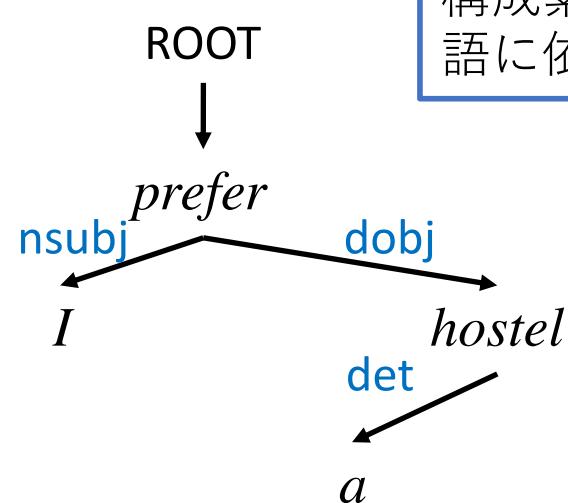
- 句構造文法 [Chomsky 1956]
 - 単語列(構成素)の階層的包含関係を解析
 - 英語など、語順が比較的固定された言語で発達
 - 形式文法 + 統計的手法



構文構造を記述する二つのアプローチ

- 依存文法 [Tesiⁿière 1959]

- 単語間の統語的依存関係を解析
- 日本語やチェコ語など、語順が自由な言語で発達
- 統計的手法

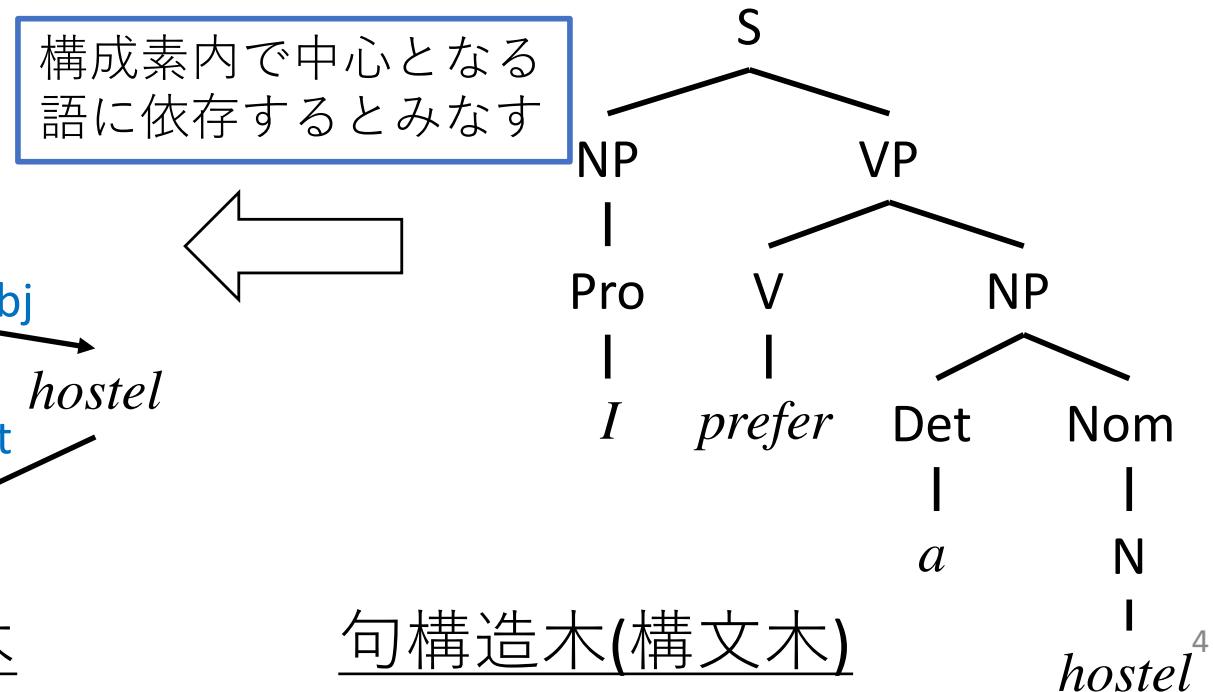


依存構造木

11/13/18

- 句構造文法 [Chomsky 1956]

- 単語列(構成素)の階層的包含関係を解析
- 英語など、語順が比較的固定された言語で発達
- 形式文法 + 統計的手法



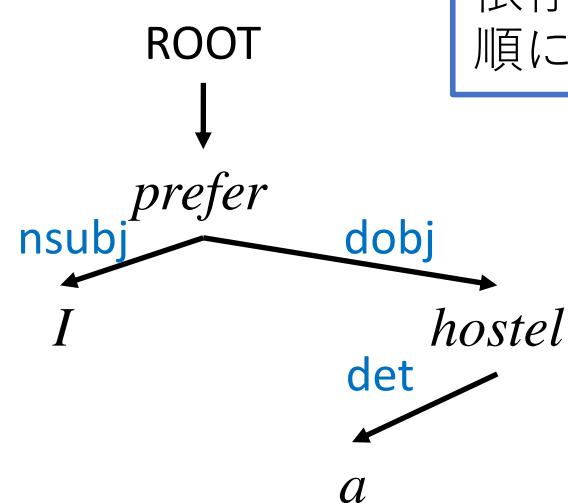
句構造木(構文木)

⁴
hostel

構文構造を記述する二つのアプローチ

- 依存文法 [Tesiⁿière 1959]

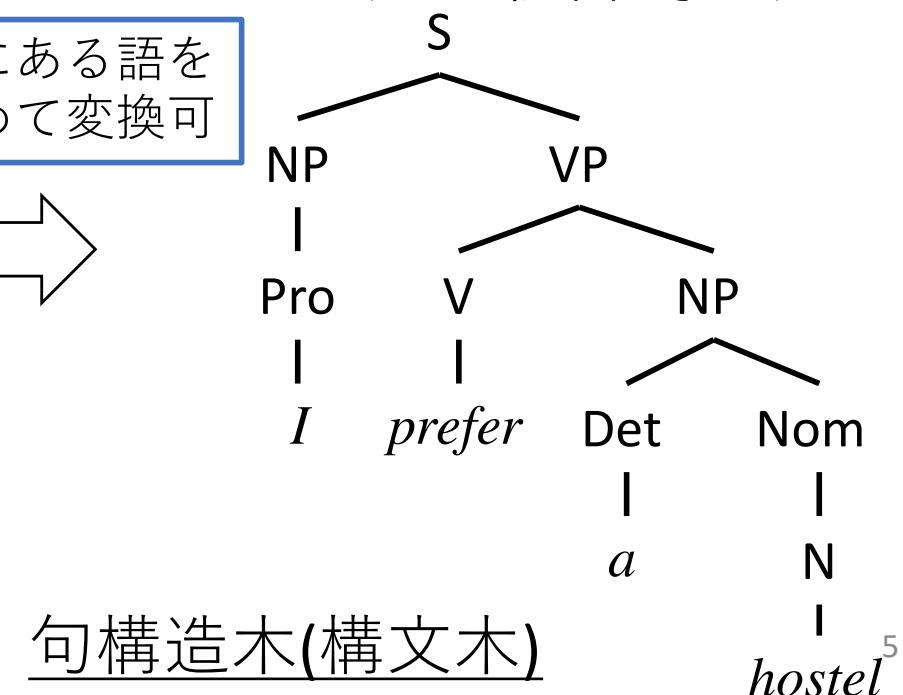
- 単語間の統語的依存関係を解析
- 日本語やチェコ語など、語順が自由な言語で発達
- 統計的手法



依存構造木

- 句構造文法 [Chomsky 1956]

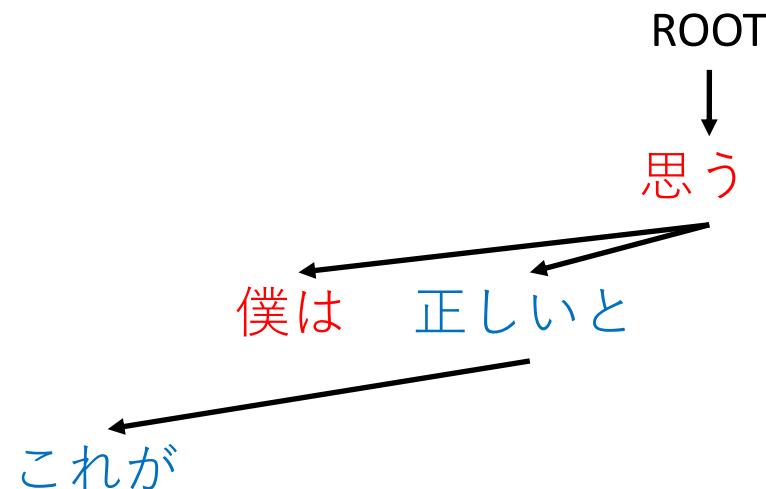
- 単語列(構成素)の階層的包含関係を解析
- 英語など、語順が比較的固定された言語で発達
- 形式文法 + 統計的手法



句構造木(構文木)

構文構造を記述する二つのアプローチ

- 依存文法 [Teschnière 1959]
 - 単語間の統語的依存関係を解析
 - 日本語やチェコ語など、語順が自由な言語で発達
 - 統計的手法



11/13/18

依存構造木

- 句構造文法 [Chomsky 1956]
 - 単語列(構成素)の階層的包含関係を解析
 - 英語など、語順が比較的固定された言語で発達
 - 形式文法 + 統計的手法

構成素中で語の連続性を仮定する句構造文法では適切に解析し難い場合も

句構造木(構文木)?

文の構造を記述する二つのアプローチ

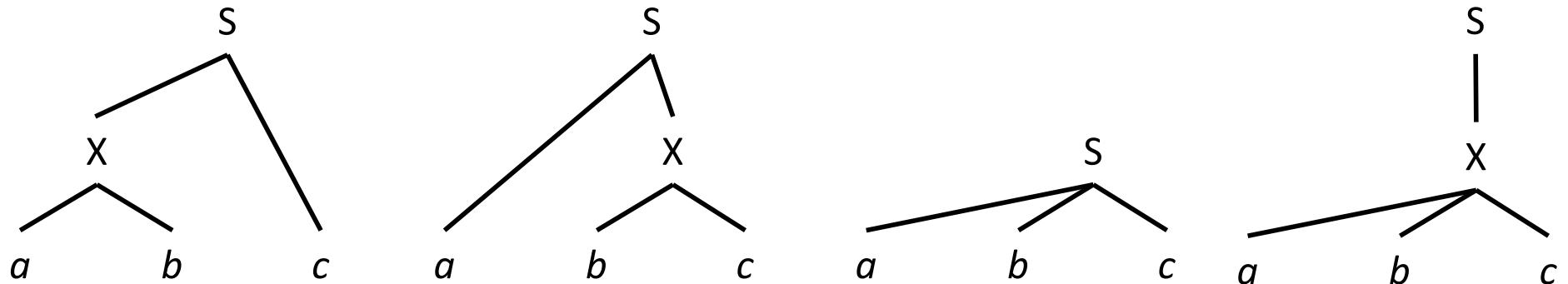
- 依存文法 [Tesanière 1959]
 - 単語間の統語的依存関係を解析
 - 日本語やチェコ語など、語順が自由な言語で発達
 - 統計的手法

- 句構造文法 [Chomsky 1956]
 - 単語列(構成素)の階層的包含関係を解析
 - 英語など、語順が比較的固定された言語で発達
 - 形式文法 + 統計的手法

句構造解析

入力文に対して、句構造木を返すタスク

- 問題: ラベルの曖昧性 \times 構造の曖昧性
 - 解空間が文長に対して指数爆発 (系列ラベリングは線形)
 - 大域的最適化は文長の多項式オーダかかる



一分岐も高頻出

- 古典的アプローチ:

- 形式文法を用いた解の絞り込み + 統計的手法による選別

今日は統計的手法を用いない句構造解析について話します

文の構成性 (constituency)

- 構成素 (constituent): 統語論的に規則的な振る舞いをする単語列 (品詞, 句 etc.)

例) 名詞句

three parties from Brooklyn arrive...

a high-class spot such as Mindy's attracts...

the Broadway coppers love...

they sit...

- 構成素の一部が欠けていたり, 構成する語が連続しないときには構成素の統語的性質は失われる

** from Brooklyn arrive...*

(言語学で)文法的に誤った文(非文)の前に付ける

** the love Broadway coppers ...*

文脈自由文法 (Context-Free Grammar; CFG) (1/2)

[Chomsky 1956]

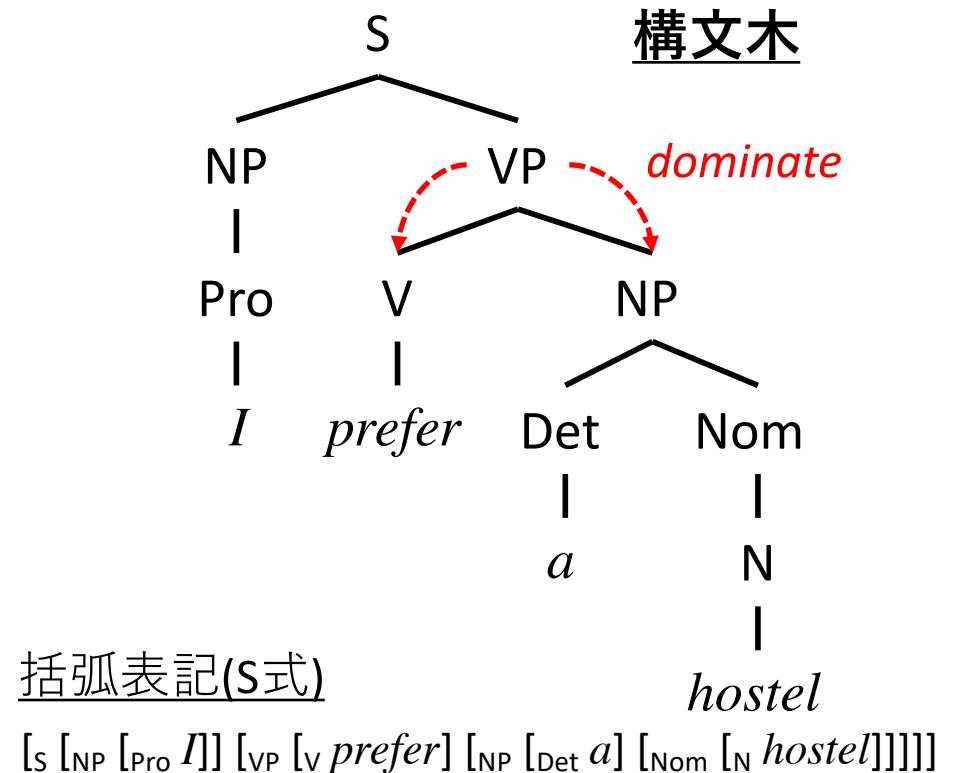
- 構成素に基づく形式文法(句構造文法)の一種
 - 句構造木の一段分の構成性を規則で表現する文法枠組
$$\boxed{\forall \alpha, \beta, \alpha A \gamma \rightarrow \alpha \beta \gamma}$$
- 文脈自由文法 G は以下で4つ組で定義される
 - N : 非終端記号 (変数) 構成素のラベル (文法カテゴリ)
 - Σ : 終端記号 語彙(単語)
 - R : 生成規則 $A \rightarrow \beta$ ($A \in N, \beta \in (\Sigma \cup N)^*$) 構成素間の包含関係
 - S : 開始記号 ($S \in N$) 文
- 形式言語: 形式文法が導出する終端記号列の集合
 - 文が与えられた形式文法から導出される = grammatical
 - 言語学では言語を記述する形式文法を生成文法と呼ぶ

文脈自由文法 (Context-Free Grammar; CFG) (2/2)

[Chomsky 1956]

- 文: 開始記号を生成規則の連続適用で書き換えて
導出される終端記号列 $S \xrightarrow{*} \alpha (\alpha \in \Sigma^+)$
- 構文木: 導出過程の木構造表現

<u>CFG G</u>	
$N : S, NP, VP, Pro, V, Det, Nom, N$	
$\Sigma : I, prefer, a, hostel$	
$R : S \rightarrow NP\ VP$	$Pro \rightarrow I$
$VP \rightarrow V\ NP$	$V \rightarrow prefer$
$NP \rightarrow Det\ Nom$	$Det \rightarrow a$
$NP \rightarrow Pro$	$N \rightarrow hostel$
$Nom \rightarrow N$	
$S : S$	



文脈自由文法で英文法を記述する (1/5)

- 文(あるいは節)を、文脈を考える上での基本単位として定義

$S \rightarrow NP\ VP$

宣言文

He runs fast.

$S \rightarrow VP$

命令文

Run fast.

$S \rightarrow Aux\ NP\ VP$

yes-no 疑問文 *Does he run fast?*

$S \rightarrow Wh-NP\ VP$

wh-subj. - *Who runs fast?*

$S \rightarrow Wh-NP\ Aux\ NP\ VP\ wh-non-subj.\ -$ *What car do you have?*

文脈自由文法で英文法を記述する (2/5)

- 名詞句: 名詞を主辞(head)とする構成素(句)
句で統語的に中心的な役割を持つ語

NP → Nominal *people, water*

NP → Det Nominal *a pencil, the people*

NP → PreDet NP *all the people*

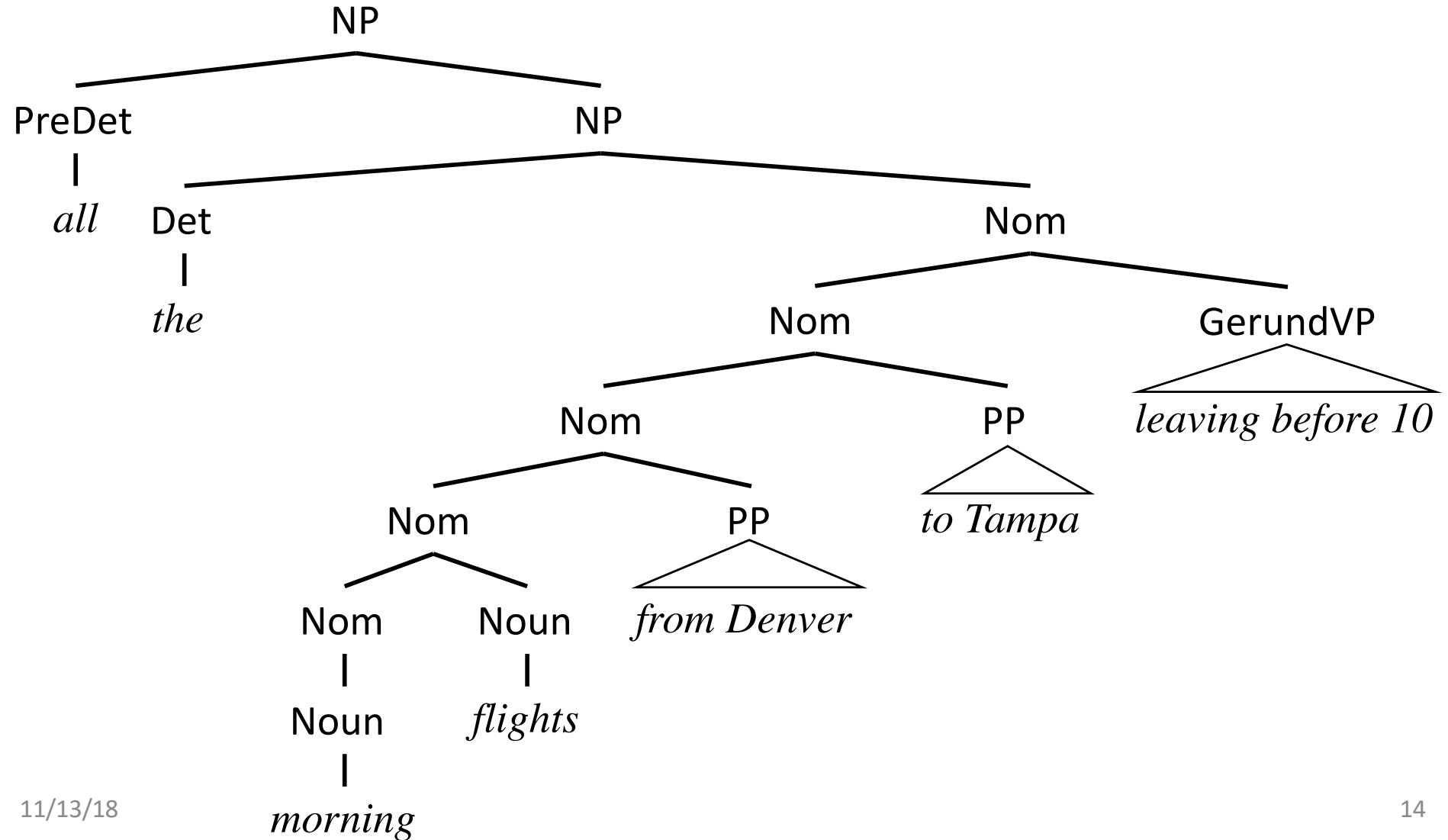
Det → *a* Nominal → Noun *pencil*

Det → *the* Nominal → Noun PP *sky with diamonds*

Det → NP 's Nominal → Noun GerundVP *bird singing in the night*
 Nominal → Noun RelClause *bird that sings in morning*

文脈自由文法で英文法を記述する (3/5)

名詞句 *All the morning flights from Denver to Tampa leaving before 10* の構文木



文脈自由文法で英文法を記述する (4/5)

- 動詞句: 主辞である動詞が様々な項(argument)を取ることで構成する句
主語, 目的語, 補語 . . .

VP → Verb *run, ate*

VP → Verb NP *prefer a hostel*

VP → Verb NP PP *help me with a flight*

VP → Verb PP *fly from Narita*

VP → Verb S *believed that she will come*

- 動詞ごとに取りうる項のパターン(下位範疇化フレーム)が異なるため、動詞・規則を細分化して対応

Verb-with-NP-complement → *find | leave | repeat | ...*

Verb-with-S-complement → *think | believe | say | ...*

Verb-with-Inf-VP-complement → *want | try | need | ...*

文脈自由文法で英文法を記述する (5/5)

- 並列句: 二つの句を等位接続詞で連結することで構成される句

$NP \rightarrow NP \ and \ NP$

rainy days and Mondays

$Nominal \rightarrow Nominal \ and \ Nominal$

the flight and costs

$VP \rightarrow VP \ and \ VP$

twist and shout

$S \rightarrow S \ and \ S$

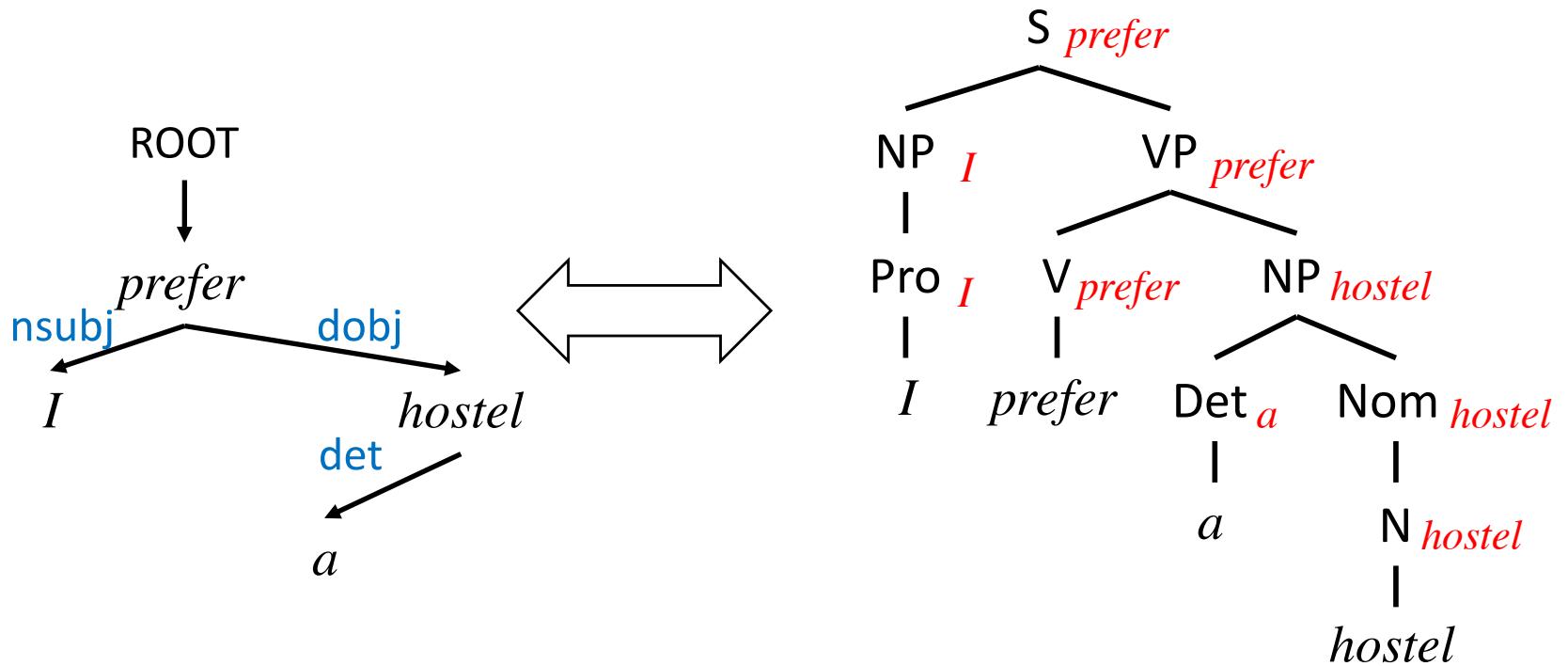
the sun sets and moon rises

- GPSG (Generalized Phrase Structure Grammar) [Gazdar+ 1985] ではメタルールによりこれらを汎化して表現

$X \rightarrow X \ and \ X$

主辞 (head)

- 句(構成素)はその統語的性質を **主辞** から継承する
 - NP (Noun), VP (Verb), PP (Prep), ADJP (ADJ), etc.
 - 句構造と依存構造の対応を考える際に重要



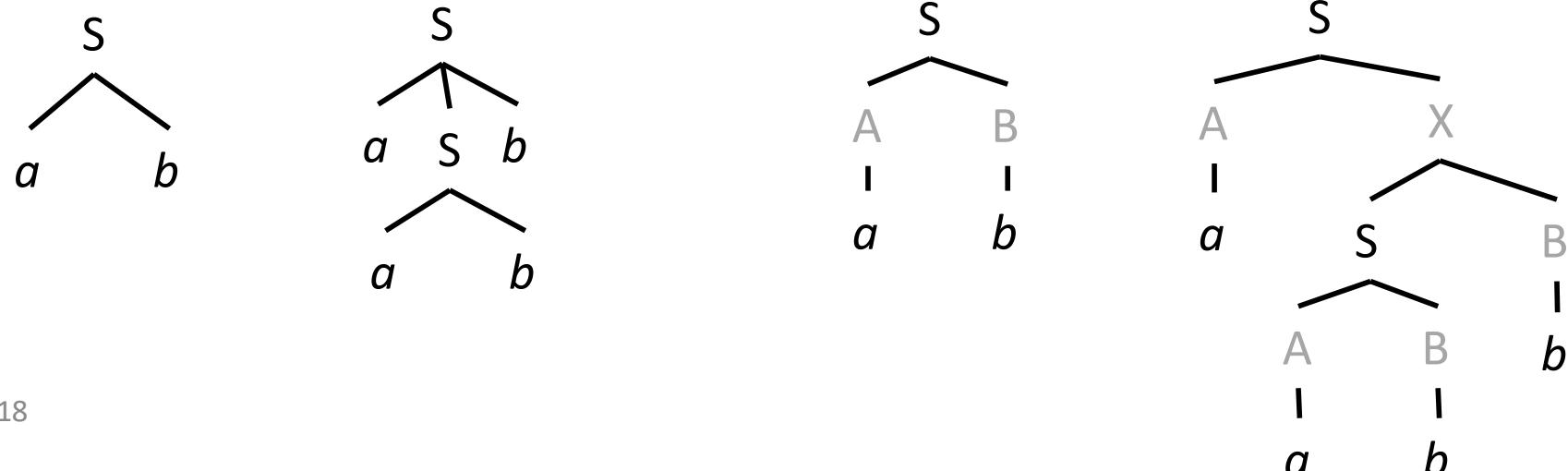
- HPSG** [Pollard & Sag 94] は主辞に注目し文法規則を抽象化
Head-Driven Phrase Structure Grammar; 主辞駆動句構造文法

文法の等価性: 形式文法間の生成能力の関係

- 形式文法 G_0, G_1 が開始記号から同じ終端記号列の集合を導出するとき G_0, G_1 は弱等価と呼ぶ

$$G_0 \quad S \rightarrow aSb \mid S \rightarrow ab \quad G_1 \quad S \rightarrow AB \quad \left| \begin{array}{l} A \rightarrow a \\ S \rightarrow AX \\ X \rightarrow SB \end{array} \right. \quad \begin{array}{l} B \rightarrow b \\ ab, aabb, aaabbb, \dots \end{array}$$

- 形式文法 G_0, G_1 が弱等価かつ同じ文字列に対して同じ導出木を与えるとき G_0, G_1 は強等価と呼ぶ
 - 導出木間に同型写像があれば強等価とする [Miller 1999]



文法の標準形

- 全ての(空文字を生成しない)文脈自由文法は弱等価な標準形に変換できる
- チョムスキー標準形
 - 全ての生成規則が $A \rightarrow B C$ or $A \rightarrow a$ ($B, C \in N, a \in \Sigma$)
 - 構文木は(非終端記号の分岐を除いて)二分木となる
 - 強等価 [Miller 1999], かつ計算機で扱う上で都合が良い
- グライバッハ標準形
 - 全ての生成規則が $A \rightarrow a\alpha$ ($\alpha \in (N \setminus \{S\})^*$)
 - 構文木深さ = 文長かつ各分岐の最左ノードは非終端記号

チョムスキ－標準形への変換

- (空文字を生成しない)任意の文脈自由文法は以下の手順でチョムスキ－標準形に変換可能

1. チョムスキ－標準形を満たす生成規則はコピー
2. 生成規則中の右辺の終端記号を非終端記号に置換

$$X \rightarrow aYb \xrightarrow{\quad} X \rightarrow A Y B, A \rightarrow a, B \rightarrow b$$

3. 単位規則 (unit production; $A \rightarrow B$) の置換

$$X \rightarrow Y \xrightarrow{\quad} X \rightarrow \alpha (X \stackrel{*}{\Rightarrow} \alpha, \alpha \in N^{2+})$$

4. 三分木以上の生成規則を二分木に変換

$$X \rightarrow Y_1 Y_2 \dots Y_N$$

$$\xrightarrow{\quad} X \rightarrow Y_1 Z_1, Z_1 \rightarrow Y_1 Z_2, \dots Z_{N-2} \rightarrow Y_{N-1} Y_N$$

発展: 文脈自由文法で全ての自然言語を記述することはできるか?

- 実はスイスドイツ語やオランダ語の cross-serial dependencies は文脈自由文法で表現できない

... mer *em Hans* es *huus* *hälfed aastriiche.*
... we *Hans (dat)* *the house (acc)* *help* *paint.*

- 原理的に $a^m b^n c^m d^n$ の形になるが、これは文脈自由文法で記述できない (反復補題により証明可能; 略)

発展: では正規文法(正規表現)で自然言語を記述することはできるか?

- 中央埋め込みすらモデル化することができない

a boy imagines a girl

a boy who loves Yoko imagines a girl

a boy who loves Yoko who loves John imagines a girl

a boy who loves Yoko who loves John who loves Yoko imagines a girl

- $A \rightarrow \alpha A \beta$ のような生成規則が必要になるが、正規文法(正規表現)で表現できるのは以下のいずれかのみ

$A \rightarrow w, A \rightarrow wA$ (left-linear grammar)

or

$A \rightarrow w, A \rightarrow Aw$ (right-linear grammar)

発展: チョムスキーフェンス [Chomsky+ 1956]

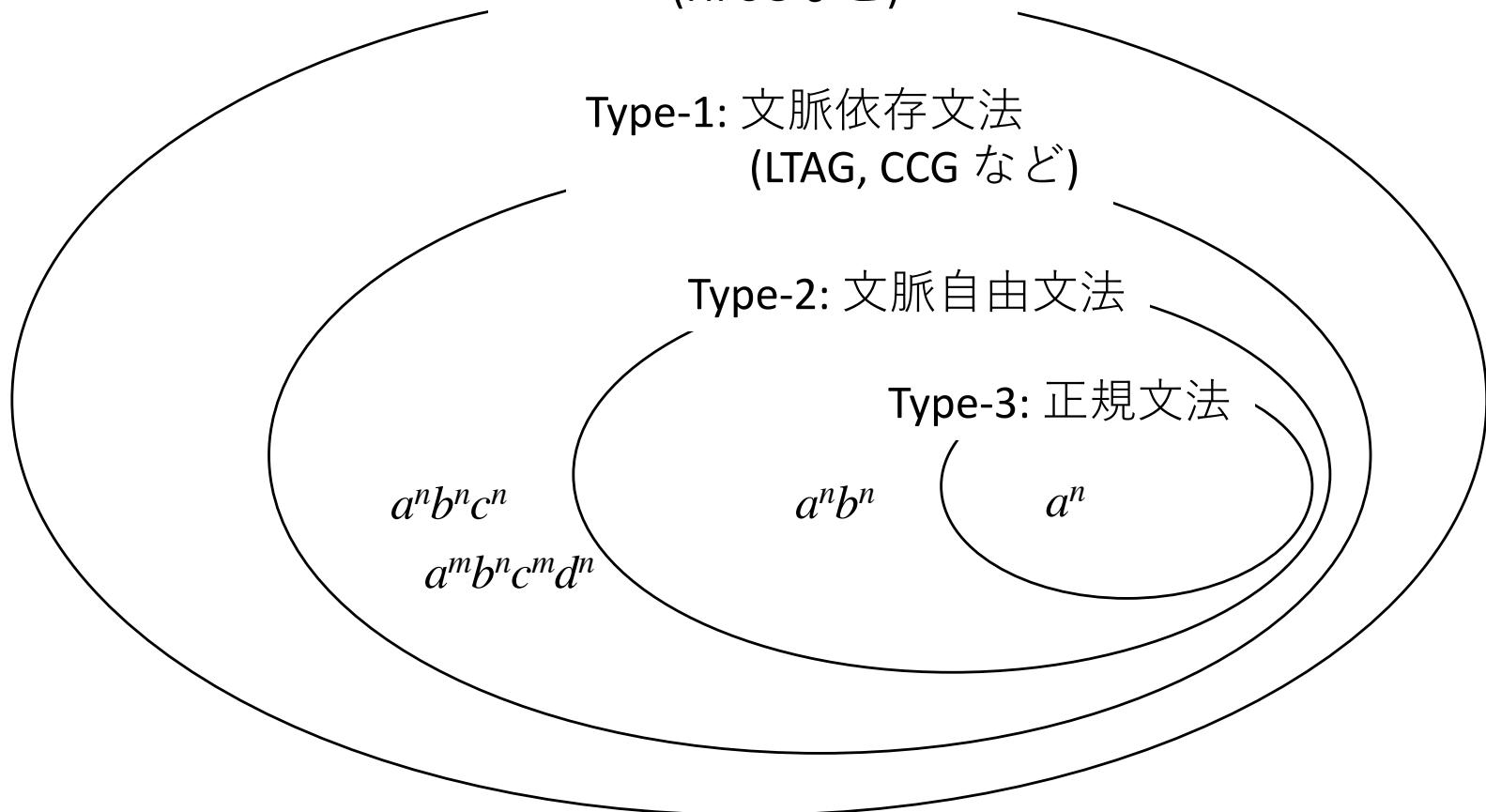
- 様々な形式文法を、対応する形式言語(生成可能な非終端記号列の集合)の包含関係で階層化したもの

Type-0: 帰納的可算言語
(HPSGなど)

Type-1: 文脈依存文法
(LTAG, CCGなど)

Type-2: 文脈自由文法

Type-3: 正規文法



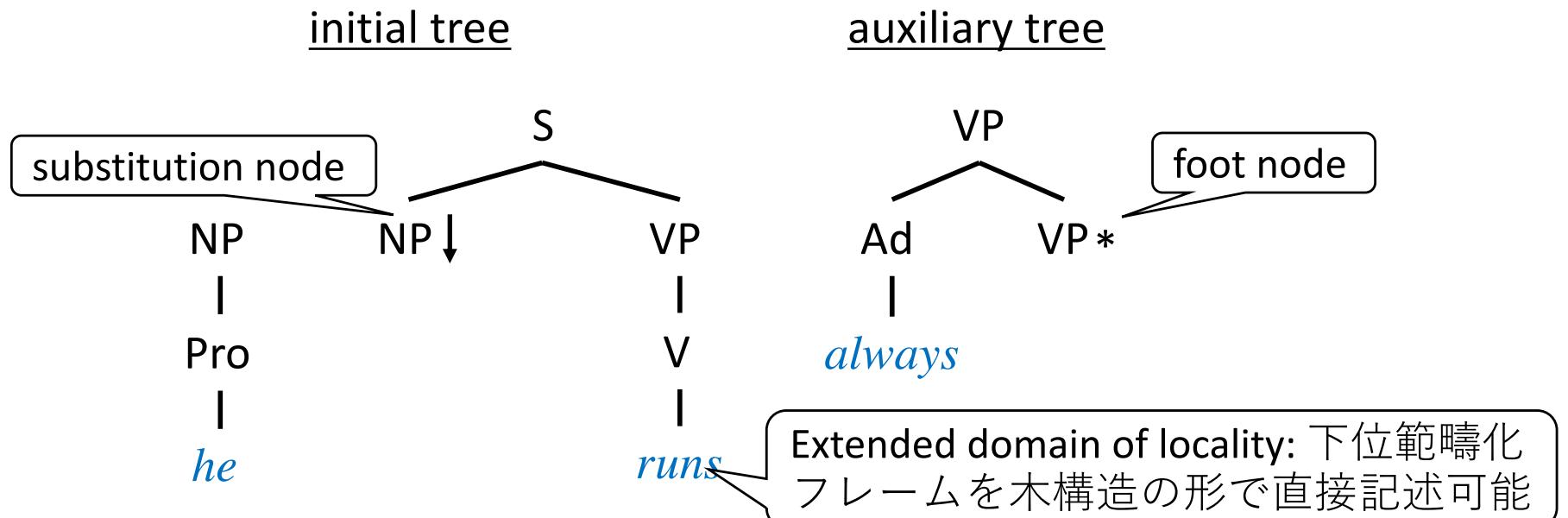
語彙化文法

- 文脈自由文法で自然言語を記述する際の問題
 - 文脈自由文法で記述できない言語現象の存在
 - (特に動詞で) 非終端記号の細分化により規則数が爆発
 - 局所的な構成性を捉えるため長距離依存はモデル化困難
- 語彙化文法: 生成規則が含む構造情報を単語の語彙に移し, 規則を単純化し語彙を複雑化
 - 多くが文脈自由文法を超える生成能力を持つ
 - 1980年代以降, 発達 (LFG [Bresnan 1982], LTAG [Schabes+ 1988], HPSG [Pollard & Sag 1994], CCG [Steedman 1996]) など

Lexicalized Tree Adjoining Grammar (LTAG; 語彙化木接合文法) [Schabes+, 1988]

- Tree Adjoining Grammar [Joshi+ 1975] を語彙化したもの
 - 基本木 (elementary tree) を代入 (substitution)・接合 (adjunction) と呼ばれる操作で組み合わせて構文木生成
 - 語彙化: 全基本木が一つ以上の終端記号(単語)を含む

語彙項目 (elementary tree)

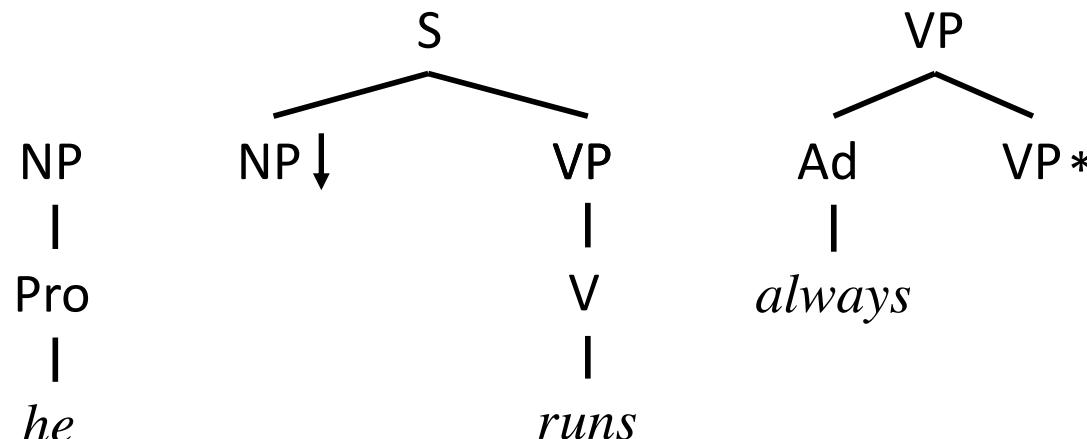


Lexicalized Tree Adjoining Grammar (LTAG; 語彙化木接合文法) [Schabes+, 1988]

- Tree Adjoining Grammar [Joshi+ 1975] を語彙化したもの
 - 基本木 (elementary tree) を代入 (substitution)・接合 (adjunction) と呼ばれる操作で組み合わせて構文木生成
 - 語彙化: 全基本木が一つ以上の終端記号(単語)を含む

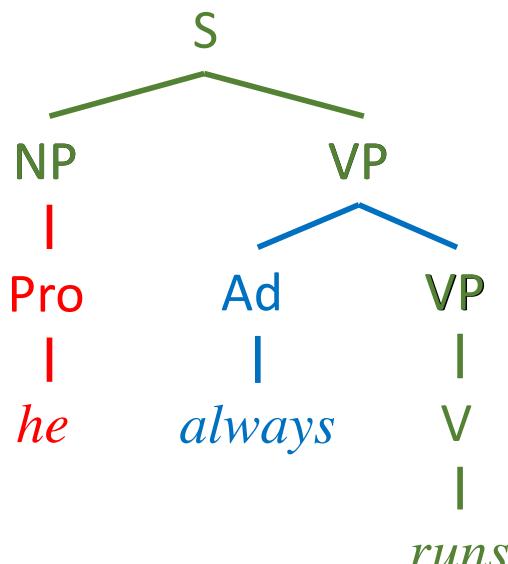
代入: substitution node を同じ非終端記号
を root に持つ基本木を代入

接合: 中間ノードに対し同じ非終端記号を
root (foot node) に持つ基本木を接合



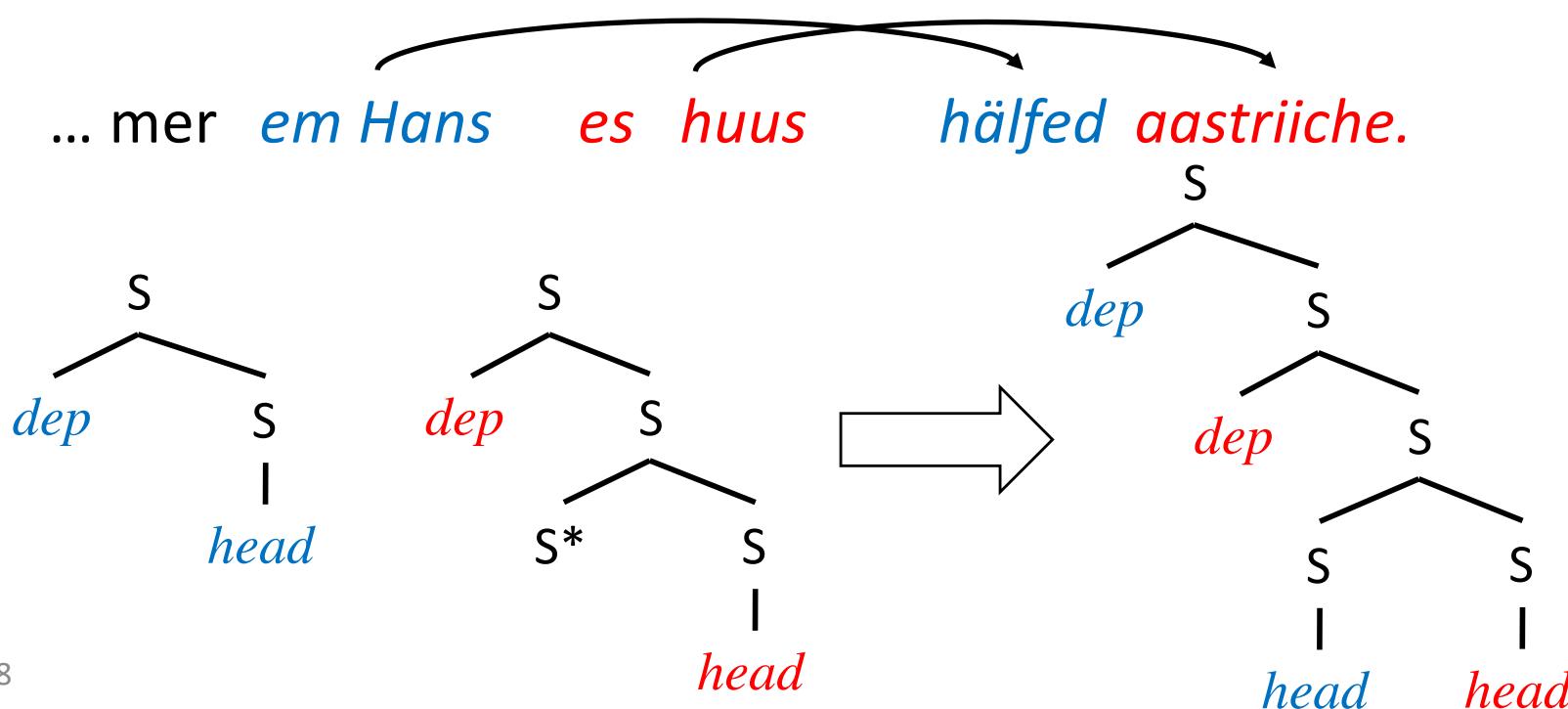
Lexicalized Tree Adjoining Grammar (LTAG; 語彙化木接合文法) [Schabes+, 1988]

- Tree Adjoining Grammar [Joshi+ 1975] を語彙化したもの
 - 基本木 (elementary tree) を代入 (substitution)・接合 (adjunction) と呼ばれる操作で組み合わせて構文木生成
 - 語彙化: 全基本木が一つ以上の終端記号(単語)を含む



Lexicalized Tree Adjoining Grammar (LTAG; 語彙化木接合文法) [Schabes+, 1988]

- Tree Adjoining Grammar [Joshi+ 1975] を語彙化したもの
 - 基本木 (elementary tree) を代入 (substitution)・接合 (adjunction) と呼ばれる操作で組み合わせて構文木生成
 - 語彙化: 全基本木が一つ以上の終端記号(単語)を含む
 - 弱文脈依存文法 (cross-serial dependencies を扱える)



CCG (Combinatorial Categorial Grammar) 組み合わせ範疇文法 [Steedman 1996]

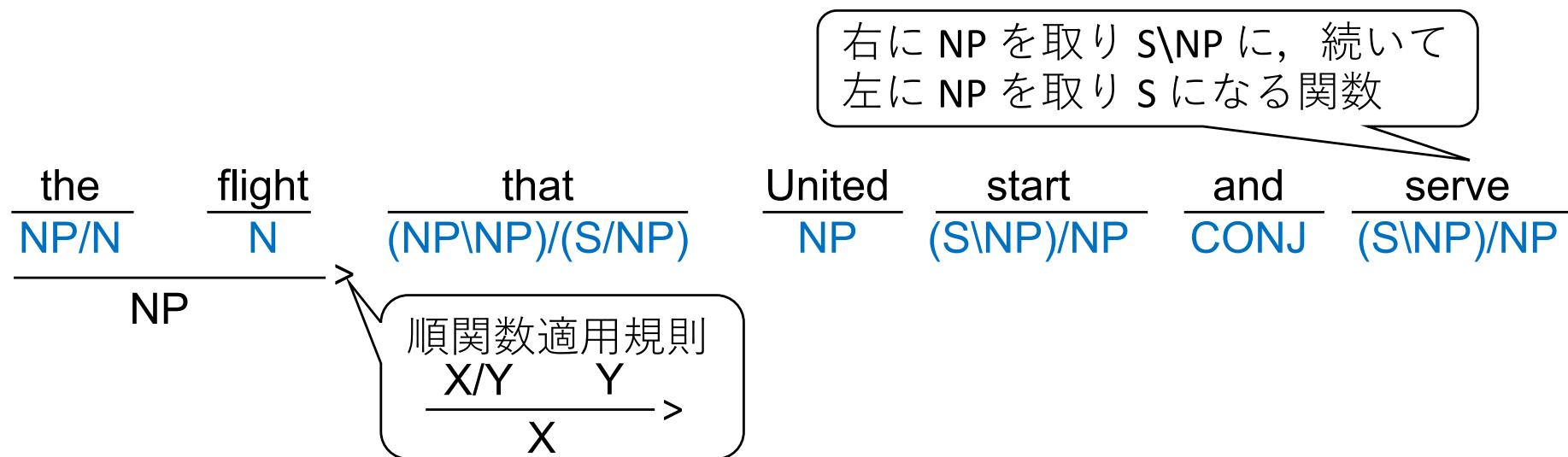
- 構造的情報を保持した統語範疇を単語に割り当て、少数の組み合わせ規則で結合して構文木を生成
 - 統語範疇: atomic な統語カテゴリまたは関数
 - 長距離依存や並列句をうまく扱うことができる

右に NP を取り S\NP に、 続いて
左に NP を取り S になる関数

the	flight	that	United	start	and	serve
NP/N	N	(NP\NP)/(S/NP)	NP	(S\NP)/NP	CONJ	(S\NP)/NP

CCG (Combinatorial Categorial Grammar) 組み合わせ範疇文法 [Steedman 1996]

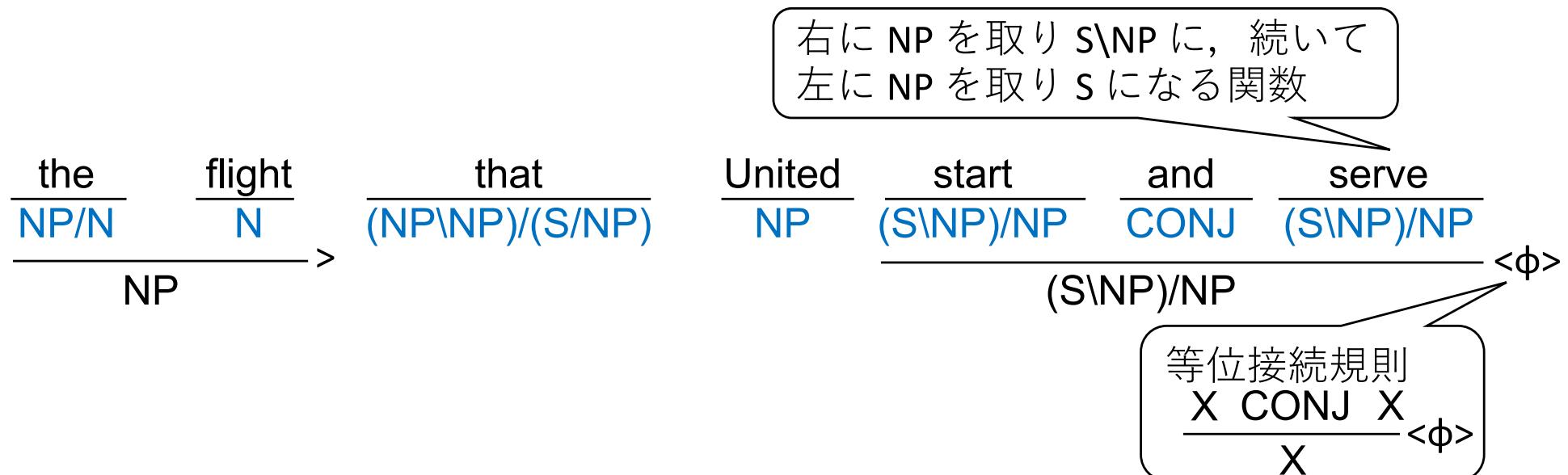
- 構造的情報を保持した統語範疇を単語に割り当て、少数の組み合わせ規則で結合して構文木を生成
 - 統語範疇: atomic な統語カテゴリまたは関数
 - 長距離依存や並列句をうまく扱うことができる



单語が右に項を取る
ときの一般規則

CCG (Combinatorial Categorial Grammar) 組み合わせ範疇文法 [Steedman 1996]

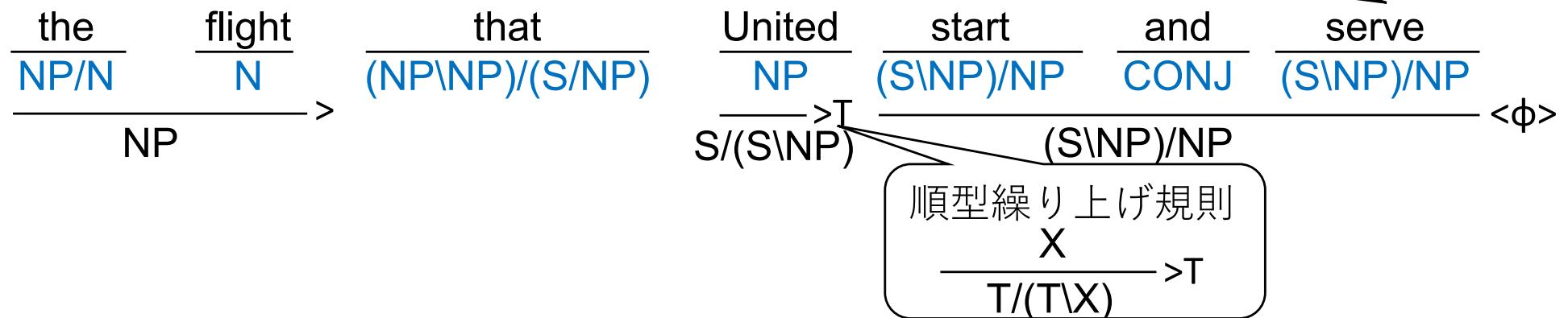
- 構造的情報を保持した統語範疇を単語に割り当て、少数の組み合わせ規則で結合して構文木を生成
 - 統語範疇: atomic な統語カテゴリまたは 関数
 - 長距離依存や並列句をうまく扱うことができる



CCG (Combinatorial Categorial Grammar) 組み合わせ範疇文法 [Steedman 1996]

- 構造的情報を保持した統語範疇を単語に割り当て、少数の組み合わせ規則で結合して構文木を生成
 - 統語範疇: atomic な統語カテゴリまたは関数
 - 長距離依存や並列句をうまく扱うことができる
 - 弱文脈依存文法

右に NP を取り S\NP に、 続いて
左に NP を取り S になる関数

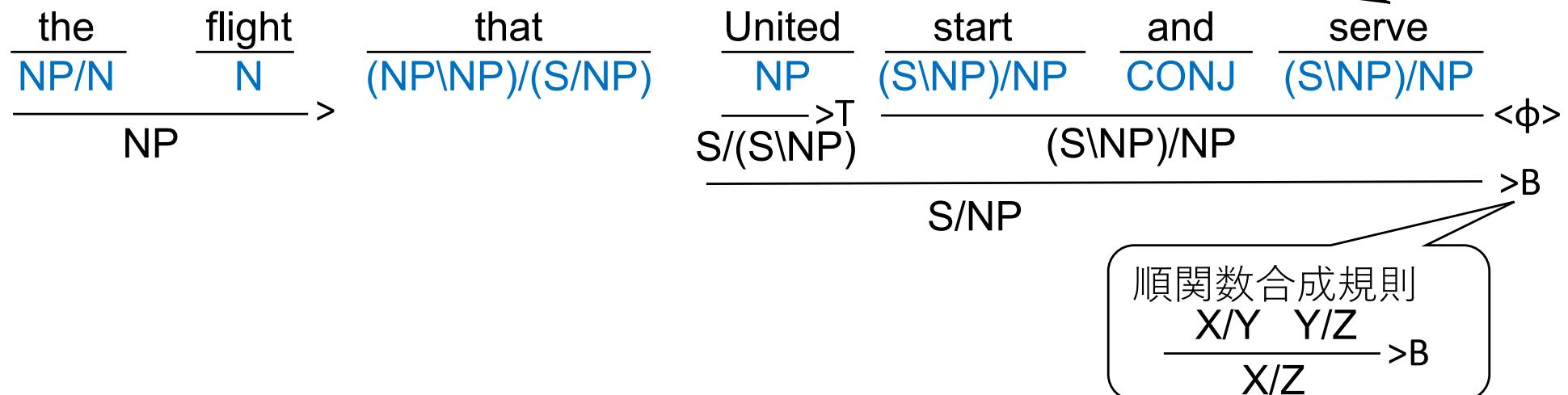


関係節で目的語が離れているため
局所的に組み上げることできない

CCG (Combinatorial Categorial Grammar) 組み合わせ範疇文法 [Steedman 1996]

- 構造的情報を保持した統語範疇を単語に割り当て、少数の組み合わせ規則で結合して構文木を生成
 - 統語範疇: atomic な統語カテゴリまたは関数
 - 長距離依存や並列句をうまく扱うことができる
 - 弱文脈依存文法

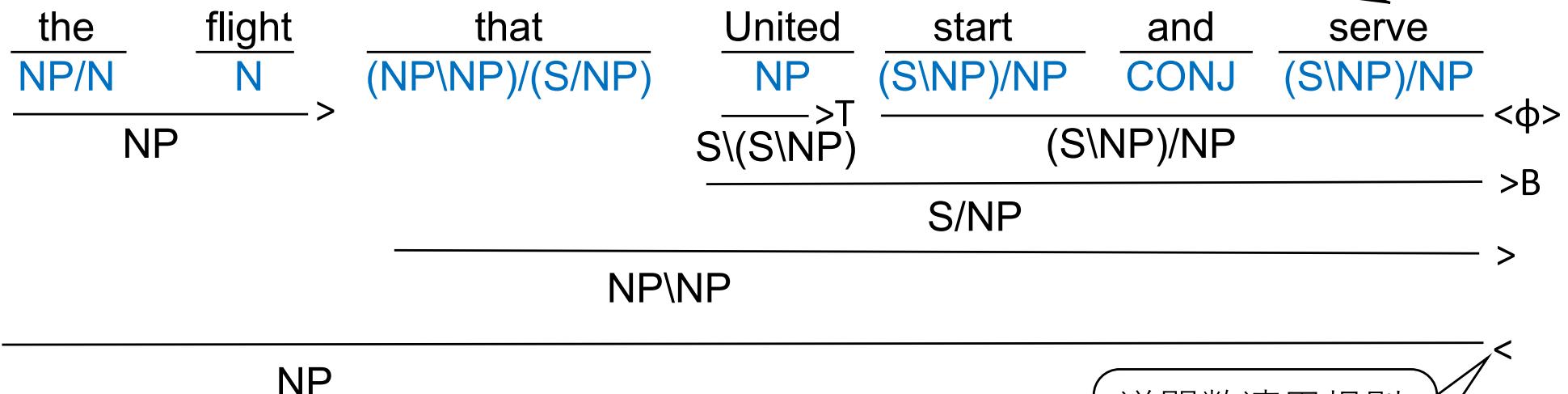
右に NP を取り S\NP に、 続いて
左に NP を取り S になる関数



CCG (Combinatorial Categorial Grammar) 組み合わせ範疇文法 [Steedman 1996]

- 構造的情報を保持した統語範疇を単語に割り当て、少数の組み合わせ規則で結合して構文木を生成
 - 統語範疇: atomic な統語カテゴリまたは関数
 - 長距離依存や並列句をうまく扱うことができる
 - 弱文脈依存文法

右に NP を取り S\NP に、 続いて
左に NP を取り S になる関数



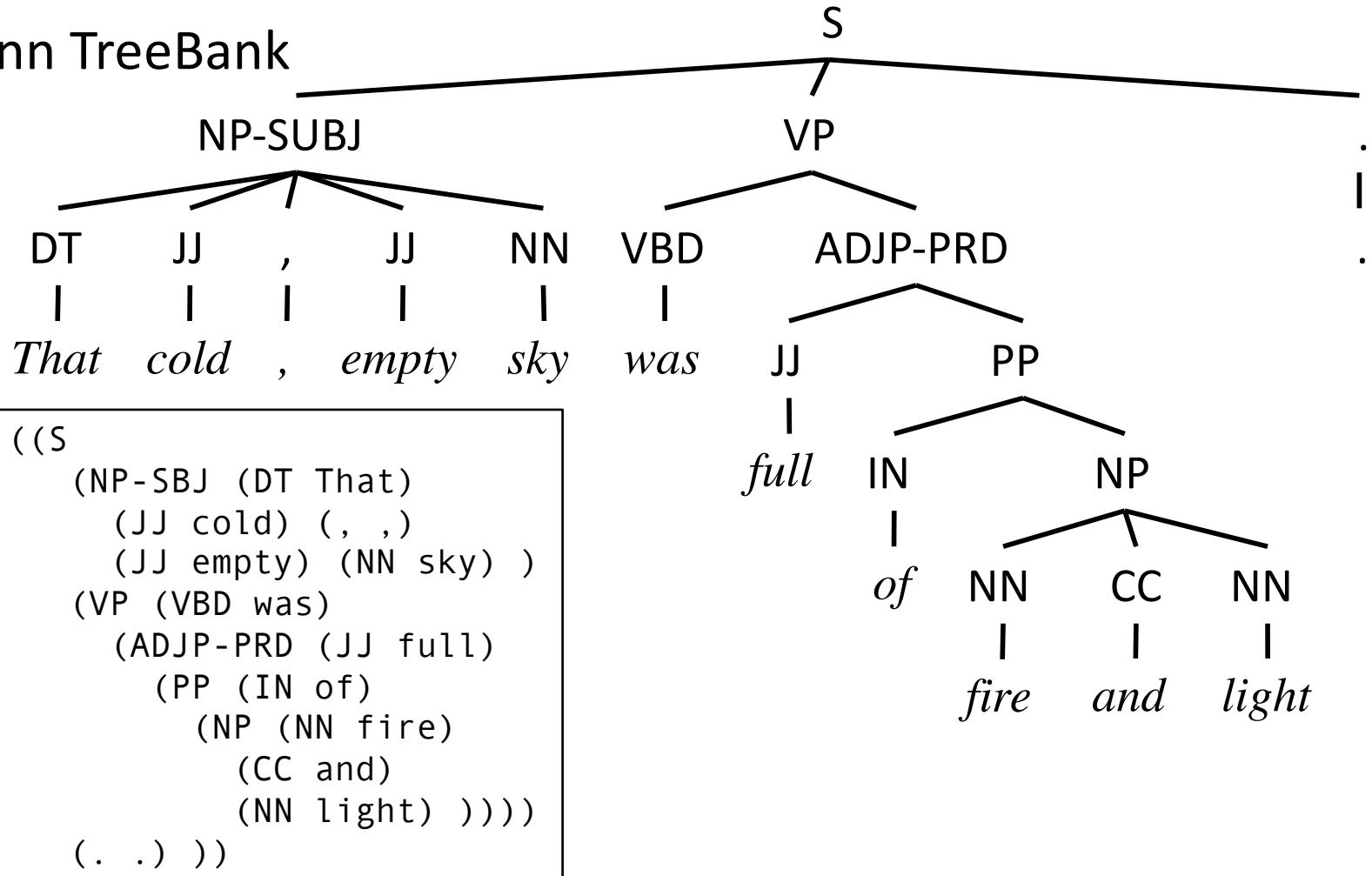
自然言語を漏れなくカバーする大規模文法を どう構築すればよいか？

- 理想的な文法
 - 適格な自然言語文のみを生成し、非文は生成しない
 - 文には構文的に許される最小限の句構造木のみを与える
- 人手の文法開発の難しさ
 - 一貫性を保って自然言語を網羅する大規模文法を人手で開発するのは困難 (生成規則 or 語彙項目が爆発)
**過剰生成を抑えつつカバレッジを上げるのには限界
(合理主義的方法論の限界)**
- 実テキストを解析できる形式文法を構築することは不可能か？

ツリーバンク文法: 純経験主義的文法開発

- CFG: ツリーバンクを分割することで文法抽出

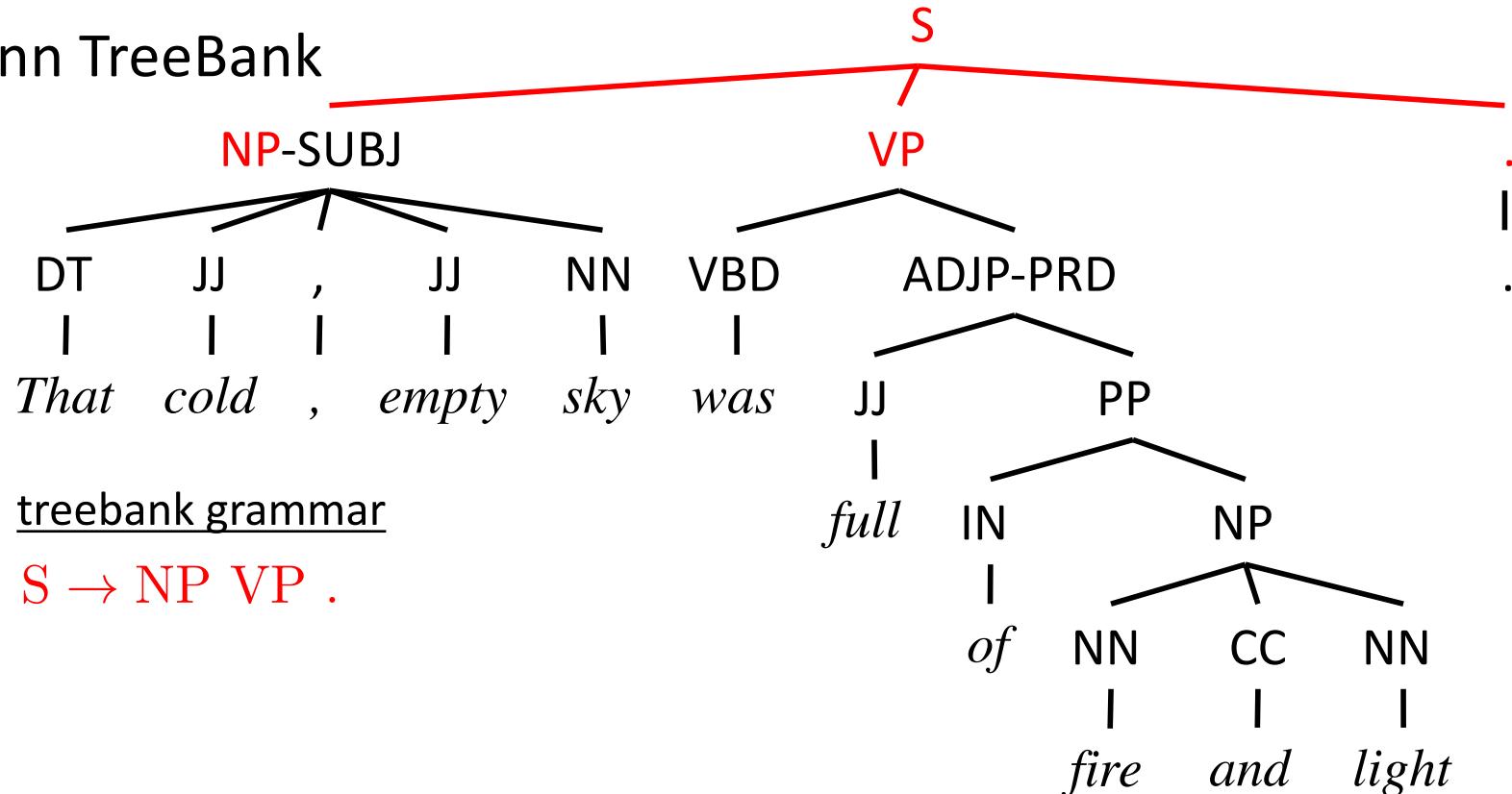
例) Penn TreeBank



ツリーバンク文法: 純経験主義的文法開発

- CFG: ツリーバンクを分割することで文法抽出

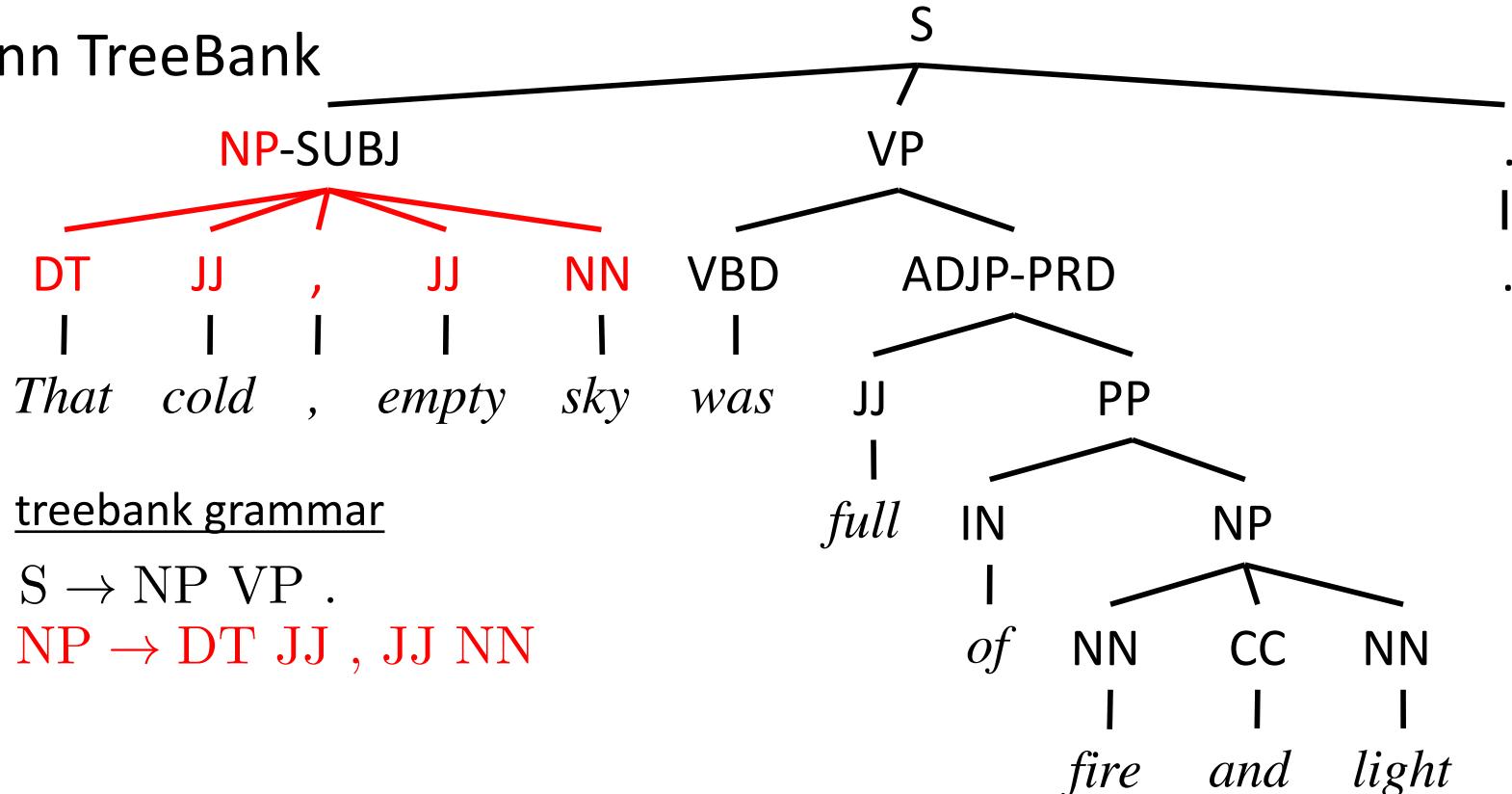
例) Penn TreeBank



ツリーバンク文法: 純経験主義的文法開発

- CFG: ツリーバンクを分割することで文法抽出

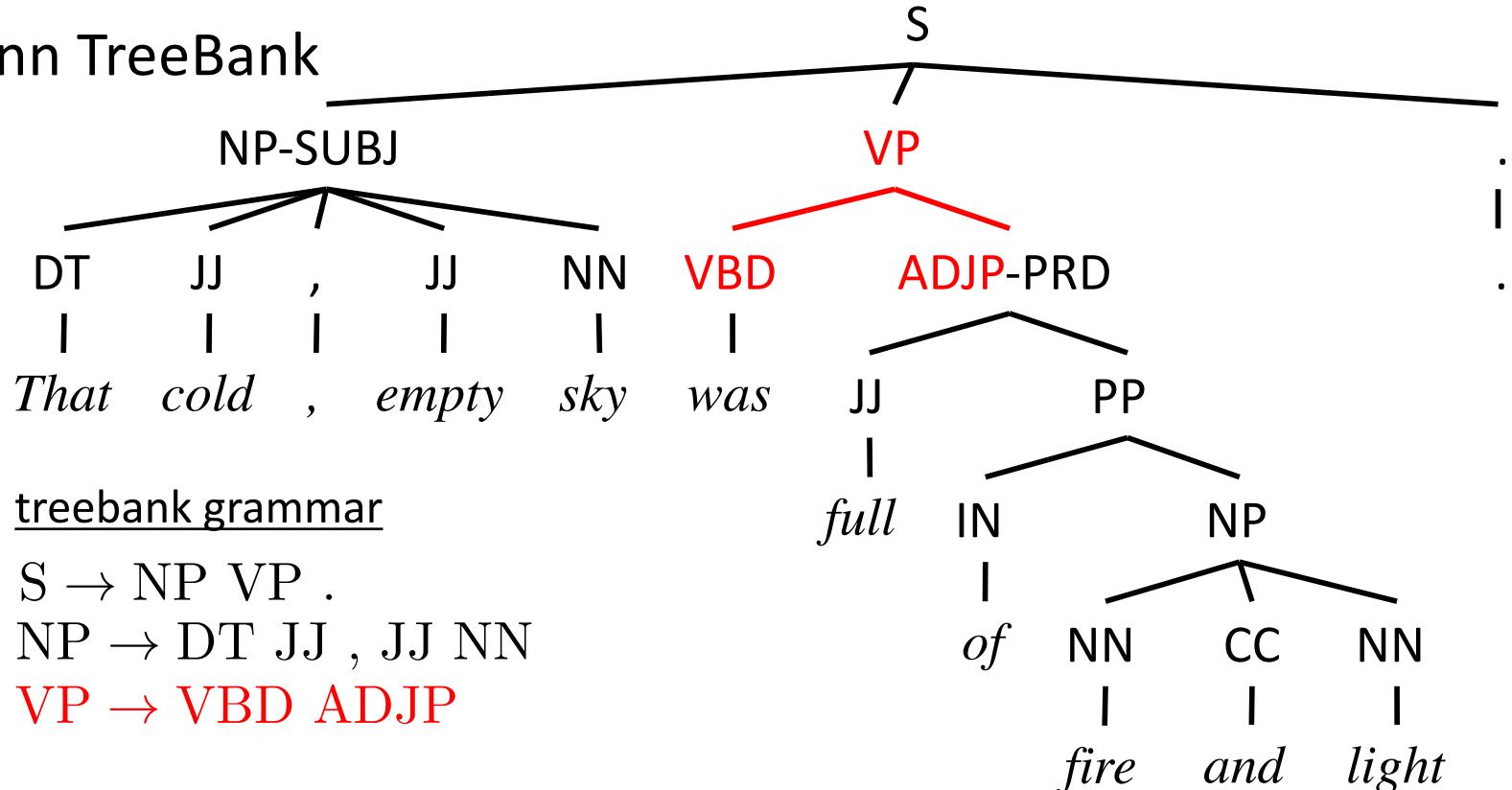
例) Penn TreeBank



ツリーバンク文法: 純経験主義的文法開発

- CFG: ツリーバンクを分割することで文法抽出

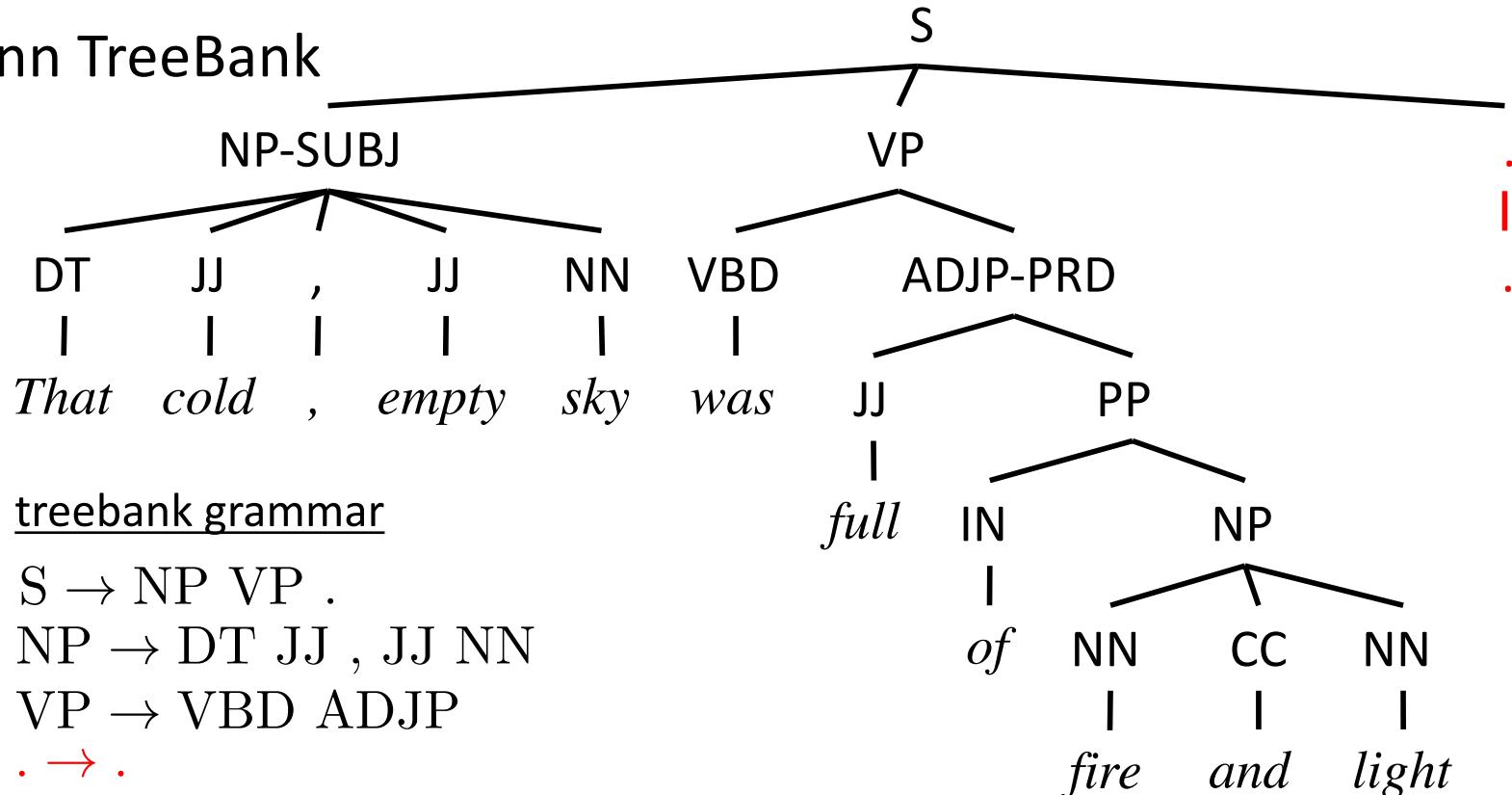
例) Penn TreeBank



ツリーバンク文法: 純経験主義的文法開発

- CFG: ツリーバンクを分割することで文法抽出

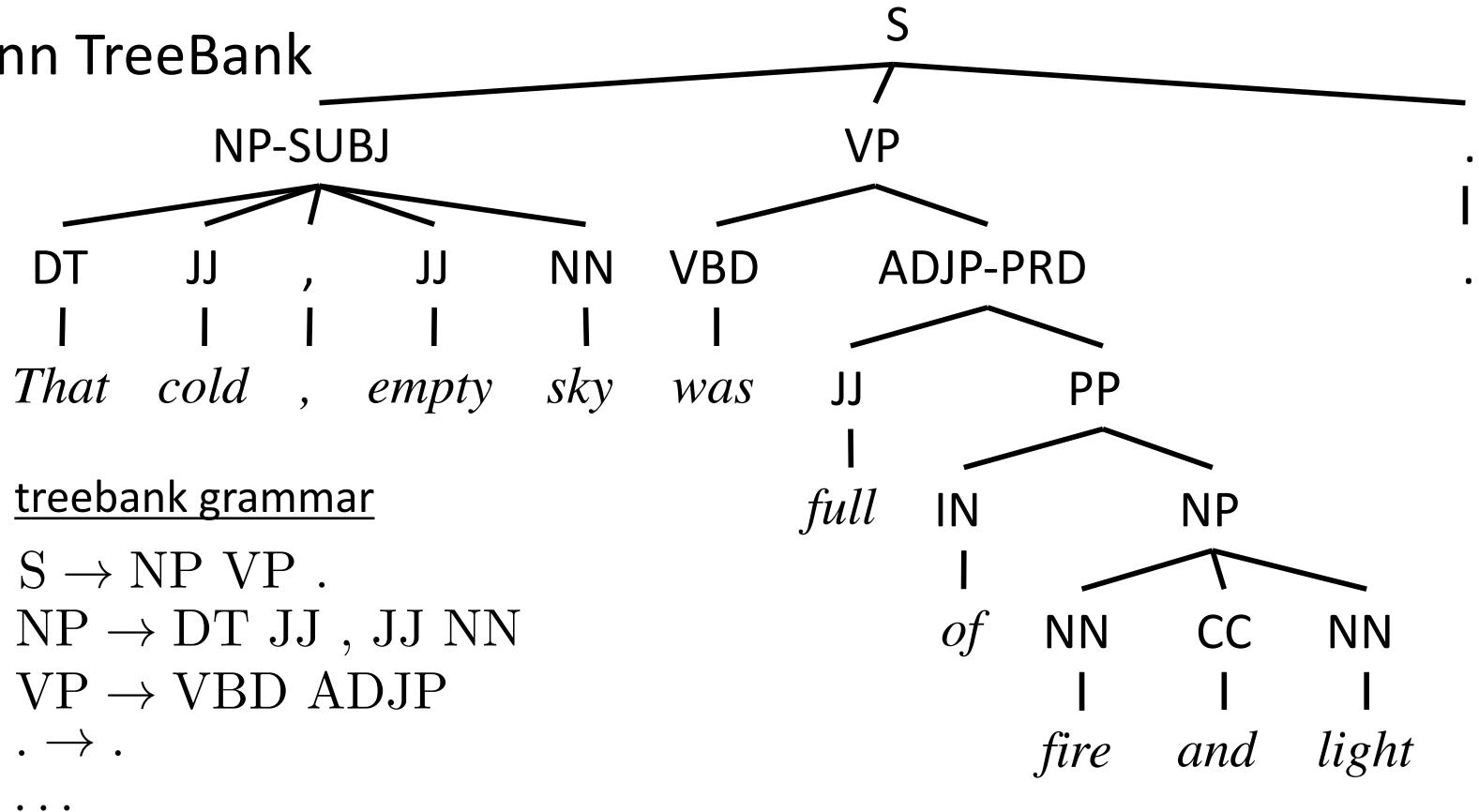
例) Penn TreeBank



ツリーバンク文法: 純経験主義的文法開発

- CFG: ツリーバンクを分割することで文法抽出

例) Penn TreeBank



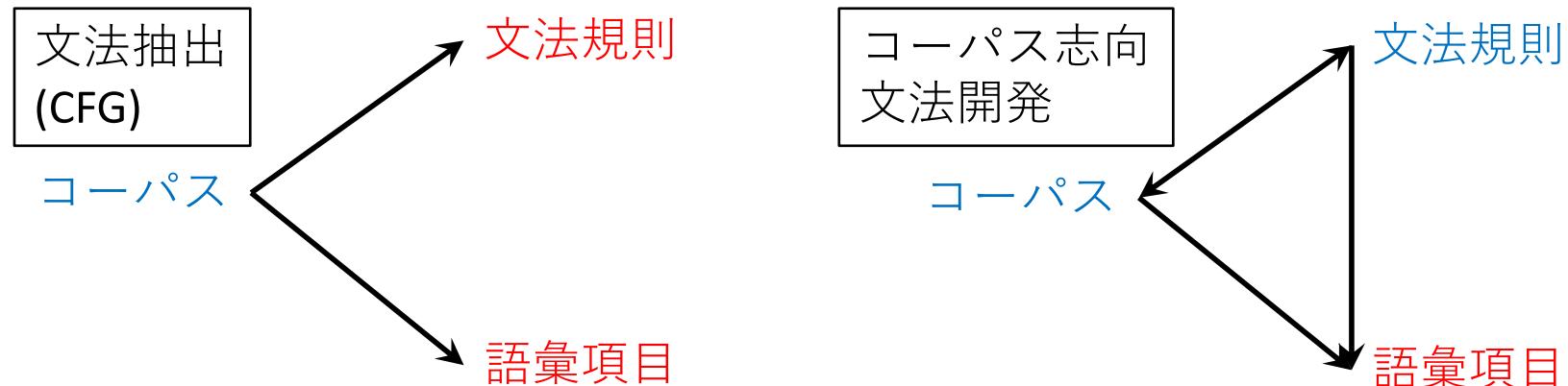
コーパスを学習・テストに分けて、前者から文法抽出した場合、後者中の文もほとんど解析可能

ツリーバンク文法: 純経験主義的文法開発

- CFG: ツリーバンクを分割することで文法抽出
- 語彙化文法: 句構造 TreeBank に(既定の)語彙化文法の文法規則を逆適用することで語彙項目を獲得
 - 構文木 = 語彙項目×文法規則 (少量, 既定)
 - 文法規則適用の曖昧性は, 語彙化文法の文法理論特有の注釈付けを補助的に行い解消

発展: コーパス志向文法開発 ～ツリーバンク文法の限界を超える～

- ツリーバンク文法の問題点
 - TreeBank 作成時に想定した言語理論が不完全で学習する形式文法の仮定する言語理論と矛盾 (特に語彙化文法)
- 語彙化文法のための **コーパス志向文法開発** [Miyao+ 2004]
抽出した語彙化文法を用いて TreeBank を検証することで、**TreeBanking** と文法理論の精緻化を繰り返す
- CCGbank [Hockenmier+ 2002], HPSG [Miyao+ 2004] など



形式文法に基づく句構造解析

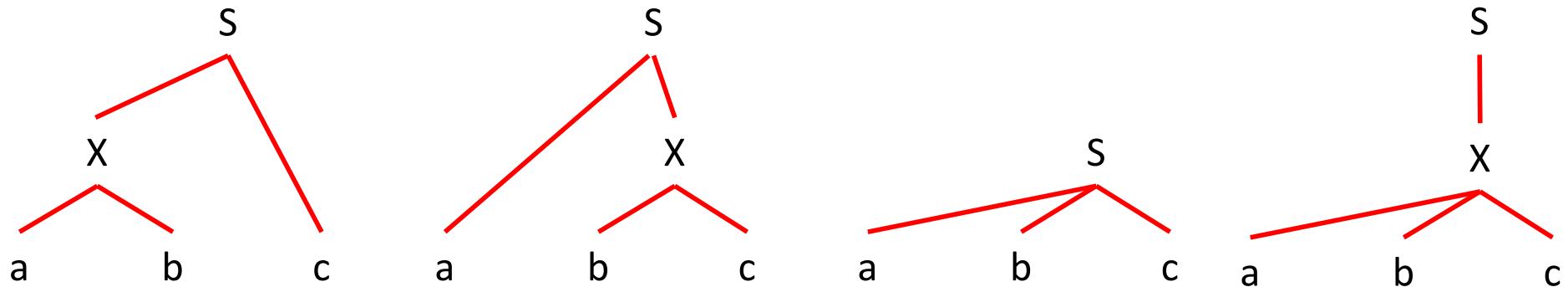
- 形式文法に基づく3つのタスク
 - 文の適格性判定 (**Recognition**): 文法を用いて、与えられた文を導出できるかを判定
 - 構文解析 (**Parsing**): 文法を用いて、与えられた文を導出する開始記号を根とする全ての構文木を出力
 - 文生成 (**Generation**): 文法を用いて、与えられた意味構造を満たすような文を生成する(本講義では扱わない)

形式文法に基づく句構造解析

- 形式文法に基づく3つのタスク
 - 文の適格性判定 (**Recognition**): 文法を用いて、与えられた文を導出できるかを判定
 - 構文解析 (**Parsing**): 文法を用いて、与えられた文を導出する開始記号を根とする全ての構文木を出力
 - どちらも与えられた文を導出する構文木が存在するかを文法を用いて計算する

構文解析の難しさ

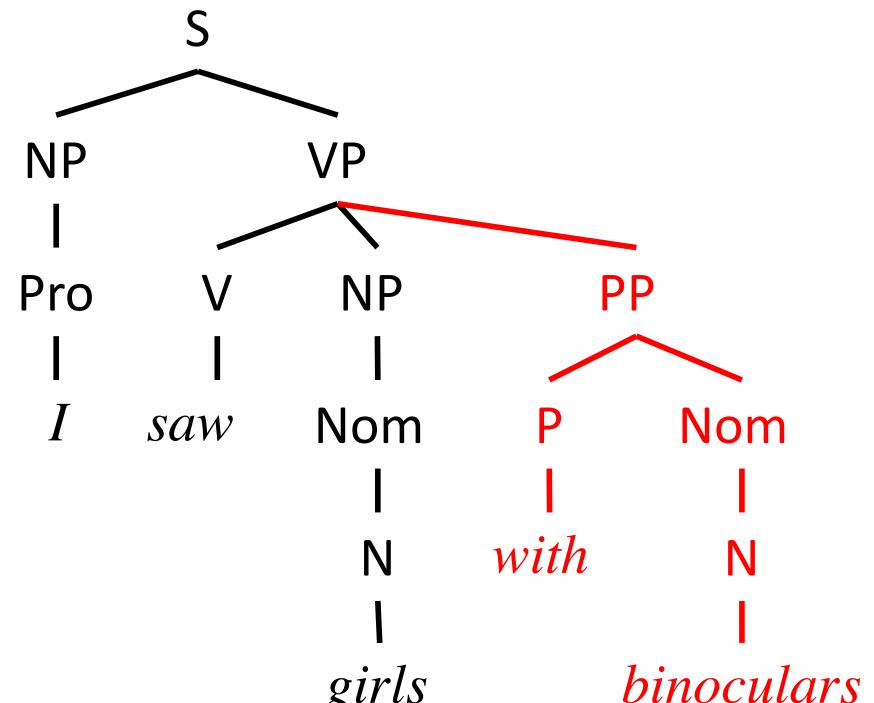
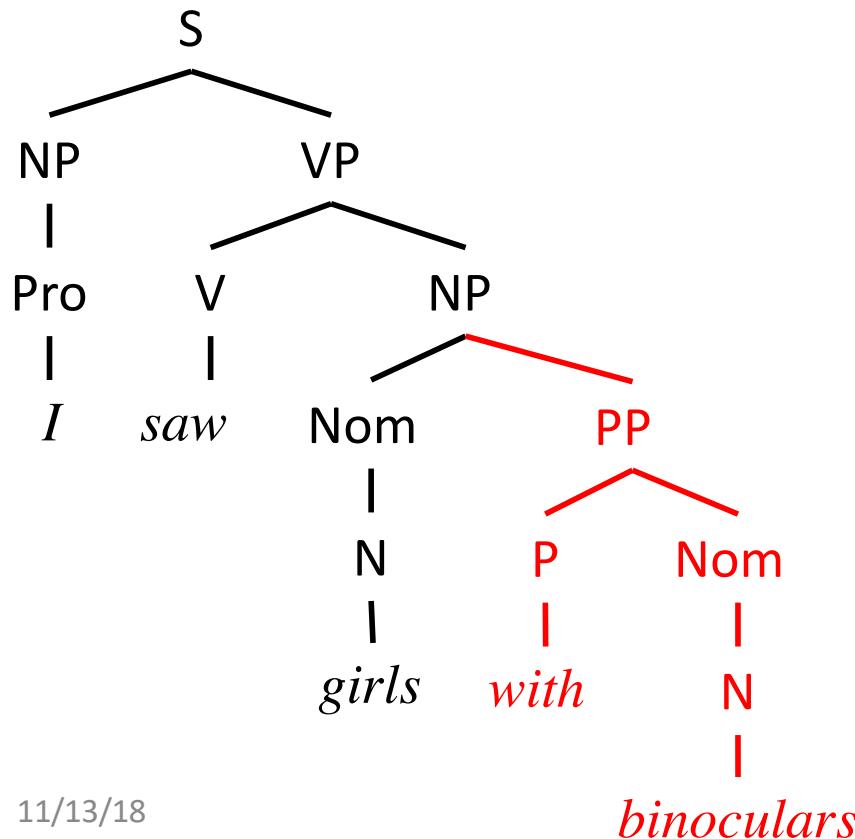
- 文法で制限したとしても、文が取りうる構文木の可能性は部分構造の曖昧性により指数的に爆発
 - 特にカバレッジを重視するツリーバンク文法では深刻



部分構造の曖昧性 (1/2): Attachment ambiguity

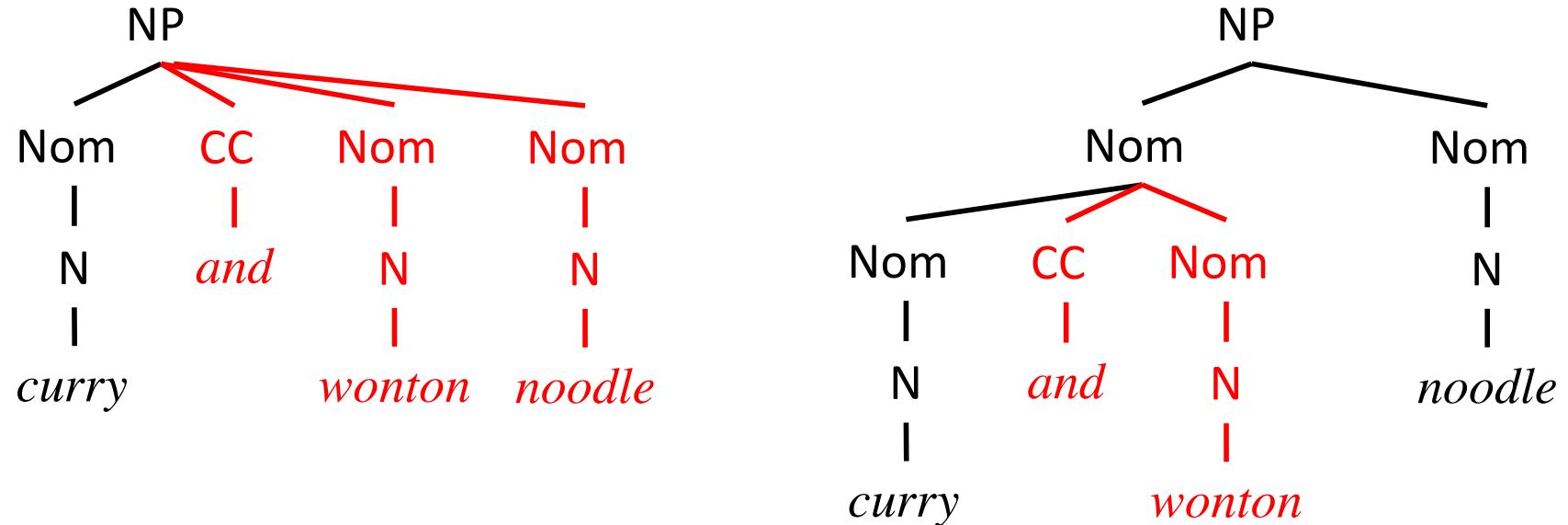
- ある句が修飾する句の候補には曖昧性がある

PP attachment (NP か VP か?)



部分構造の曖昧性 (1/2): Coordination ambiguity

- 並列句の並列範囲の曖昧性



これらの曖昧性は、必ずしも統語的には解消できない場合があり最終的には統計的手法により曖昧性解消する必要がある

Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

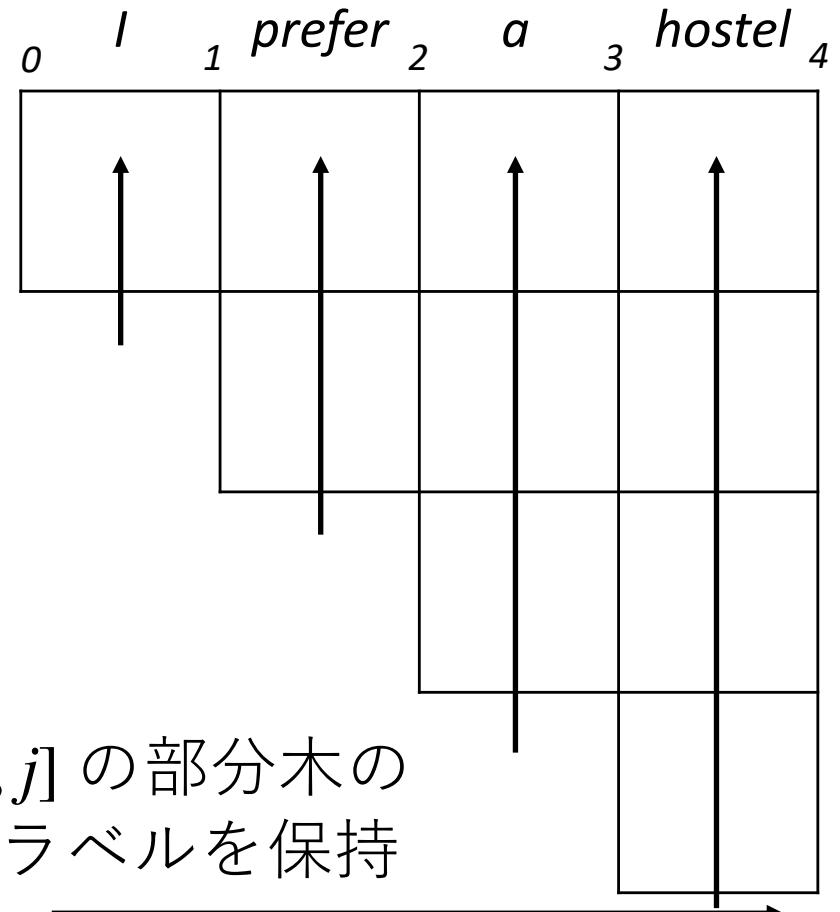
$NP \rightarrow Det\ Nom$

$NP \rightarrow Pro$

$Nom \rightarrow N$

$S : S$

$Pro \rightarrow I$
 $V \rightarrow prefer$
 $Det \rightarrow a$
 $N \rightarrow hostel$



Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法は チョムスキイ標準形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

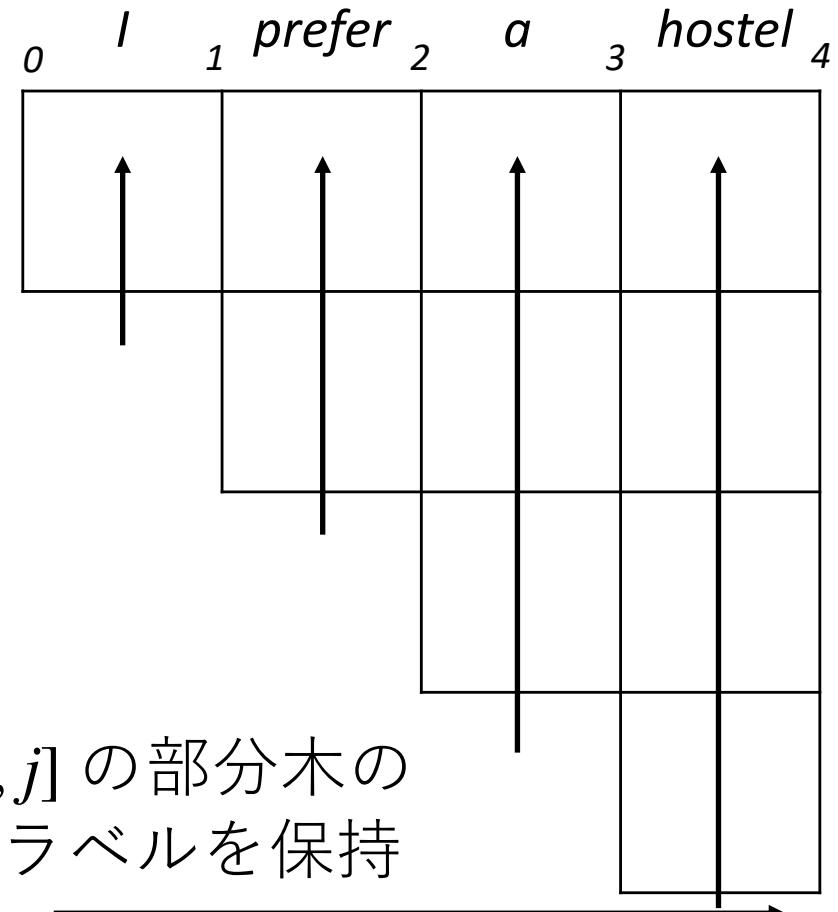
$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

$Pro \rightarrow I$
 $V \rightarrow prefer$
 $Det \rightarrow a$
 $N \rightarrow hostel$



Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーライントークン標準形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

N : S, NP, VP, Pro, V, Det, Nom, N

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

VP → V NP

NP \Rightarrow Det. Nom

NP \rightarrow I

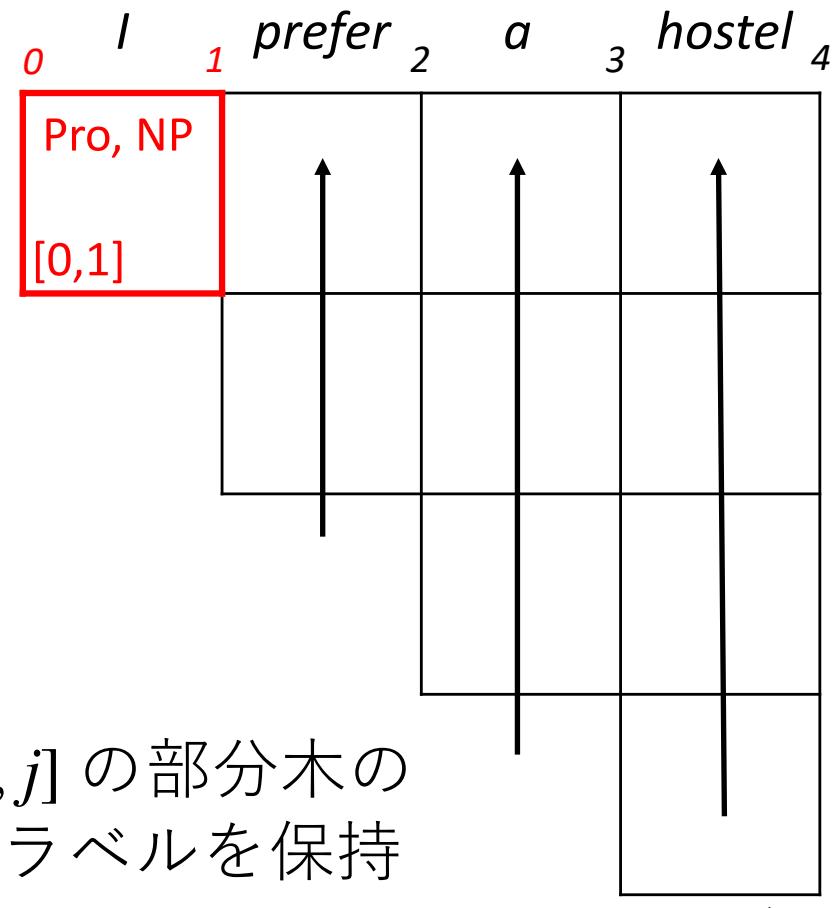
Nom $\rightarrow hostell$

S. S.

Pro → *I*
V → *prefer*
Det → *a*
N → *hostel*

CKY 表

各セルは $[i,j]$ の部分木の根ノードのラベルを保持



Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

$Pro \rightarrow I$

$V \rightarrow prefer$

$Det \rightarrow a$

$N \rightarrow hostel$

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

I	$prefer$	a	$hostel$
0	Pro, NP [0,1]		
1		V [1,2]	
2			
3			
4			

Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

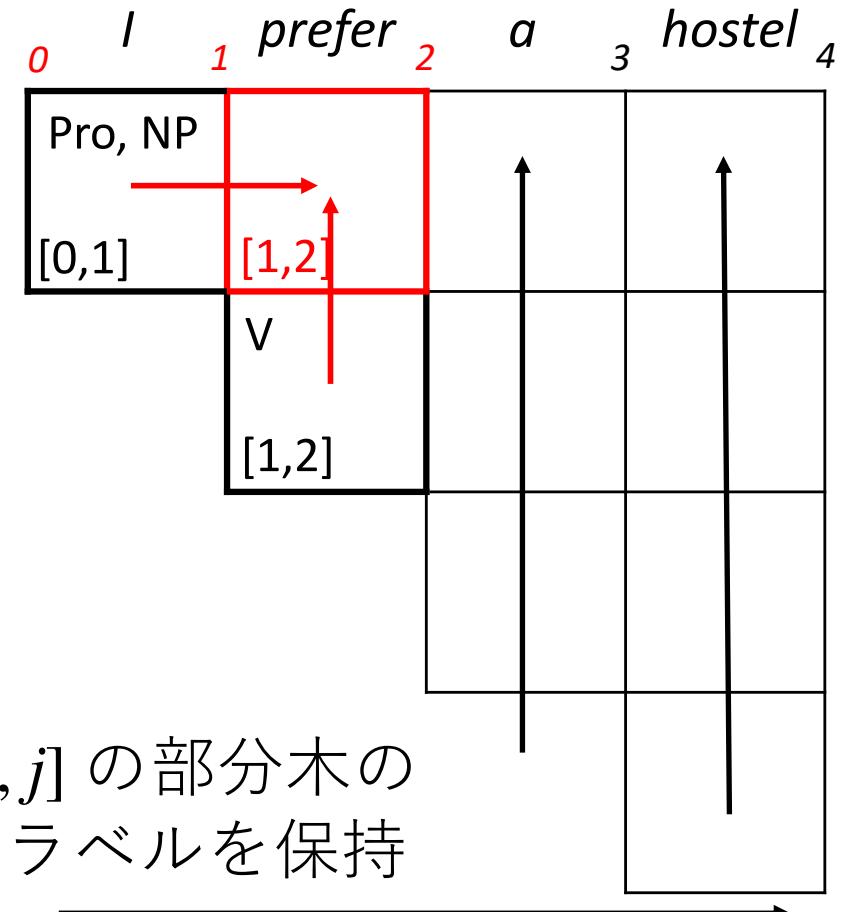
$Nom \rightarrow hostel$

$S : S$

$Pro \rightarrow I$ $V \rightarrow prefer$ $Det \rightarrow a$ $N \rightarrow hostel$
--

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持



Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

$Pro \rightarrow I$ $V \rightarrow prefer$ $Det \rightarrow a$ $N \rightarrow hostel$	I $prefer$ a $hostel$
--	------------------------------------

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

I	$prefer$	a	$hostel$
Pro, NP			
$[0,1]$	$[0,2]$		
	V		
	$[1,2]$		
		Det	
		$[2,3]$	

Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

$Pro \rightarrow I$ $V \rightarrow prefer$ $Det \rightarrow a$ $N \rightarrow hostel$
--

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

I	$prefer$	a	$hostel$
Pro, NP [0,1]			
	V [1,2]		
		Det [1,3]	
			[2,3]

Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

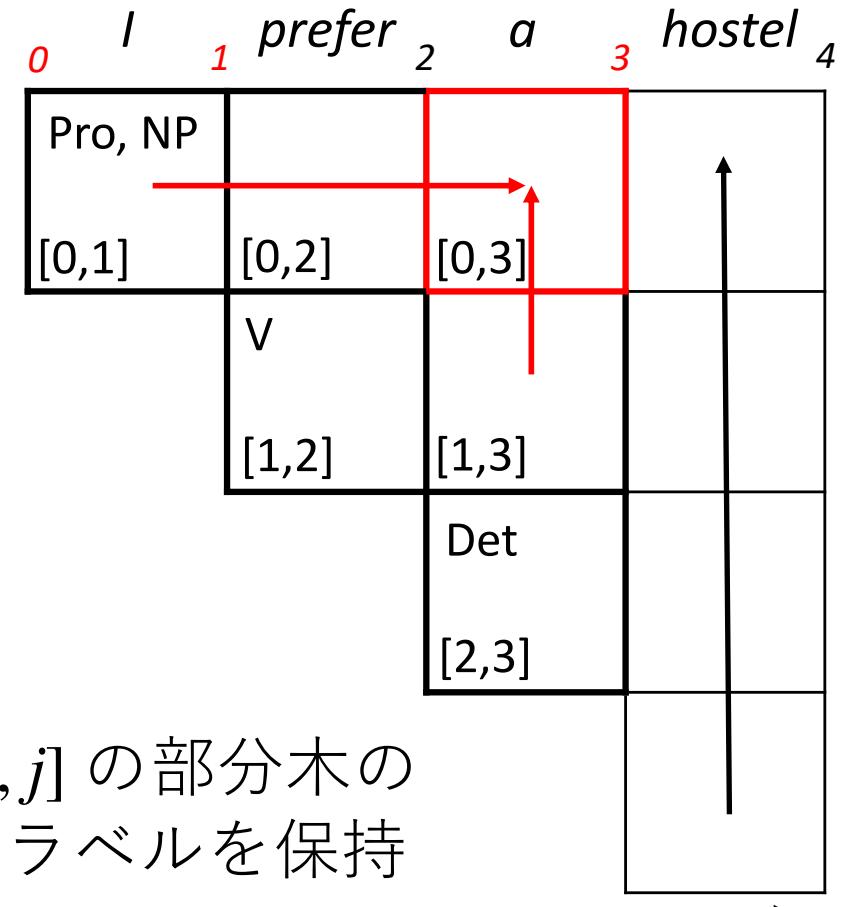
$S : S$

$Pro \rightarrow I$ $V \rightarrow prefer$ $Det \rightarrow a$ $N \rightarrow hostel$	Pro, NP $[0,1]$ V $[1,2]$ Det $[2,3]$
--	--

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

I 0	$prefer$ 1	a 2	$hostel$ 3	
Pro, NP				
$[0,1]$	$[0,2]$	$[0,3]$		
	V			
	$[1,2]$	$[1,3]$		
			Det	
			$[2,3]$	



Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

$Pro \rightarrow I$ $V \rightarrow prefer$ $Det \rightarrow a$ $N \rightarrow hostel$	I $prefer$ a $hostel$
--	------------------------------------

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

I	$prefer$	a	$hostel$
Pro, NP			
$[0,1]$	$[0,2]$	$[0,3]$	
V			
$[1,2]$	$[1,3]$		
Det			
$[2,3]$			

Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

$Pro \rightarrow I$
 $V \rightarrow prefer$
 $Det \rightarrow a$
 $N \rightarrow hostel$

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

I $[0,1]$	$prefer$ $[0,2]$	a $[0,3]$	$hostel$ $[1,2]$	$[1,3]$	Det $[2,3]$	N, Nom $[3,4]$
Pro, NP			V			

Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーライントルク形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

$Pro \rightarrow I$ $V \rightarrow prefer$ $Det \rightarrow a$ $N \rightarrow hostel$	I $prefer$ a $hostel$
--	------------------------------------

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

I $[0,1]$	$prefer$ $[0,2]$	a $[0,3]$	$hostel$
Pro, NP $[0,1]$			
V $[1,2]$		$[1,3]$	
Det $[2,3]$	NP $[2,4]$		N, Nom $[3,4]$

Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

$Pro \rightarrow I$
 $V \rightarrow prefer$
 $Det \rightarrow a$
 $N \rightarrow hostel$

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

I	$prefer$	a	$hostel$
Pro, NP			
$[0,1]$	$[0,2]$	$[0,3]$	
V			VP
$[1,2]$	$[1,3]$		$[1,4]$
	Det		NP
	$[2,3]$		N, Nom
			$[3,4]$

Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

$Pro \rightarrow I$ $V \rightarrow prefer$ $Det \rightarrow a$ $N \rightarrow hostel$	I $prefer$ a $hostel$
--	------------------------------------

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

I $[0,1]$	$prefer$ $[0,2]$	a $[0,3]$	$hostel$ $[1,4]$
Pro, NP [0,1]			
	V [1,2]		VP [1,3]
		Det [2,3]	NP [2,4]
			N, Nom [3,4]

Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

Pro $\rightarrow I$
V $\rightarrow prefer$
Det $\rightarrow a$
N $\rightarrow hostel$

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

I 0	$prefer$ 1	a 2	$hostel$ 3	
Pro, NP			S	
[0,1]	[0,2]	[0,3]	[0,4]	
	V		VP	
	[1,2]	[1,3]	[1,4]	
		Det	NP	
		[2,3]	[2,4]	
			N,Nom	
			[3,4]	

Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

Pro $\rightarrow I$	V $\rightarrow prefer$	Det $\rightarrow a$	N $\rightarrow hostel$
---------------------	------------------------	---------------------	------------------------

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

I $[0,1]$	$prefer$ $[0,2]$	a $[0,3]$	$hostel$ $[0,4]$
Pro, NP [0,1]			S [0,4]
	V [1,2]		VP [1,4]
		Det [2,3]	NP [2,4]
			N, Nom [3,4]

Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

Pro $\rightarrow I$	V $\rightarrow prefer$	Det $\rightarrow a$	N $\rightarrow hostel$
---------------------	------------------------	---------------------	------------------------

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

I 0	$prefer$ 1	a 2	$hostel$ 3	
Pro, NP [0,1]			S [0,4]	
	V [1,2]		VP [1,4]	
		Det [2,3]	NP [2,4]	
			N, Nom [3,4]	

Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

$N : S, NP, VP, Pro, V, Det, Nom, N$

$\Sigma : I, prefer, a, hostel$

$R : S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow Det\ Nom$

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

$Pro \rightarrow I$

$V \rightarrow prefer$

$Det \rightarrow a$

$N \rightarrow hostel$

時間計算量: $O(T^3)$

CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

I	$prefer$	a	$hostel$
Pro, NP			S
$[0,1]$	$[0,2]$	$[0,3]$	$[0,4]$
V			VP
	$[1,2]$	$[1,3]$	$[1,4]$
		Det	NP
		$[2,3]$	$[2,4]$
			N, Nom
			$[3,4]$

Cocke-Kasami-Younger (CKY) アルゴリズム CFG に基づく句構造解析

- 動的計画法に基づき部分構文木の計算を纏め上げ
 - 文法はチョムスキーラベル形に変換して処理
 - CKY 表と呼ばれる2次元配列に部分結果を格納

N : S, NP, VP, Pro, V, Det, Nom, N

Σ : $I, prefer, a, hostel$

R : $S \rightarrow NP\ VP \quad | \quad Pro \rightarrow I$

VP - バックポインタを保持する

NP - ことで構文木を復元可能

$NP \rightarrow I$

$Nom \rightarrow hostel$

$S : S$

時間計算量: $O(T^3)$

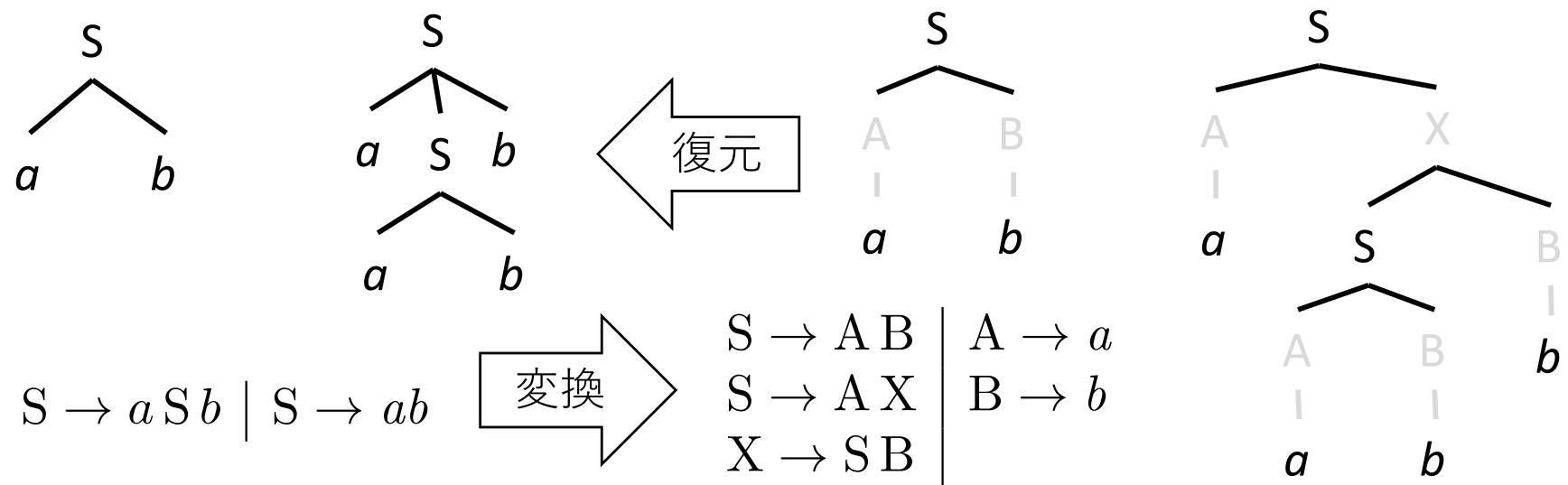
CKY 表

各セルは $[i, j]$ の部分木の根ノードのラベルを保持

0	I	1	$prefer$	2	a	3	$hostel$	4
Pro, NP [0,1]						S [0,4]		
					V [1,2]	VP [1,3]	VP [1,4]	
						Det [2,3]	NP [2,4]	
							N, Nom [3,4]	

CNF変換前の文法の構文木の復元

- ・ チョムスキ－標準形変換後の文法による解析結果は容易に元の文法の解析結果に変換可能
 - ・ 変換時のダミー シンボルに関する分岐を削除



- ・ 単位規則は CKY アルゴリズムで直接扱えるようになるのが都合が良い

形式文法の句構造解析器としての評価

- 文に対して正解の構文木が与えられた注釈付きコーパスを利用して、以下の指標で評価
 - 被覆率: 構文木を生成できた文の割合
 - 被覆率(強意): 正解の構文木を生成できた文の割合
 - 語彙力バレッジ: 正解の語彙項目が得られた単語の割合
 - 平均構文木数: 一文辺りの構文木の数
(曖昧性をどこまで絞っているか)
- 近年は形式文法の性能を単体で評価するのはまれ
 - 統計的曖昧性解消と組み合わせた構文解析評価を行う

本日のまとめ

- 構文解析
 - 依存構造解析
 - **句構造解析**
- 句構造解析のための形式文法
 - 形式文法: 文脈自由文法, 語彙化文法
 - 文法獲得: ツリーバンク文法, コーパス志向文法開発
- 形式文法を用いた句構造解析
 - CKY 構文解析