

Syntaktisk parsning (Jurafsky & Martin kapitel 13)

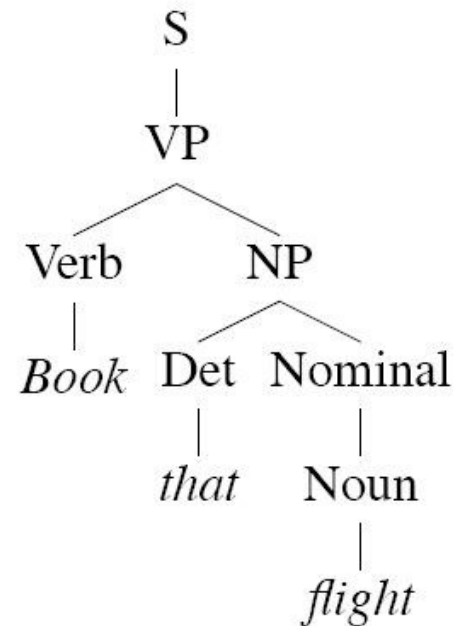
Mats Wirén
Institutionen för lingvistik
Stockholms universitet
mats.wiren@ling.su.se

DH2418 Språkteknologi
DA3010 Språkteknologi för datorlingvister
Föreläsning 7, 13/9 2011

Syntaktisk parsning

- Indata: sträng (mening) och grammatik
 - Här: Kontextfri grammatik
- Utdata: trädstruktur
 - Härledningen ska följa grammatikreglerna
 - Rotad i S, ska täcka indatasträngen
- Ytterligare antaganden:
 - Morfologisk analys är gjord
 - Alla ord finns i lexikonet

Resultat av parsning



”Book that flight”

Syntaktisk parsning

- Vad ska man ha det till?
 - Ger predikat–argument-struktur
 - “Vem gör vad med vem?”
 - Indata till semantisk analys

Parsningsalgoritmer

- Parsning kan ses som sökproblem
- Två grundläggande (naiva) algoritmer:
 - Toppen–nedåt-sökning (top-down)
 - Botten–uppåt-sökning (bottom-up)

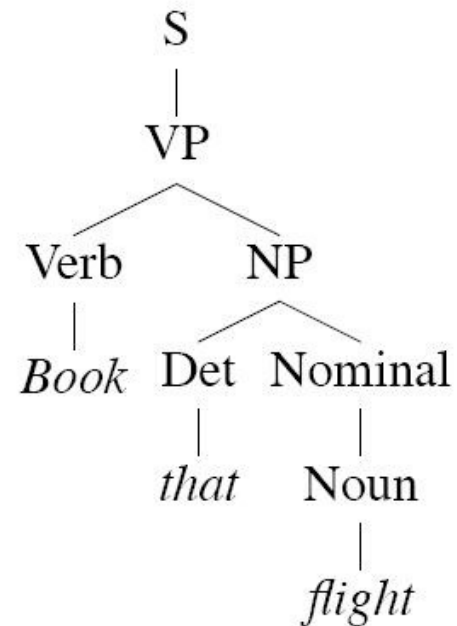
Toppen-nedåt

- Observation: träd måste vara rotade i en S-nod
- Parsningsstrategi:
 - Starta från toppen med en S-nod
 - Tillämpa grammatikregler för att bygga delträd nedåt
 - Arbeta ner till löven

Exempelgrammatik L1

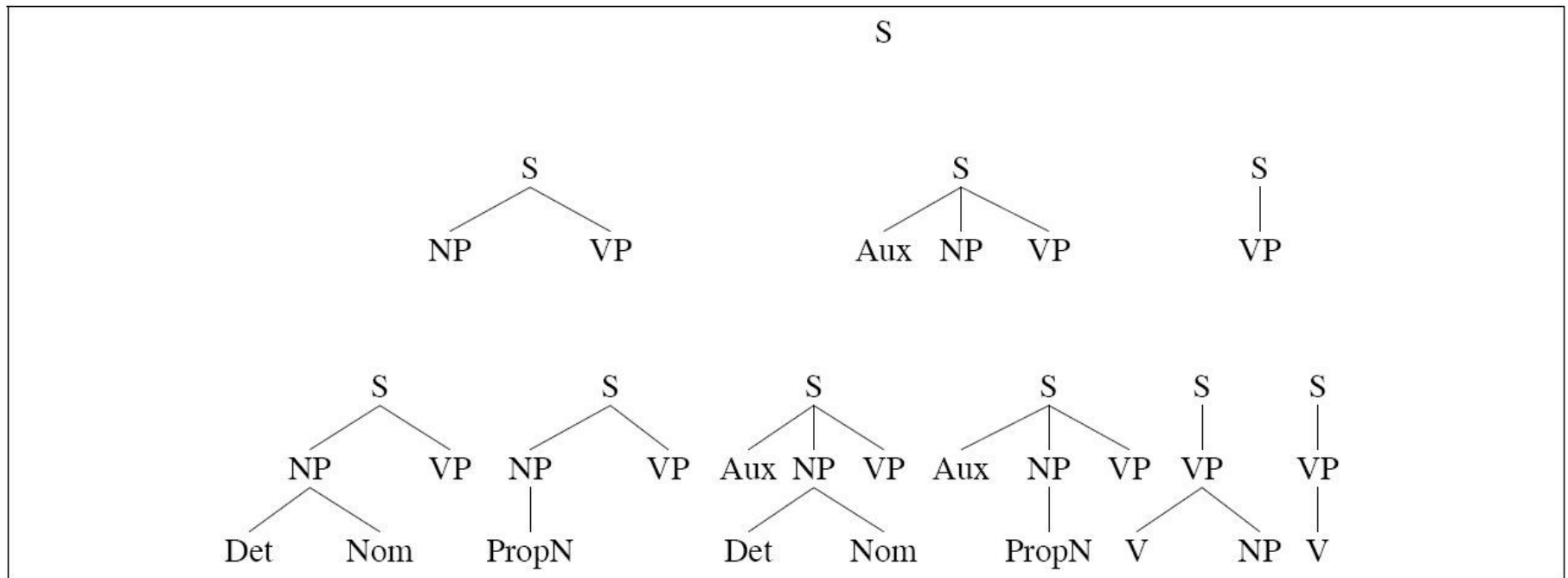
Grammar	Lexicon
$S \rightarrow NP VP$	$Det \rightarrow that \mid this \mid a$
$S \rightarrow Aux NP VP$	$Noun \rightarrow book \mid flight \mid meal \mid money$
$S \rightarrow VP$	$Verb \rightarrow book \mid include \mid prefer$
$NP \rightarrow Pronoun$	$Pronoun \rightarrow I \mid she \mid me$
$NP \rightarrow Proper-Noun$	$Proper-Noun \rightarrow Houston \mid NWA$
$NP \rightarrow Det Nominal$	$Aux \rightarrow does$
$Nominal \rightarrow Noun$	$Preposition \rightarrow from \mid to \mid on \mid near \mid through$
$Nominal \rightarrow Nominal Noun$	
$Nominal \rightarrow Nominal PP$	
$VP \rightarrow Verb$	
$VP \rightarrow Verb NP$	
$VP \rightarrow Verb NP PP$	
$VP \rightarrow Verb PP$	
$VP \rightarrow VP PP$	
$PP \rightarrow Preposition NP$	

Resultat av parsning



”Book that flight”

Toppen-nedåt

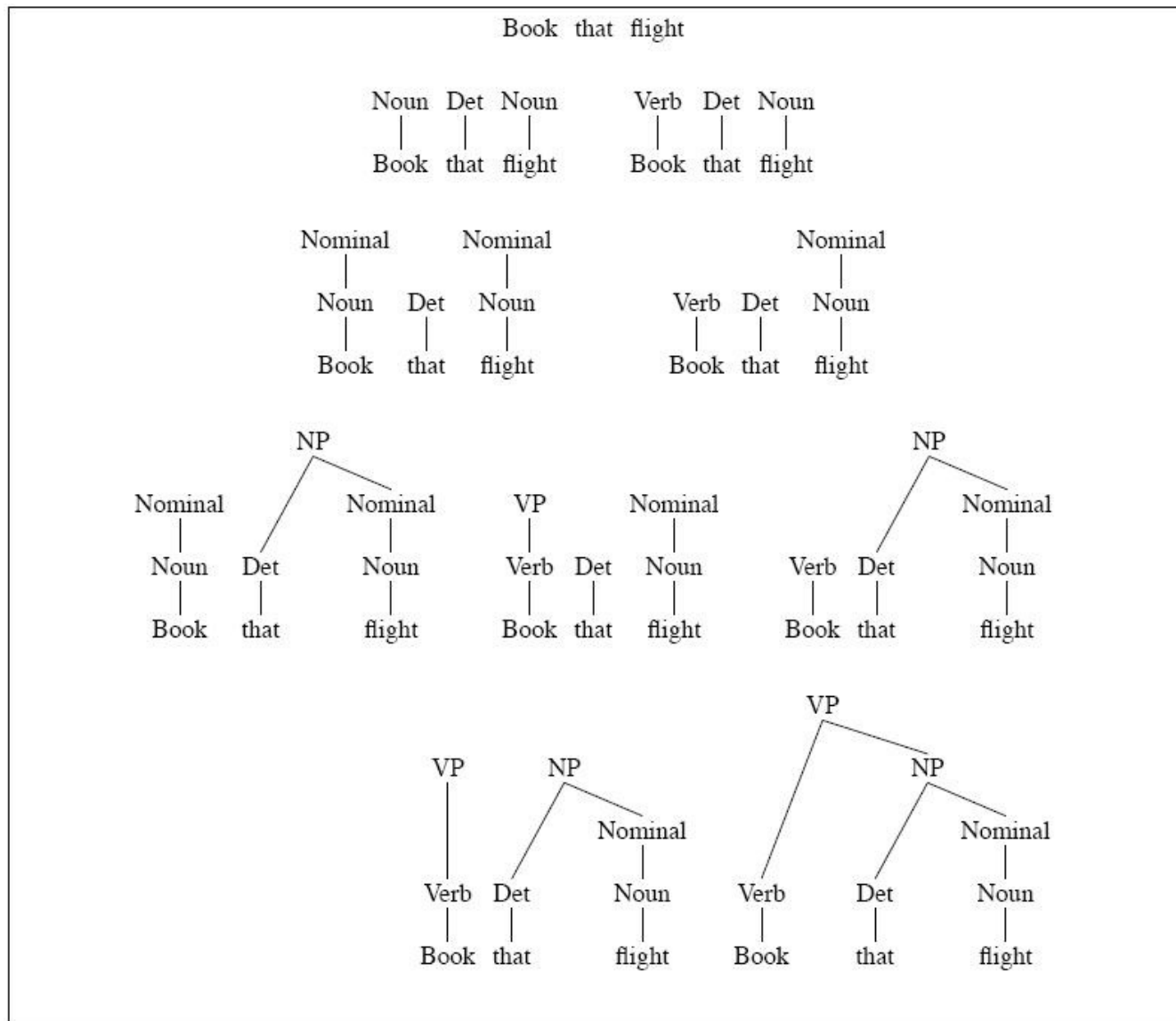


Sökrymden kan expanderas bredden först eller djupet först

Botten-uppåt

- Observation: träd måste täcka alla orden i indatasträngen
- Parsningsstrategi:
 - Starta från botten med orden
 - Tillämpa grammatikregler för att bygga delträd
 - Arbeta upp mot roten S

Botten-uppåt



Figur 13.4 från Jurafsky & Martin (2009). Copyright ©2009 by Pearson Education, Inc.

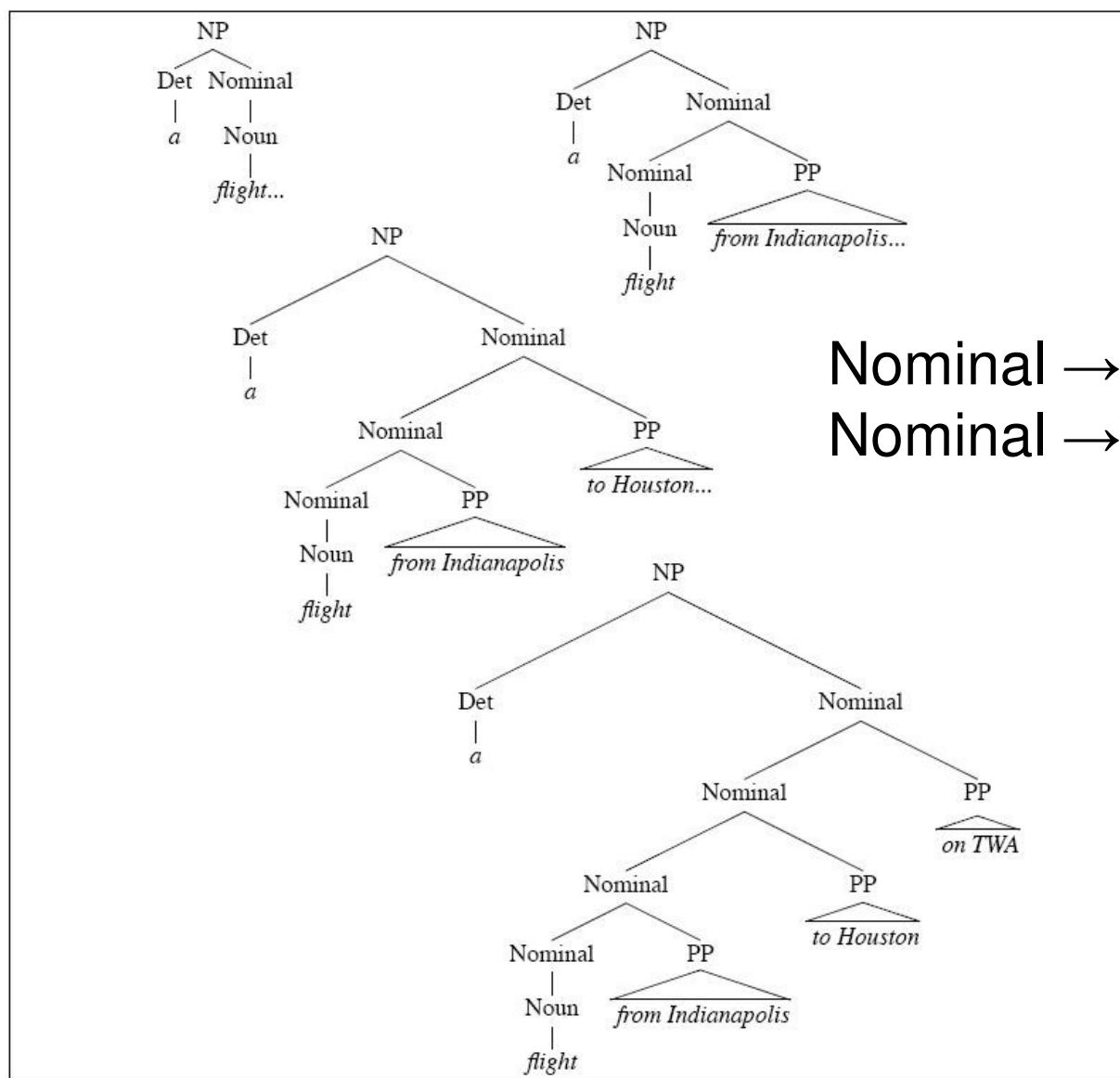
Jämförelse

- Toppen-nedåt
 - Söker efter giltiga träd utifrån roten S
 - Bygger en del träd som inte är konsistenta med orden i indata
- Botten-uppåt
 - Bygger bara träd som är konsistenta med indata
 - Bygger en del träd som inte leder till S

Parsning som sökproblem

- Behövs strategi för att hantera icke-determinism
- Exempel: Backtracking (återspårning)
 - Välj ett alternativ, om det fungerar så fortsätt
 - Annars, backa tillbaka till senaste valpunkt
- Men: Skilda globala analyser har ofta många delanalyser som är lika
 - Dvs backtracking leder till onödigt arbete

Samma delanalys görs om

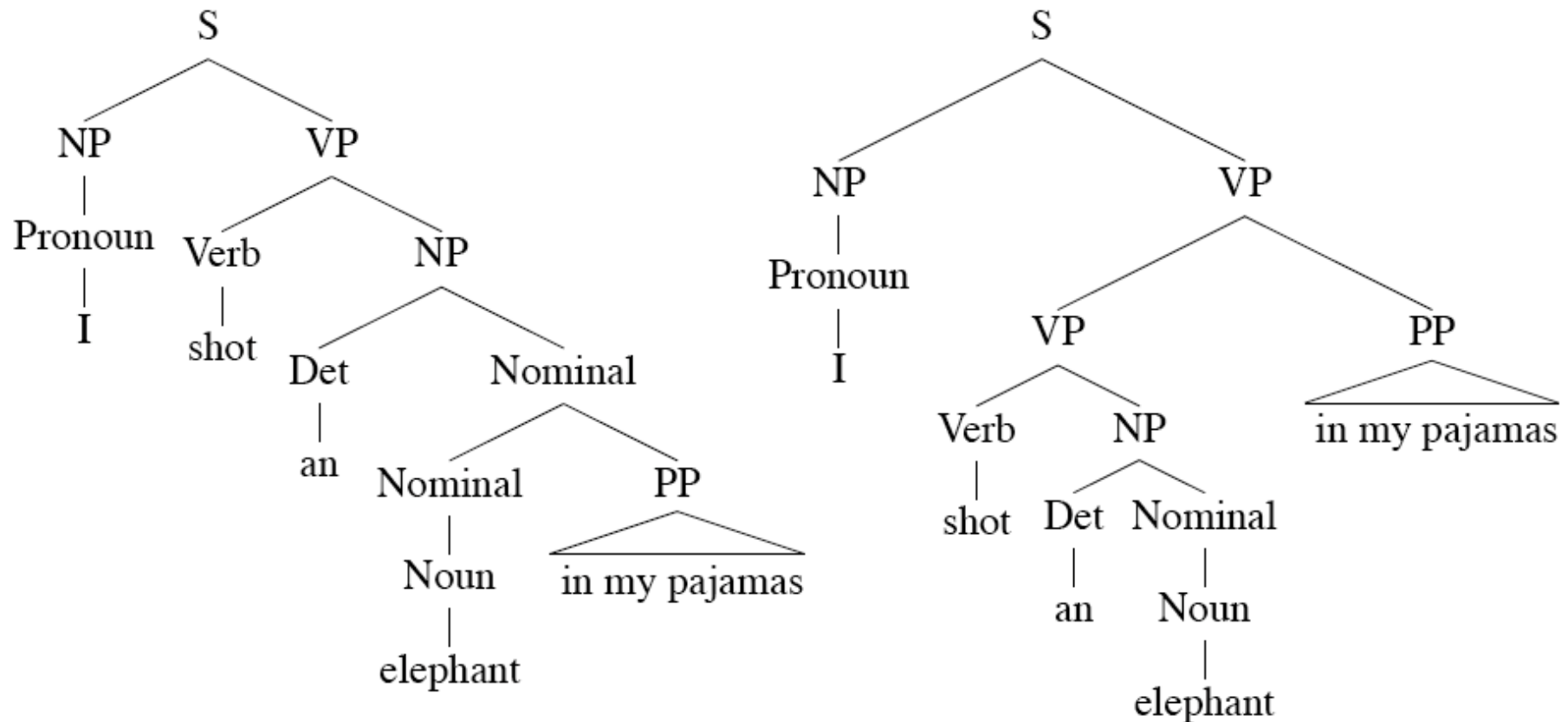


Nominal → Noun

Nominal → Nominal PP

Figur 13.7 från Jurafsky & Martin (2009). Copyright ©2009 by Pearson Education, Inc.

Underliggande problem: Lokal/global ambiguitet*



*Flertydighet

Ambiguitet (1)

- 2 prepositionsfraser:
 - “Put the block in the box on the table”
- Anta subkategorisering: “Put X in Y”
- 2 analyser:
 1. Put the block [in the box on the table]
 2. Put [the block in the box] on the table

Ambiguitet (2)

- 3 prepositionsfraser ger 5 analyser:
 1. Put the block **[in the box on the table]** in the kitchen
 2. Put the block **[in the box [on the table in the kitchen]]**
 3. Put **[the block in the box]** on the table in the kitchen
 4. Put **[the block [in the box on the table]]** in the kitchen
 5. Put **[the block in the box]** **[on the table in the kitchen]**

Ambiguitet (3)

- Antalet analyser växer *exponentiellt* med antalet prepositionsfraser*
 - Catalan-tal = $C_n = 1, 2, 5, 14, 42, 132, 469, 1430, 4862, 16796, \dots$ **

*Church & Patil (1982), “Coping with Syntactic Ambiguity or How to Put the Block in the Box on the Table”, *AJCL* 8(3–4)

** http://en.wikipedia.org/wiki/Catalan_number

Ambiguitet (4)

- Analoga problem
 - Koordinering: “Jurafsky och Martin och Manning och Schütze”
 - Sammansättningar: “Skolbokhylla”
- I matematik: prioritering och parenteser

$$2 + 3 \times 4 = 14$$

$$(2 + 3) \times 4 = 20$$

Ambiguitet (5)

- Analoga problem
 - Koordinering: “Jurafsky & Martin och Manning & Schütze”
 - Sammansättningar: “Skolbokhylla”
- I matematik: prioritering och parenteser
$$2 + 3 \times 4 = 14$$
$$(2 + 3) \times 4 = 20$$

Fråga

- *Hur kan parsning göras effektiv om det finns exponentiellt många analyser?*
- Svar: Dynamisk programmering
 - Spara alla delanalyser så att de delas av överordnade analyser i vilka de ingår
 - Inget arbete görs om
 - Ger polynomisk algoritm
- CKY (Cocke–Kasami–Younger*)

*D. H. Younger (1967). Recognition and parsing of context-free languages in time n^3 . *Information and Control* 10(2): 189–208

Grammatikformat

- CKY kräver att grammatiken är på Chomsky Normal Form (CNF)
 - Alla regler har formen:
$$A \rightarrow B C$$
$$D \rightarrow w$$
- Om min grammatik inte är i CNF?

CKY med generell CFG

- Problem: Anta att grammatiken har regel $VP \rightarrow NP PP PP$
- Skriv om grammatiken till CNF
 - Introducera nya mellanliggande icke-terminaler

$$A \rightarrow B C D \quad \longrightarrow \quad \begin{array}{l} A \rightarrow X D \\ X \rightarrow B C \end{array}$$

- Vad innebär detta?
 - Svag ekvivalens: Accepterar samma mängd strängar, men ger andra trädstrukturer

Exempelgrammatik L1

Grammar	Lexicon
$S \rightarrow NP VP$	$Det \rightarrow that \mid this \mid a$
$S \rightarrow Aux NP VP$	$Noun \rightarrow book \mid flight \mid meal \mid money$
$S \rightarrow VP$	$Verb \rightarrow book \mid include \mid prefer$
$NP \rightarrow Pronoun$	$Pronoun \rightarrow I \mid she \mid me$
$NP \rightarrow Proper-Noun$	$Proper-Noun \rightarrow Houston \mid NWA$
$NP \rightarrow Det Nominal$	$Aux \rightarrow does$
$Nominal \rightarrow Noun$	$Preposition \rightarrow from \mid to \mid on \mid near \mid through$
$Nominal \rightarrow Nominal Noun$	
$Nominal \rightarrow Nominal PP$	
$VP \rightarrow Verb$	
$VP \rightarrow Verb NP$	
$VP \rightarrow Verb NP PP$	
$VP \rightarrow Verb PP$	
$VP \rightarrow VP PP$	
$PP \rightarrow Preposition NP$	

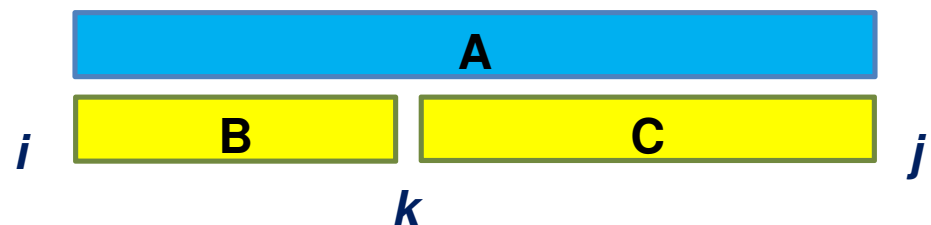
L1 efter CNF-konversion

$\mathcal{L}_1$ Grammar	$\mathcal{L}_1$ in CNF
$S \rightarrow NP VP$	$S \rightarrow NP VP$
$S \rightarrow Aux NP VP$	$S \rightarrow X1 VP$
	$X1 \rightarrow Aux NP$
$S \rightarrow VP$	$S \rightarrow book \mid include \mid prefer$
	$S \rightarrow Verb NP$
	$S \rightarrow X2 PP$
	$S \rightarrow Verb PP$
	$S \rightarrow VP PP$
$NP \rightarrow Pronoun$	$NP \rightarrow I \mid she \mid me$
$NP \rightarrow Proper-Noun$	$NP \rightarrow TWA \mid Houston$
$NP \rightarrow Det Nominal$	$NP \rightarrow Det Nominal$
$Nominal \rightarrow Noun$	$Nominal \rightarrow book \mid flight \mid meal \mid money$
$Nominal \rightarrow Nominal Noun$	$Nominal \rightarrow Nominal Noun$
$Nominal \rightarrow Nominal PP$	$Nominal \rightarrow Nominal PP$
$VP \rightarrow Verb$	$VP \rightarrow book \mid include \mid prefer$
$VP \rightarrow Verb NP$	$VP \rightarrow Verb NP$
$VP \rightarrow Verb NP PP$	$VP \rightarrow X2 PP$
	$X2 \rightarrow Verb NP$
$VP \rightarrow Verb PP$	$VP \rightarrow Verb PP$
$VP \rightarrow VP PP$	$VP \rightarrow VP PP$
$PP \rightarrow Preposition NP$	$PP \rightarrow Preposition NP$

Lexikala definitioner visas inte, men ingår här också

Två typer av regler i CKY

- $D \rightarrow w$
 - Terminalen (ordet) utgör en konstituent
- $A \rightarrow B C$
 - Om det finns ett A någonstans i indata så måste det finnas ett B följt av ett C
 - Intervall $[i, j]$ där i, j är positioner mellan ord
 - Om A går från i till j så måste det finnas ett k så att $i < k < j$



Delanalyser lagras i tabell

- Konstituent kan täcka $[i, j]$ för alla $0 \leq i < j \leq N$, där N = längd av indatasträng
 - Behövs (halv) $N \times N$ -tabell för att representera alla konstituenten
- Cell $[i, j]$ innehåller A omm A täcker i till j i indatasträngen

		TO: <i>j</i>					
		1	2	3	4	5	6
FROM: <i>i</i>	0	0-1	0-2	0-3	0-4	0-5	0-6
	1		1-2	1-3	1-4	1-5	1-6
	2			2-3	2-4	2-5	2-6
	3				3-4	3-5	3-6
	4					4-5	4-6
	5						5-6

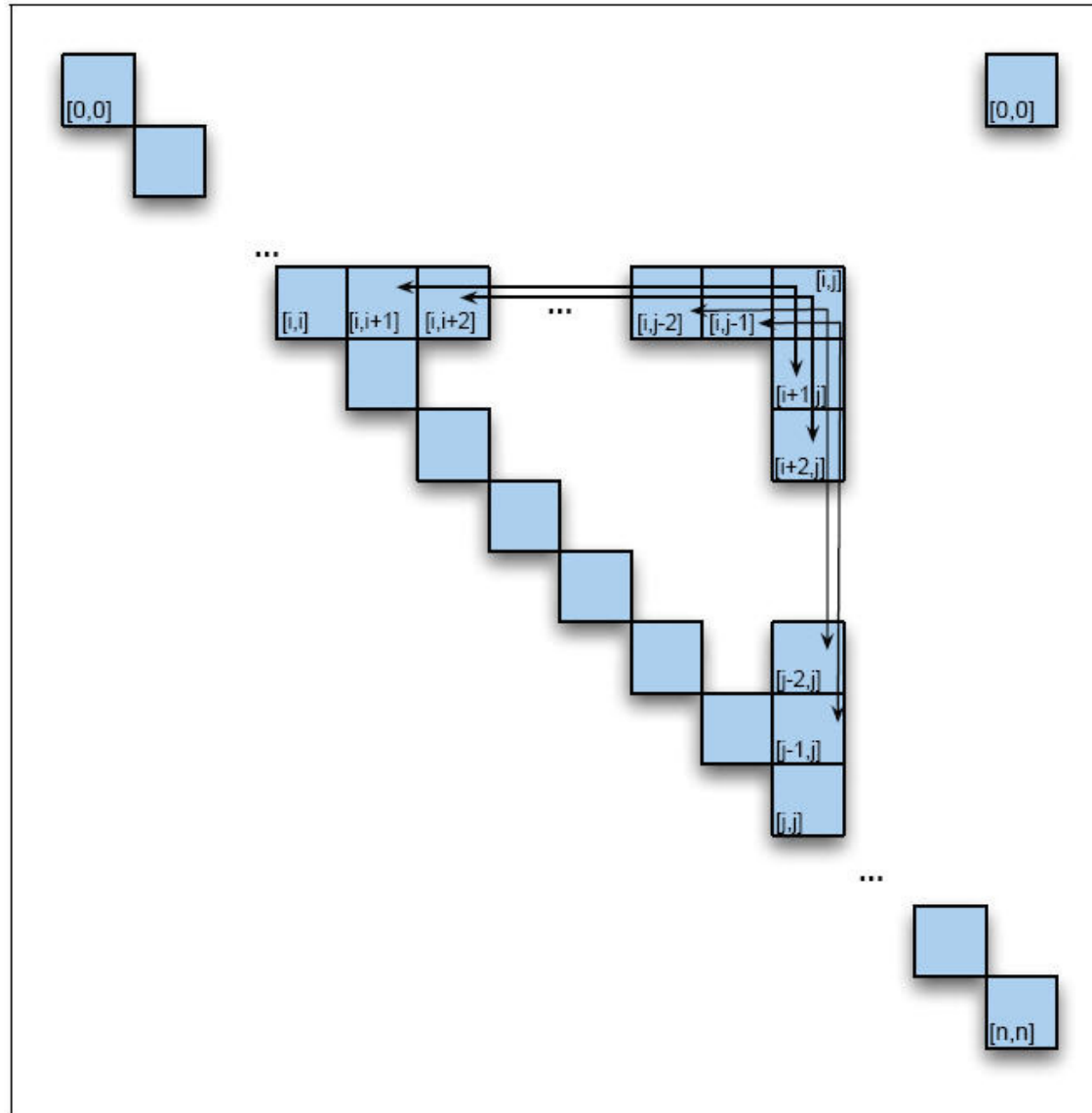
Delanalyser lagras i tabell

- För att A ska täcka $[i, j]$:
 - $A \rightarrow B C$ är regel i grammatiken
 - Det finns ett B i $[i, k]$ och ett C i $[k, j]$ för något $i < k < j$
 - För att applicera $A \rightarrow B C$, sök efter ett B i $[i, k]$ och ett C i $[k, j]$

		TO: <i>j</i>					
		1	2	3	4	5	6
FROM: <i>i</i>	0	0-1	0-2	0-3	0-4	0-5	0-6
	1		1-2	1-3	1-4	1-5	1-6
	2			2-3	2-4	2-5	2-6
	3				3-4	3-5	3-6
	4					4-5	4-6
	5						5-6

CKY: Regelapplicering

Observera trycktel i boken, figur 13.11, s. 475: ska vara $[0,n]$



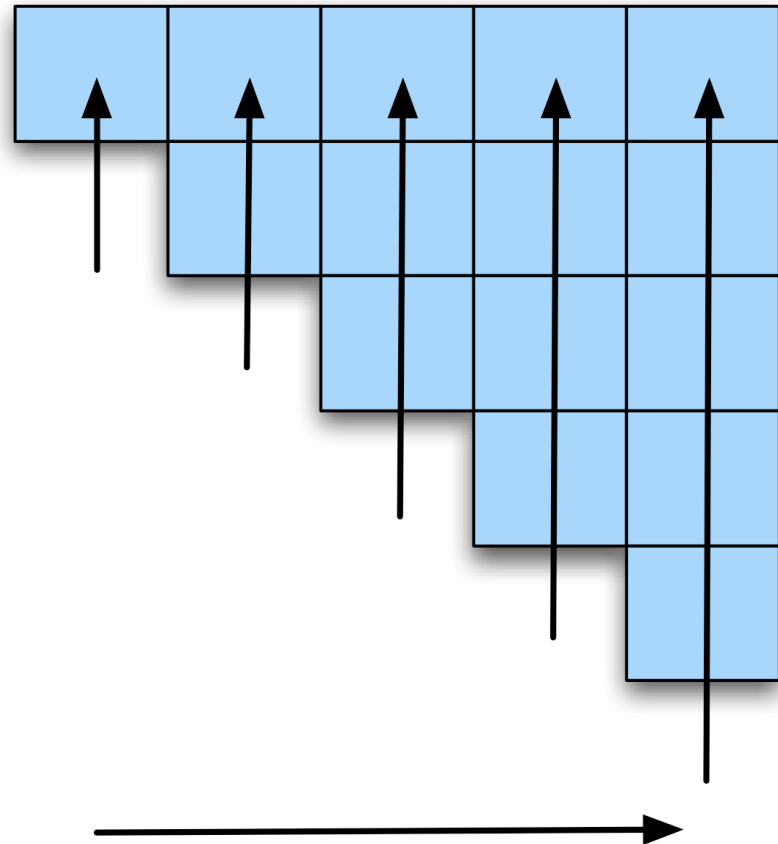
Figur 13.11 från Jurafsky & Martin (2009). Copyright ©2009 by Pearson Education, Inc.

CKY: Ordning

- Behövs kanonisk ordning:
 - Fyll tabellen en kolumn i taget, vänster till höger, nedifrån och upp
 - När vi då ska fylla en cell, så finns allt som förutsätts redan i tabellen (till vänster och nedanför)
- Bra egenskap: analyserar indata vänster–höger, ett ord i taget

CKY: Ordning

<i>Book</i>	<i>the</i>	<i>flight</i>	<i>through</i>	<i>Houston</i>
S, VP, Verb Nominal, Noun [0,1]		S,VP,X2 [0,3]		S,VP,X2 [0,5]
	Det [1,2]	NP [1,3]		NP [1,5]
		Nominal, Noun [2,3]		Nominal [2,5]
			Prep [3,4]	PP [3,5]
				NP, Proper- Noun [4,5]



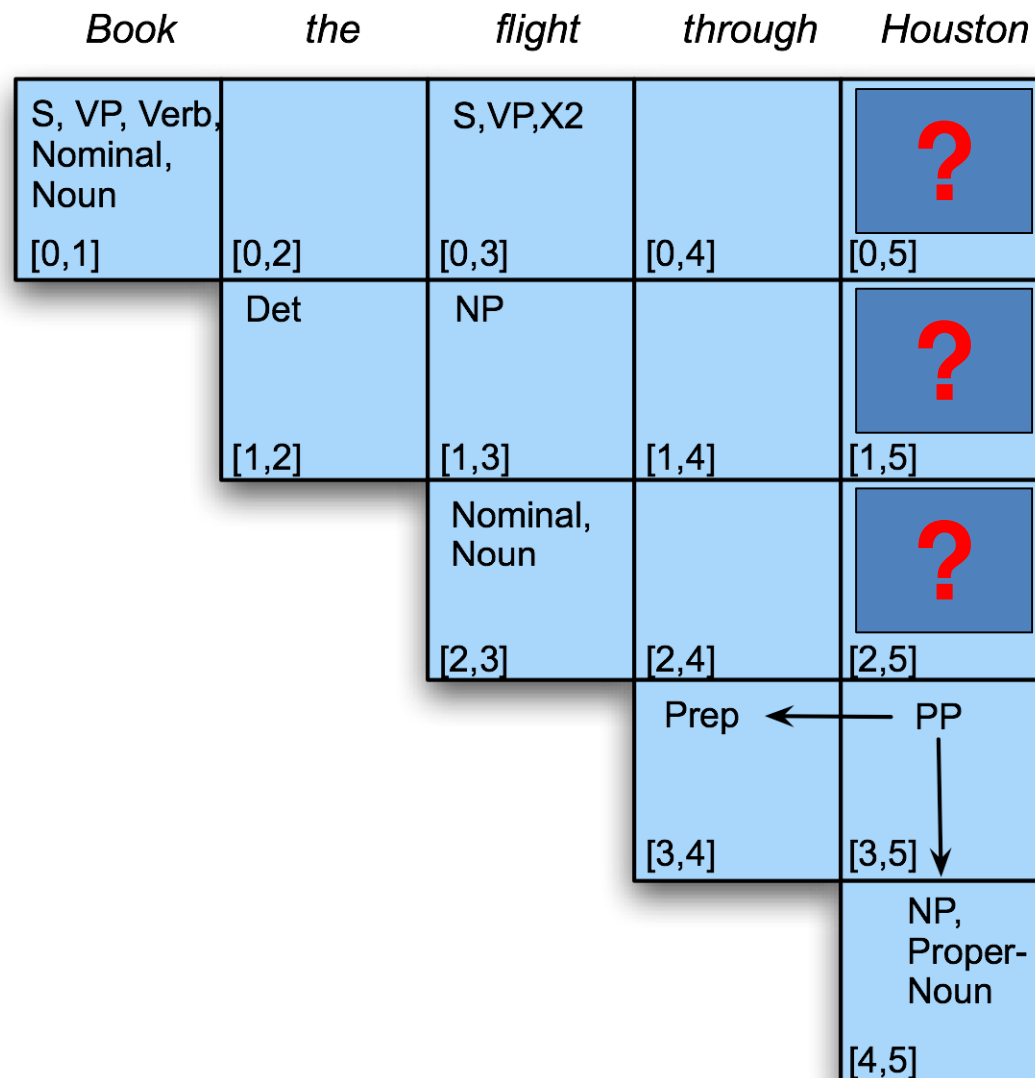
<i>Book</i>	<i>the</i>	<i>flight</i>	<i>through</i>	<i>Houston</i>
S, VP, Verb, Nominal, Noun [0,1]	[0,2]	S,VP,X2 [0,3]	[0,4]	?
	Det [1,2]	NP [1,3]	[1,4]	?
		Nominal, Noun [2,3]	[2,4]	?
			Prep [3,4]	?
				?

Exempel:
Läs "Houston",
fyll kolumn 5

Figur 13.12 från Jurafsky & Martin (2009).). Copyright ©2009 by Pearson Education, Inc.
Adapterad från Jimmy Lin

<i>Book</i>	<i>the</i>	<i>flight</i>	<i>through</i>	<i>Houston</i>
S, VP, Verb, Nominal, Noun [0,1]	[0,2]	S,VP,X2 [0,3]	[0,4]	[0,5]
	Det [1,2]	NP [1,3]	[1,4]	[1,5]
		Nominal, Noun [2,3]	[2,4]	[2,5]
			Prep [3,4]	[3,5]
				NP, Proper- Noun [4,5]

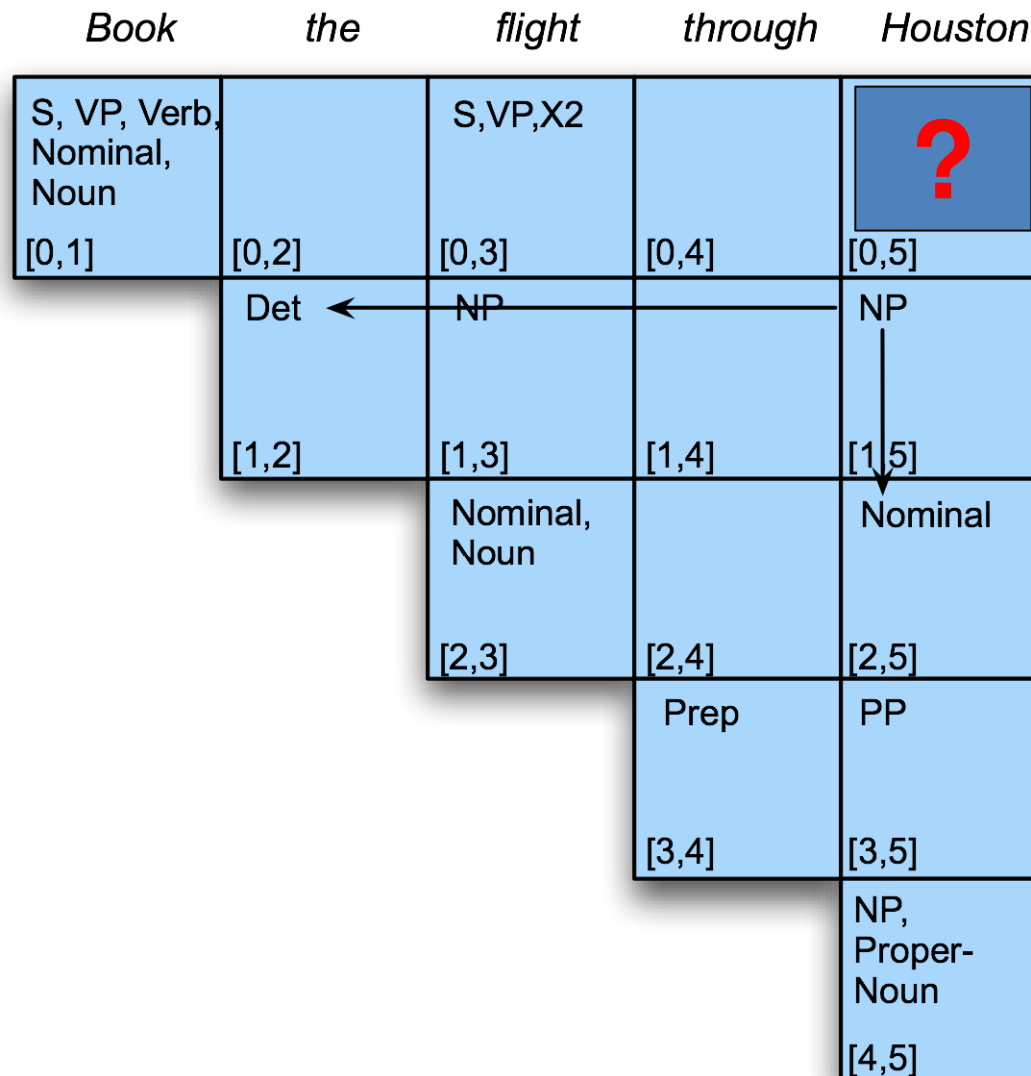
Figur 13.12 från Jurafsky & Martin (2009).). Copyright ©2009 by Pearson Education, Inc.
 Adapterad från Jimmy Lin



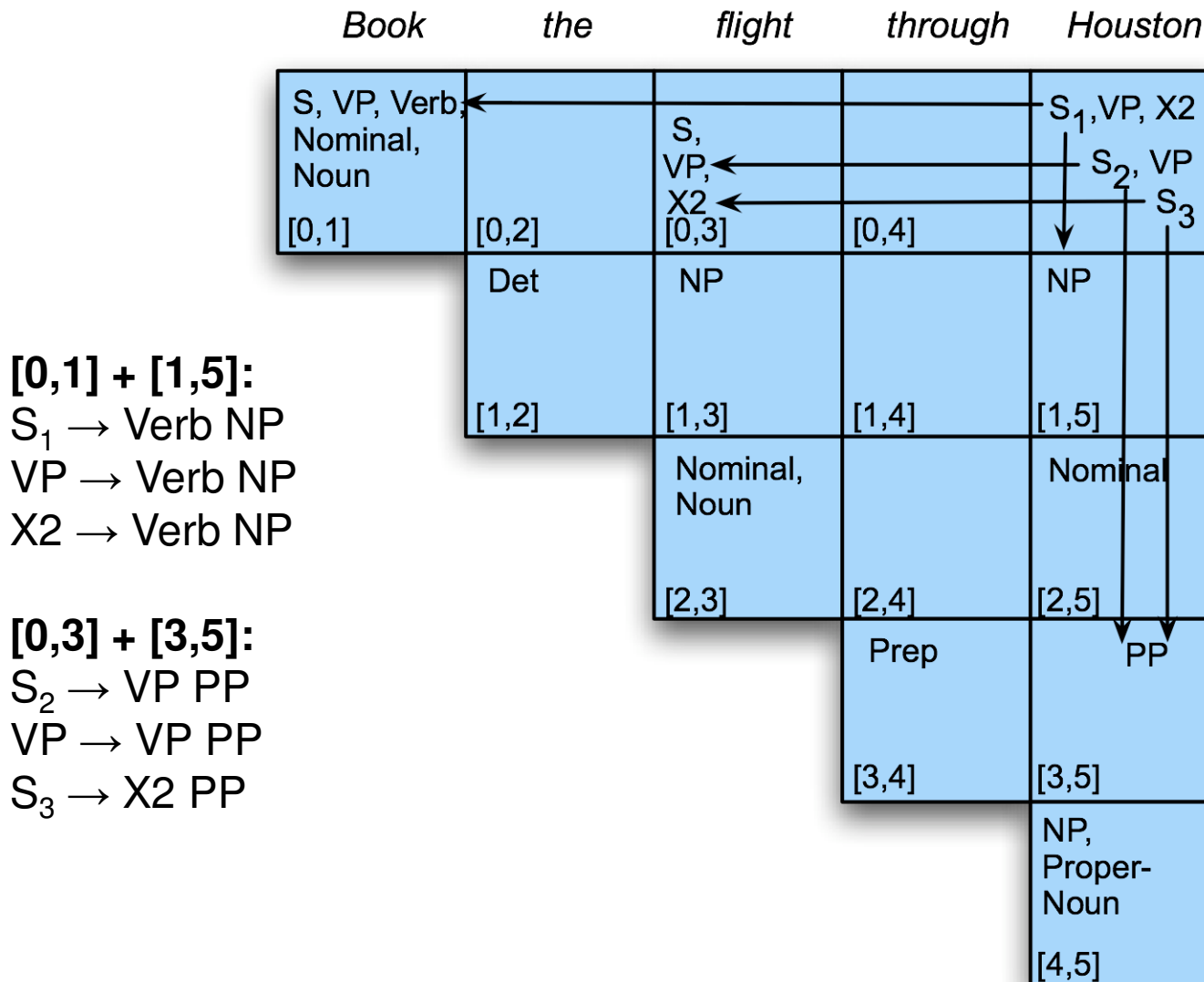
Figur 13.12 från Jurafsky & Martin (2009).). Copyright ©2009 by Pearson Education, Inc.
 Adapterad från Jimmy Lin

<i>Book</i>	<i>the</i>	<i>flight</i>	<i>through</i>	<i>Houston</i>
S, VP, Verb, Nominal, Noun [0,1]	[0,2]	S,VP,X2 [0,3]	[0,4]	[0,5]
	Det [1,2]	NP [1,3]	[1,4]	[1,5]
		Nominal, Noun [2,3]	[2,4]	Nominal [2,5]
			Prep [3,4]	PP [3,5]
				NP, Proper- Noun [4,5]

Figur 13.12 från Jurafsky & Martin (2009).). Copyright ©2009 by Pearson Education, Inc.
 Adapterad från Jimmy Lin



Figur 13.12 från Jurafsky & Martin (2009).). Copyright ©2009 by Pearson Education, Inc.
 Adapterad från Jimmy Lin



Figur 13.12 från Jurafsky & Martin (2009).). Copyright ©2009 by Pearson Education, Inc.
 Adapterad från Jimmy Lin

CKY: Analys

- Tidskomplexitet: $O(n^3)$
- Botten-uppåt: skapar en del onödiga konstituenten i tabellen
- Konversion till CNF ger ökat antal icke-terminala noder
 - Svagt ekvivalent med källgrammatik
 - Kan städas upp med post-processning

Earleys algoritm

- Avsnitt 13.4.2 i Jurafsky & Martin
- Klarar generell CFG
- Samma komplexitet som CKY

Probabilistisk CFG

- Avsnitt 14.1 i Jurafsky & Martin
- PCFG: Varje regel i CFG utvidgas med sannolikhet
 - $A \rightarrow \beta [p]$
 - Betingad sannolikhet för β givet A
 - $P(A \rightarrow \beta \mid A)$
 - Summan av sannolikheterna för alla omskrivningar av en icke-terminal A är 1

PCFG baserad på *L1*

Grammar		Lexicon
$S \rightarrow NP VP$	[.80]	$Det \rightarrow that [.10] \mid a [.30] \mid the [.60]$
$S \rightarrow Aux NP VP$	[.15]	$Noun \rightarrow book [.10] \mid flight [.30]$
$S \rightarrow VP$	[.05]	$\mid meal [.15] \mid money [.05]$
$NP \rightarrow Pronoun$	[.35]	$\mid flights [.40] \mid dinner [.10]$
$NP \rightarrow Proper-Noun$	[.30]	$Verb \rightarrow book [.30] \mid include [.30]$
$NP \rightarrow Det Nominal$	[.20]	$\mid prefer; [.40]$
$NP \rightarrow Nominal$	[.15]	$Pronoun \rightarrow I [.40] \mid she [.05]$
$Nominal \rightarrow Noun$	[.75]	$\mid me [.15] \mid you [.40]$
$Nominal \rightarrow Nominal Noun$	[.20]	$Proper-Noun \rightarrow Houston [.60]$
$Nominal \rightarrow Nominal PP$	[.05]	$\mid NWA [.40]$
$VP \rightarrow Verb$	[.35]	$Aux \rightarrow does [.60] \mid can [.40]$
$VP \rightarrow Verb NP$	[.20]	$Preposition \rightarrow from [.30] \mid to [.30]$
$VP \rightarrow Verb NP PP$	[.10]	$\mid on [.20] \mid near [.15]$
$VP \rightarrow Verb PP$	[.15]	$\mid through [.05]$
$VP \rightarrow Verb NP NP$	[.05]	
$VP \rightarrow VP PP$	[.15]	
$PP \rightarrow Preposition NP$	[1.0]	

Statistisk CKY-parsning

- Avsnitt 14.2 i Jurafsky & Martin
- Utvidgning av CKY för att hantera PCFG-regler
- Hur lär man sig sannolikheterna?
 - Tex med trädbank: beräkna sannolikhet för varje expansion av varje icke-terminal

Partiell parsning

- Avsnitt 13.5 i Jurafsky & Martin
- Ibland räcker det med en ytlig analys, eller analys som inte täcker allt indata
 - Till exempel informationsextraktion