Natural Language Processing

Syntactic parsing

Yulia Tsvetkov

yuliats@cs.washington.edu

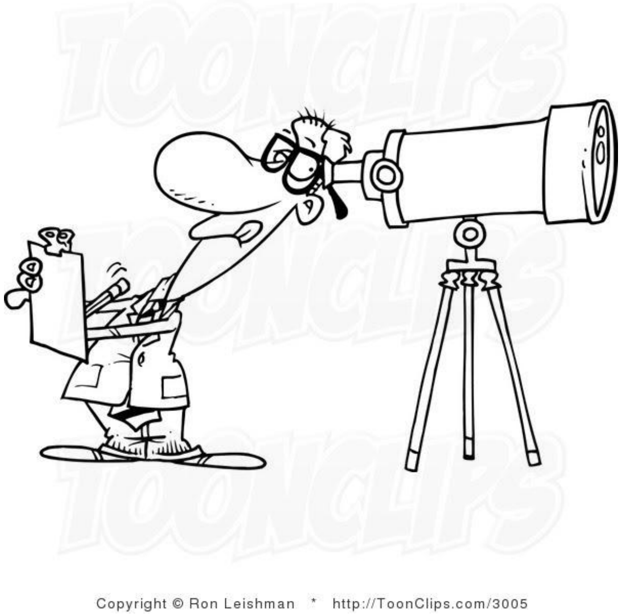
PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Announcements

-

Ambiguity

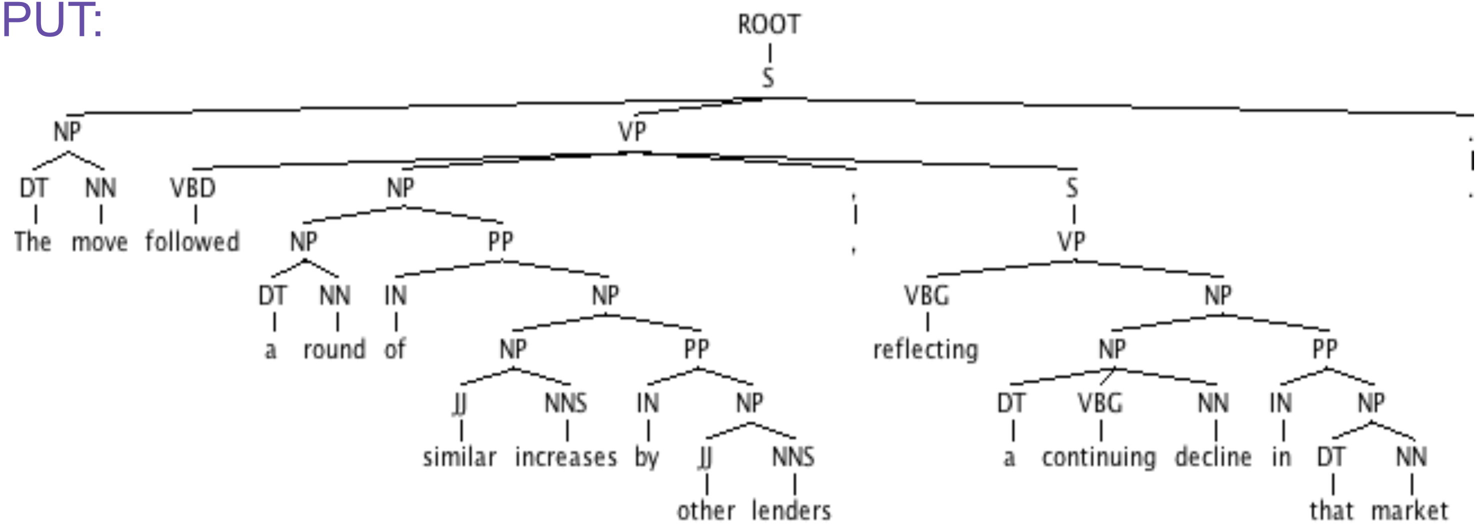
- I saw a girl with a telescope



Syntactic Parsing

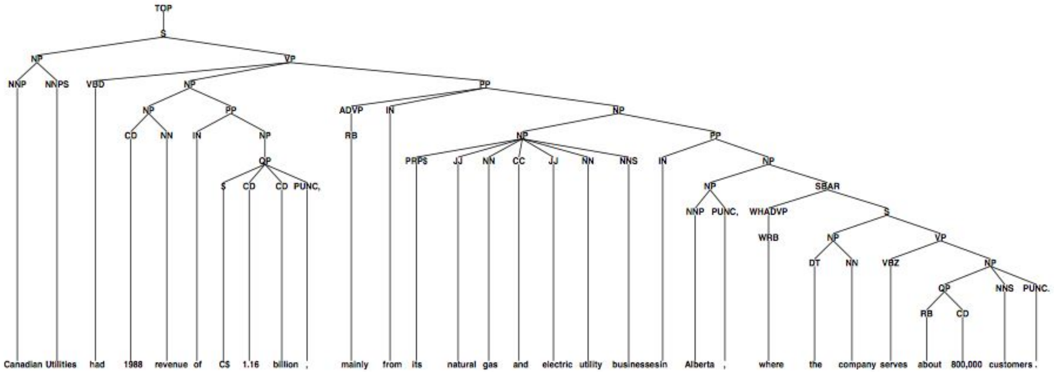
- INPUT:
  - The move followed a round of similar increases by other lenders, reflecting a continuing decline in that market

- OUTPUT:



A Supervised ML Problem

- Data for parsing experiments:
  - Penn WSJ Treebank = 50,000 sentences with associated trees
  - Usual set-up: 40,000 training, 2,400 test



Canadian Utilities had 1988 revenue of $ 1.16 billion , mainly from its natural gas and electric utility businesses in Alberta , where the company serves about 800,000 customers *[from Michael Collins slides]*

Syntax

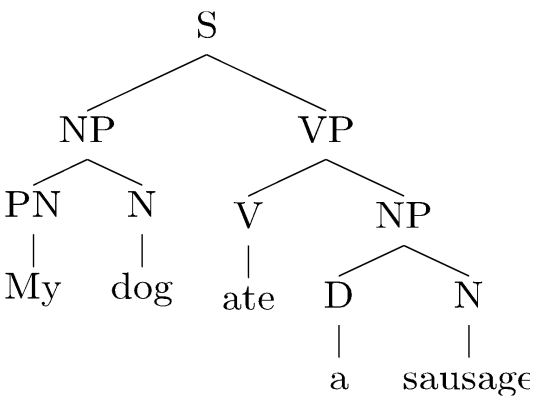
Syntax

- The study of the patterns of formation of sentences and phrases from words

my dog	Pron N
the dog	Det N
the cat	Det N
and	Conj
the large cat	Det Adj N
the black cat	Det Adj N
ate a sausage	V Det N

Parsing

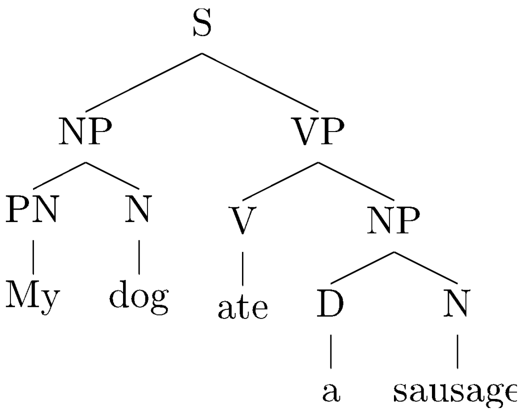
- The process of predicting syntactic representations
- Different types of syntactic representations are possible, for example:



Constituent (a.k.a. phrase-structure) tree

Constituent trees

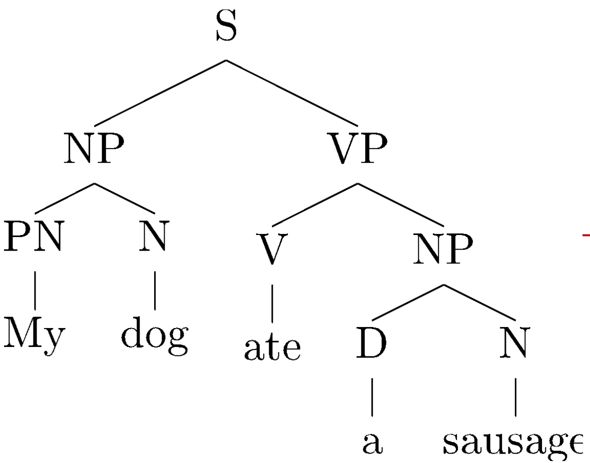
- Internal nodes correspond to phrases
  - S – a sentence
  - NP – Noun Phrase: My dog, a sandwich, lakes,..
  - VP – Verb Phrase: ate a sausage, barked, …
  - PP – Prepositional phrases: with a friend, in a car, …



- Nodes immediately above words are PoS tags (aka preterminals)
  - PN – pronoun
  - D – determiner
  - V – verb
  - N – noun
  - P – preposition

Bracketing notation

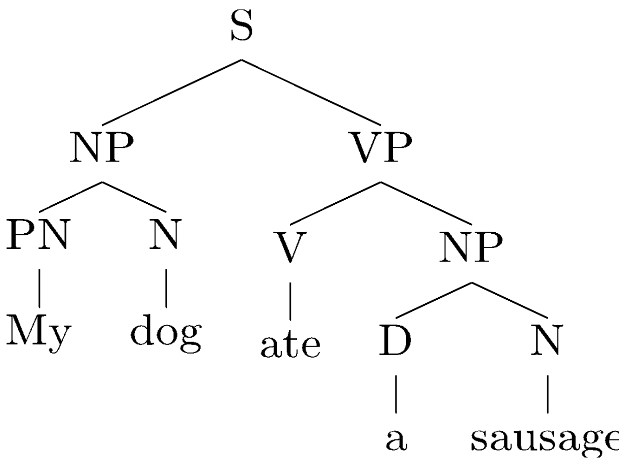
- It is often convenient to represent a tree as a bracketed sequence



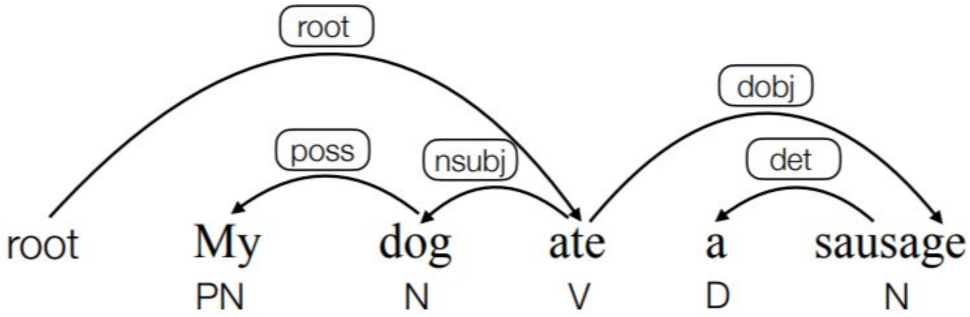

```
(S
   (NP (PN My) (N dog) )
   (VP (V ate)
      (NP (D a) (N sausage) )
   )
)
```

Parsing

- The process of predicting syntactic representations
- Different types of syntactic representations are possible, for example:



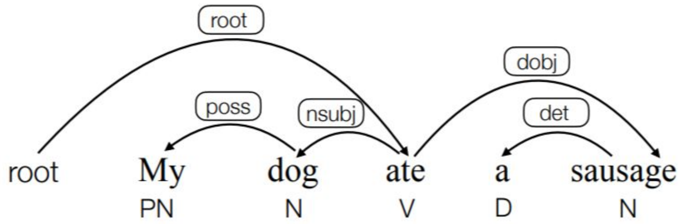
Constituent (a.k.a. phrase-structure) tree



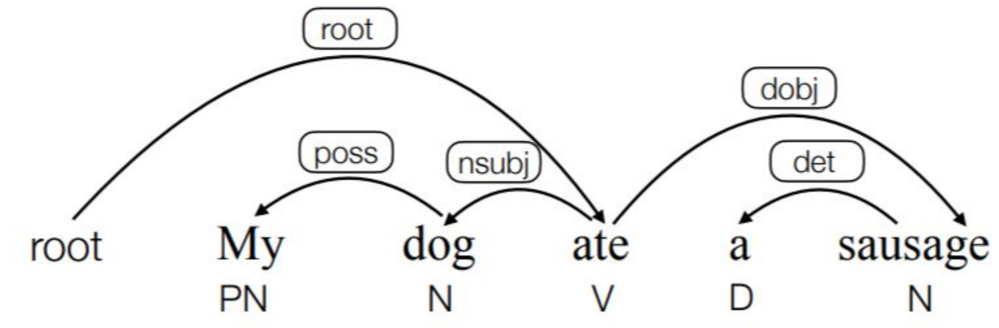
Dependency tree

Dependency trees

- Nodes are words (along with part-of-speech tags)
- Directed arcs encode syntactic dependencies between them
- Labels are types of relations between the words
  - poss – possessive
  - dobj – direct object
  - nsub - subject
  - det - determiner

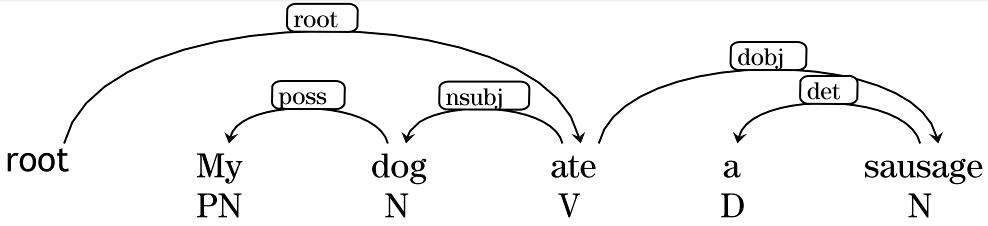


Recovering shallow semantics



- Some semantic information can be (approximately) derived from syntactic information
  - Subjects (nsubj) are (often) agents ("initiator / doers for an action")
  - Direct objects (dobj) are (often) patients ("affected entities")

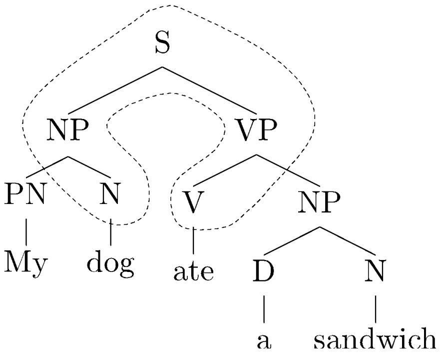
Recovering shallow semantics



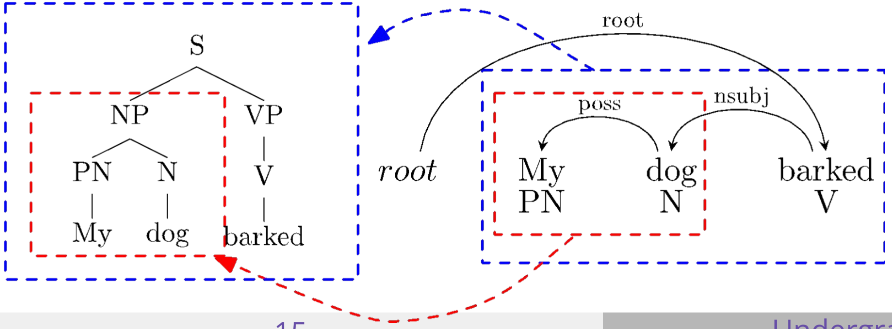
- Some semantic information can be (approximately) derived from syntactic information
  - Subjects (nsubj) are (often) agents ("initiator / doers for an action")
  - Direct objects (dobj) are (often) patients ("affected entities")
- But even for agents and patients consider:
  - Mary is baking a cake in the oven
  - A cake is baking in the oven
- In general it is not trivial even for the most shallow forms of semantics
  - E.g., consider prepositions: *in* can encode direction, position, temporal information, …

Constituent and dependency representations

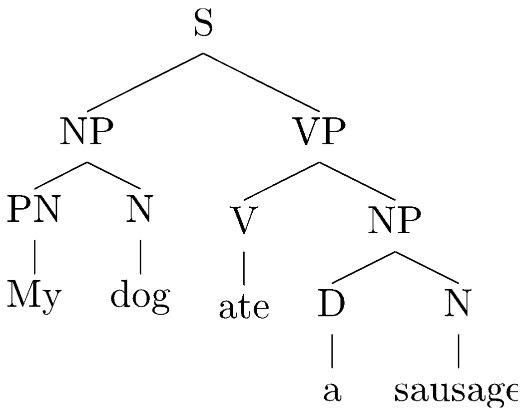
- Constituent trees can (potentially) be converted to dependency trees



- Dependency trees can (potentially) be converted to constituent trees



Constituent trees

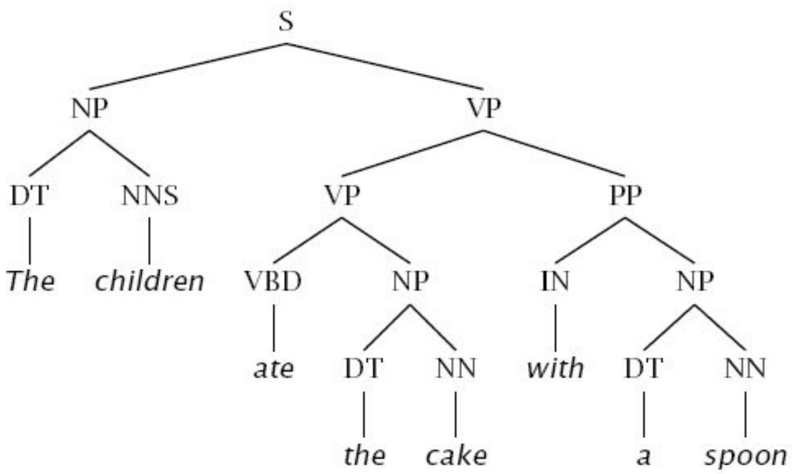


- Internal nodes correspond to phrases
  - S – a sentence
  - NP (Noun Phrase): My dog, a sandwich, lakes,..
  - VP (Verb Phrase): ate a sausage, barked, …
  - PP (Prepositional phrases): with a friend, in a car, …
- Nodes immediately above words are PoS tags (aka preterminals)
  - PN – pronoun
  - D – determiner
  - V – verb
  - N – noun
  - P – preposition

Constituency Tests

- How do we know what nodes go in the tree?

- Classic constituency tests:
  - Replacement
  - Movement
    - Passive
    - Clefting
    - Preposing
  - Substitution by *proform*
  - Modification
  - Coordination/Conjunction
  - Ellipsis/Deletion



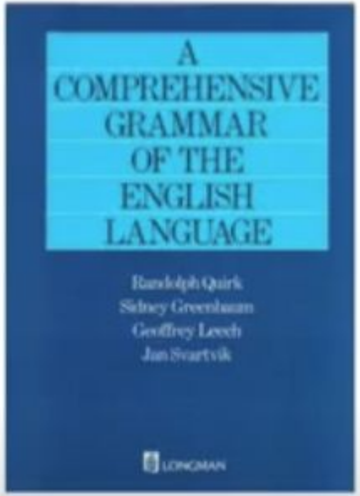
Morphology/Syntax/Semantics

- Syntax: The study of the patterns of formation of sentences and phrases from word
  - Borders with semantics and morphology sometimes blurred

Afyonkarahisarlılaştırabildiklerimizdenmişsinizcesinee

in Turkish means "as if you are one of the people that we thought to be originating from Afyonkarahisar" [wikipedia]

English grammar



Product Details (from Amazon)
Hardcover: 1779 pages
Publisher: Longman; 2nd Revised edition
Language: English
ISBN-10: 0582517346
ISBN-13: 978-0582517349
Product Dimensions: 8.4 x 2.4 x 10 inches
Shipping Weight: 4.6 pounds

Context Free Grammar (CFG)

Context Free Grammar (CFG)

Grammar (CFG)

ROOT → S
S → NP VP
NP → DT NN
NP → NN NNS
NP → NP PP
VP → VBP NP
VP → VBP NP PP
PP → IN NP

Lexicon

NN → interest
NNS → raises
VBP → interest
VBZ → raises
…

Other grammar formalisms: LFG, HPSG, TAG, CCG…

CFGs

S

$$S \rightarrow NP\ VP$$

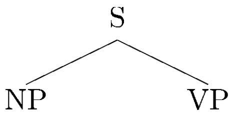
$$VP \rightarrow V$$
$$VP \rightarrow V\ NP$$
$$VP \rightarrow VP\ PP$$

$$NP \rightarrow NP\ PP$$
$$NP \rightarrow D\ N$$
$$NP \rightarrow PN$$

$$PP \rightarrow P\ NP$$

$$N \rightarrow girl$$
$$N \rightarrow telescope$$
$$N \rightarrow sandwich$$
$$PN \rightarrow I$$
$$V \rightarrow saw$$
$$V \rightarrow ate$$
$$P \rightarrow with$$
$$P \rightarrow in$$
$$D \rightarrow a$$
$$D \rightarrow the$$

CFGs



$$S \rightarrow NP \;\; VP$$

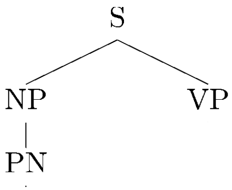
$$VP \rightarrow V$$
$$VP \rightarrow V \;\; NP$$
$$VP \rightarrow VP \;\; PP$$

$$NP \rightarrow NP \;\; PP$$
$$NP \rightarrow D \;\; N$$
$$NP \rightarrow PN$$

$$PP \rightarrow P \;\; NP$$

$$N \rightarrow girl$$
$$N \rightarrow telescope$$
$$N \rightarrow sandwich$$
$$PN \rightarrow I$$
$$V \rightarrow saw$$
$$V \rightarrow ate$$
$$P \rightarrow with$$
$$P \rightarrow in$$
$$D \rightarrow a$$
$$D \rightarrow the$$

CFGs



$$S \rightarrow NP\ \ VP$$

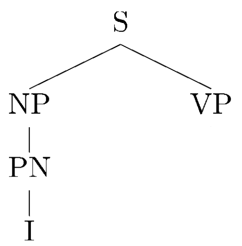
$$VP \rightarrow V$$
$$VP \rightarrow V\ \ NP$$
$$VP \rightarrow VP\ \ PP$$

$$NP \rightarrow NP\ \ PP$$
$$NP \rightarrow D\ \ N$$
$$NP \rightarrow PN$$

$$PP \rightarrow P\ \ NP$$

$$N \rightarrow girl$$
$$N \rightarrow telescope$$
$$N \rightarrow sandwich$$
$$PN \rightarrow I$$
$$V \rightarrow saw$$
$$V \rightarrow ate$$
$$P \rightarrow with$$
$$P \rightarrow in$$
$$D \rightarrow a$$
$$D \rightarrow the$$

CFGs



$S \rightarrow NP\ \ VP$

$VP \rightarrow V$

$VP \rightarrow V\ \ NP$

$VP \rightarrow VP\ \ PP$

$NP \rightarrow NP\ \ PP$

$NP \rightarrow D\ \ N$

$NP \rightarrow PN$

$PP \rightarrow P\ \ NP$

$N \rightarrow girl$

$N \rightarrow telescope$

$N \rightarrow sandwich$

$PN \rightarrow I$

$V \rightarrow saw$

$V \rightarrow ate$

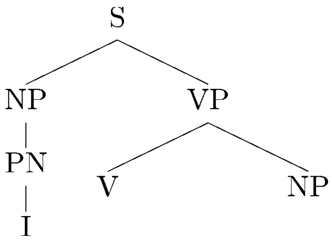
$P \rightarrow with$

$P \rightarrow in$

$D \rightarrow a$

$D \rightarrow the$

CFGs



$$S \rightarrow NP\ \ VP$$

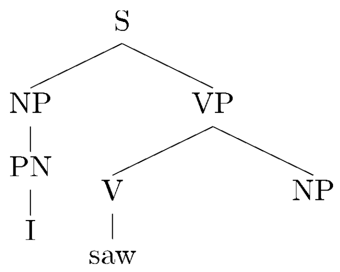
$$VP \rightarrow V$$
$$VP \rightarrow V\ \ NP$$
$$VP \rightarrow VP\ \ PP$$

$$NP \rightarrow NP\ \ PP$$
$$NP \rightarrow D\ \ N$$
$$NP \rightarrow PN$$

$$PP \rightarrow P\ \ NP$$

$$N \rightarrow girl$$
$$N \rightarrow telescope$$
$$N \rightarrow sandwich$$
$$PN \rightarrow I$$
$$V \rightarrow saw$$
$$V \rightarrow ate$$
$$P \rightarrow with$$
$$P \rightarrow in$$
$$D \rightarrow a$$
$$D \rightarrow the$$

CFGs



$$S \rightarrow NP\ \ VP$$

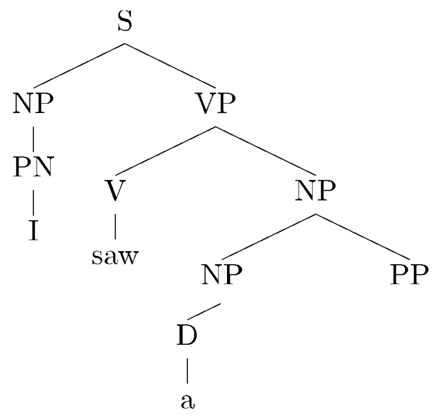
$$VP \rightarrow V$$
$$VP \rightarrow V\ \ NP$$
$$VP \rightarrow VP\ \ PP$$

$$NP \rightarrow NP\ \ PP$$
$$NP \rightarrow D\ \ N$$
$$NP \rightarrow PN$$

$$PP \rightarrow P\ \ NP$$

$$N \rightarrow girl$$
$$N \rightarrow telescope$$
$$N \rightarrow sandwich$$
$$PN \rightarrow I$$
$$V \rightarrow saw$$
$$V \rightarrow ate$$
$$P \rightarrow with$$
$$P \rightarrow in$$
$$D \rightarrow a$$
$$D \rightarrow the$$

CFGs



$$S \rightarrow NP \;\; VP$$

$$VP \rightarrow V$$

$$VP \rightarrow V \;\; NP$$

$$VP \rightarrow VP \;\; PP$$

$$NP \rightarrow NP \;\; PP$$

$$NP \rightarrow D \;\; N$$

$$NP \rightarrow PN$$

$$PP \rightarrow P \;\; NP$$

$$N \rightarrow girl$$

$$N \rightarrow telescope$$

$$N \rightarrow sandwich$$

$$PN \rightarrow I$$

$$V \rightarrow saw$$

$$V \rightarrow ate$$

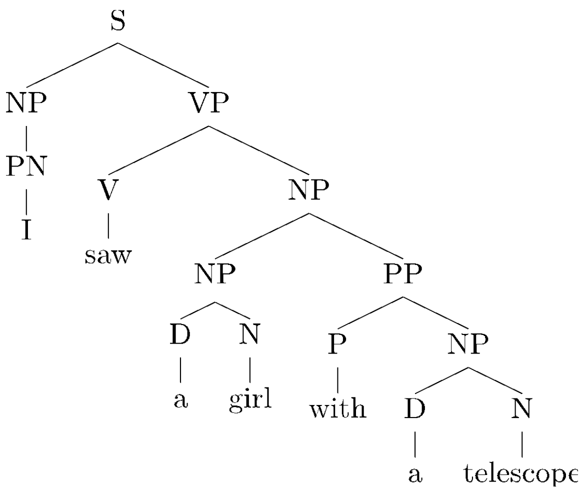
$$P \rightarrow with$$

$$P \rightarrow in$$

$$D \rightarrow a$$

$$D \rightarrow the$$

CFGs



$$S \rightarrow NP \;\; VP$$

$$VP \rightarrow V$$
$$VP \rightarrow V \;\; NP$$
$$VP \rightarrow VP \;\; PP$$

$$NP \rightarrow NP \;\; PP$$
$$NP \rightarrow D \;\; N$$
$$NP \rightarrow PN$$

$$PP \rightarrow P \;\; NP$$

$$N \rightarrow girl$$
$$N \rightarrow telescope$$
$$N \rightarrow sandwich$$
$$PN \rightarrow I$$
$$V \rightarrow saw$$
$$V \rightarrow ate$$
$$P \rightarrow with$$
$$P \rightarrow in$$
$$D \rightarrow a$$
$$D \rightarrow the$$

Treebank Sentences

```
( (S (NP-SBJ The move)
     (VP followed
         (NP (NP a round)
             (PP of
                 (NP (NP similar increases)
                     (PP by
                         (NP other lenders))
                     (PP against
                         (NP Arizona real estate loans)))))
         ,
         (S-ADV (NP-SBJ *)
                (VP reflecting
                    (NP (NP a continuing decline)
                        (PP-LOC in
                                (NP that market))))))
     .))
```

Context-Free Grammars

- A context-free grammar is a 4-tuple <N, T, S, R>
  - N : the set of non-terminals
    - **Phrasal categories**: S, NP, VP, ADJP, etc.
    - **Parts-of-speech** (pre-terminals): NN, JJ, DT, VB
  - T : the set of terminals (the words)
  - S : the start symbol
    - Often written as ROOT or TOP
    - Not usually the sentence non-terminal S
  - R : the set of rules
    - Of the form $X \rightarrow Y_1\ Y_2 \ldots Y_k$, with $X, Y_i \in N$
    - Examples: $S \rightarrow NP\ VP$, $VP \rightarrow VP\ CC\ VP$
    - Also called rewrites, productions, or local trees

An example grammar

$N = \{S, VP, NP, PP, N, V, PN, P\}$

$T = \{girl, telescope, sandwich, I, saw, ate, with, in, a, the\}$

$S = \{S\}$

$R:$

Called **Inner rules**

$S \rightarrow NP\ VP$ (NP A girl) (VP ate a sandwich)

$VP \rightarrow V$

$VP \rightarrow V\ NP$ (V ate) (NP a sandwich)

$VP \rightarrow VP\ PP$ (VP saw a girl) (PP with a telescope)

$NP \rightarrow NP\ PP$ (NP a girl) (PP with a sandwich)

$NP \rightarrow D\ N$ (D a) (N sandwich)

$NP \rightarrow PN$

$PP \rightarrow P\ NP$ (P with) (NP with a sandwich)

Preterminal rules

$N \rightarrow girl$

$N \rightarrow telescope$

$N \rightarrow sandwich$

$PN \rightarrow I$

$V \rightarrow saw$

$V \rightarrow ate$

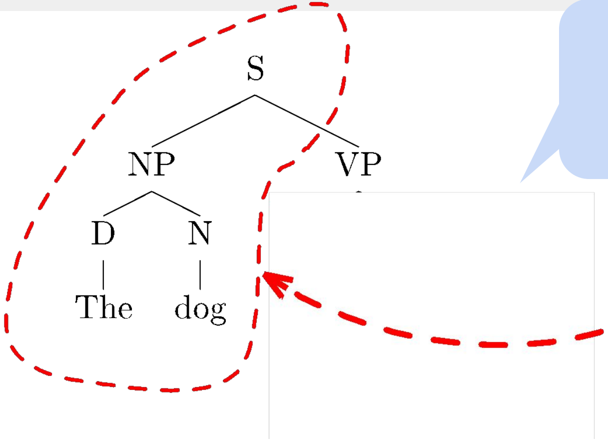
$P \rightarrow with$

$P \rightarrow in$

$D \rightarrow a$

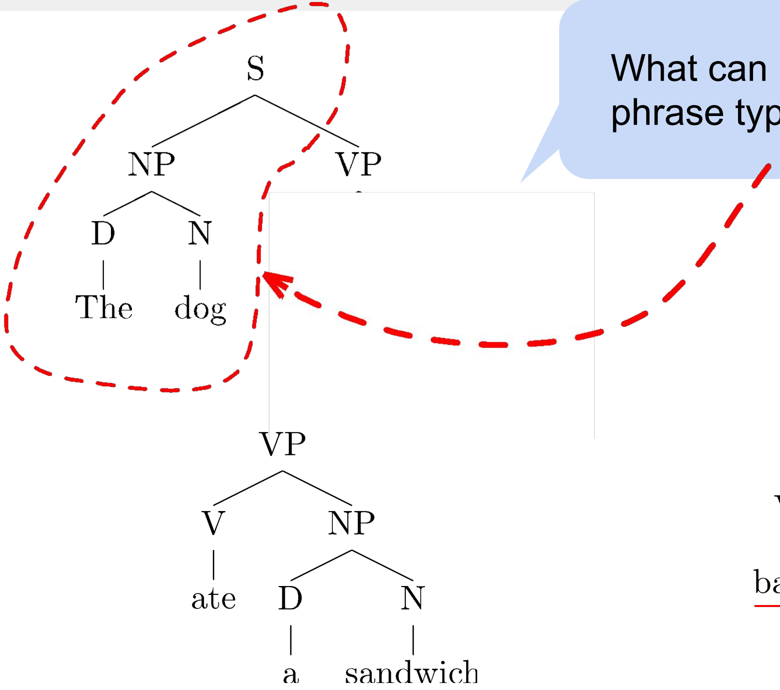
$D \rightarrow the$

Why context-free?

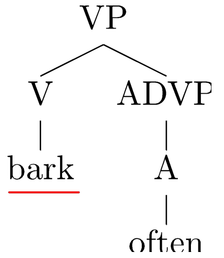


What can be a sub-tree is only affected by what the phrase type is (VP) but not the context

Why context-free?



What can be a sub-tree is only affected by what the phrase type is (VP) but not the context

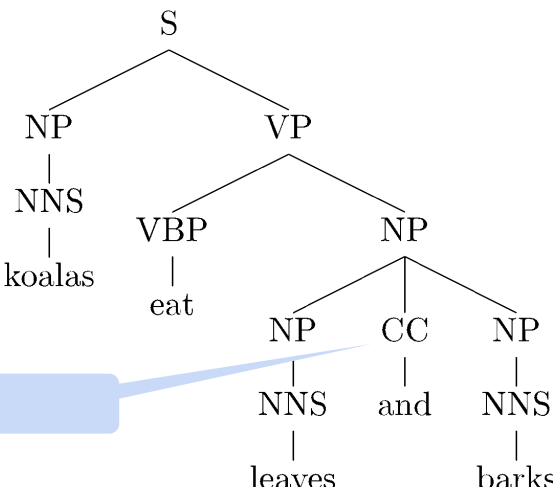


Not grammatical

Ambiguities

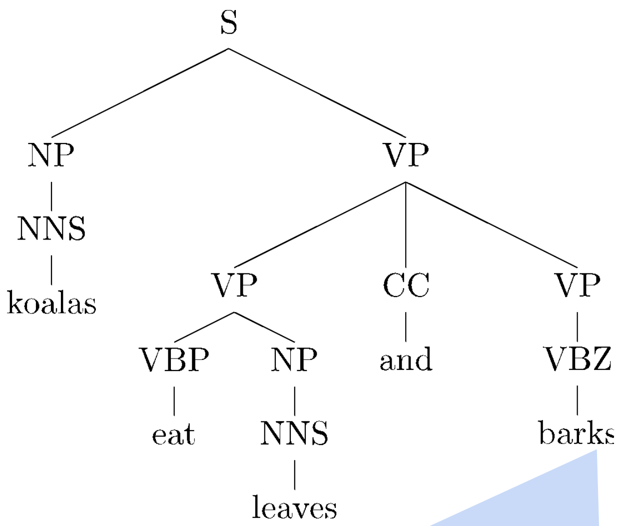
Coordination ambiguity

- Here, the coarse VP and NP categories cannot enforce subject-verb agreement in number resulting in the coordination ambiguity



Coordination

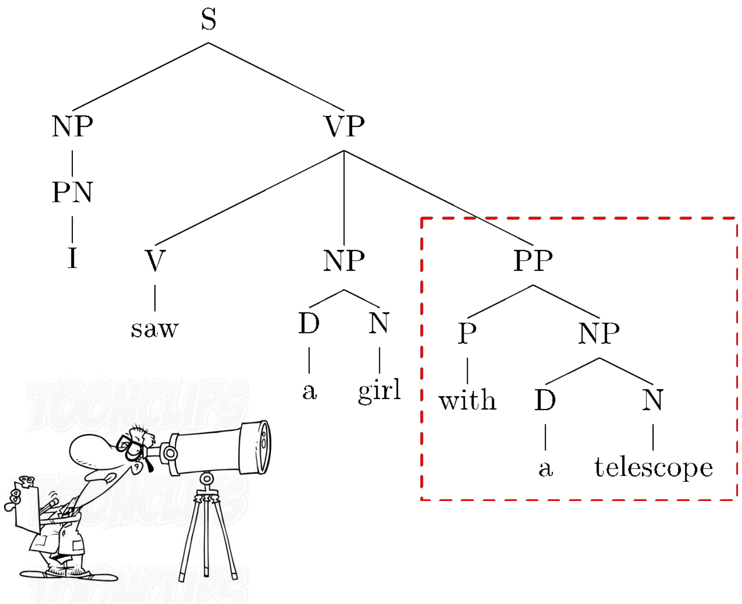
"Bark" can refer both to a noun or a verb

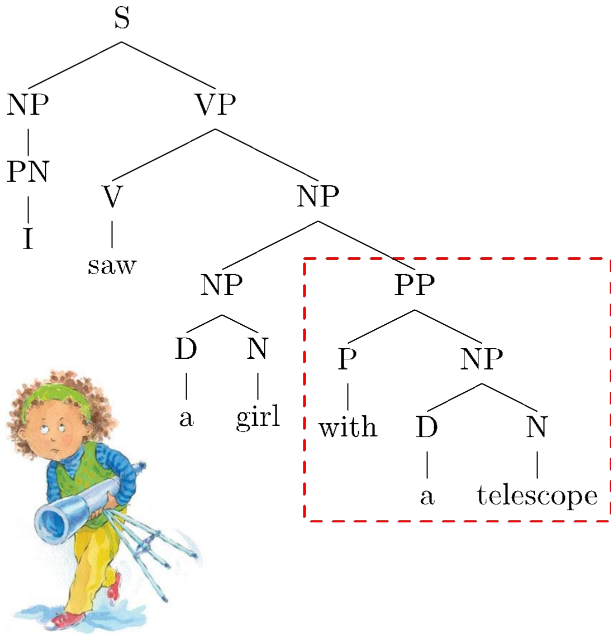


This tree would be ruled out if the context would be somehow captured (subject-verb agreement)

Why parsing is hard? Ambiguity

- Prepositional phrase attachment ambiguity





PP Ambiguity

Put the block in the box on the table in the kitchen

3 prepositional phrases, 5 interpretations:

- Put the block ((in the box on the table) in the kitchen)
- Put the block (in the box (on the table in the kitchen))
- Put ((the block in the box) on the table) in the kitchen.
- Put (the block (in the box on the table)) in the kitchen.
- Put (the block in the box) (on the table in the kitchen)

PP Ambiguity

Put the block in the box on the table in the kitchen

3 prepositional phrases, 5 interpretations:

- Put the block ((in the box on the table) in the kitchen)
- Put the block (in the box (on the table in the kitchen))
- …

A general case:

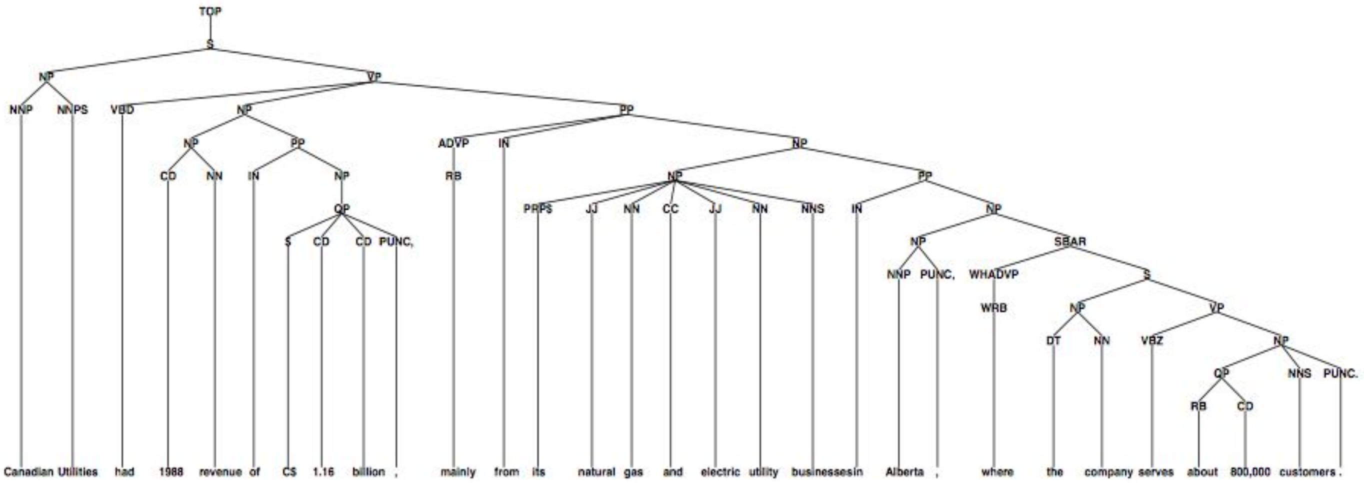
- ((())) ()(()) ()()() (())() (()())

$$Cat_n = \binom{2n}{n} - \binom{2n}{n-1} \sim \frac{4^n}{n^{3/2}\sqrt{\pi}}$$

Catalan numbers

$$1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, \ldots$$

A typical tree from a standard dataset (Penn treebank WSJ)



Canadian Utilities had 1988 revenue of $ 1.16 billion , mainly from its natural gas and electric utility businesses in Alberta , where the company serves about 800,000 customers .

[from Michael Collins slides]

Syntactic Ambiguities I

- Prepositional phrases:
  - They cooked the beans in the pot on the stove with handles.
- Particle vs. preposition:
  - The puppy tore up the staircase.
- Complement structures
  - The tourists objected to the guide that they couldn't hear.
    She knows you like the back of her hand.
- Gerund vs. participial adjective
  - Visiting relatives can be boring.
    Changing schedules frequently confused passengers.

Syntactic Ambiguities II

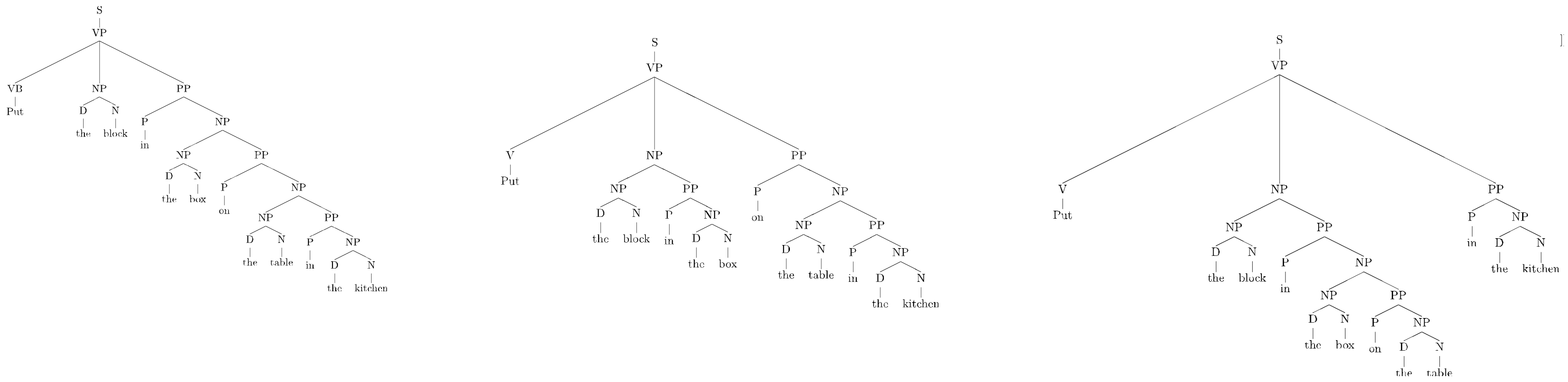
- Modifier scope within NPs
  - impractical design requirements
    plastic cup holder

- Multiple gap constructions
  - The chicken is ready to eat.
    The contractors are rich enough to sue.

- Coordination scope:
  - Small rats and mice can squeeze into holes or cracks in the wall.

How to Deal with Ambiguity?

- We want to score all the derivations to encode how plausible they are



Put the block in the box on the table in the kitchen

Probabilistic Context Free Grammar (PCFG)

Probabilistic Context-Free Grammars

- A context-free grammar is a 4-tuple <N, T, S, R>
  - N : the set of non-terminals
    - **Phrasal categories**: S, NP, VP, ADJP, etc.
    - **Parts-of-speech** (pre-terminals): NN, JJ, DT, VB
  - T : the set of terminals (the words)
  - S : the start symbol
    - Often written as ROOT or TOP
    - Not usually the sentence non-terminal S
  - R : the set of rules
    - Of the form $X \rightarrow Y_1\ Y_2 \ldots Y_k$, with $X, Y_i \in N$
    - Examples: $S \rightarrow NP\ VP$, $VP \rightarrow VP\ CC\ VP$
    - Also called rewrites, productions, or local trees

- A PCFG adds:
  - A top-down production probability per rule $P(Y_1\ Y_2 \ldots Y_k \mid X)$

PCFGs

Associate probabilities with the rules : $p(X \to \alpha)$

$$\forall \; X \to \alpha \in R: \quad 0 \leq p(X \to \alpha) \leq 1$$

$$\forall X \in N: \quad \sum_{\alpha : X \to \alpha \in R} p(X \to \alpha) = 1$$

Now we can score a tree as a product of probabilities corresponding to the used rules

$S \to NP \;\; VP$	1.0	(NP A girl) (VP ate a sandwich)
$VP \to V$	0.2	
$VP \to V \;\; NP$	0.4	(VP ate) (NP a sandwich)
$VP \to VP \;\; PP$	0.4	(VP saw a girl) (PP with …)
$NP \to NP \;\; PP$	0.3	(NP a girl) (PP with ….)
$NP \to D \;\; N$	0.5	(D a) (N sandwich)
$NP \to PN$	0.2	
$PP \to P \;\; NP$	1.0	(P with) (NP with a sandwich)

$N \to girl$	0.2
$N \to telescope$	0.7
$N \to sandwich$	0.1
$PN \to I$	1.0
$V \to saw$	0.5
$V \to ate$	0.5
$P \to with$	0.6
$P \to in$	0.4
$D \to a$	0.3
$D \to the$	0.7

PCFGs

S

$$S \to NP \;\; VP \; 1.0$$

$$VP \to V \; 0.2$$
$$VP \to V \;\; NP \; 0.4$$
$$VP \to VP \;\; PP \; 0.4$$

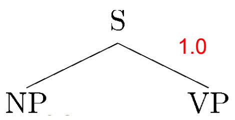
$$NP \to NP \;\; PP \; 0.3$$
$$NP \to D \;\; N \; 0.5$$
$$NP \to PN \; 0.2$$

$$PP \to P \;\; NP \; 1.0$$

$$N \to girl \; 0.2$$
$$N \to telescope \; 0.7$$
$$N \to sandwich \; 0.1$$
$$PN \to I \; 1.0$$
$$V \to saw \; 0.5$$
$$V \to ate \; 0.5$$
$$P \to with \; 0.6$$
$$P \to in \; 0.4$$
$$D \to a \; 0.3$$
$$D \to the \; 0.7$$

$$p(T) =$$

PCFGs



$p(T) = 1.0 \times$

$S \rightarrow NP\ VP$ 1.0

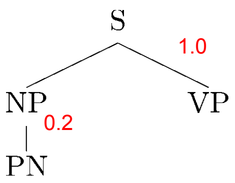
$VP \rightarrow V$ 0.2
$VP \rightarrow V\ NP$ 0.4
$VP \rightarrow VP\ PP$ 0.4

$NP \rightarrow NP\ PP$ 0.3
$NP \rightarrow D\ N$ 0.5
$NP \rightarrow PN$ 0.2

$PP \rightarrow P\ NP$ 1.0

$N \rightarrow girl$ 0.2
$N \rightarrow telescope$ 0.7
$N \rightarrow sandwich$ 0.1
$PN \rightarrow I$ 1.0
$V \rightarrow saw$ 0.5
$V \rightarrow ate$ 0.5
$P \rightarrow with$ 0.6
$P \rightarrow in$ 0.4
$D \rightarrow a$ 0.3
$D \rightarrow the$ 0.7

PCFGs



$S \rightarrow NP \;\; VP$ 1.0

$VP \rightarrow V$ 0.2
$VP \rightarrow V \;\; NP$ 0.4
$VP \rightarrow VP \;\; PP$ 0.4

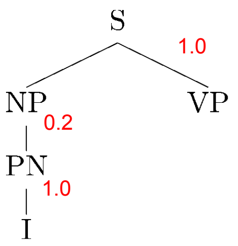
$NP \rightarrow NP \;\; PP$ 0.3
$NP \rightarrow D \;\; N$ 0.5
$NP \rightarrow PN$ 0.2

$PP \rightarrow P \;\; NP$ 1.0

$N \rightarrow girl$ 0.2
$N \rightarrow telescope$ 0.7
$N \rightarrow sandwich$ 0.1
$PN \rightarrow I$ 1.0
$V \rightarrow saw$ 0.5
$V \rightarrow ate$ 0.5
$P \rightarrow with$ 0.6
$P \rightarrow in$ 0.4
$D \rightarrow a$ 0.3
$D \rightarrow the$ 0.7

$$p(T) = 1.0 \times 0.2 \times$$

PCFGs



$S \rightarrow NP \ \ VP$ 1.0

$VP \rightarrow V$ 0.2
$VP \rightarrow V \ \ NP$ 0.4
$VP \rightarrow VP \ \ PP$ 0.4

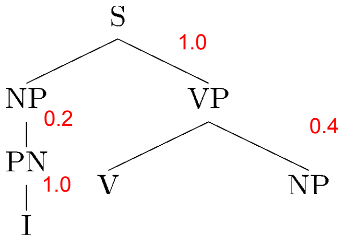
$NP \rightarrow NP \ \ PP$ 0.3
$NP \rightarrow D \ \ N$ 0.5
$NP \rightarrow PN$ 0.2

$PP \rightarrow P \ \ NP$ 1.0

$N \rightarrow girl$ 0.2
$N \rightarrow telescope$ 0.7
$N \rightarrow sandwich$ 0.1
$PN \rightarrow I$ 1.0
$V \rightarrow saw$ 0.5
$V \rightarrow ate$ 0.5
$P \rightarrow with$ 0.6
$P \rightarrow in$ 0.4
$D \rightarrow a$ 0.3
$D \rightarrow the$ 0.7

$$p(T) = 1.0 \times 0.2 \times 1.0 \times$$

PCFGs



$S \rightarrow NP \;\; VP$ 1.0

$VP \rightarrow V$ 0.2

$VP \rightarrow V \;\; NP$ 0.4

$VP \rightarrow VP \;\; PP$ 0.4

$NP \rightarrow NP \;\; PP$ 0.3

$NP \rightarrow D \;\; N$ 0.5

$NP \rightarrow PN$ 0.2

$PP \rightarrow P \;\; NP$ 1.0

$N \rightarrow girl$ 0.2

$N \rightarrow telescope$ 0.7

$N \rightarrow sandwich$ 0.1

$PN \rightarrow I$ 1.0

$V \rightarrow saw$ 0.5

$V \rightarrow ate$ 0.5

$P \rightarrow with$ 0.6

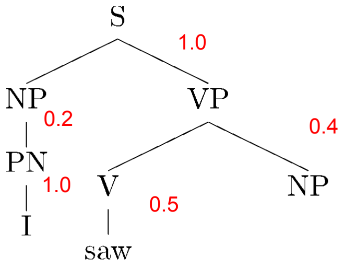
$P \rightarrow in$ 0.4

$D \rightarrow a$ 0.3

$D \rightarrow the$ 0.7

$$p(T) = 1.0 \times 0.2 \times 1.0 \times 0.4 \times$$

PCFGs



$S \to NP \ \ VP$ 1.0

$VP \to V$ 0.2

$VP \to V \ \ NP$ 0.4

$VP \to VP \ \ PP$ 0.4

$NP \to NP \ \ PP$ 0.3

$NP \to D \ \ N$ 0.5

$NP \to PN$ 0.2

$PP \to P \ \ NP$ 1.0

$N \to girl$ 0.2

$N \to telescope$ 0.7

$N \to sandwich$ 0.1

$PN \to I$ 1.0

$V \to saw$ 0.5

$V \to ate$ 0.5

$P \to with$ 0.6

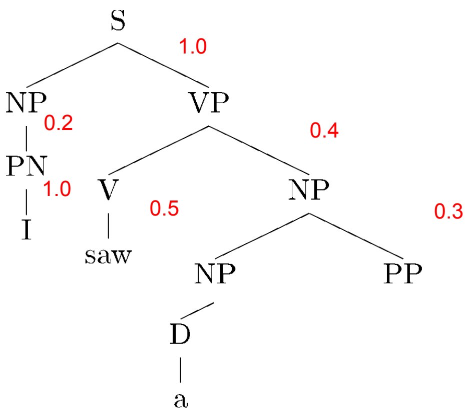
$P \to in$ 0.4

$D \to a$ 0.3

$D \to the$ 0.7

$$p(T) = 1.0 \times 0.2 \times 1.0 \times 0.4 \times 0.5 \times$$

PCFGs



$$S \rightarrow NP\ VP\ 1.0$$

$$VP \rightarrow V\ 0.2$$
$$VP \rightarrow V\ NP\ 0.4$$
$$VP \rightarrow VP\ PP\ 0.4$$

$$NP \rightarrow NP\ PP\ 0.3$$
$$NP \rightarrow D\ N\ 0.5$$
$$NP \rightarrow PN\ 0.2$$

$$PP \rightarrow P\ NP\ 1.0$$

$$N \rightarrow girl\ 0.2$$
$$N \rightarrow telescope\ 0.7$$
$$N \rightarrow sandwich\ 0.1$$
$$PN \rightarrow I\ 1.0$$
$$V \rightarrow saw\ 0.5$$
$$V \rightarrow ate\ 0.5$$
$$P \rightarrow with\ 0.6$$
$$P \rightarrow in\ 0.4$$
$$D \rightarrow a\ 0.3$$
$$D \rightarrow the\ 0.7$$

$$p(T) = 1.0 \times 0.2 \times 1.0 \times 0.4 \times 0.5 \times 0.3 \times$$

PCFGs

$S \to NP\ \ VP$ 1.0

$VP \to V$ 0.2

$VP \to V\ \ NP$ 0.4

$VP \to VP\ \ PP$ 0.4

$NP \to NP\ \ PP$ 0.3

$NP \to D\ \ N$ 0.5

$NP \to PN$ 0.2

$PP \to P\ \ NP$ 1.0

$N \to girl$ 0.2

$N \to telescope$ 0.7

$N \to sandwich$ 0.1

$PN \to I$ 1.0

$V \to saw$ 0.5

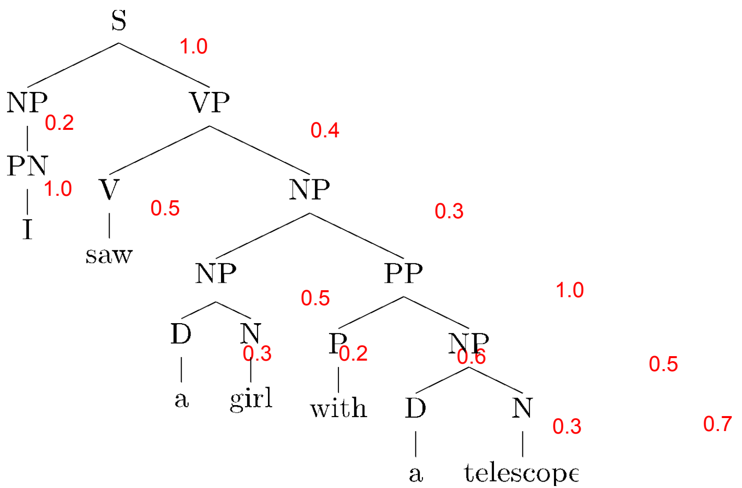
$V \to ate$ 0.5

$P \to with$ 0.6

$P \to in$ 0.4

$D \to a$ 0.3

$D \to the$ 0.7

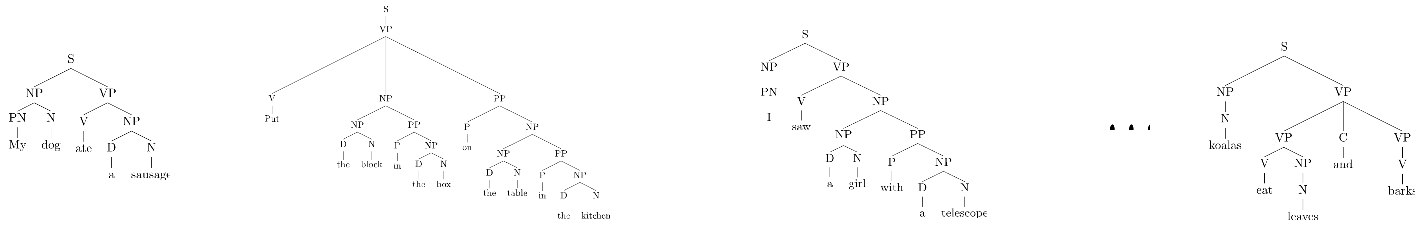


$$p(T) = 1.0 \times 0.2 \times 1.0 \times 0.4 \times 0.5 \times 0.3 \times$$
$$0.5 \times 0.3 \times 0.2 \times 1.0 \times 0.6 \times 0.5 \times 0.3 \times 0.7 = 2.26 \times 10^{-5}$$

PCFG Estimation

ML estimation

- A treebank: a collection sentences annotated with constituent trees



- An estimated probability of a rule (maximum likelihood estimates)

$$p(X \to \alpha) = \frac{C(X \to \alpha)}{C(X)}$$

The number of times the rule used in the corpus

The number of times the nonterminal X appears in the treebank

- Smoothing is helpful
  - Especially important for preterminal rules

CKY Parsing

Parsing

- Parsing is search through the space of all possible parses
  - e.g., we may want either any parse, all parses or the highest scoring parse (if PCFG):

$$\arg\max_{T \in G(x)} P(T)$$

- Bottom-up:
  - One starts from words and attempt to construct the full tree
- Top-down
  - Start from the start symbol and attempt to expand to get the sentence

CKY algorithm (aka CYK)

- Cocke-Kasami-Younger algorithm
  - Independently discovered in late 60s / early 70s

- An efficient bottom up parsing algorithm for (P)CFGs
  - can be used both for the recognition and parsing problems
  - Very important in NLP (and beyond)

- We will start with the non-probabilistic version

Constraints on the grammar

- The basic CKY algorithm supports only rules in the Chomsky Normal Form (CNF):

$$C \to x$$

$$C \to C_1 C_2$$

Unary preterminal rules (generation of words given PoS tags)

$N \to telescope$ $\quad D \to the$

Binary inner rules $S \to NP\ VP$ $\quad NP \to D\ N$

Constraints on the grammar

- The basic CKY algorithm supports only rules in the Chomsky Normal Form (CNF):

$$C \to x$$

$$C \to C_1 C_2$$

- Any CFG can be converted to an equivalent CNF
  - Equivalent means that they define the same language
  - However (syntactic) trees will look differently
  - It is possible to address it by defining such transformations that allows for easy reverse transformation

Transformation to CNF form

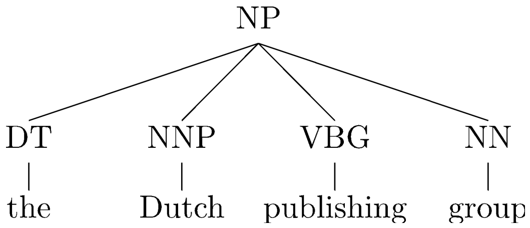
- What one need to do to convert to CNF form
  - Get rid of rules that mix terminals and non-terminals
  - Get rid of unary rules: $C \to C_1$
  - Get rid of N-ary rules: $C \to C_1\ C_2 \ldots C_n \quad (n > 2)$

Crucial to process them, as required for efficient parsing

Transformation to CNF form: binarization

- Consider

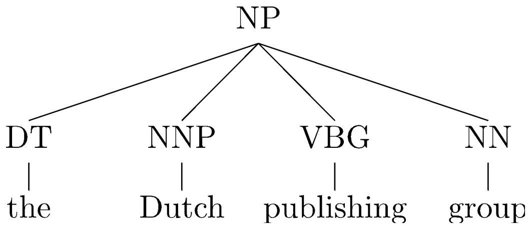
$$NP \rightarrow DT \;\; NNP \;\; VBG \;\; NN$$



- How do we get a set of binary rules which are equivalent?

Transformation to CNF form: binarization

- Consider $NP \rightarrow DT \;\; NNP \;\; VBG \;\; NN$

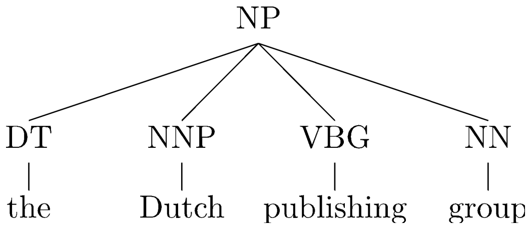


- How do we get a set of binary rules which are equivalent?

$$NP \rightarrow DT \;\; X$$
$$X \rightarrow NNP \;\; Y$$
$$Y \rightarrow VBG \;\; NN$$

Transformation to CNF form: binarization

- Consider $NP \rightarrow DT \;\; NNP \;\; VBG \;\; NN$



- How do we get a set of binary rules which are equivalent?

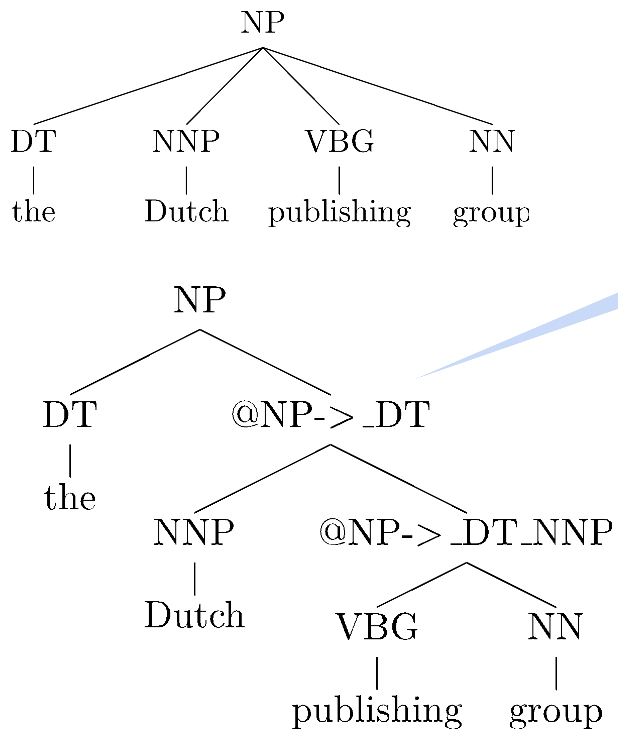
$$NP \rightarrow DT \;\; X$$
$$X \rightarrow NNP \;\; Y$$
$$Y \rightarrow VBG \;\; NN$$

- A more systematic way to refer to new non-terminals

$$NP \rightarrow DT \;\; @NP|DT$$
$$@NP|DT \rightarrow NNP \;\; @NP|DT\_NNP$$
$$@NP|DT\_NNP \rightarrow VBG \;\; NN$$

Transformation to CNF form: binarization

- Instead of binarizing tuples we can binarize trees on preprocessing:



Also known as **lossless Markovization** in the context of PCFGs

Can be easily reversed on postprocessing

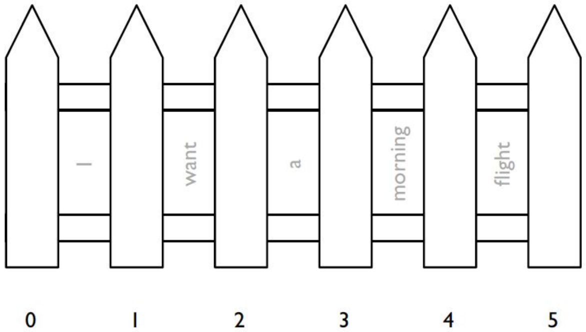
CKY: Parsing task

- We are given
  - a grammar <N, T, S, R>
  - a sequence of words $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)$

- Our goal is to produce a parse tree for w

CKY: Parsing task

- We a given
  - a grammar <N, T, S, R>
  - a sequence of words $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)$
- Our goal is to produce a parse tree for w
- We need an easy way to refer to substrings of w



indices refer to fenceposts

span (i, j) refers to words between fenceposts i and j

Parsing one word

$C \rightarrow w_i$

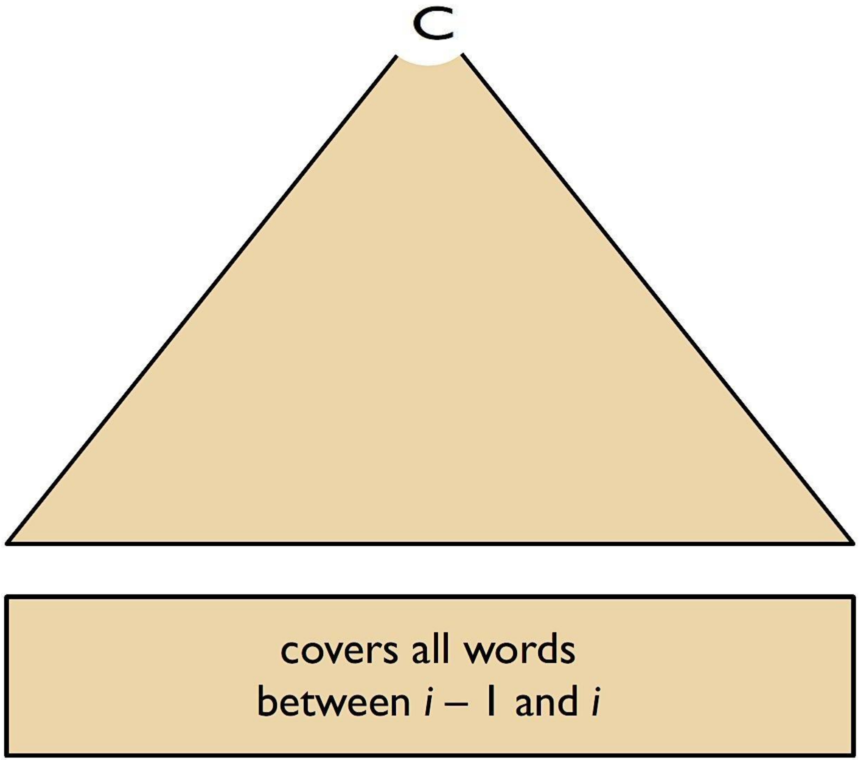
w_i

Parsing one word

$$C \rightarrow w_i$$

Parsing one word

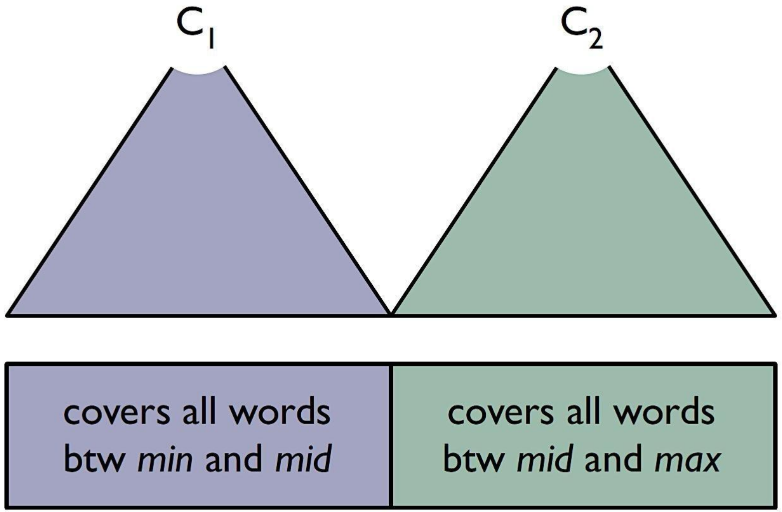
$C \rightarrow w_i$



Parsing longer spans

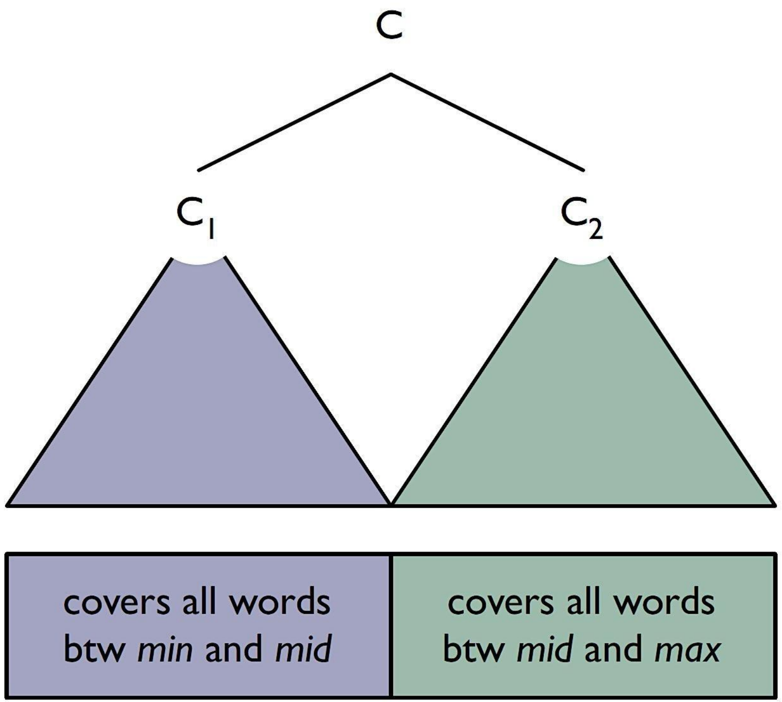
$$C \rightarrow C_1 \;\; C_2$$

Check through all C1, C2, mid



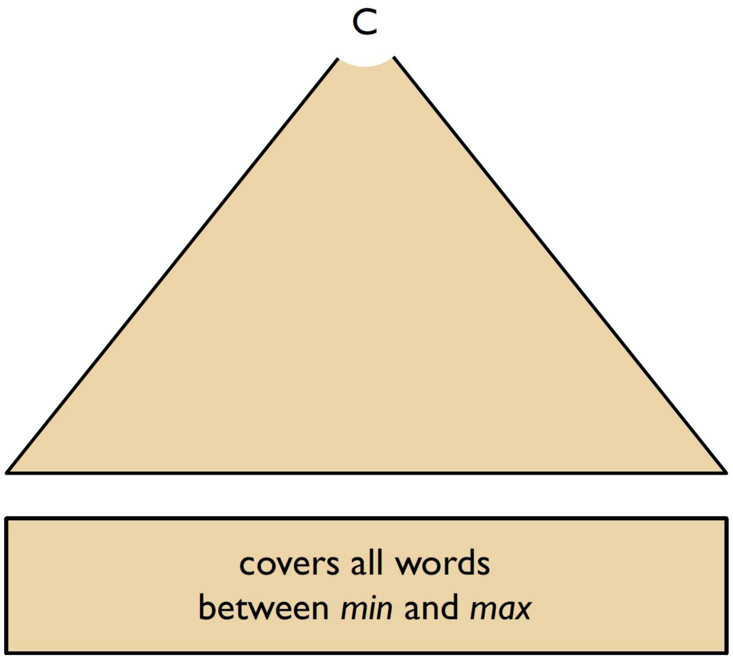
Parsing longer spans

$$C \rightarrow C_1 \ \ C_2$$



Check through all C1, C2, mid

Parsing longer spans



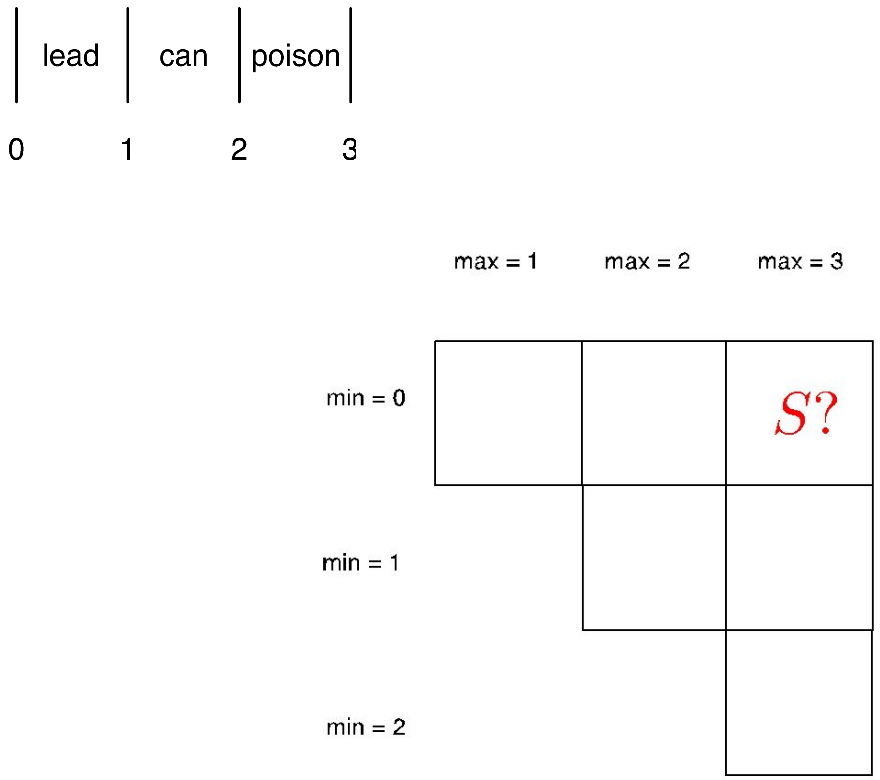


Chart (aka parsing triangle)

Inner rules

$$S \rightarrow NP\ VP$$

$$VP \rightarrow M\ V$$

$$VP \rightarrow V$$

$$NP \rightarrow N$$

$$NP \rightarrow N\ NP$$

Preterminal rules

$$N \rightarrow can$$

$$N \rightarrow lead$$

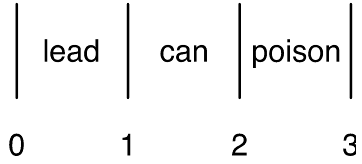
$$N \rightarrow poison$$

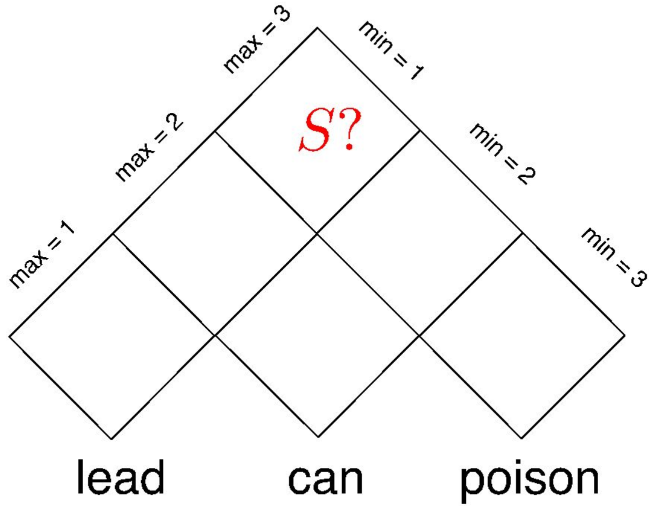
$$M \rightarrow can$$

$$M \rightarrow must$$

$$V \rightarrow poison$$

$$V \rightarrow lead$$





$$S \rightarrow NP\ VP$$

$$VP \rightarrow M\ V$$
$$VP \rightarrow V$$

$$NP \rightarrow N$$
$$NP \rightarrow N\ NP$$

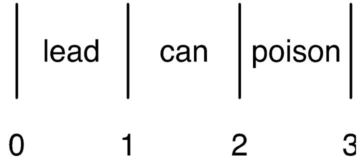
Inner rules

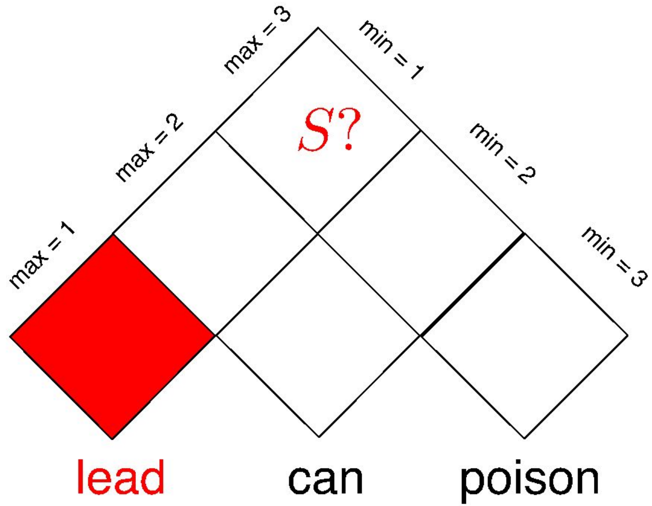
$$N \rightarrow can$$
$$N \rightarrow lead$$
$$N \rightarrow poison$$

$$M \rightarrow can$$
$$M \rightarrow must$$

$$V \rightarrow poison$$
$$V \rightarrow lead$$

Preterminal rules





$$S \to NP\ VP$$

$$VP \to M\ V$$
$$VP \to V$$

$$NP \to N$$
$$NP \to N\ NP$$

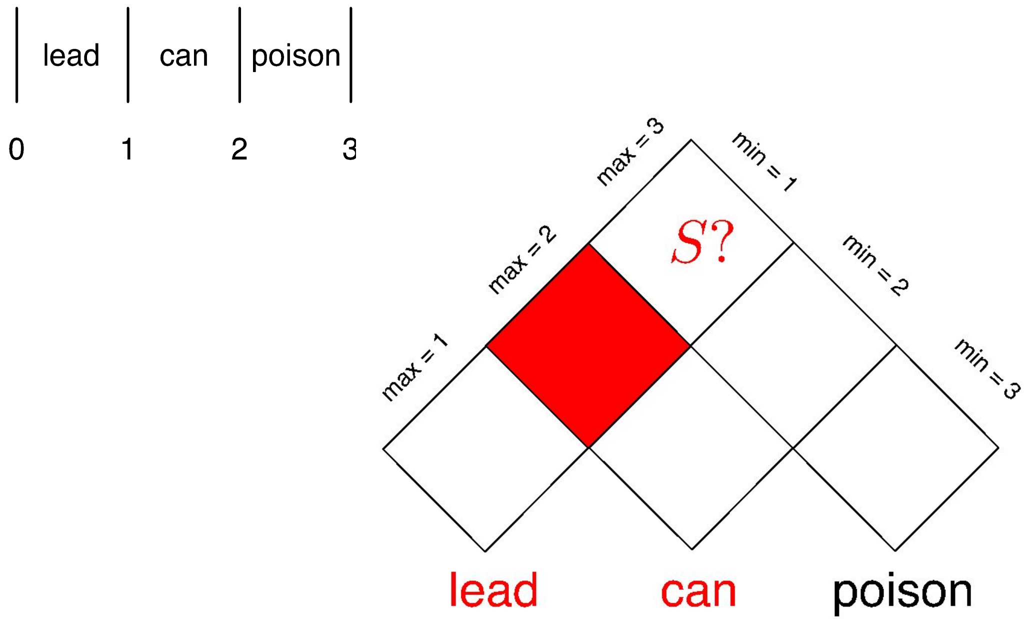
Inner rules

$$N \to can$$
$$N \to lead$$
$$N \to poison$$

$$M \to can$$
$$M \to must$$

$$V \to poison$$
$$V \to lead$$

Preterminal rules



$S \to NP\ VP$

$VP \to M\ V$

$VP \to V$

$NP \to N$

$NP \to N\ NP$

Inner rules

$N \to can$

$N \to lead$

$N \to poison$

$M \to can$

$M \to must$

$V \to poison$

$V \to lead$

Preterminal rules

	lead		can		poison	
0		1		2		3

	max = 1	max = 2	max = 3
min = 0			$S?$
min = 1			
min = 2			

$$S \to NP\ VP$$

$$VP \to M\ \ V$$

$$VP \to V$$

$$NP \to N$$

$$NP \to N\ NP$$

Inner rules

$$N \to can$$

$$N \to lead$$

$$N \to poison$$

$$M \to can$$

$$M \to must$$

$$V \to poison$$

$$V \to lead$$

Preterminal rules

lead	can	poison

0 1 2 3

	max = 1	max = 2	max = 3
min = 0	1	4	6 $S?$
min = 1		2	5
min = 2			3

Inner rules

$$S \rightarrow NP\ VP$$

$$VP \rightarrow M\ V$$

$$VP \rightarrow V$$

$$NP \rightarrow N$$

$$NP \rightarrow N\ NP$$

Preterminal rules

$$N \rightarrow can$$

$$N \rightarrow lead$$

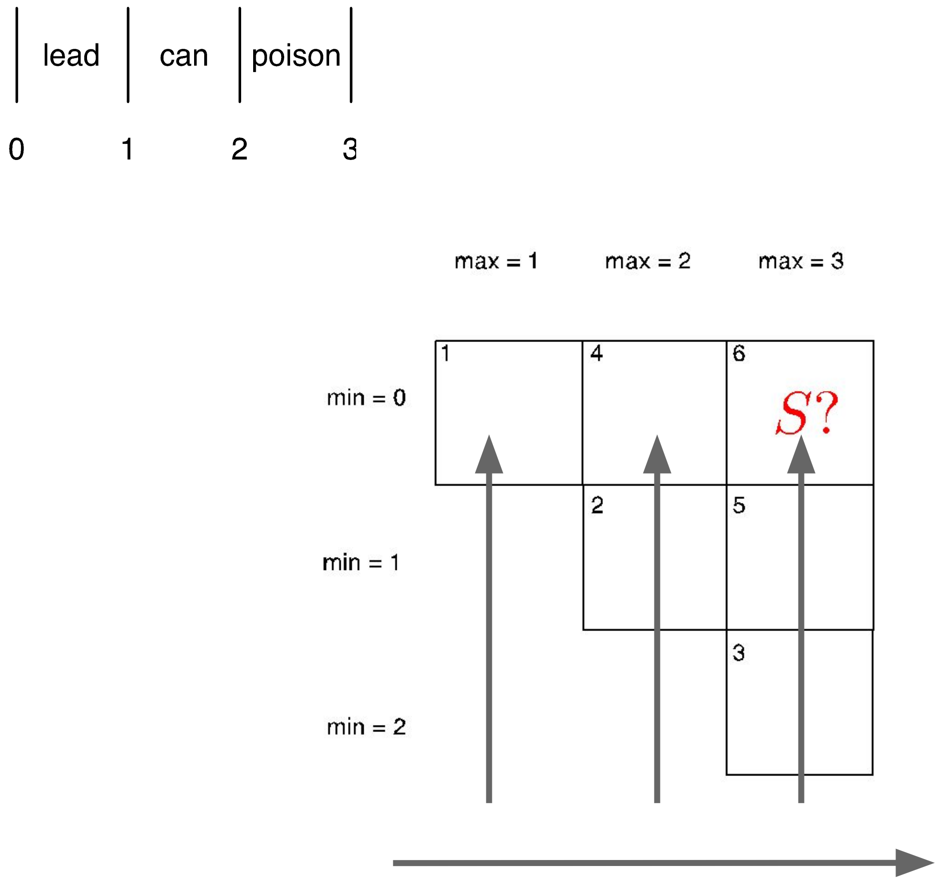
$$N \rightarrow poison$$

$$M \rightarrow can$$

$$M \rightarrow must$$

$$V \rightarrow poison$$

$$V \rightarrow lead$$



Inner rules

$$S \rightarrow NP\ VP$$

$$VP \rightarrow M\ V$$

$$VP \rightarrow V$$

$$NP \rightarrow N$$

$$NP \rightarrow N\ NP$$

Preterminal rules

$$N \rightarrow can$$

$$N \rightarrow lead$$

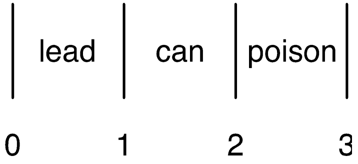
$$N \rightarrow poison$$

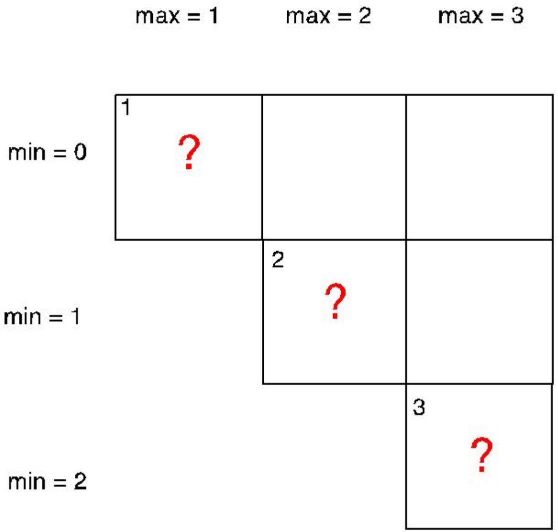
$$M \rightarrow can$$

$$M \rightarrow must$$

$$V \rightarrow poison$$

$$V \rightarrow lead$$





Inner rules

$$S \rightarrow NP\ VP$$

$$VP \rightarrow M\ V$$

$$VP \rightarrow V$$

$$NP \rightarrow N$$

$$NP \rightarrow N\ NP$$

Preterminal rules

$$N \rightarrow can$$

$$N \rightarrow lead$$

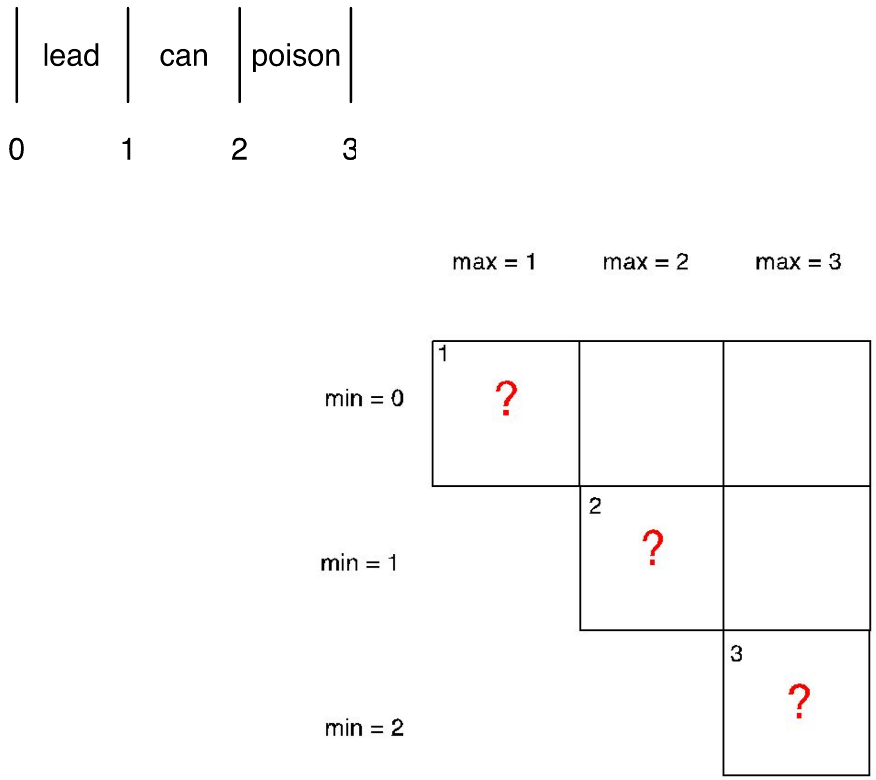
$$N \rightarrow poison$$

$$M \rightarrow can$$

$$M \rightarrow must$$

$$V \rightarrow poison$$

$$V \rightarrow lead$$



Inner rules

$$S \rightarrow NP\ VP$$

$$VP \rightarrow M\ V$$

$$VP \rightarrow V$$

$$NP \rightarrow N$$

$$NP \rightarrow N\ NP$$

Preterminal rules

$$N \rightarrow can$$

$$N \rightarrow lead$$

$$N \rightarrow poison$$

$$M \rightarrow can$$

$$M \rightarrow must$$

$$V \rightarrow poison$$

$$V \rightarrow lead$$

lead	can	poison

0 1 2 3

	max = 1	max = 2	max = 3
min = 0	1 N, V		
min = 1		2 N, M	
min = 2			3 N, V

Inner rules

$$S \rightarrow NP\ VP$$

$$VP \rightarrow M\ V$$

$$VP \rightarrow V$$

$$NP \rightarrow N$$

$$NP \rightarrow N\ NP$$

Preterminal rules

$$N \rightarrow can$$

$$N \rightarrow lead$$

$$N \rightarrow poison$$

$$M \rightarrow can$$

$$M \rightarrow must$$

$$V \rightarrow poison$$

$$V \rightarrow lead$$

lead	can	poison

0 1 2 3

	max = 1	max = 2	max = 3
min = 0	1 N, V NP, VP	4 **?**	
min = 1		2 N, M NP	
min = 2			3 N, V NP, VP

Inner rules

$$S \rightarrow NP\ VP$$

$$VP \rightarrow M\ V$$

$$VP \rightarrow V$$

$$NP \rightarrow N$$

$$NP \rightarrow N\ NP$$

Preterminal rules

$$N \rightarrow can$$

$$N \rightarrow lead$$

$$N \rightarrow poison$$

$$M \rightarrow can$$

$$M \rightarrow must$$

$$V \rightarrow poison$$

$$V \rightarrow lead$$

| lead | can | poison |

0 1 2 3

	max = 1	max = 2	max = 3
min = 0	1 N, V NP, VP	4 ?	
min = 1		2 N, M NP	
min = 2			3 N, V NP, VP

Inner rules

$S \rightarrow NP\ VP$

$VP \rightarrow M\ V$

$VP \rightarrow V$

$NP \rightarrow N$

$NP \rightarrow N\ NP$

Preterminal rules

$N \rightarrow can$

$N \rightarrow lead$

$N \rightarrow poison$

$M \rightarrow can$

$M \rightarrow must$

$V \rightarrow poison$

$V \rightarrow lead$

lead	can	poison

0 1 2 3

	max = 1	max = 2	max = 3
min = 0	1 N, V NP, VP	4 NP	
min = 1		2 N, M NP	
min = 2			3 N, V NP, VP

$S \rightarrow NP\ VP$

$VP \rightarrow M\ V$

$VP \rightarrow V$

$NP \rightarrow N$

$NP \rightarrow N\ NP$

Inner rules

$N \rightarrow can$

$N \rightarrow lead$

$N \rightarrow poison$

$M \rightarrow can$

$M \rightarrow must$

$V \rightarrow poison$

$V \rightarrow lead$

Preterminal rules

| lead | can | poison |

0 1 2 3

	max = 1	max = 2	max = 3
min = 0	1 N, V NP, VP	4 NP	
min = 1		2 N, M NP	5 ?
min = 2			3 N, V NP, VP

Inner rules

$$S \rightarrow NP\ VP$$

$$VP \rightarrow M\ V$$
$$VP \rightarrow V$$

$$NP \rightarrow N$$
$$NP \rightarrow N\ NP$$

Preterminal rules

$$N \rightarrow can$$
$$N \rightarrow lead$$
$$N \rightarrow poison$$

$$M \rightarrow can$$
$$M \rightarrow must$$

$$V \rightarrow poison$$
$$V \rightarrow lead$$

lead	can	poison

0 1 2 3

	max = 1	max = 2	max = 3
min = 0	1 N, V NP, VP	4 NP	
min = 1		2 N, M NP	5 $S, VP,$ NP
min = 2			3 N, V NP, VP

Inner rules

$S \rightarrow NP\ VP$

$VP \rightarrow M\ V$

$VP \rightarrow V$

$NP \rightarrow N$

$NP \rightarrow N\ NP$

Preterminal rules

$N \rightarrow can$

$N \rightarrow lead$

$N \rightarrow poison$

$M \rightarrow can$

$M \rightarrow must$

$V \rightarrow poison$

$V \rightarrow lead$

lead	can	poison

0 1 2 3

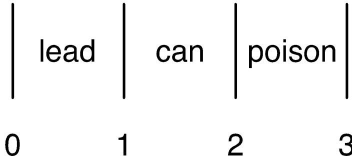
	max = 1	max = 2	max = 3
min = 0	1 N, V NP, VP	4 NP	6 ?
min = 1		2 N, M NP	5 $S, VP,$ NP
min = 2			3 N, V NP, VP

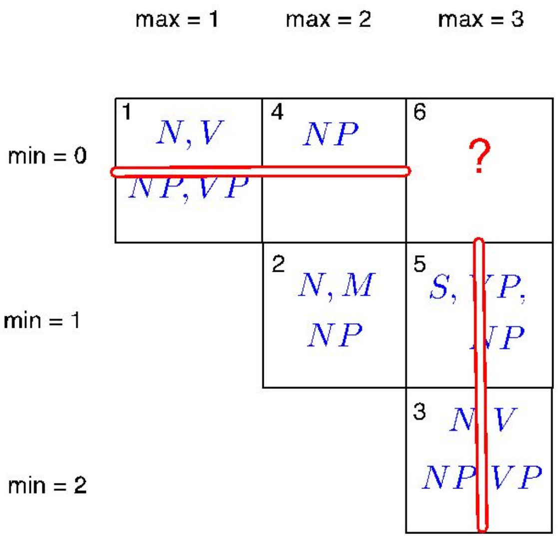
Inner rules

$$S \to NP\ VP$$
$$VP \to M\ V$$
$$VP \to V$$
$$NP \to N$$
$$NP \to N\ NP$$

Preterminal rules

$$N \to can$$
$$N \to lead$$
$$N \to poison$$
$$M \to can$$
$$M \to must$$
$$V \to poison$$
$$V \to lead$$





Inner rules

$$S \rightarrow NP\ VP$$

$$VP \rightarrow M\ V$$

$$VP \rightarrow V$$

$$NP \rightarrow N$$

$$NP \rightarrow N\ NP$$

Preterminal rules

$$N \rightarrow can$$

$$N \rightarrow lead$$

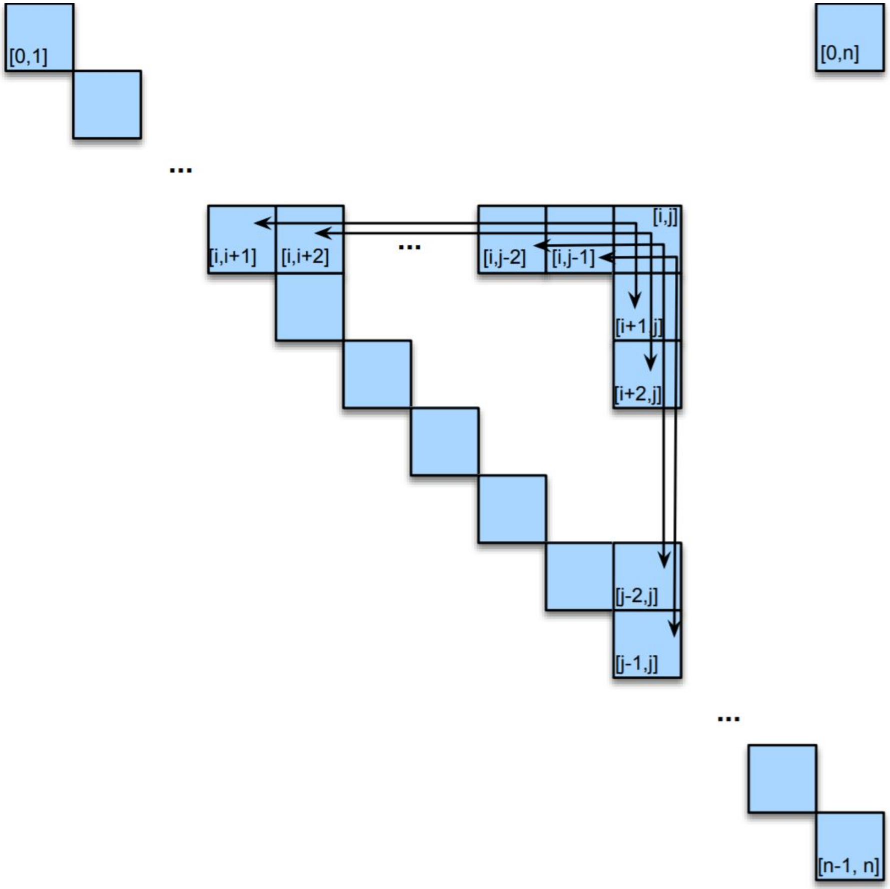
$$N \rightarrow poison$$

$$M \rightarrow can$$

$$M \rightarrow must$$

$$V \rightarrow poison$$

$$V \rightarrow lead$$


[0,1]
[0,n]
...
[i,i+1]
[i,i+2]
...
[i,j-2]
[i,j-1]
[i,j]
[i+1,j]
[i+2,j]
[j-2,j]
[j-1,j]
...
[n-1, n]

	lead		can		poison	
0		1		2		3

mid=1

	max = 1	max = 2	max = 3
min = 0	1 N, V NP, VP	4 NP	6 S, NP
min = 1		2 N, M NP	5 $S, VP,$ NP
min = 2			3 N, V NP, VP

Inner rules

$$S \rightarrow NP\ VP$$

$$VP \rightarrow M\ V$$

$$VP \rightarrow V$$

$$NP \rightarrow N$$

$$NP \rightarrow N\ NP$$

Preterminal rules

$$N \rightarrow can$$

$$N \rightarrow lead$$

$$N \rightarrow poison$$

$$M \rightarrow can$$

$$M \rightarrow must$$

$$V \rightarrow poison$$

$$V \rightarrow lead$$

| lead | can | poison |
0 1 2 3

mid=2

	max = 1	max = 2	max = 3
min = 0	1 N, V NP, VP	4 NP	6 S, NP $S(?!)$
min = 1		2 N, M NP	5 $S, VP,$ NP
min = 2			3 N, V NP, VP

$S \rightarrow NP\ VP$

$VP \rightarrow M\ V$

$VP \rightarrow V$

$NP \rightarrow N$

$NP \rightarrow N\ NP$

Inner rules

$N \rightarrow can$

$N \rightarrow lead$

$N \rightarrow poison$

$M \rightarrow can$

$M \rightarrow must$

$V \rightarrow poison$

$V \rightarrow lead$

Preterminal rules

$S \rightarrow NP\ VP$

lead | can | poison

0 1 2 3

	max = 1	max = 2	max = 3
min = 0	1 N, V NP, VP	4 NP	6 S, NP S(?!)
min = 1		2 N, M NP	5 S, VP, NP
min = 2			3 N, V NP, VP

Apparently the sentence is ambiguous for the grammar: (as the grammar overgenerates)

Inner rules

$VP \rightarrow M\ V$

$VP \rightarrow V$

$NP \rightarrow N$

$NP \rightarrow N\ NP$

Preterminal rules

$N \rightarrow can$

$N \rightarrow lead$

$N \rightarrow poison$

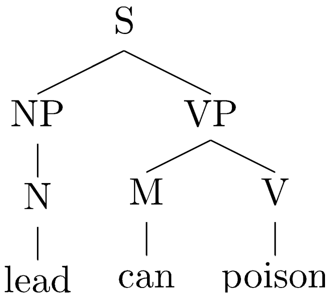
$M \rightarrow can$

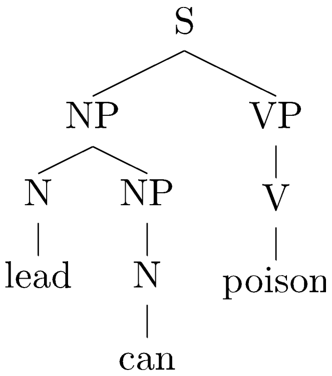
$M \rightarrow must$

$V \rightarrow poison$

$V \rightarrow lead$

Ambiguity





No subject-verb agreement, and *poison* used as an intransitive verb