

LING/C SC 581:

Advanced Computational Linguistics

Lecture 24

Today's Topic

- Last Time:
 - `t.productions()` (= *phrase structure rules*)
 - `p.lhs()/p.rhs()`
 - `len(p.rhs())` (= # children)
 - `p.is_nonlexical()/p.is_lexical()`
 - `p.lhs().symbol()` (nonterminal symbol)
- A CFG experiment with the ptb:
 - *CFG = Context-free Grammar*
 - *How many common ptb rules do we need to parse a sentence?*

Distribution of Productions

```
>>> import matplotlib.pyplot as plt
>>> import numpy as np
>>> from scipy.stats import norm
>>> from statistics import mean, stdev
```

```
>>> from nltk.corpus import ptb
>>> trees = ptb.parsed_sents()
>>> len(trees)
```

73451

```
>>> lrulec = [len(t.productions()) for t in trees]
>>> sum(lrulec)
```

3131242



lrulec: list of rule counts

Distribution of Productions

```
>>> mean(lrulec)
42.63035220759418
>>> stdev(lrulec)
23.671107782972243
>>> m = mean(lrulec)
>>> sd = stdev(lrulec)
>>> x = np.linspace(m-2*sd,m+2*sd,100)

>>> plt.plot(x,norm.pdf(x,m,sd))

>>> plt.grid()
>>> plt.show()
```

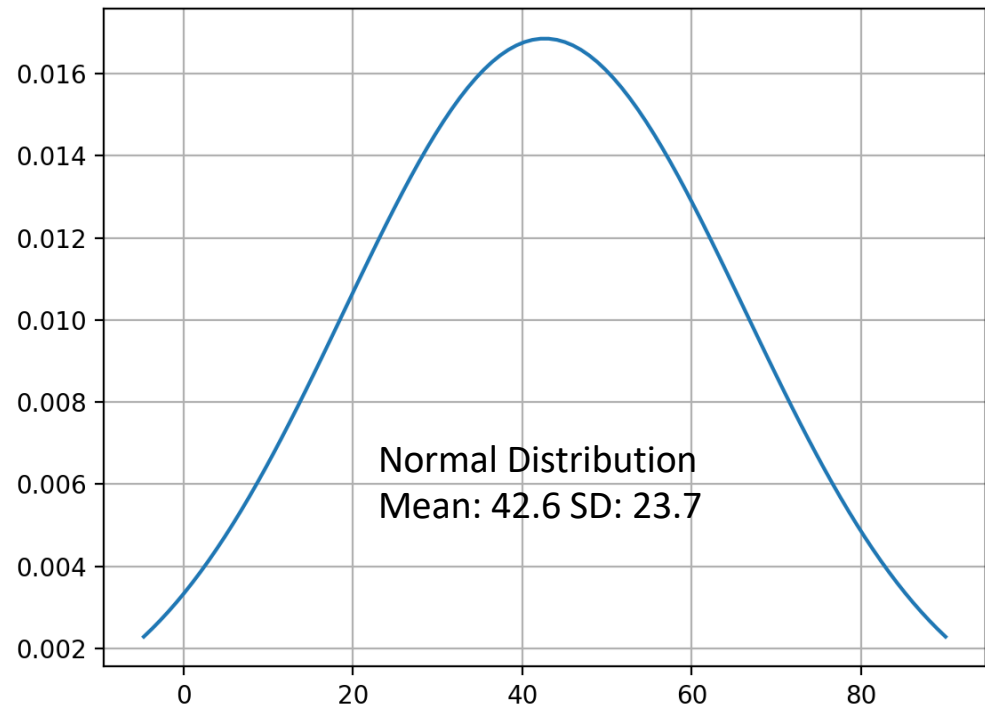
```
numpy.linspace(start, stop, num, dtype=None, axis=0)
Return evenly spaced values within a given interval.
Returns num evenly spaced values, including endpoints.
```

The probability density function of a normal distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

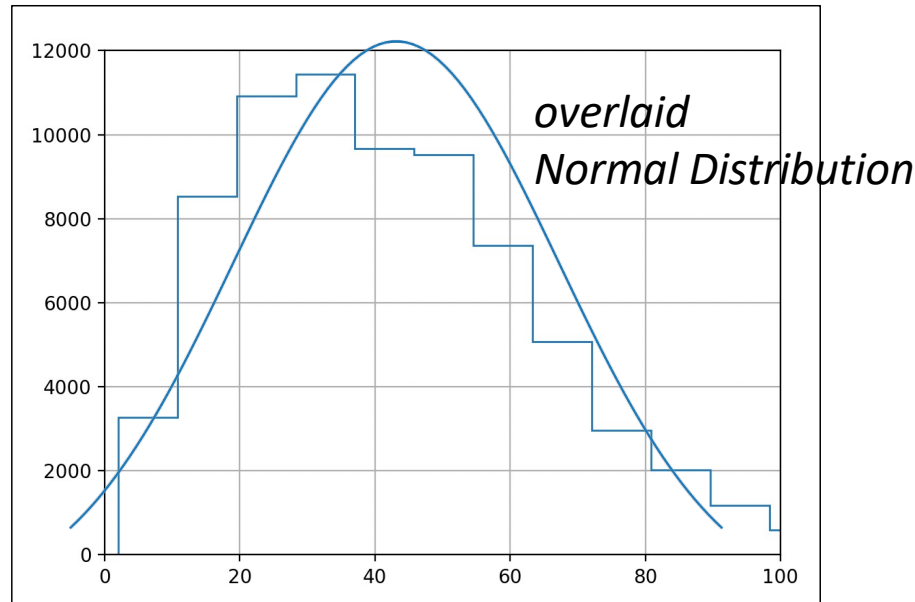
for a real number x .

The probability density function of a normal distribution use the `loc` and `scale` parameters. Specifically, `norm.pdf(x, loc, scale)` is



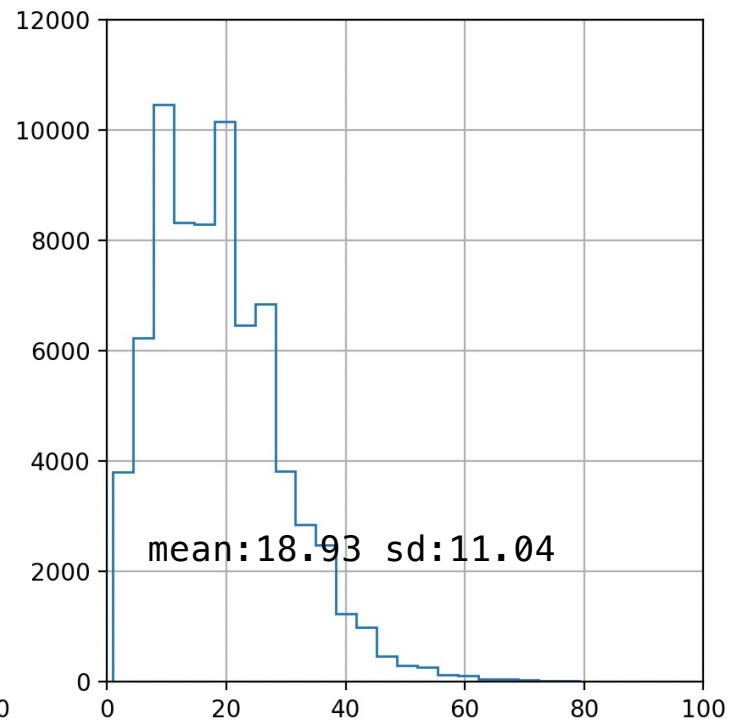
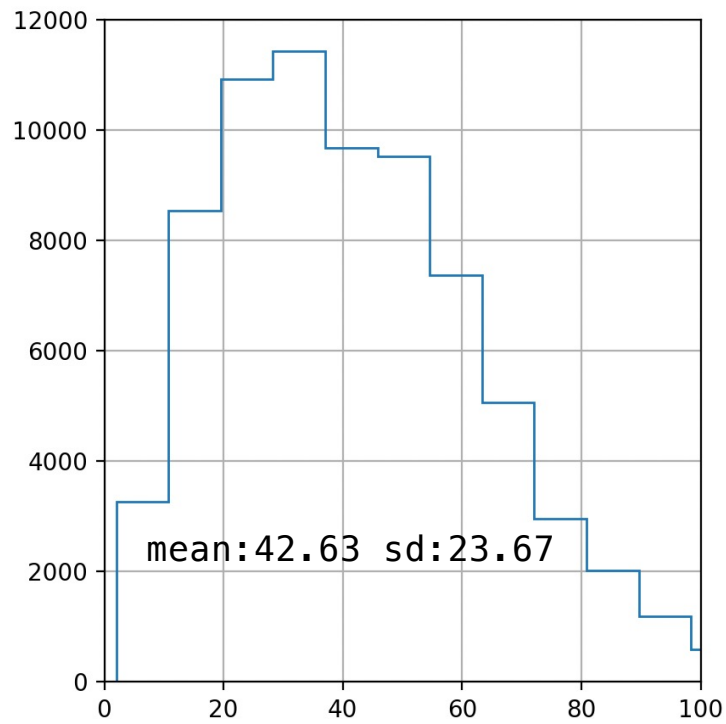
Distribution of Productions

```
>>> plt.hist(lrulec,histtype='step',bins=50)  
>>> plt.xlim(0,100)  
>>> plt.grid()  
>>> plt.show()
```

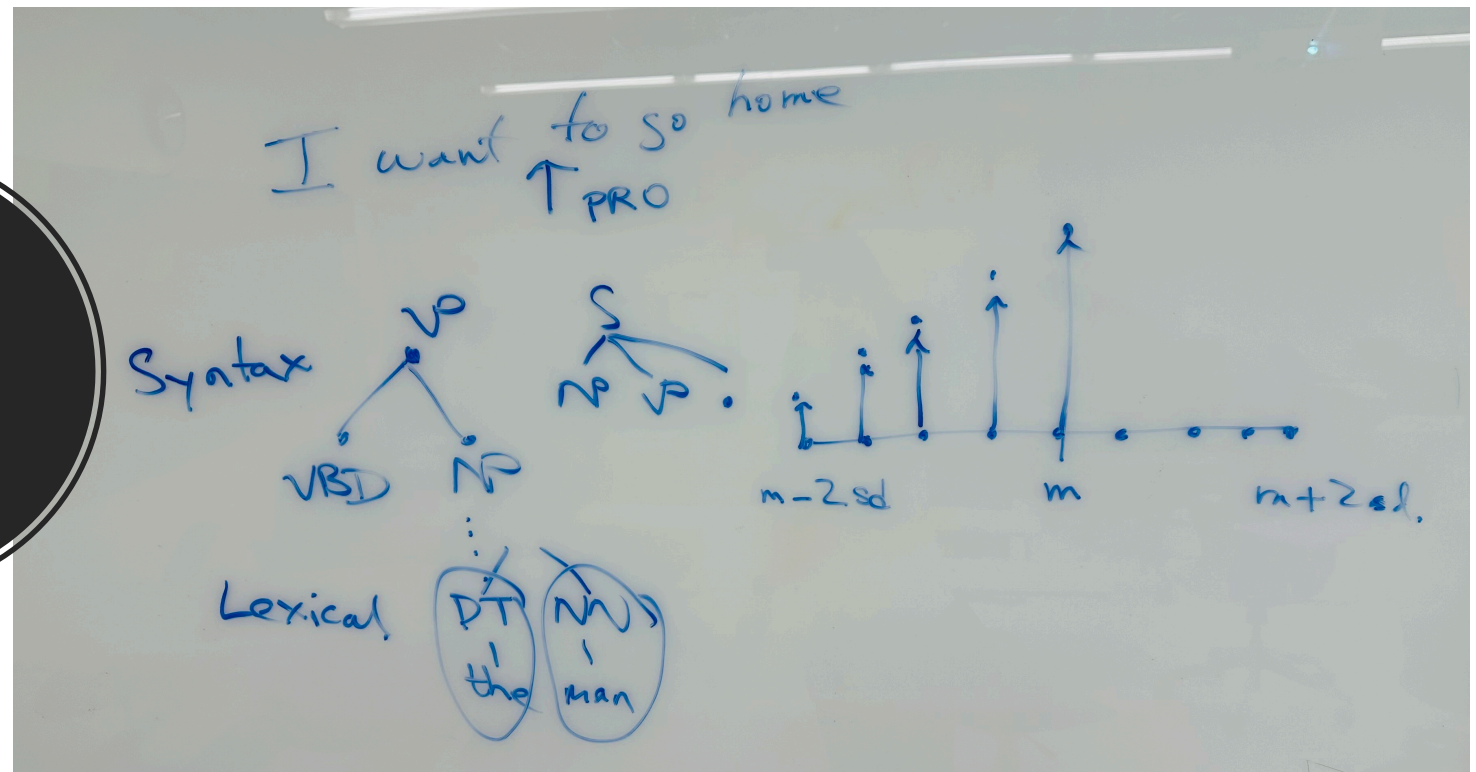


Distribution of Productions

with (*left*) and without (*right*) (word POS) nodes



Distribution
of
Productions



of Productions

Suppose list `rules` be all the rules in the ptb.

```
>>> rules = [p for t in trees for p in t productions()]  
>>> len(rules)
```

3131242

- Then let `srules` be all the (syntax), i.e. *nonlexical*, rules:

```
>>> srules = [r for r in rules if not(len(r.rhs()) == 1  
    and isinstance(r.rhs()[0], str))]  
>>> len(srules)
```

1390347

of Productions

```
>>> fd = nltk.FreqDist(srules)
```

top-5 highlighted there: *interesting!*



```
>>> fd
```

```
FreqDist({PP -> IN NP: 78040, S -> NP-SBJ VP: 63335, NP  
-> DT NN: 40876, NP-SBJ -> -NONE-: 39712, NP -> NP PP:  
35819, NP-SBJ -> PRP: 31272, S -> NP-SBJ VP .: 24467, VP  
-> TO VP: 21899, NP -> NN: 20798, NP -> -NONE-: 20312,  
...})
```

```
>>> len(fd)
```

```
52134
```

← *Is this # cognitively plausible?*



```
>>> fd.N()
```

```
1390347
```

of Productions

```
• >>> fd.most_common(20)
1. [(PP -> IN NP, 78040),
2. (S -> NP-SBJ VP, 63335),
3. (NP -> DT NN, 40876),
4. (NP-SBJ -> -NONE-, 39712),
5. (NP -> NP PP, 35819),
6. (NP-SBJ -> PRP, 31272),
7. (S -> NP-SBJ VP ., 24467),
8. (VP -> TO VP, 21899),
9. (NP -> NN, 20798),
10. (NP -> -NONE-, 20312),
11. (PP-LOC -> IN NP, 18021),
12. (ADVP -> RB, 15449),
13. (NP -> DT JJ NN, 14898),
14. (NP -> NNS, 14875),
15. (VP -> MD VP, 13714),
16. (NP -> NNP, 12767),
17. (VP -> VB NP, 12730),
18. (PP-TMP -> IN NP, 11032),
19. (NP -> PRP, 10988),
20. (SBAR -> -NONE- S, 10774)]
```

of Productions

- Most syntax rules only occur once:

```
>>> len(fd)
```

```
52134
```

```
>>> fd.N()
```

```
1390347
```

```
>>> len(fd.hapaxes())
```

```
33631
```

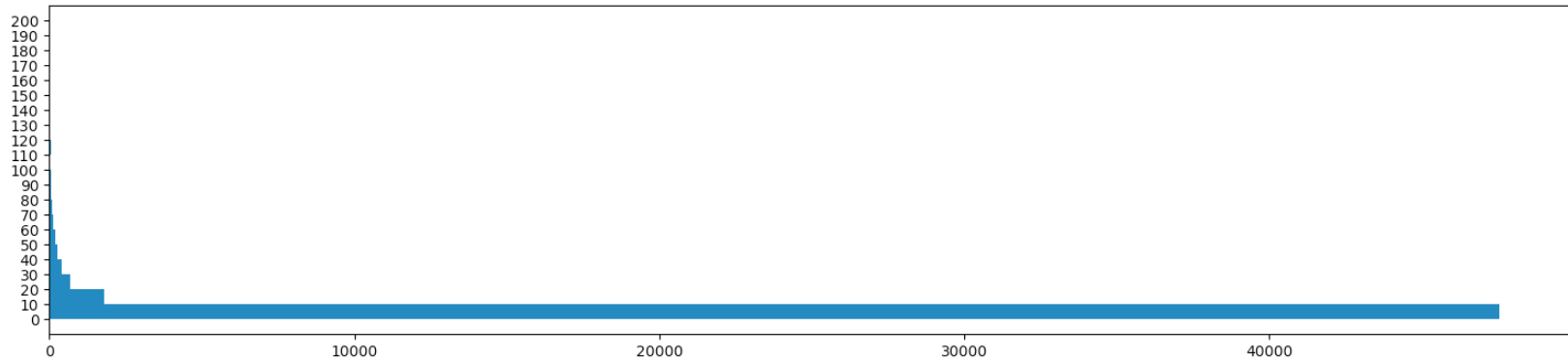
```
hapaxes()
```

```
Return a list of all samples that occur once (hapax legomena)
```

How do they get learnt?

of Productions: histogram

binned: # of times
rule occurs



of rules

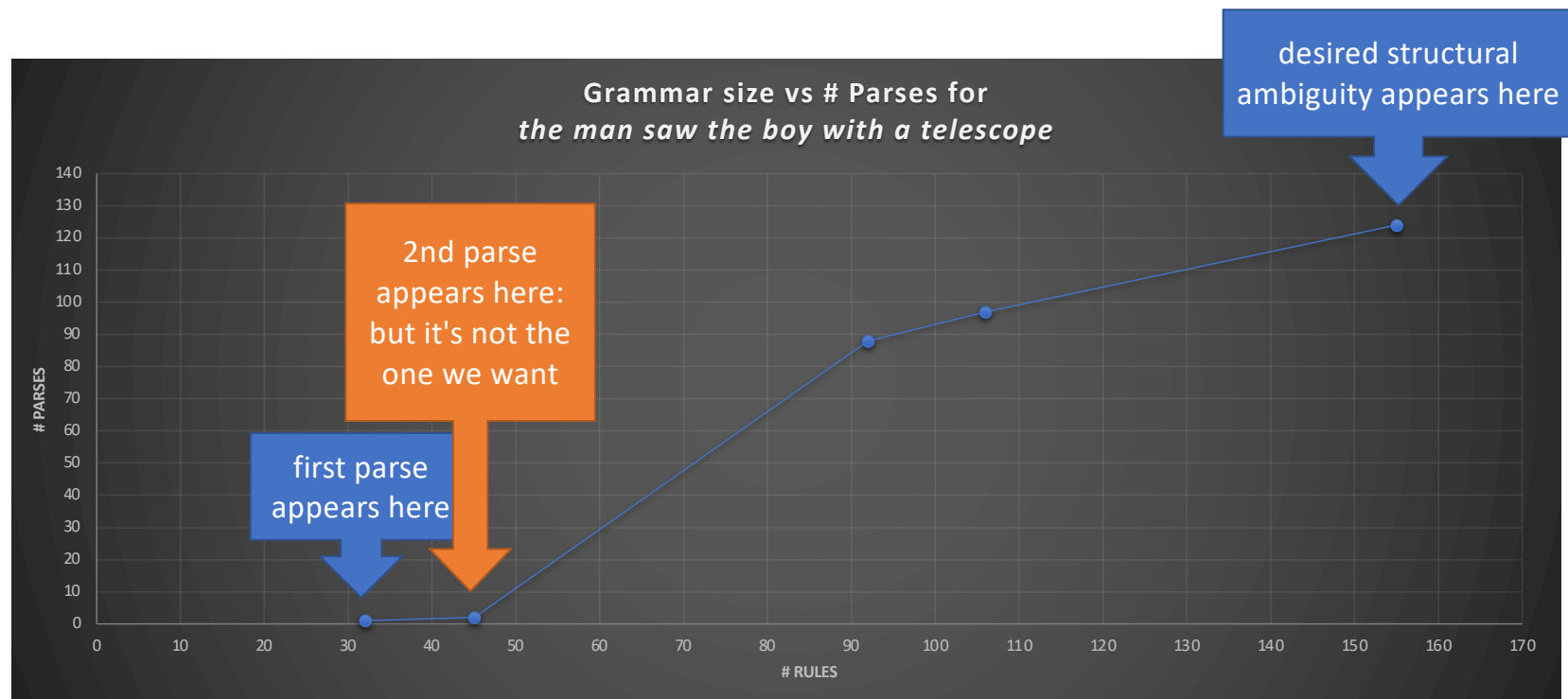
Smallest grammar for a sentence

- **Propose an Experiment:**

- suppose we start with an empty CFG (no rules)
- take PTB rules in order of frequency (*highest first*)
- add them one at a time to the grammar
- how many PTB rules before we can parse this sentence?^{*}
 - *the man saw the boy with a telescope*
- how many PTB rules do we need to obtain the structural ambiguity?

^{*}some simplifications applied, explained later.

Smallest grammar for a sentence



Smallest grammar for a sentence

```
>>> find_scfg(s)
```

```
the DT 73202
```

```
man NN 886
```

```
saw VBD 329
```

```
the DT 73202
```

```
boy NN 191
```

```
with IN 7953
```

```
a DT 32606
```

```
telescope NN 3
```

```
(S
```

```
  (NP-SBJ (DT the) (NN man))
```

```
  (VP
```

```
    (VBD saw)
```

```
    (NP
```

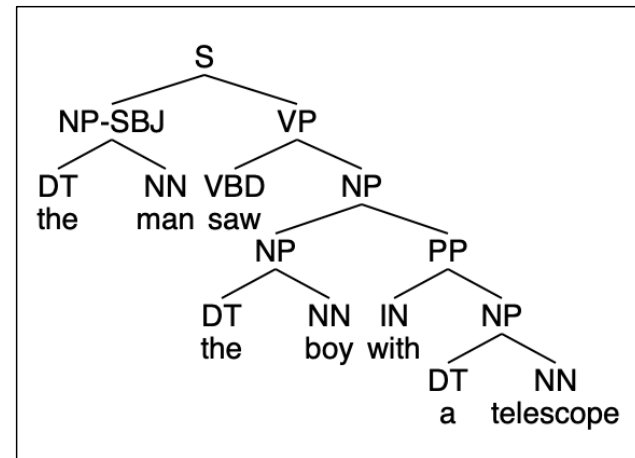
```
      (NP (DT the) (NN boy))
```

```
      (PP (IN with) (NP (DT a) (NN telescope))))))
```

```
Parses: 1,
```

```
# rules: 32,
```

```
# lexical rules: 7
```



Smallest grammar for a sentence


Parses: 1, # rules: 32, # lexical rules: 7

- | | | |
|---------------------|---------------------|-----------------------|
| 1. S -> NP-SBJ VP | 14. VP -> MD VP | 27. NP-SBJ -> NP PP |
| 2. PP -> IN NP | 15. SBAR -> WHNP S | 28. VP -> VBD VP |
| 3. NP-SBJ -> NONE | 16. NP -> NNP | 29. NP-SBJ -> NNP |
| 4. NP -> DT NN | 17. VP -> VB NP | 30. VP -> VBD SBAR |
| 5. NP-SBJ -> PRP | 18. NP -> PRP | 31. ADVP-TMP -> RB |
| 6. NP -> NP PP | 19. PP-TMP -> IN NP | 32. VP -> VBD NP |
| 7. VP -> TO VP | 20. SBAR -> NONE S | 33. DT -> 'a' |
| 8. NP -> NN | 21. PP-CLR -> IN NP | 34. NN -> 'telescope' |
| 9. NP -> NONE | 22. NP -> NNP NNP | 35. IN -> 'with' |
| 10. PP-LOC -> IN NP | 23. NP-SBJ -> DT NN | 36. DT -> 'the' |
| 11. ADVP -> RB | 24. NP -> JJ NNS | 37. NN -> 'man' |
| 12. NP -> DT JJ NN | 25. NP -> NP SBAR | 38. NN -> 'boy' |
| 13. NP -> NNS | 26. SBAR -> IN S | 39. VBD -> 'saw' |

Smallest grammar for a sentence

Parses: 2, # rules: 45, # lexical rules: 7

- | | | |
|---------------------|---------------------|-----------------------|
| 1. S -> NP-SBJ VP | 19. PP-TMP -> IN NP | 37. NP -> DT NNS |
| 2. PP -> IN NP | 20. SBAR -> NONE S | 38. VP -> VBZ VP |
| 3. NP-SBJ -> NONE | 21. PP-CLR -> IN NP | 39. VP -> VBG NP |
| 4. NP -> DT NN | 22. NP -> NNP NNP | 40. NP-SBJ -> NNP NNP |
| 5. NP-SBJ -> PRP | 23. NP-SBJ -> DT NN | 41. NP -> NP CC NP |
| 6. NP -> NP PP | 24. NP -> JJ NNS | 42. NP -> JJ NN |
| 7. VP -> TO VP | 25. NP -> NP SBAR | 43. NP -> PRPS NN |
| 8. NP -> NN | 26. SBAR -> IN S | 44. VP -> VP CC VP |
| 9. NP -> NONE | 27. NP-SBJ -> NP PP | 45. NP -> NP PP-LOC |
| 10. PP-LOC -> IN NP | 28. VP -> VBD VP | 46. NN -> 'man' |
| 11. ADVP -> RB | 29. NP-SBJ -> NNP | 47. IN -> 'with' |
| 12. NP -> DT JJ NN | 30. VP -> VBD SBAR | 48. DT -> 'a' |
| 13. NP -> NNS | 31. ADVP-TMP -> RB | 49. DT -> 'the' |
| 14. VP -> MD VP | 32. VP -> VBD NP | 50. NN -> 'boy' |
| 15. SBAR -> WHNP S | 33. PP -> TO NP | 51. VBD -> 'saw' |
| 16. NP -> NNP | 34. QP -> CD CD | 52. NN -> 'telescope' |
| 17. VP -> VB NP | 35. S -> NONE | |
| 18. NP -> PRP | 36. WHNP -> WDT | |



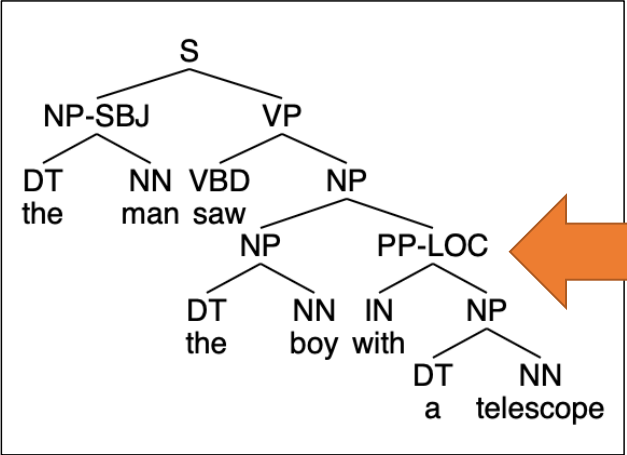
always look
at the last
rule added!

Smallest grammar for a sentence

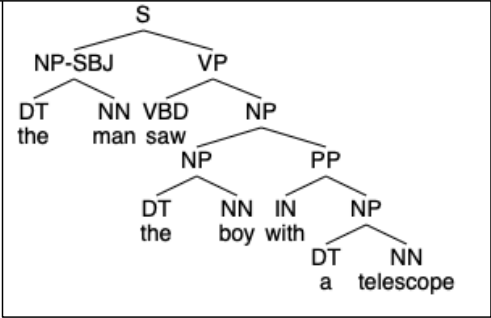
```
(S
  (NP-SBJ (DT the) (NN man))
  (VP
    (VBD saw)
    (NP
      (NP (DT the) (NN boy))
      (PP-LOC (IN with) (NP (DT a) (NN
telescope))))))
```

```
(S
  (NP-SBJ (DT the) (NN man))
  (VP
    (VBD saw)
    (NP
      (NP (DT the) (NN boy))
      (PP (IN with) (NP (DT a) (NN
telescope))))))
```

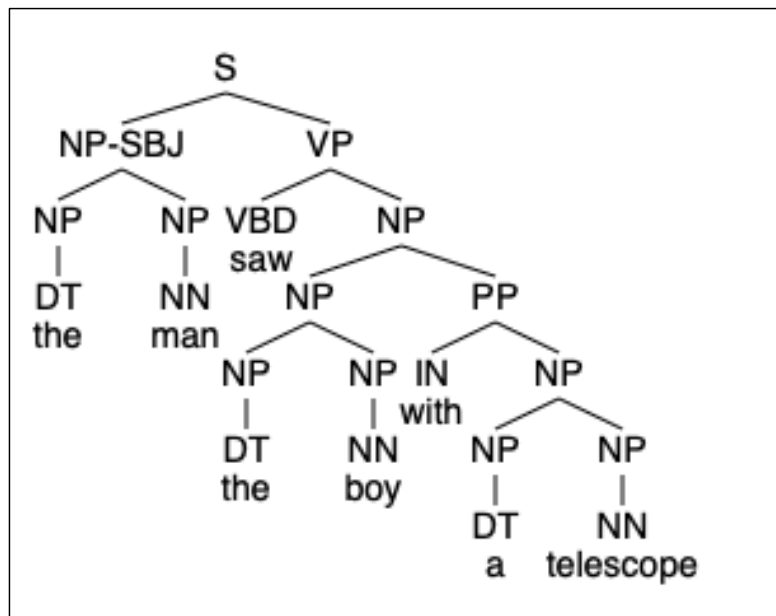
Parses: 2, # rules: 45, # lexical rules: 7



not what you'd expect for the 2nd parse!



Smallest grammar for a sentence



- After two parses, it explodes combinatorially:
 - many more than 3 parses
 - *can you spot why?*

Smallest grammar for a sentence

Parses: 88, # rules: 92, # lexical rules: 7

- | | | | | |
|--------------------|----------------------|-----------------------|-------------------------|----------------------|
| 1. S → NP-SBJ VP | 22. NP → NNP NNP | 43. NP → PRPS NN | 64. SBAR-ADV → IN S | 85. NP-SBJ → NP SBAR |
| 2. PP → IN NP | 23. NP-SBJ → DT NN | 44. VP → VP CC VP | 65. NP-SBJ → NP NP | 86. SBAR → WHADVP S |
| 3. NP-SBJ → NONE | 24. NP → JJ NNS | 45. NP → NP PP-LOC | 66. S-ADV → NP-SBJ VP | 87. NP → PRPS NNS |
| 4. NP → DT NN | 25. NP → NP SBAR | 46. NP → NP NP | 67. NP → CD NNS | 88. NP-SBJ → DT NNS |
| 5. NP-SBJ → PRP | 26. SBAR → IN S | 47. VP → VBD S | 68. VP → VBN NP PP | 89. S → PP NP-SBJ VP |
| 6. NP → NP PP | 27. NP-SBJ → NP PP | 48. NP → QP NONE | 69. PP → IN NP-LGS | 90. PP → IN S-NOM |
| 7. VP → TO VP | 28. VP → VBD VP | 49. NP → DT NN NN | 70. VP → VBZ NP | 91. PP-DIR → IN NP |
| 8. NP → NN | 29. NP-SBJ → NNP | 50. S-NOM → NP-SBJ VP | 71. ADVP-MNR → RB | 92. NP → DT |
| 9. NP → NONE | 30. VP → VBD SBAR | 51. PRT → RP | 72. PP-DIR → TO NP | 93. IN → 'with' |
| 10. PP-LOC → IN NP | 31. ADVP-TMP → RB | 52. VP → VBP VP | 73. WHNP → NONE | 94. DT → 'a' |
| 11. ADVP → RB | 32. VP → VBD NP | 53. WHNP → WP | 74. S-TPC → NP-SBJ VP | 95. VBD → 'saw' |
| 12. NP → DT JJ NN | 33. PP → TO NP | 54. NP → CD | 75. NP-PRD → NP PP | 96. DT → 'the' |
| 13. NP → NNS | 34. QP → CD CD | 55. NP → NP VP | 76. NP → CD NN | 97. NN → 'boy' |
| 14. VP → MD VP | 35. S → NONE | 56. ADJP-PRD → JJ | 77. NP → NP NN | 98. NN → 'man' |
| 15. SBAR → WHNP S | 36. WHNP → WDT | 57. VP → VB VP | 78. VP → VBZ NP-PRD | 99. NN → 'telescope' |
| 16. NP → NNP | 37. NP → DT NNS | 58. NP → NNP POS | 79. NP → NNP NNP NNP | |
| 17. VP → VB NP | 38. VP → VBZ VP | 59. S → S CC S | 80. VP → VBZ S | |
| 18. NP → PRP | 39. VP → VBG NP | 60. WHADVP → WRB | 81. VP → VB S | |
| 19. PP-TMP → IN NP | 40. NP-SBJ → NNP NNP | 61. VP → VBN NP | 82. ADVP-TMP → NONE | |
| 20. SBAR → NONE S | 41. NP → NP CC NP | 62. NP-SBJ → NNS | 83. S → S-TPC NP-SBJ VP | |
| 21. PP-CLR → IN NP | 42. NP → JJ NN | 63. NP → NN NNS | 84. NP → DT NN POS | |

always
look at
the last
rule
added!

Smallest grammar for a sentence

Parses: 106, # rules: 97, # lexical rules: 7

S → NP-SBJ VP	NP-SBJ → DT NN	NP → NP PP-LOC	NP → CD NNS	S → PP NP-SBJ VP
PP → IN NP	NP → JJ NNS	NP → NP NP	VP → VBN NP PP	PP → IN S-NOM
NP-SBJ → NONE	NP → NP SBAR	VP → VBD S	PP → IN NP-LGS	PP-DIR → IN NP
NP → DT NN	SBAR → IN S	NP → QP NONE	VP → VBZ NP	NP → DT
NP-SBJ → PRP	NP-SBJ → NP PP	NP → DT NN NN	ADVP-MNR → RB	PP-CLR → TO NP
NP → NP PP	VP → VBD VP	S-NOM → NP-SBJ VP	PP-DIR → TO NP	NP → NN NN
VP → TO VP	NP-SBJ → NNP	PRT → RP	WHNP → NONE	S → NP-SBJ ADVP VP
NP → NN	VP → VBD SBAR	VP → VBP VP	S-TPC → NP-SBJ VP	VP → VBP NP
NP → NONE	ADVP-TMP → RB	WHNP → WP	NP-PRD → NP PP	VP → VBD NP-PRD
PP-LOC → IN NP	VP → VBD NP	NP → CD	NP → CD NN	IN → 'with'
ADVP → RB	PP → TO NP	NP → NP VP	NP → NP NN	DT → 'a'
NP → DT JJ NN	QP → CD CD	ADJP-PRD → JJ	VP → VBZ NP-PRD	VBD → 'saw'
NP → NNS	S → NONE	VP → VB VP	NP → NNP NNP NNP	DT → 'the'
VP → MD VP	WHNP → WDT	NP → NNP POS	VP → VBZ S	NN → 'boy'
SBAR → WHNP S	NP → DT NNS	S → S CC S	VP → VB S	NN → 'man'
NP → NNP	VP → VBZ VP	WHADVP → WRB	ADVP-TMP → NONE	NN → 'telescope'
VP → VB NP	VP → VBG NP	VP → VBN NP	S → S-TPC NP-SBJ VP	
NP → PRP	NP-SBJ → NNP NNP	NP-SBJ → NNS	NP → DT NN POS	
PP-TMP → IN NP	NP → NP CC NP	NP → NN NNS	NP-SBJ → NP SBAR	
SBAR → NONE S	NP → JJ NN	SBAR-ADV → IN S	SBAR → WHADVP S	
PP-CLR → IN NP	NP → PRPS NN	NP-SBJ → NP NP	NP → PRPS NNS	
NP → NNP NNP	VP → VP CC VP	S-ADV → NP-SBJ VP	NP-SBJ → DT NNS	

Smallest grammar for a sentence

Parses: 124, # rules: 155, # lexical rules: 7

S -> NP-SBJ VP	NP -> NP SBAR	NP -> DT NN NN	WHNP -> NONE	VP -> VBD NP-PRD	S -> ADVP NP-SBJ VP	NP -> DT JJ JJ NN
PP -> IN NP	SBAR -> IN S	S-NOM -> NP-SBJ VP	S-TPC -> NP-SBJ VP	VP -> VBN S	NP-SBJ -> DT JJ NN	ADJP-PRD -> JJ PP
NP-SBJ -> NONE	NP-SBJ -> NP PP	PRT -> RP	NP-PRD -> NP PP	VP -> VBZ SBAR	SINV -> S-TPC VP NP-SBJ	NP -> NP PP PP
NP -> DT NN	VP -> VBD VP	VP -> VBP VP	NP -> CD NN	NP-SBJ -> NN	VP -> VBG S	VP -> VB ADJP-PRD
NP-SBJ -> PRP	NP-SBJ -> NNP	WHNP -> WP	NP -> NP NN	NP -> DT NNP	NP -> NP NP-ADV	SBAR-PRP -> IN S
NP -> NP PP	VP -> VBD SBAR	NP -> CD	VP -> VBZ NP-PRD	S-PRP -> NP-SBJ VP	S -> S S	VP -> VB PP-CLR
VP -> TO VP	ADVP-TMP -> RB	NP -> NP VP	NP -> NNP NNP NNP	SBAR-TMP -> IN S	VP -> VBN VP	VP -> VBD
NP -> NN	VP -> VBD NP	ADJP-PRD -> JJ	VP -> VBZ S	S -> CC NP-SBJ VP	NP -> PRPS JJ NN	VP -> VBP RB VP
NP -> NONE	PP -> TO NP	VP -> VB VP	VP -> VB S	NP-SBJ -> DT	NP -> DT JJ NN NN	NP-SBJ -> NN NNS
PP-LOC -> IN NP	QP -> CD CD	NP -> NNP POS	ADVP-TMP -> NONE	NP -> DT JJ NNS	NP-SBJ -> JJ NNS	VP -> VBP S
ADVP -> RB	S -> NONE	S -> S CC S	S -> S-TPC NP-SBJ VP	VP -> VBD ADJP-PRD	S -> NP-SBJ ADVP-TMP VP	VP -> VBD NP PP
NP -> DT JJ NN	WHNP -> WDT	WHADVP -> WRB	NP -> DT NN POS	NP -> CD NONE	NP -> DT NNP NNP	DT -> 'a'
NP -> NNS	NP -> DT NNS	VP -> VBN NP	NP-SBJ -> NP SBAR	VP -> VBN NP PP-CLR	NP -> DT NNP NN	NN -> 'boy'
VP -> MD VP	VP -> VBZ VP	NP-SBJ -> NNS	SBAR -> WHADVP S	NP-SBJ -> EX	NP -> QP NNS	VBD -> 'saw'
SBAR -> WHNP S	VP -> VBG NP	NP -> NN NNS	NP -> PRPS NNS	VP -> VBP SBAR	VP -> VB NP PP	DT -> 'the'
NP -> NNP	NP-SBJ -> NNP NNP	SBAR-ADV -> IN S	NP-SBJ -> DT NNS	S -> PP-TMP NP-SBJ VP	VP -> VB SBAR	IN -> 'with'
VP -> VB NP	NP -> NP CC NP	NP-SBJ -> NP NP	S -> PP NP-SBJ VP	SBAR-TMP -> WHADVP S	ADJP -> JJ	NN -> 'man'
NP -> PRP	NP -> JJ NN	S-ADV -> NP-SBJ VP	PP -> IN S-NOM	NP-ADV -> DT NN	VP -> VBD RB VP	NN -> 'telescope'
PP-TMP -> IN NP	NP -> PRPS NN	NP -> CD NNS	PP-DIR -> IN NP	VP -> VB	S -> SBAR-ADV NP-SBJ VP	
SBAR -> NONE S	VP -> VP CC VP	VP -> VBN NP PP	NP -> DT	NP -> DT ADJP NN	NP -> JJ NN NNS	
PP-CLR -> IN NP	NP -> NP PP-LOC	PP -> IN NP-LGS	PP-CLR -> TO NP	VP -> MD RB VP	S -> PP-LOC NP-SBJ VP	
NP -> NNP NNP	NP -> NP NP	VP -> VBZ NP	NP -> NN NN	ADJP -> RB JJ	NP-SBJ -> NP PP-LOC	
NP-SBJ -> DT NN	VP -> VBD S	ADVP-MNR -> RB	S -> NP-SBJ ADVP VP	VP -> VBZ ADJP-PRD	NP-SBJ -> PRPS NN	
NP -> JJ NNS	NP -> QP NONE	PP-DIR -> TO NP	VP -> VBP NP	NP -> NNP NNP POS	PP-PRD -> IN NP	

Smallest grammar for a sentence

Finally, we get the parse we wanted!

