LING/C SC 581:
Advanced Computational Linguistics

Lecture 29

Adminstrivia

- Optional Homeworks 11 and 12
  - due on Thursday
  - I'll grade them on Friday

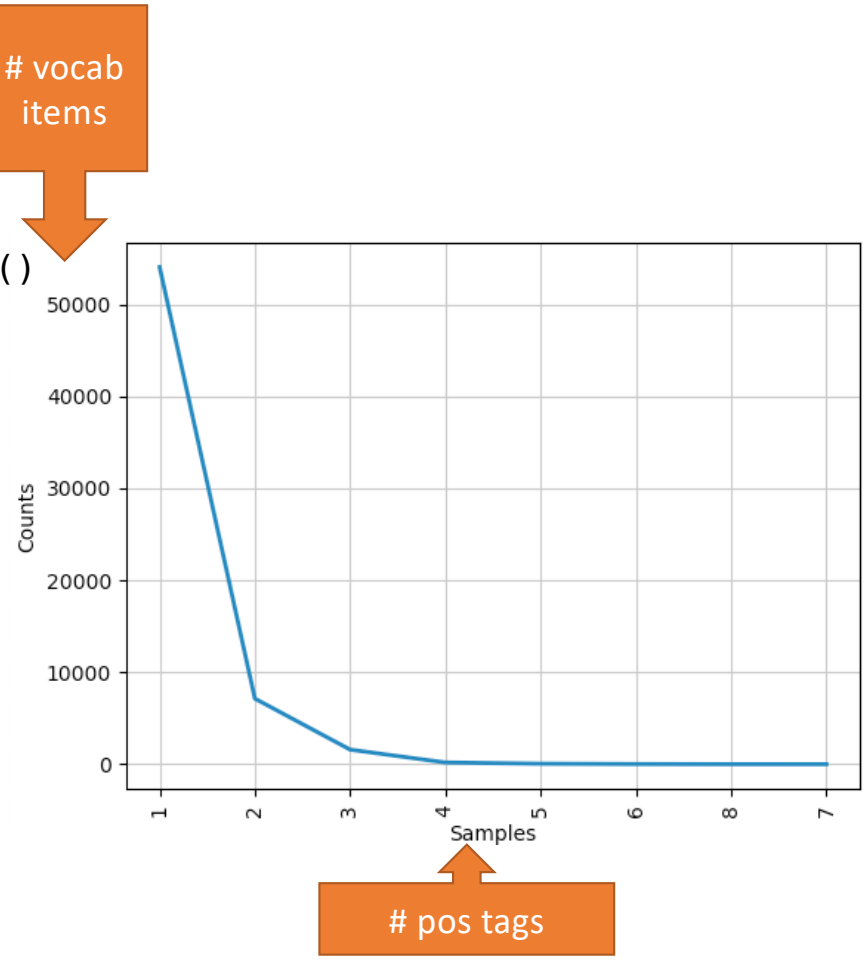
Today's Topic

- Another CFG experiment with the ptb:
  - *CFG = Context-free Grammar*
  - *how many ptb rules do we need to parse a sentence*?

Two lectures ago

- Frequency distribution by # of pos tags:

```
>>> wts = set([tuple for tree in ptb.parsed_sents()
for tuple in tree.pos()])
>>> d = {}
>>> for item in wts:
...     d.setdefault(item[0], []).append(item[1])
...
>>> fd = nltk.FreqDist([len(d[k]) for k in d])
>>> fd.most_common()
[(1, 54075), (2, 7137), (3, 1588), (4, 188), (5,
60), (6, 20), (8, 3), (7, 2)]
>>> fd.plot()
>>> [(k, d[k]) for k in d if len(d[k]) == n_tags]
```



Lots of POS tags for these words

- 8 POS tags:
  - `[('a', ['IN', 'JJ', 'SYM', 'LS', 'NNP', 'DT', ',', 'FW']),`
    `('in', ['NN', 'RBR', 'RB', 'IN|RP', 'NNP', 'RP', 'FW', 'IN']),`
    `('that', ['WP', 'RB', 'UH', 'VBP', 'DT', 'WDT', 'NN', 'IN'])]`
- 7 POS tags:

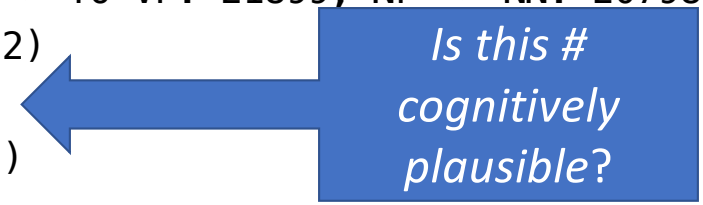
```
[('down', ['IN', 'RB', 'RP', 'RBR', 'JJ', 'NN', 'VBP']),
 ("'s", ['PRP', 'NNP', 'VBZ', 'VBP', 'IN', 'NNS', 'POS'])
]
```

- 6 POS tags (*a selection*):
  - `('less', ['RB', 'RBR', 'CC', 'NN', 'JJR', 'JJS'])`
  - `('Japanese', ['JJ', 'NNP', 'NN', 'VBP', 'NNPS', 'NNS'])`
  - `('put', ['VB', 'VBP', 'JJ', 'VBN', 'NN', 'VBD'])`

of Productions

- Let ps be all rules in the ptb. Then:

```
>>> ps2 = [p for p in ps if not(len(p.rhs()) == 1 and type(p.rhs()[0]) == str)]
>>> len(ps2)
1390347
>>> fd2 = nltk.FreqDist(ps2)
>>> fd2
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(fd2)
52134
>>> fd2.N()
1390347
```



of Productions

- >>> fd2.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)]
```

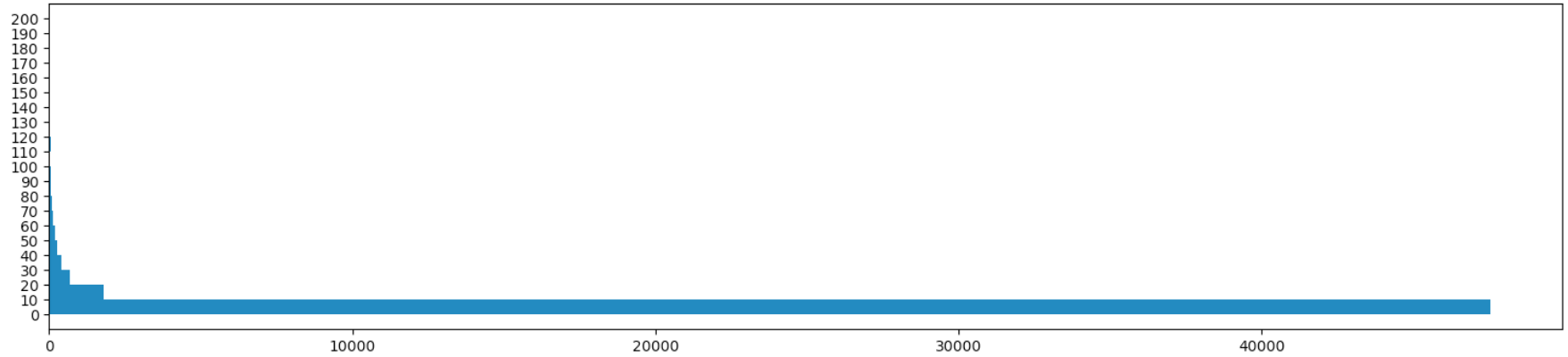
of Productions

- Most syntax rules only occur once:

```
>>> len(fd2)
52134
>>> fd2.N()
1390347
>>> len(fd2.hapaxes())
33631
```

of Productions: histogram

binned: # of times rule occurs



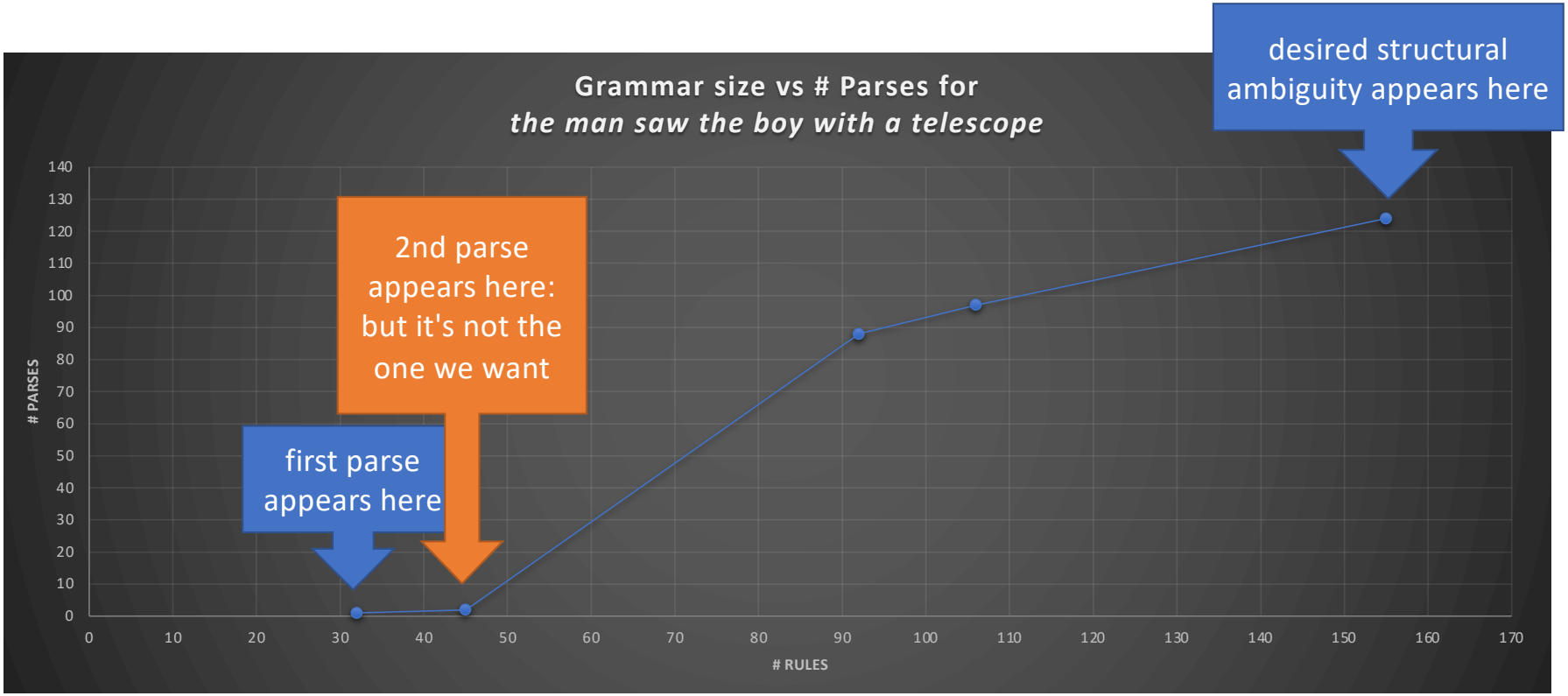
# of rules

Smallest grammar for a sentence

- **Experiment:**
  - suppose we start with an empty context-free grammar
  - 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?

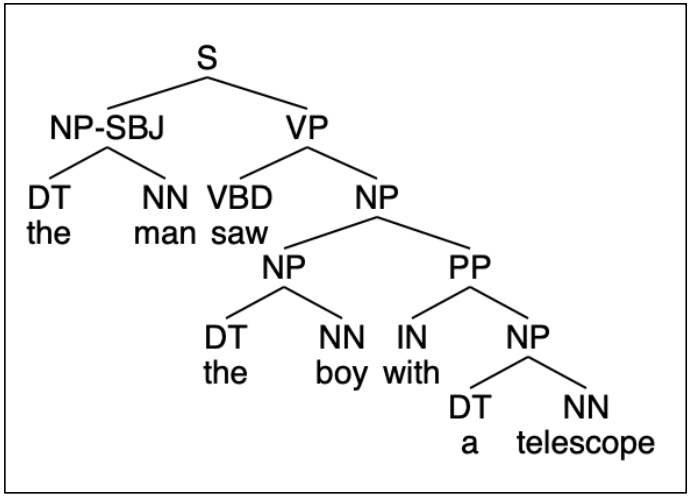
*some simplifications applied, explained later.

Smallest grammar for a sentence



Smallest grammar for a sentence

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

Smallest grammar for a sentence

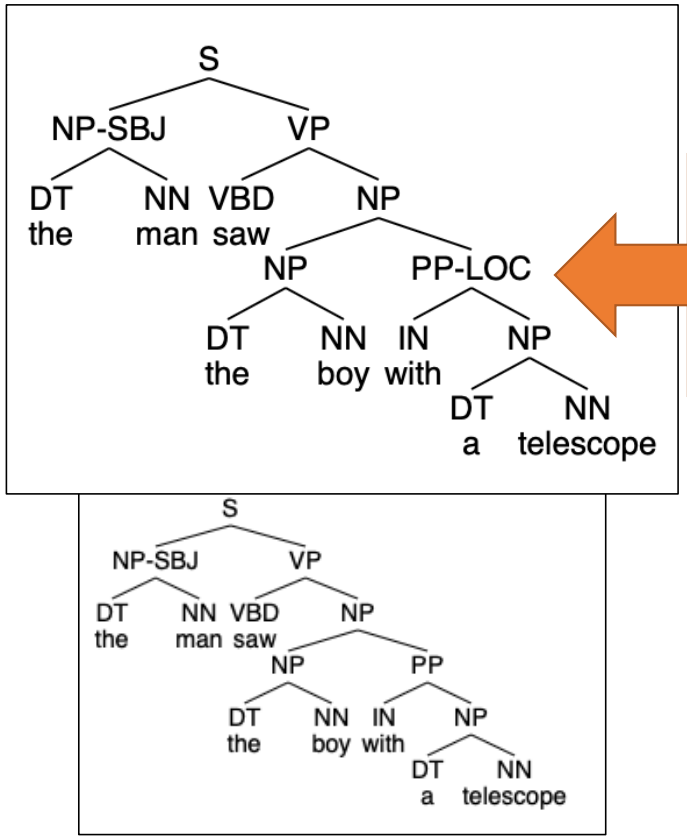
Parses: 2, # rules: 45, # lexical rules: 7

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

always look at the last rule added!

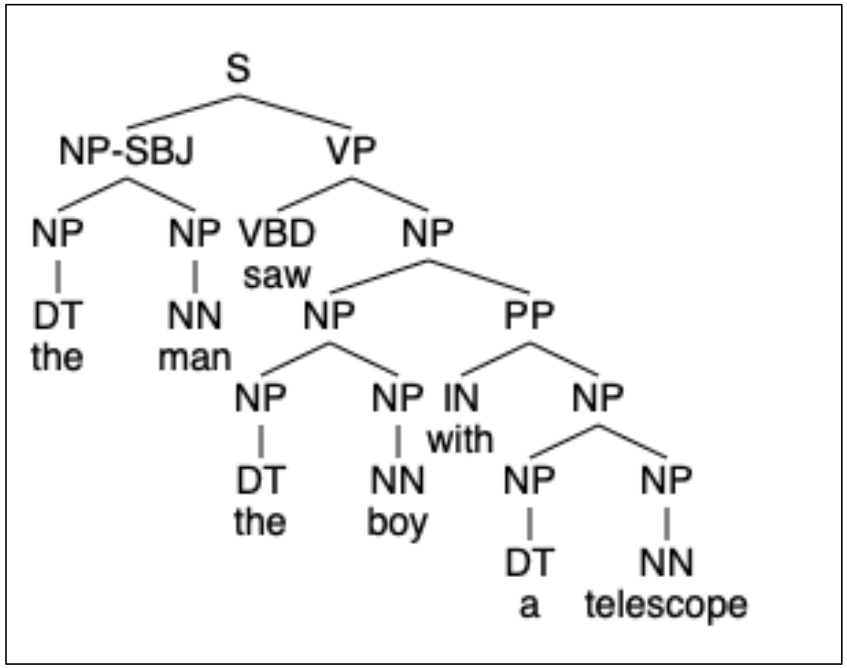
Smallest grammar for a sentence

```
(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

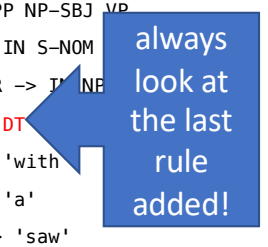


- After two parses, it explodes combinatorially:
  - many more than 3 parses
  - *can you spot why*?

Smallest grammar for a sentence

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

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



93. IN -> 'with'
94. DT -> 'a'
95. VBD -> 'saw'
96. DT -> 'the'
97. NN -> 'boy'
98. NN -> 'man'
99. NN -> 'telescope'

Smallest grammar for a sentence

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

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

Smallest grammar for a sentence

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

S -> NP-SBJ VP
PP -> IN NP
NP-SBJ -> NONE
NP -> DT NN
NP-SBJ -> PRP
NP -> NP PP
VP -> TO VP
NP -> NN
NP -> NONE
PP-LOC -> IN NP
ADVP -> RB
NP -> DT JJ NN
NP -> NNS
VP -> MD VP
SBAR -> WHNP S
NP -> NNP
VP -> VB NP
NP -> PRP
PP-TMP -> IN NP
SBAR -> NONE S
PP-CLR -> IN NP
NP -> NNP NNP
NP-SBJ -> DT NN
NP -> JJ NNS

NP -> NP SBAR
SBAR -> IN S
NP-SBJ -> NP PP
VP -> VBD VP
NP-SBJ -> NNP
VP -> VBD SBAR
ADVP-TMP -> RB
VP -> VBD NP
PP -> TO NP
QP -> CD CD
S -> NONE
WHNP -> WDT
NP -> DT NNS
VP -> VBZ VP
VP -> VBG NP
NP-SBJ -> NNP NNP
NP -> NP CC NP
NP -> JJ NN
NP -> PRPS NN
VP -> VP CC VP
NP -> NP PP-LOC
NP -> NP NP
VP -> VBD S
NP -> QP NONE

NP -> DT NN NN
S-NOM -> NP-SBJ VP
PRT -> RP
VP -> VBP VP
WHNP -> WP
NP -> CD
NP -> NP VP
ADJP-PRD -> JJ
VP -> VB VP
NP -> NNP POS
S -> S CC S
WHADVP -> WRB
VP -> VBN NP
NP-SBJ -> NNS
NP -> NN NNS
SBAR-ADV -> IN S
NP-SBJ -> NP NP
S-ADV -> NP-SBJ VP
NP -> CD NNS
VP -> VBN NP PP
PP -> IN NP-LGS
VP -> VBZ NP
ADVP-MNR -> RB
PP-DIR -> TO NP

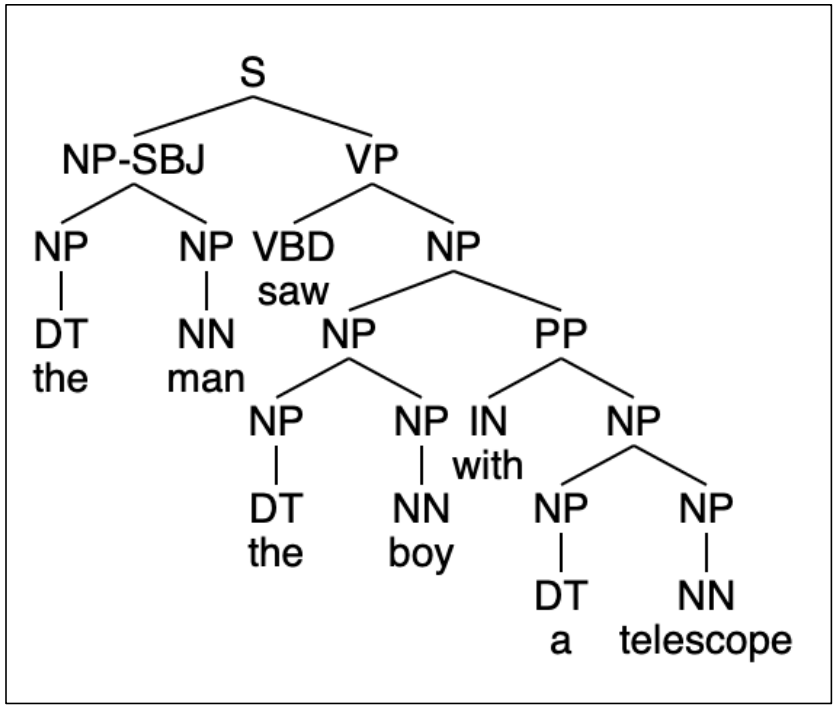
WHNP -> NONE
S-TPC -> NP-SBJ VP
NP-PRD -> NP PP
NP -> CD NN
NP -> NP NN
VP -> VBZ NP-PRD
NP -> NNP NNP NNP
VP -> VBZ S
VP -> VB S
ADVP-TMP -> NONE
S -> S-TPC NP-SBJ VP
NP -> DT NN POS
NP-SBJ -> NP SBAR
SBAR -> WHADVP S
NP -> PRPS NNS
NP-SBJ -> DT NNS
S -> PP NP-SBJ VP
PP -> IN S-NOM
PP-DIR -> IN NP
NP -> DT
PP-CLR -> TO NP
NP -> NN NN
S -> NP-SBJ ADVP VP
VP -> VBP NP

VP -> VBD NP-PRD
VP -> VBN S
VP -> VBZ SBAR
NP-SBJ -> NN
NP -> DT NNP
S-PRP -> NP-SBJ VP
SBAR-TMP -> IN S
S -> CC NP-SBJ VP
NP-SBJ -> DT
NP -> DT JJ NNS
VP -> VBD ADJP-PRD
NP -> CD NONE
VP -> VBN NP PP-CLR
NP-SBJ -> EX
VP -> VBP SBAR
S -> PP-TMP NP-SBJ VP
SBAR-TMP -> WHADVP S
NP-ADV -> DT NN
VP -> VB
NP -> DT ADJP NN
VP -> MD RB VP
ADJP -> RB JJ
VP -> VBZ ADJP-PRD
NP -> NNP NNP POS

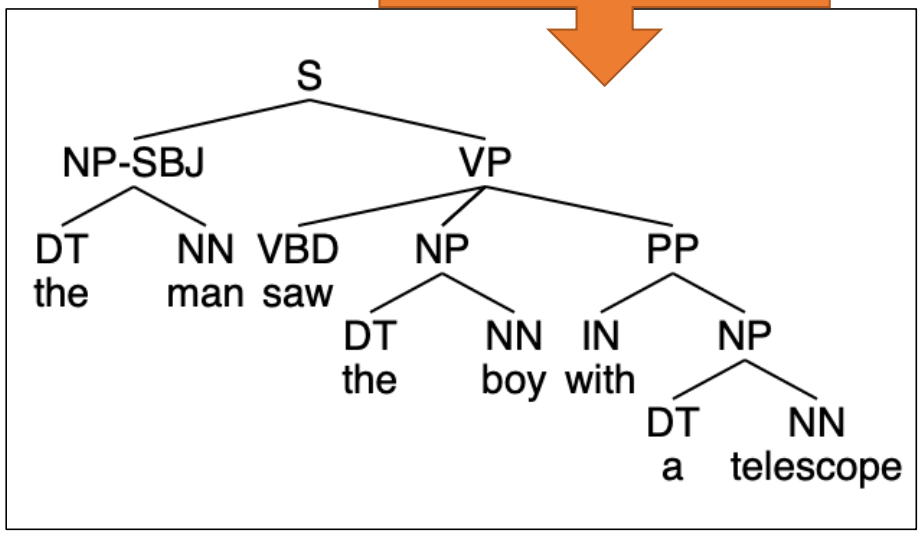
S -> ADVP NP-SBJ VP
NP-SBJ -> DT JJ NN
SINV -> S-TPC VP NP-SBJ
VP -> VBG S
NP -> NP NP-ADV
S -> S S
VP -> VBN VP
NP -> PRPS JJ NN
NP -> DT JJ NN NN
NP-SBJ -> JJ NNS
S -> NP-SBJ ADVP-TMP VP
NP -> DT NNP NNP
NP -> DT NNP NN
NP -> QP NNS
VP -> VB NP PP
VP -> VB SBAR
ADJP -> JJ
VP -> VBD RB VP
S -> SBAR-ADV NP-SBJ VP
NP -> JJ NN NNS
S -> PP-LOC NP-SBJ VP
NP-SBJ -> NP PP-LOC
NP-SBJ -> PRPS NN
PP-PRD -> IN NP

NP -> DT JJ JJ NN
ADJP-PRD -> JJ PP
NP -> NP PP PP
VP -> VB ADJP-PRD
SBAR-PRP -> IN S
VP -> VB PP-CLR
VP -> VBD
VP -> VBP RB VP
NP-SBJ -> NN NNS
VP -> VBP S
VP -> VBD NP PP
DT -> 'a'
NN -> 'boy'
VBD -> 'saw'
DT -> 'the'
IN -> 'with'
NN -> 'man'
NN -> 'telescope'

Smallest grammar for a sentence



Finally, we get the parse we wanted!



Smallest grammar for a sentence

plus a whole bunch of similar ones …

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

Some Rule Filtering

```
>>> ps2strs[:10]
1.  ['SBAR-PRP -> IN , S',
2.  'WHNP-1 -> WP$ NNP NNP NN',
3.  'SINV -> PP-LOC-TPC-1 VBD VP NP-SBJ',
4.  'NX -> JJ NN',
5.  'S -> NP-SBJ , PP-TMP VP',
6.  'VP -> S-ADV , VBD ADJP-PRD',
7.  'S -> S CC S ADVP',
8.  'VP -> ADVP VBD SBAR-NOM',
9.  'S-TPC-1 -> ADVP-TMP , S , CC S',
10. 'S-IMP -> INTJ , NP-SBJ VP .']
ValueError: Unable to parse line 2: WHNP -> WP$ NNP NNP NN
Expected a nonterminal, found: $ NNP NNP NN
ValueError: Unable to parse line 8: NP-SBJ -> -NONE-
Expected a nonterminal, found: -NONE-
```

- Eliminate:
  - numeric suffixes (indexing),
    - e.g. -1
  - punctuation (terminals)*
    - e.g. , .
- Replace:
  - $ in POS tag names, signifies possessive form, by S*
    - e.g. WP$ *whose* PRP$ *his*
  - -NONE- by NONE*
    - also -LRB- -RRB-*

*due to nltk CFG limitations