

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

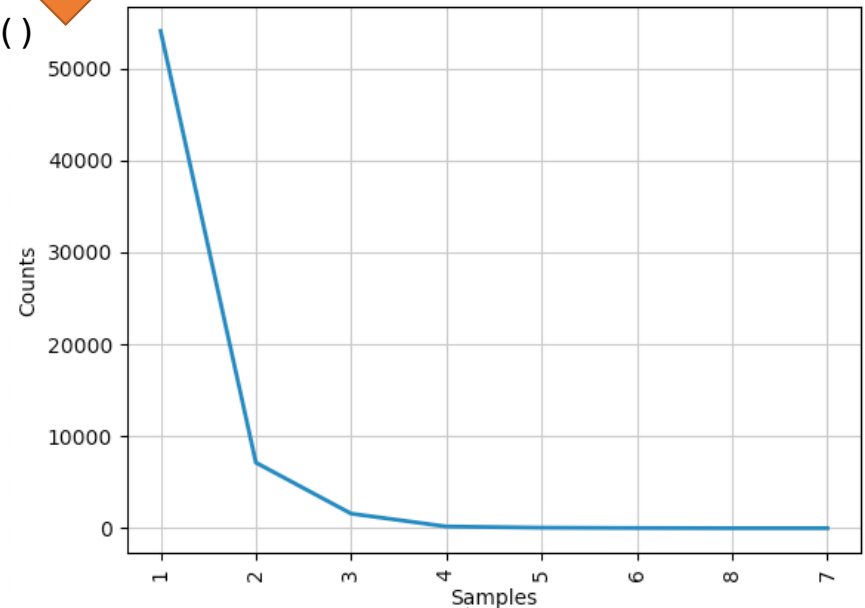
- Another CFG experiment with the ptb:
 - *CFG = Context-free Grammar*
 - *how many ptb rules do we need to parse a sentence?*

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]
```

vocab
items



pos 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
```

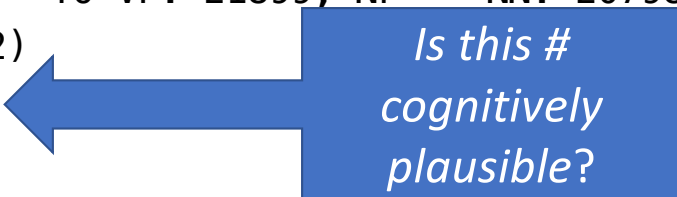
```
FreqDist({PP -> IN NP: 78040, S -> NP-SBJ VP: 63335, NP -> DT NN: 40876, NP-SBJ  
-> -NONE-: 39712, NP -> NP PP: 35819, NP-SBJ -> PRP: 31272, S -> NP-SBJ VP .:  
24467, VP -> TO VP: 21899, NP -> NN: 20798, NP -> -NONE-: 20312, ...})
```

```
>>> len(fd2)
```

```
52134
```

```
>>> fd2.N()
```

```
1390347
```



*Is this #
cognitively
plausible?*

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()
```

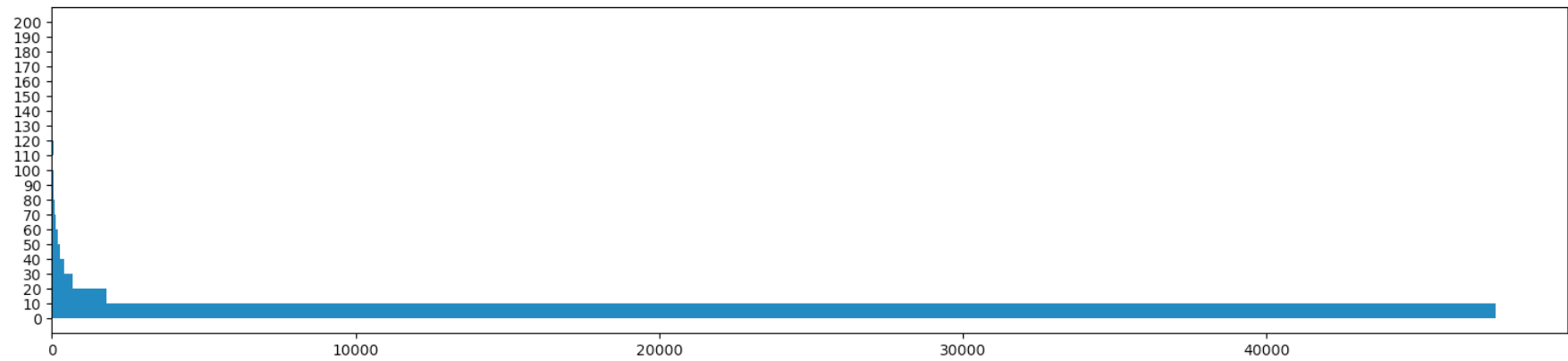
```
1390347
```

```
>>> len(fd2.hapaxes())
```

```
33631
```


of Productions: histogram

binned: # of times
rule occurs



of rules

Smallest grammar for a sentence

- **Experiment:**

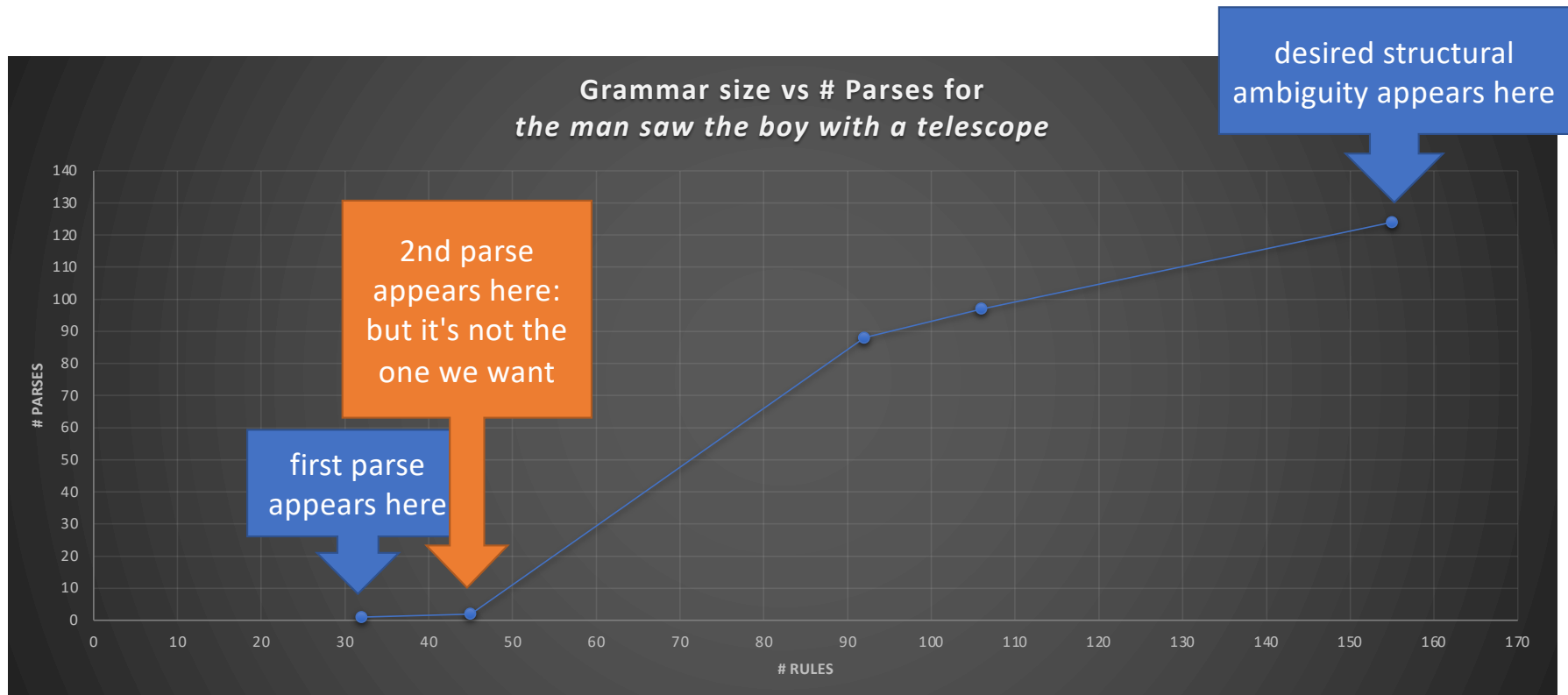
- suppose we start with an empty context-free grammar
- take PTB rules in order of frequency (*highest first*)
- add them one at a time to the grammar
- how many PTB rules before we can parse this sentence?*

 - *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.

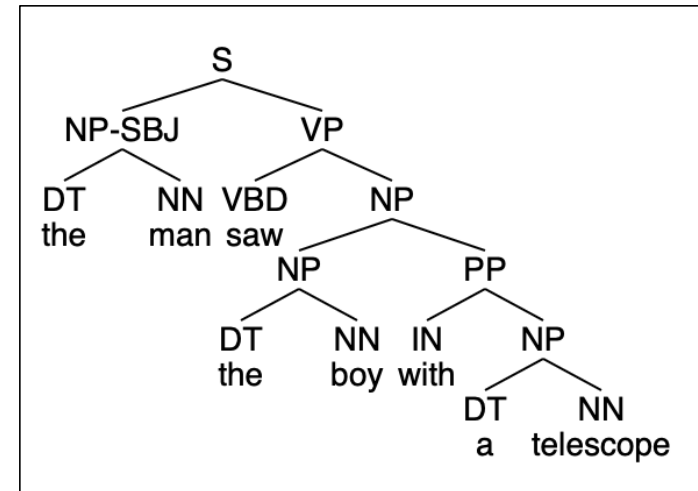
Smallest grammar for a sentence



Smallest grammar for a sentence

```
>>> find_smallestcfg(s, True)      (NP
the DT 73202                        (NP (DT the) (NN boy))
man NN 886                          (PP (IN with) (NP (DT a)
saw VBD 329                          (NN telescope))))))
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 | 15. SBAR → WHNP S | 29. NP-SBJ → NNP |
| 2. PP → IN NP | 16. NP → NNP | 30. VP → VBD SBAR |
| 3. NP-SBJ → NONE | 17. VP → VB NP | 31. ADVP-TMP → RB |
| 4. NP → DT NN | 18. NP → PRP | 32. VP → VBD NP |
| 5. NP-SBJ → PRP | 19. PP-TMP → IN NP | 33. DT → 'a' |
| 6. NP → NP PP | 20. SBAR → NONE S | 34. NN → 'telescope' |
| 7. VP → TO VP | 21. PP-CLR → IN NP | 35. IN → 'with' |
| 8. NP → NN | 22. NP → NNP NNP | 36. DT → 'the' |
| 9. NP → NONE | 23. NP-SBJ → DT NN | 37. NN → 'man' |
| 10. PP-LOC → IN NP | 24. NP → JJ NNS | 38. NN → 'boy' |
| 11. ADVP → RB | 25. NP → NP SBAR | 39. VBD → 'saw' |
| 12. NP → DT JJ NN | 26. SBAR → IN S | |
| 13. NP → NNS | 27. NP-SBJ → NP PP | |
| 14. VP → MD VP | 28. VP → VBD VP | |

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'

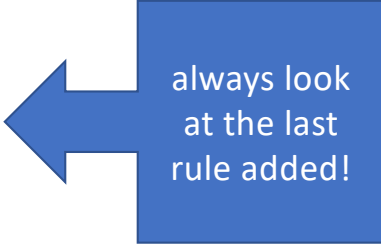
48. DT -> 'a'

49. DT -> 'the'

50. NN -> 'boy'

51. VBD -> 'saw'

52. NN -> 'telescope'



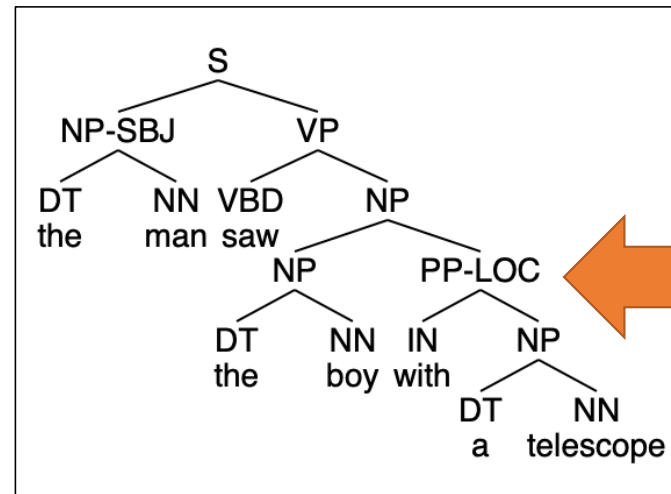
always look
at the last
rule added!

Smallest grammar for a sentence

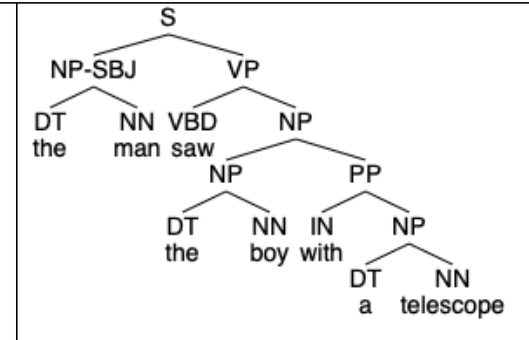
```
(S
  (NP-SBJ (DT the) (NN man))
  (VP
    (VBD saw)
    (NP
      (NP (DT the) (NN boy))
      (PP-LOC (IN with) (NP (DT a) (NN
telescope))))))
```

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

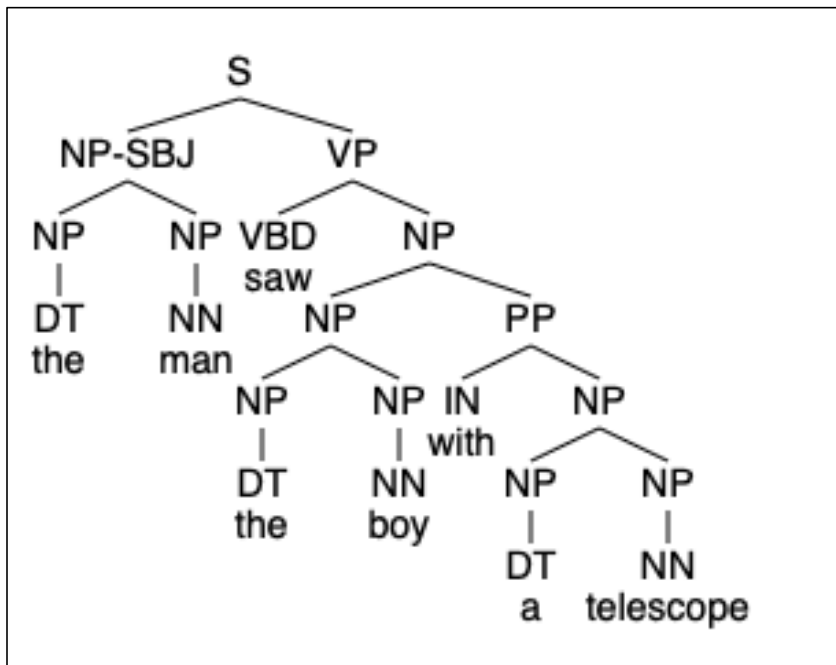
Parses: 2, # rules: 45, # lexical rules: 7



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 | 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 NONE |
| 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	

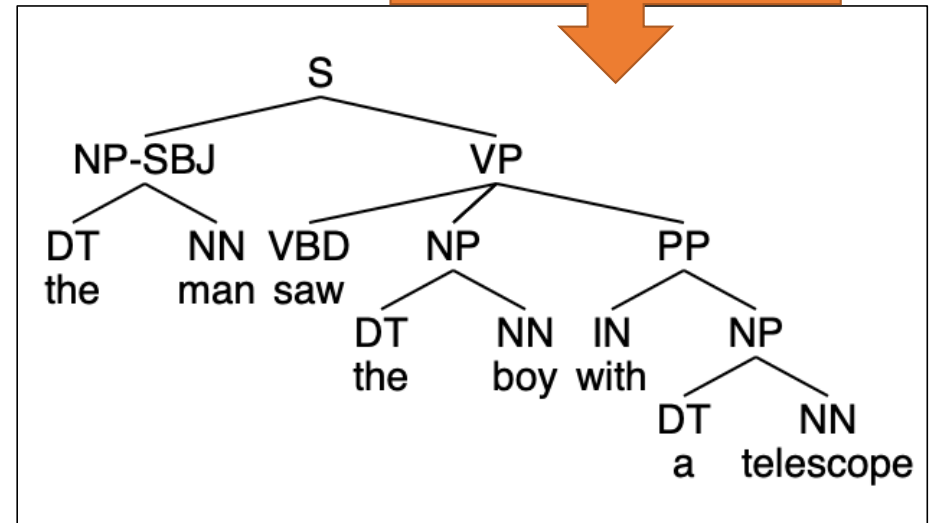
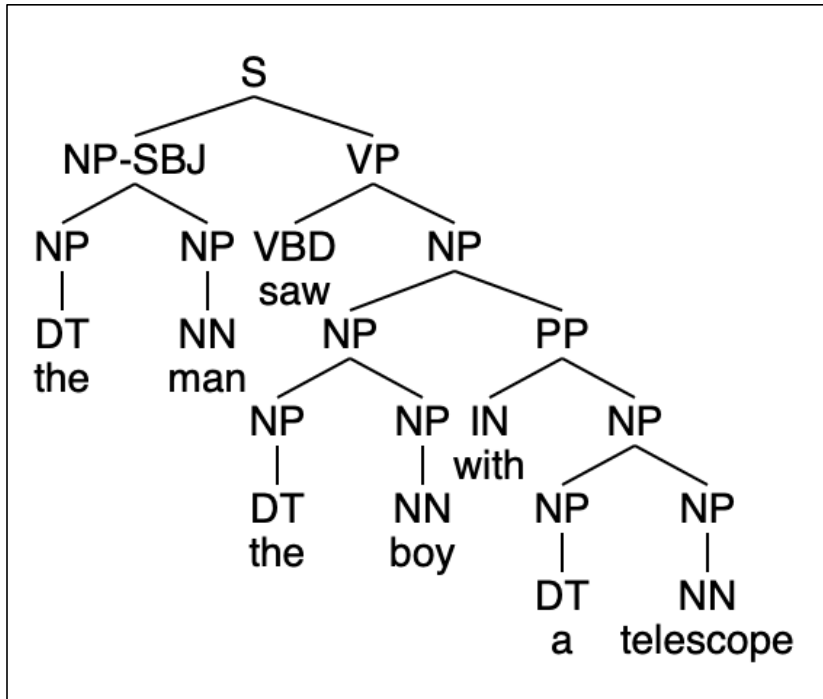
Smallest grammar for a sentence

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!



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)) (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)) (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)) (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)) (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