HW #4

Probabilistic Parsing

- Goals:
  - Learn about PCFGs
  - Implement PCKY
  - Analyze Parsing Evaluation
  - Assess improvements to PCFG Parsing

Tasks

1. Train a PCFG
    1. Estimate rule probabilities from treebank
    2. Treebank is already in CNF
    3. More ATIS data from Penn Treebank
2. Build PCKY Parser
    1. Modify (your) existing CKY implementation

Tasks

3. Evaluation
    1. Evaluate your parser using standard metric
    2. We will provide **evalb** program and gold standard
4. Improvement
    1. Improve your parser in some way:
        1. Coverage
        2. Accuracy
        3. Speed
    2. Evaluate new parser

Improvement Possibilities

- Coverage:
  - Some test sentences won't parse as is!
    - Lexical gaps (aka out-of-vocabulary [OOV] tokens)
      - …remember to model the probabilities, too
- Better context modeling
  - e.g. — Parent Annotation
- Better Efficiency
  - e.g. — Heuristic Filtering, Beam Search
- No "cheating" improvements:
  - improvement can't change training by looking at test data

`evalb`

- `evalb` available on dropbox in `hw4/tools`
- `evalb […] <gold-file> <test-file>`
- `evalb --help` for more info
- NB: specify **full/absolute path** to evalb when invoking in your scripts

HW #4 Notes

HW4 Notes

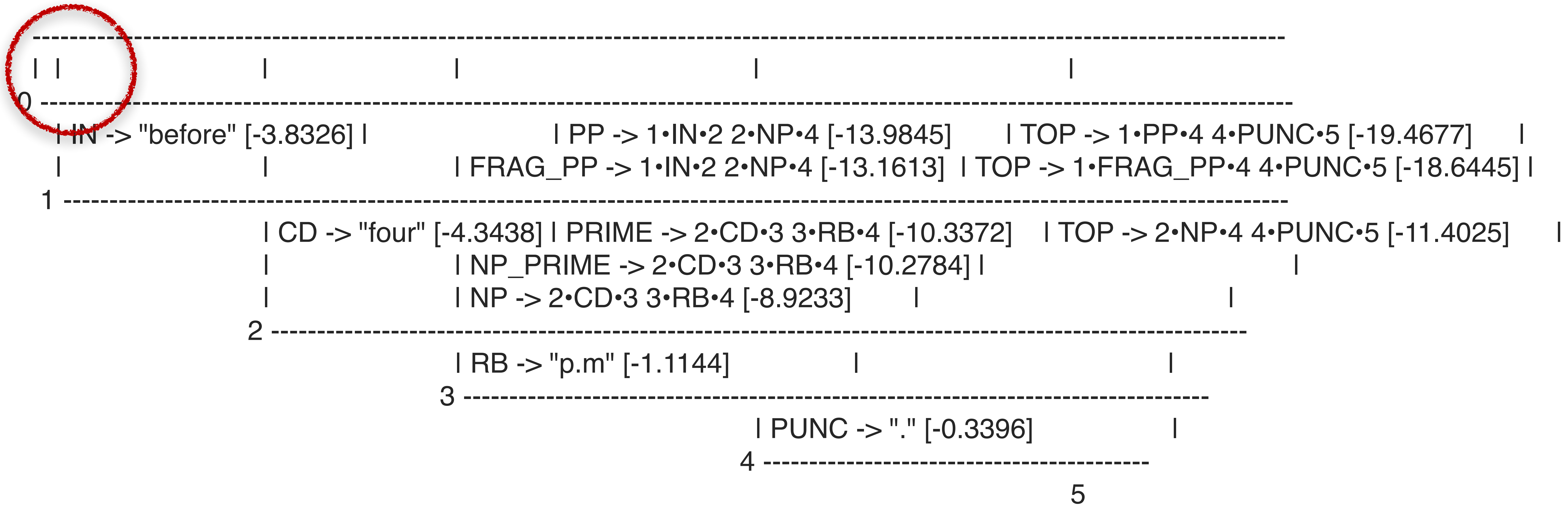
- If your improvement is along a dimension not measured by evalb (e.g. runtime):
  - Still run evalb on both old and improved code and report both results
    - NB: improved runtime cannot occur at “drastic” reduction in accuracy
  - Write code to measure your performance, and report before/after results in the readme

HW #4: OOV Handling

- As we discussed previously, you will find OOV tokens
- Sometimes this as as simple as case-sensitivity:

OOV: Case Sensitivity

Sentence #23: "*Arriving before four p.m .*"

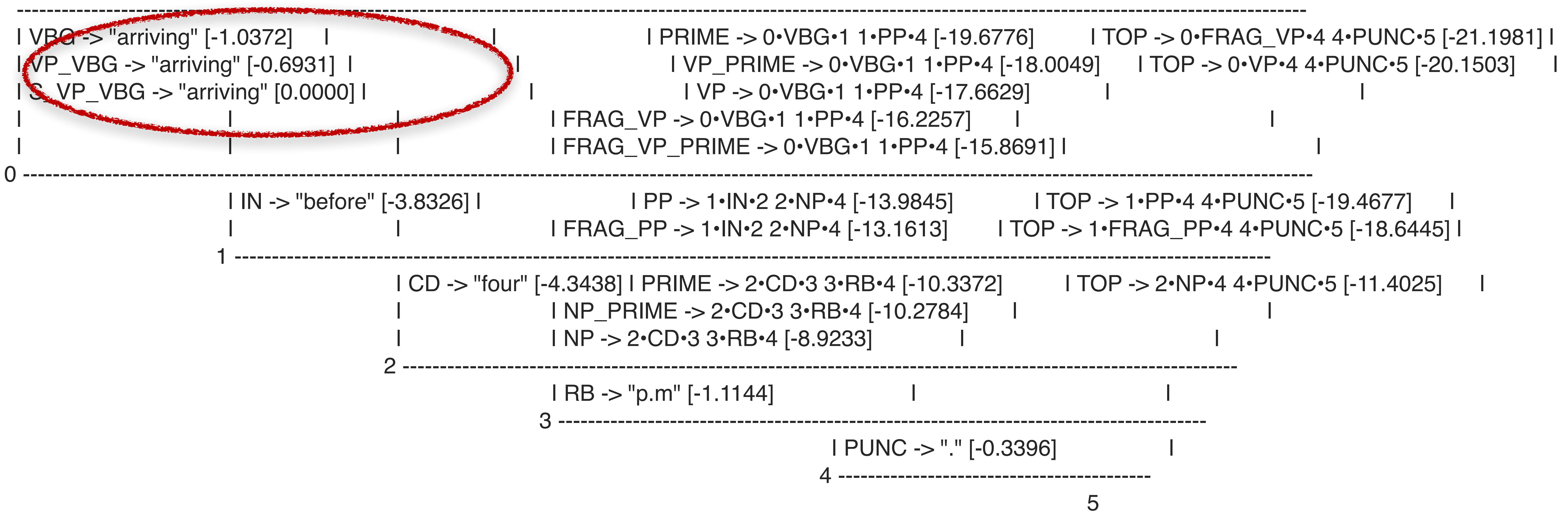


```
 -----------------------------------------------------------------------------------------------------------
 | |                 |                |                          |                          |
0 ----------------------------------------------------------------------------------------------------------
   | IN -> "before" [-3.8326] |                | PP -> 1•IN•2 2•NP•4 [-13.9845]      | TOP -> 1•PP•4 4•PUNC•5 [-19.4677]      |
   |                 |                | FRAG_PP -> 1•IN•2 2•NP•4 [-13.1613]  | TOP -> 1•FRAG_PP•4 4•PUNC•5 [-18.6445] |
  1 --------------------------------------------------------------------------------------------------------
                     | CD -> "four" [-4.3438] | PRIME -> 2•CD•3 3•RB•4 [-10.3372]    | TOP -> 2•NP•4 4•PUNC•5 [-11.4025]      |
                     |                | NP_PRIME -> 2•CD•3 3•RB•4 [-10.2784] |                                  |
                     |                | NP -> 2•CD•3 3•RB•4 [-8.9233]        |                          |
                    2 --------------------------------------------------------------------------------------
                                      | RB -> "p.m" [-1.1144]            |                          |
                                     3 ---------------------------------------------------------------
                                                                 | PUNC -> "." [-0.3396]              |
                                                                4 -----------------------------------
                                                                                            5
```

"**a**rriving" is in our grammar, but not "**A**rriving"

OOV: Case Sensitivity

Sentence #23:“*Arriving before four p.m .*”



```
  ----------------------------------------------------------------------------------------------------------------------------------------------
  | VBG -> "arriving" [-1.0372]     |                |                 | PRIME -> 0•VBG•1 1•PP•4 [-19.6776]        | TOP -> 0•FRAG_VP•4 4•PUNC•5 [-21.1981] |
  | VP_VBG -> "arriving" [-0.6931]  |                 |                  | VP_PRIME -> 0•VBG•1 1•PP•4 [-18.0049]      | TOP -> 0•VP•4 4•PUNC•5 [-20.1503]      |
  | S_VP_VBG -> "arriving" [0.0000] |                 |                  | VP -> 0•VBG•1 1•PP•4 [-17.6629]           |                            |
  |                      |                  |          | FRAG_VP -> 0•VBG•1 1•PP•4 [-16.2257]       |                            |
  |                      |                  |          | FRAG_VP_PRIME -> 0•VBG•1 1•PP•4 [-15.8691] |                            |
0 ----------------------------------------------------------------------------------------------------------------------------------------------
                         | IN -> "before" [-3.8326] |                 | PP -> 1•IN•2 2•NP•4 [-13.9845]           | TOP -> 1•PP•4 4•PUNC•5 [-19.4677]      |
                         |                  |          | FRAG_PP -> 1•IN•2 2•NP•4 [-13.1613]       | TOP -> 1•FRAG_PP•4 4•PUNC•5 [-18.6445] |
                       1 -----------------------------------------------------------------------------------------------------------------
                                            | CD -> "four" [-4.3438] | PRIME -> 2•CD•3 3•RB•4 [-10.3372]        | TOP -> 2•NP•4 4•PUNC•5 [-11.4025]      |
                                            |          | NP_PRIME -> 2•CD•3 3•RB•4 [-10.2784]      |                            |
                                            |          | NP -> 2•CD•3 3•RB•4 [-8.9233]           |                            |
                                          2 --------------------------------------------------------------------------------------------
                                                       | RB -> "p.m" [-1.1144]                |                            |
                                                     3 -----------------------------------------------------------------------
                                                                                | PUNC -> "." [-0.3396]                |
                                                                              4 -----------------------------------
                                                                                                              5
```

HW #4: OOV Handling

- Propose some number of N most likely tags at runtime…

OOV: Propose POS Tags

“Show me Ground transportation in Denver during weekdays .” — No “during”!

	FRAG_NP_PRIME → 2FRAG_NP_PRIME 4 PP 6[-21.810] FRAG_NP → 2FRAG_NP_PRIME 4 PP 6[-20.858]			
	NP_PRIME → 3 NN 4 PP 6[-16.296] PRIME → 3 NN 4 PP 6[-15.949]			
IN → "in" [-2.4018]	PP → 4 IN 5 NP_NNP 6[-7.505] FRAG_PP → 4 IN 5NP_NNP 6 [-6.828]			
5	NNP → "Denver" [-4.4002] NP_NNP → "Denver" [-3.3280]			
	6	🟥		
		7	NNS → "weekdays" [-5.5759] NP_NNS → "weekdays" [-3.7257]	TOP → 7NP_NNS 8PUNC 9[-11.001]
			8	PUNC → "." [-0.3396]
				9

OOV: Propose POS Tags

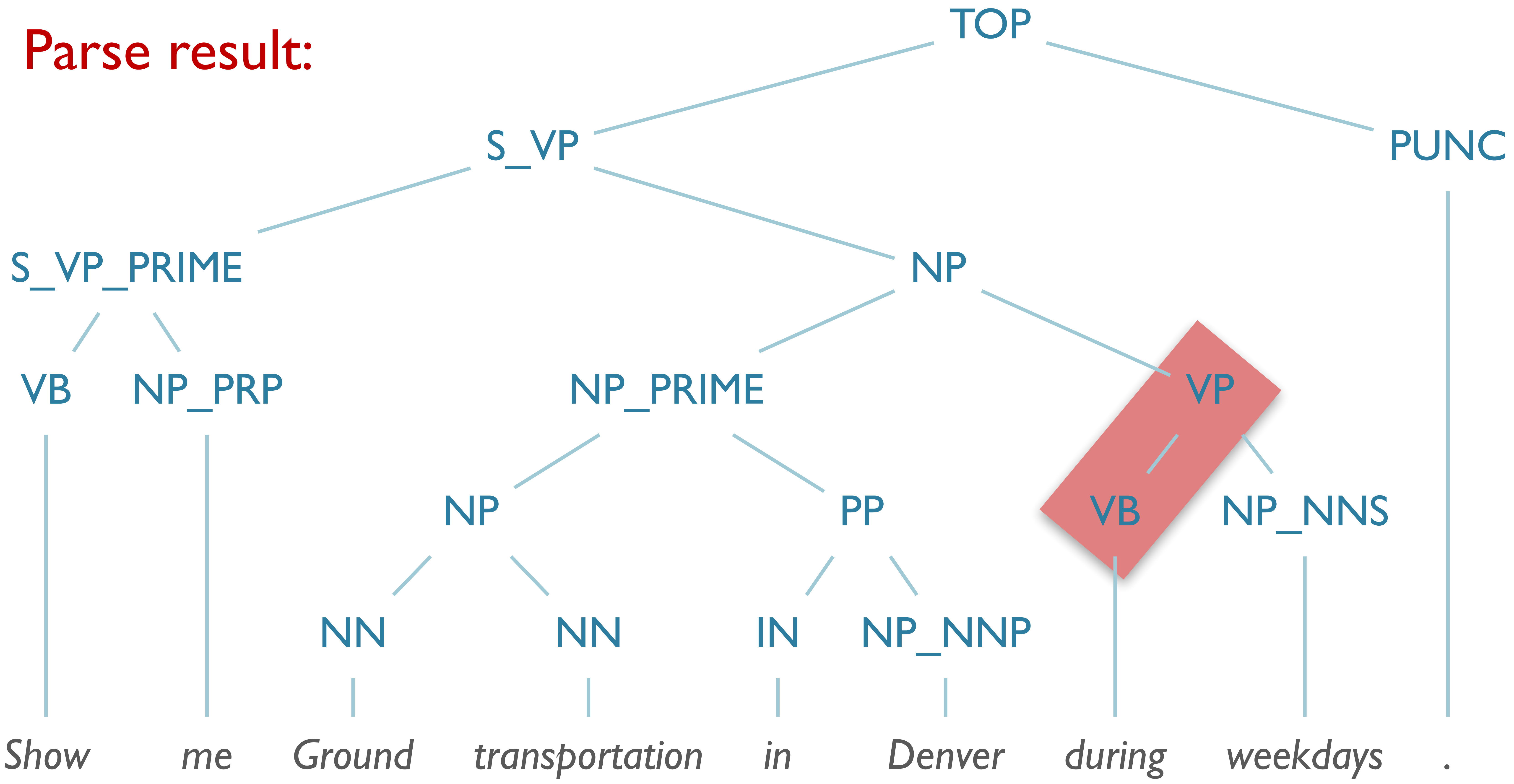
"Show me Ground transportation in Denver during weekdays ." — No "during"!

FRAG_NP_PRIME → ... FRAG_NP → ...	FRAG_NP_PRIME → ... FRAG_NP → ...	FRAG_NP →... FRAG_NP → ...	TOP → 2FRAG_NP 8 PUNC 9[-34.939] TOP → 2FRAG_NP 8 PUNC 9[-34.006]
NP_PRIME → ... PRIME → ...	PRIME → 3 NN 4PP 7 [-17.145] QP → 3 PRIME 6CD 7 [-15.930]	NP → 3 PRIME 7NNS 8 [-26.542] NP → 3 QP 7 NNS 8 [-26.398]	TOP → 3NP 8PUNC 9[-29.022] TOP → 3NP 8PUNC 9[-28.877]
PP → ... FRAG_PP → ...	PP → 4 IN 5 NP 7[-8.701] FRAG_PP → 4 IN 5NP 7 [-7.878]	PP → 4 IN 5 NP 8[-19.056] FRAG_PP → 4 IN 5NP 8 [-18.233]	TOP → 4PP 8PUNC 9[-24.540] TOP → 4FRAG_PP 8 PUNC 9[-23.716]
NNP → "Denver" [-4.4002] NP_NNP → "Denver" [-3.3280]	NP_PRIME → 5NNP 6 NNP 7[-6.110] NP → 5 NNP 6NNP 7 [-5.070]	NP → 5 NP 7 NNS 8 [-17.330] NP → 5NP_PRIME 7 NNS 8 [-15.426]	TOP → 5NP 8PUNC 9[-19.809] TOP → 5NP 8PUNC 9[-17.905]
6	NNP → "during" [1.0000] NN → "during" [1.0000] NP_NNP → "during" [1.0000] VB → "during" [1.0000] CD → "during" [1.0000]	VP → 6 VB 7NP_NNS 8[-8.922] S_VP → 6 VB 7NP_NNS 8[-6.611]	TOP → 6VP 8PUNC 9[-11.410] TOP → 6S_VP 8PUNC 9[-9.176]
	7	NNS → "weekdays" [-5.5759] NP_NNS → "weekdays" [-3.7257]	TOP → 7NP_NNS 8 PUNC 9[-11.001]
		8	PUNC → "." [-0.3396]
			9

OOV: Propose POS Tags

"Show me Ground transportation in Denver during weekdays ." — No "during"!

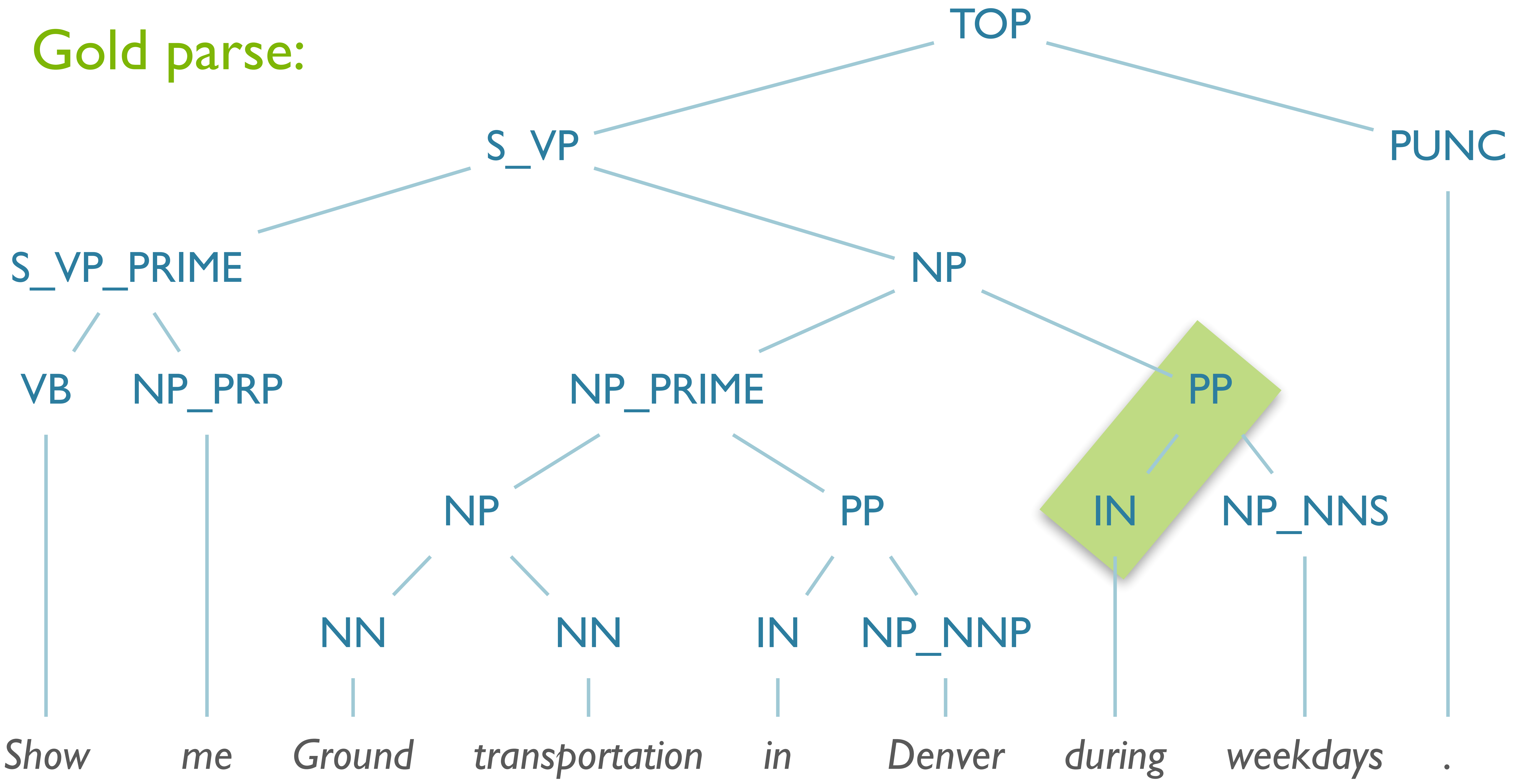
Parse result:



OOV: Propose POS Tags

"Show me Ground transportation in Denver during weekdays ." — No "during"!

Gold parse:



Problems with this approach?

Handling OOV

- **Option #1**:
  - Choose subset of training data vocab to be hidden
  - Hidden words replaced by `<UNK>`
  - Run induction as usual, but some words are now `'<UNK>'`
- **Option #2**:
  - Implicit vocab creation:
    - Replace all words occurring less than *n* times with <UNK>
    - Fix size of V (e.g. 50,000), anything not among |V| most frequent is <UNK>
- (See J&M 2nd ed 4.3.2 — 3rd ed, 3.3.1)

Problems with These Approaches?

- **Option #1**
  - May sample “closed-class” words
  - Closed-class words are disproportionately more common
    - ∴ Approximation will be worse the more data there is, *because Zipf*
- **Option #2**
  - **Con**: Requires a lot more data
  - **Pros**: Samples from all word classes
    - Will only count closed-class words once