CZ4045 Natural Language Processing

Tutorial 7: Statistical Parsing

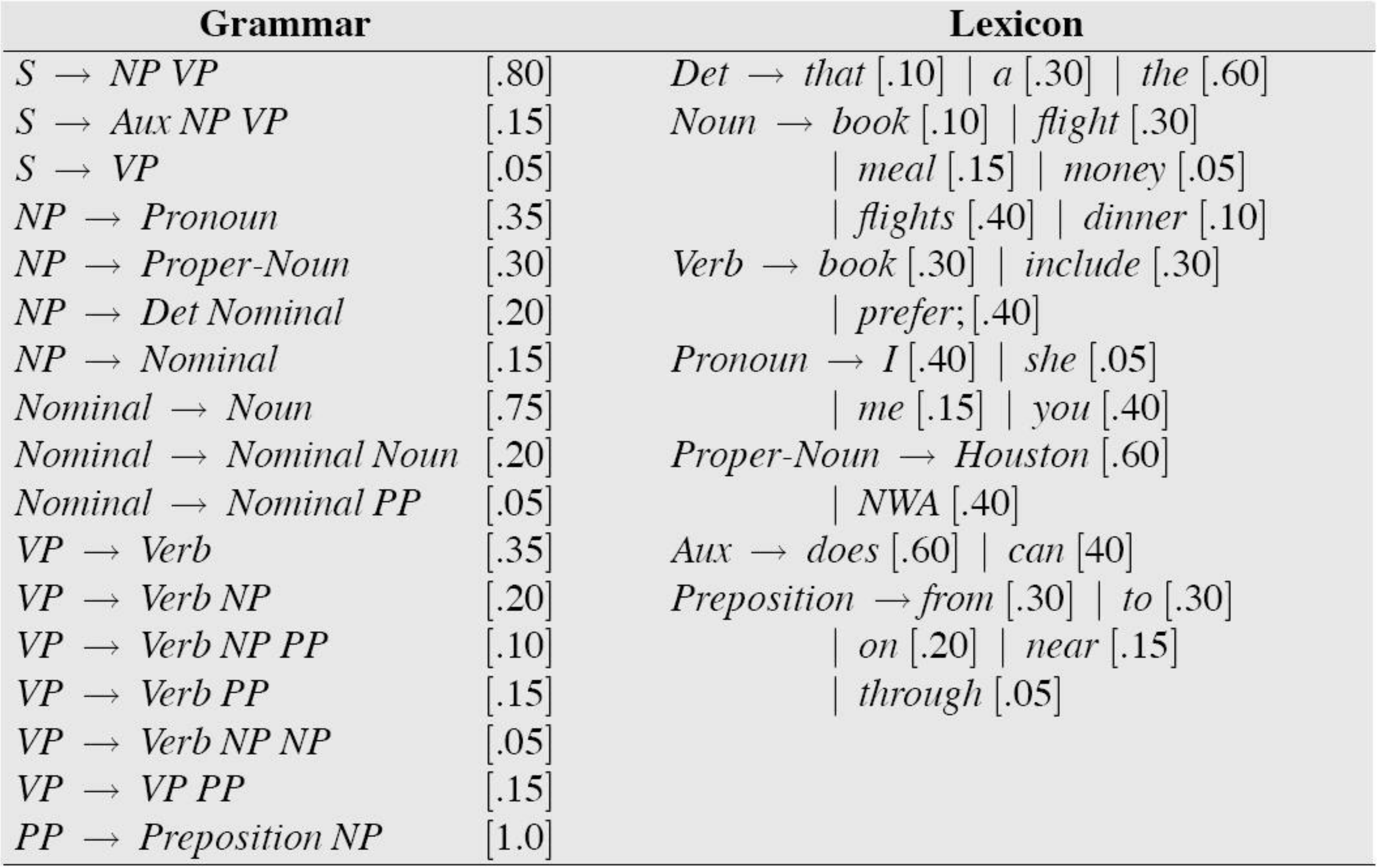
Q1. Derive a PCFG for the following corpus

```
( (S
    (NP-SBJ
      (NP (NNP Pierre) (NNP Vinken) )
      (, ,)
      (ADJP
        (NP (CD 61) (NNS years) )
        (JJ old) )
      (, ,) )
    (VP (MD will)
      (VP (VB join)
        (NP (DT the) (NN board) )
        (PP-CLR (IN as)
          (NP (DT a) (JJ nonexecutive)
             (NN director) ))
        (NP-TMP (NNP Nov.) (CD 29) )))
    (. .) ))
```

```
( (S
    (NP-SBJ (NNP Mr.) (NNP Vinken) )
    (VP (VBZ is)
      (NP-PRD
        (NP (NN chairman) )
        (PP (IN of)
          (NP
            (NP (NNP Elsevier) (NNP N.V.) )
            (, ,)
            (NP (DT the) (NNP Dutch)
              (VBG publishing) (NN group) )))))
    (. .) ))
```

Review

PCFG: Example



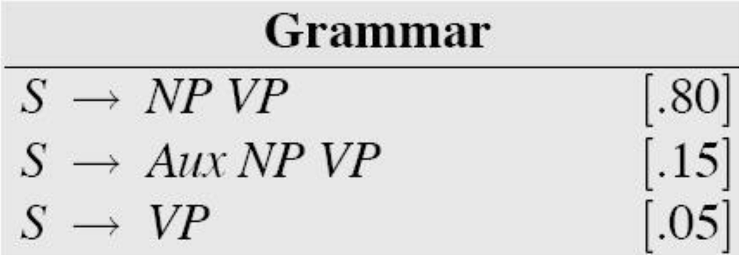
Grammar		Lexicon
S → *NP VP*	[.80]	*Det* → *that* [.10] | *a* [.30] | *the* [.60]
S → *Aux NP VP*	[.15]	*Noun* → *book* [.10] | *flight* [.30]
S → *VP*	[.05]	| *meal* [.15] | *money* [.05]
NP → *Pronoun*	[.35]	| *flights* [.40] | *dinner* [.10]
NP → *Proper-Noun*	[.30]	*Verb* → *book* [.30] | *include* [.30]
NP → *Det Nominal*	[.20]	| *prefer;* [.40]
NP → *Nominal*	[.15]	*Pronoun* → *I* [.40] | *she* [.05]
Nominal → *Noun*	[.75]	| *me* [.15] | *you* [.40]
Nominal → *Nominal Noun*	[.20]	*Proper-Noun* → *Houston* [.60]
Nominal → *Nominal PP*	[.05]	| *NWA* [.40]
VP → *Verb*	[.35]	*Aux* → *does* [.60] | *can* [40]
VP → *Verb NP*	[.20]	*Preposition* → *from* [.30] | *to* [.30]
VP → *Verb NP PP*	[.10]	| *on* [.20] | *near* [.15]
VP → *Verb PP*	[.15]	| *through* [.05]
VP → *Verb NP NP*	[.05]	
VP → *VP PP*	[.15]	
PP → *Preposition NP*	[1.0]	

Review

Probabilistic Context-free grammar (PCFG)

- G = (T, N, S, R, P)
  - T: a set of terminals (e.g. ‘boy’)
  - N: a set of nonterminals (e.g. Noun)
  - S: the start symbol, a nonterminal
  - R: rules of the form X →γ
  - P(R) gives the probability of each rule

$$\forall X \in N, \sum_{X \to \gamma \in R} P(X \to \gamma) = 1$$



Grammar			
S	$\rightarrow$	$NP\ VP$	[.80]
S	$\rightarrow$	$Aux\ NP\ VP$	[.15]
S	$\rightarrow$	VP	[.05]

Q1. Derive a PCFG for the following corpus

- You can make your own simplification
  - e.g. NNS → NN
  - ignore punctuation marks

```
( (S
    (NP-SBJ
      (NP (NNP Pierre) (NNP Vinken) )
      (, ,)
      (ADJP
        (NP (CD 61) (NNS years) )
        (JJ old) )
      (, ,) )
    (VP (MD will)
      (VP (VB join)
        (NP (DT the) (NN board) )
        (PP-CLR (IN as)
          (NP (DT a) (JJ nonexecutive)
             (NN director) ))
        (NP-TMP (NNP Nov.) (CD 29) )))
    (. .) ))
```

```
( (S
    (NP-SBJ (NNP Mr.) (NNP Vinken) )
    (VP (VBZ is)
      (NP-PRD
        (NP (NN chairman) )
        (PP (IN of)
          (NP
            (NP (NNP Elsevier) (NNP N.V.) )
            (, ,)
            (NP (DT the) (NNP Dutch)
              (VBG publishing) (NN group) )))))
    (. .) ))
```

Q1. Derive a PCFG for the following corpus

- You can make your own simplification
  - e.g. NNS → NN
  - ignore punctuation marks

```
( (S
    (NP
      (NP (NN Pierre) (NN Vinken) )
      (ADJP
        (NP (CD 61) (NN years) )
        (JJ old) )
      )
    (VP (MD will)
      (VP (VB join)
        (NP (DT the) (NN board) )
        (PP (IN as)
          (NP (DT a) (JJ nonexecutive)
             (NN director) ))
        (NP (NN Nov.) (CD 29) )))
))
```

```
( (S
    (NP (NN Mr.) (NN Vinken) )
    (VP (VB is)
      (NP
        (NP (NN chairman) )
        (PP (IN of)
          (NP
            (NP (NN Elsevier) (NN N.V.) )
            (NP (DT the) (NN Dutch)
              (VB publishing) (NN group) )))))
 ))
```

A1.

- NP -> NN NN
- NP -> CD NN
- ADJP -> NP JJ
- NP -> NP ADJP
- NP -> DT NN
- NP -> DT JJ NN
- PP -> IN NP
- NP -> NN CD
- VP -> VB NP PP NP

- VP -> MD VP
- S -> NP VP
- NP -> NN
- NP -> DT NN VB NN
- NP -> NP NP
- NP -> NP PP
- VP -> VB NP
- NP -> JJ NN

A1.

- NP -> CD NN (1/12)
- NP -> DT JJ NN (1/12)
- NP -> DT NN (1/12)
- NP -> DT NN VB NN (1/12)
- NP -> NN (1/12)
- NP -> NN CD (1/12)
- NP -> NN NN (3/12)
- NP -> NP ADJP (1/12)
- NP -> NP NP (1/12)
- NP -> NP PP (2/12)

- VP -> VB NP (1/3)
- VP -> VB NP PP NP (1/3)
- VP -> MD VP (1/3)

- S -> NP VP (2/2)

- ADJP -> NP JJ (1/1)

- PP -> IN NP (2/2)

Q2. Probability of a parse tree

- Assign arbitrary probabilities ($0 < P < 1$) to the rules of the revised L1 grammar from Q3 of Tutorial 6. Based on them, calculate the probability of the phrase structure of the following sentence

- Please repeat that.

Review

Probability of parse trees

- A derivation (parse tree) consists of the bag of grammar rules that are in the tree
  - The probability of a tree is the product of the probabilities of the rules in the derivation.

$$P(T,S) = \prod_{node \in T} P(rule(n))$$

1. S → NP VP
2. NP → Pro
   Pro → I
3. VP → Verb NP
   Verb → prefer
4. NP → Det Nom
   Det → a
5. Nom → Nom Noun
   Noun → morning
6. Nom → Noun
   Noun → flight

Tutorial 6, Answer 3.

- S → NP VP
- S → Aux NP VP
- S → VP
- NP → Pronoun
- NP → ProperNoun
- NP → Det Nominal
- NP → NP Conj NP
- Nominal → Noun
- Nominal → Nominal Noun
- Nominal → Nominal PP
- VP → Verb
- VP → Verb NP
- VP → Verb NP PP
- VP → Verb PP
- VP → VP PP
- VP → Aux VP
- VP → Verb VP
- VP → Inf Verb PP
- VP → Adv Verb NP
- PP → Preposition NP
- -----------------------------
- Det → the
- Noun → fare
- Verb → like | fly | repeat | need | is
- Pronoun → I | that | what
- ProperNoun → American airlines | Philadelphia | Atlanta | Denver
- Aux → would
- Preposition → from | to | on | between
- Conj → and
- Inf → to
- Adv → please

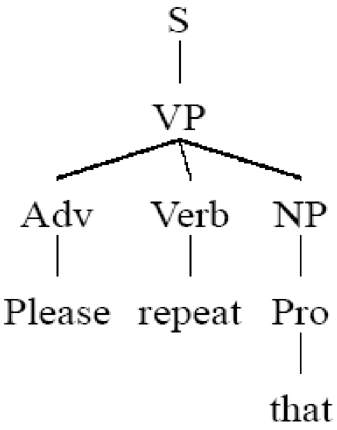
Assign random values

- S → NP VP [.40]
- S → Aux NP VP [.30]
- S → VP [.30]
- NP → Pronoun [.10]
- NP → ProperNoun [.10]
- NP → Det Nominal [.10]
- NP → NP Conj NP [.70]
- Nominal → Noun [.30]
- Nominal → Nominal Noun [.40]
- Nominal → Nominal PP [.30]
- VP → Verb [.10]
- VP → Verb NP [.10]
- VP → Verb NP PP [.10]
- VP → Verb PP [.10]
- VP → VP PP [.10]
- VP → Aux VP [.10]
- VP → Verb VP [.10]
- VP → Inf Verb PP [.10]
- VP → Adv Verb NP [.20]
- PP → Preposition NP [1.00]
- ----------------------------
- Det → the [1.00]
- Noun → fare [1.00]
- Verb → like [.30] | fly [.30] | repeat [.10] | need [.10] | is [.20]
- Pronoun → I [.20] | that [.60] | what [.20]
- ProperNoun → American airlines [.10] | Philadelphia [.10] | Atlanta [.10] | Denver [.10]
- Aux → would [1.00]
- Preposition → from [.10] | to [.10] | on [.10] | between [.70]
- Conj → and [1.00]
- Inf → to [1.00]
- Adv → please [1.00]

Sentence: Please repeat that

S → NP VP [.40]
S → Aux NP VP [.30]
S → VP [.30]

NP → Pronoun [.10]
NP → ProperNoun [.10]
NP → Det Nominal [.10]
NP → NP Conj NP [.70]



VP → Verb [.10]
VP → Verb NP [.10]
VP → Verb NP PP [.10]
VP → Verb PP [.10]
VP → VP PP [.10]
VP → Aux VP [.10]
VP → Verb VP [.10]
VP → Inf Verb PP [.10]
VP → Adv Verb NP [.20]

Verb → like [.30] | fly [.30] | repeat [.10] | need [.10] | is [.20]

Pronoun → I [.20] | that [.60] | what [.20]

Adv → please [1.00]

Sentence: Please repeat that

S → NP VP [.40]
S → Aux NP VP [.30]
S → VP [.30]

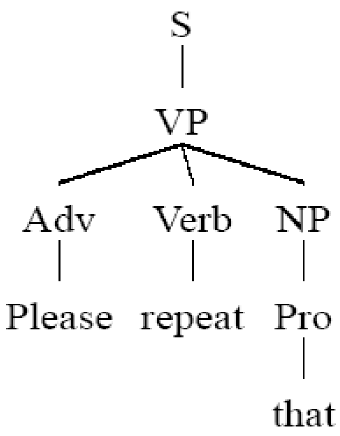
NP → Pronoun [.10]
NP → ProperNoun [.10]
NP → Det Nominal [.10]
NP → NP Conj NP [.70]

VP → Verb [.10]
VP → Verb NP [.10]
VP → Verb NP PP [.10]
VP → Verb PP [.10]
VP → VP PP [.10]
VP → Aux VP [.10]
VP → Verb VP [.10]
VP → Inf Verb PP [.10]
VP → Adv Verb NP [.20]

Verb → like [.30] | fly [.30] | **repeat [.10]** | need [.10] | is [.20]

Pronoun → I [.20] **| that [.60]** | what [.20]

Adv → please [1.00]



$P = 3.6*10^{-5}$

Q3.

- Use two sentences from Q3 of Tutorial 6 as examples to show the improved version of the revised L1 grammar

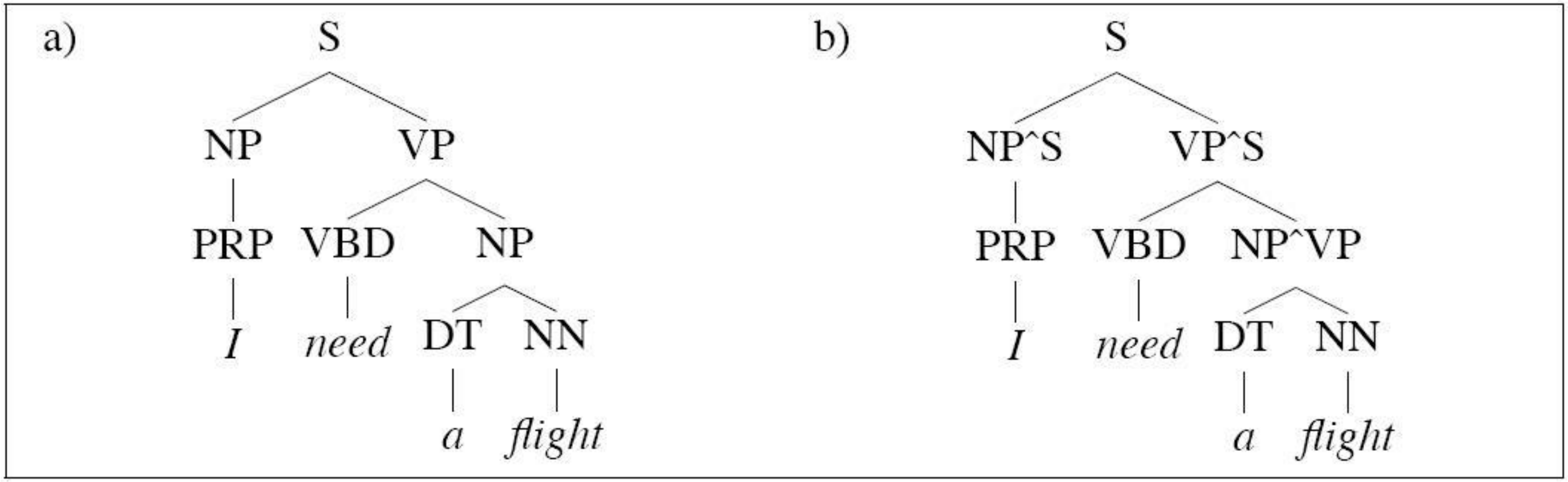
  – Splitting non-terminals

  – Lexicalization

Improving PCFG: Splitting Non-Terminals

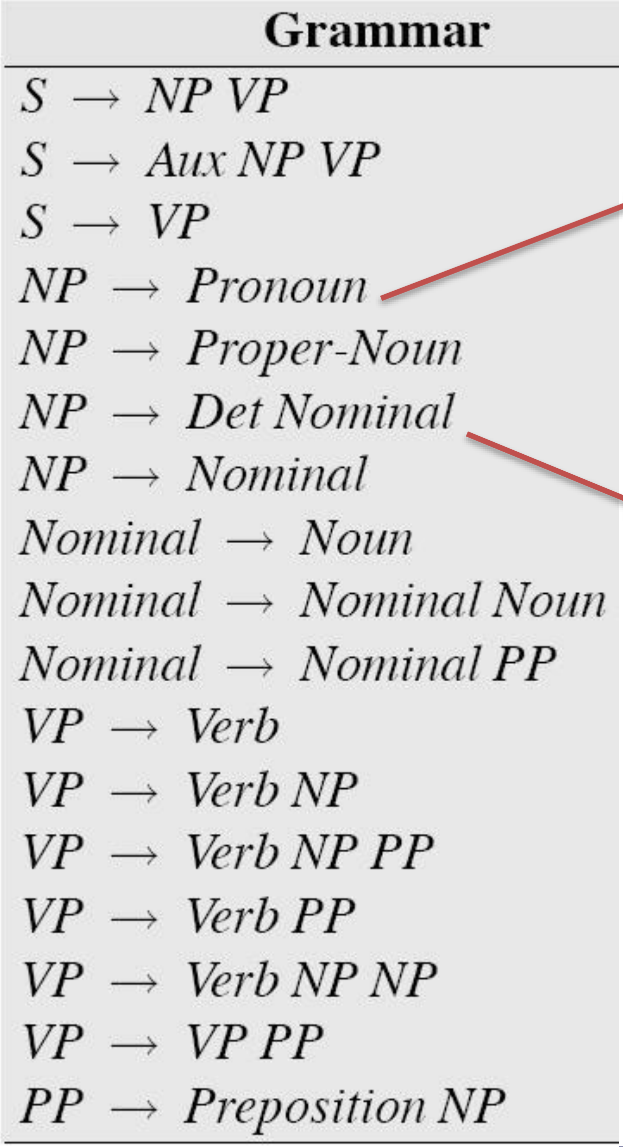
Review

- Encoding contextual dependencies into PCFG symbols



Review

Improving PCFG: Splitting Non-Terminals

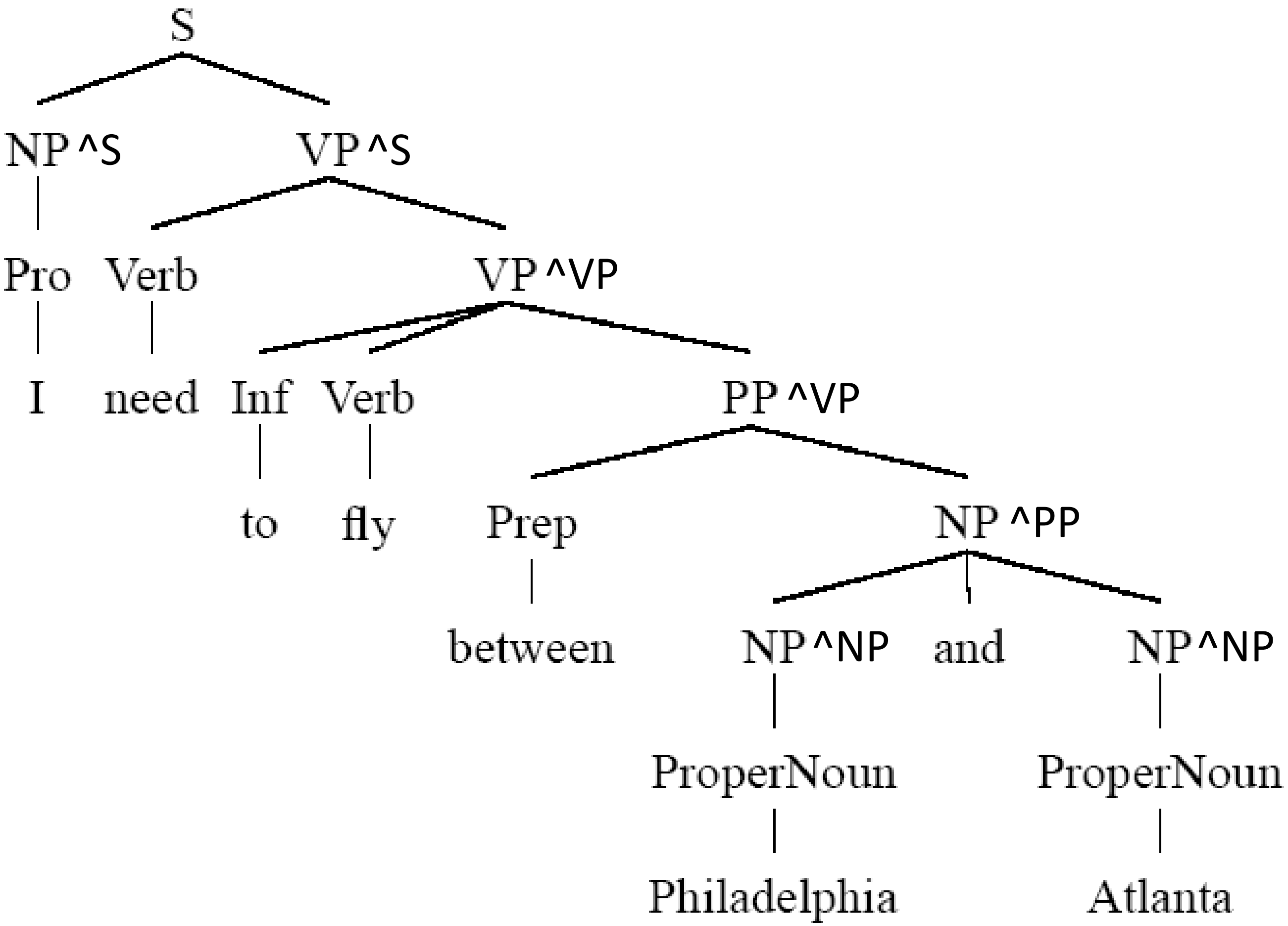


Grammar
S → *NP VP*
S → *Aux NP VP*
S → *VP*
NP → *Pronoun*
NP → *Proper-Noun*
NP → *Det Nominal*
NP → *Nominal*
Nominal → *Noun*
Nominal → *Nominal Noun*
Nominal → *Nominal PP*
VP → *Verb*
VP → *Verb NP*
VP → *Verb NP PP*
VP → *Verb PP*
VP → *Verb NP NP*
VP → *VP PP*
PP → *Preposition NP*

NP^S → Pronoun
NP^VP → Pronoun
NP^PP → Pronoun

NP^S → Det Nominal^NP
NP^VP → Det Nominal^NP
NP^PP → Det Nominal^NP

A3-a.


S
NP ^S
VP ^S
Pro
Verb
VP ^VP
I
need
Inf
Verb
PP ^VP
to
fly
Prep
NP ^PP
between
NP ^NP
and
NP ^NP
ProperNoun
ProperNoun
Philadelphia
Atlanta

A3-a.

- S → NP VP [.40]
- S → Aux NP VP [.30]
- S → VP [.30]
- NP → Pronoun [.50]
- NP → ProperNoun [.30]
- NP → Det Nominal [.10]
- NP → NP Conj NP [.10]
- Nominal → Noun [.30]
- Nominal → Nominal Noun [.40]
- Nominal → Nominal PP [.30]
- VP → Verb [.10]
- VP → Verb NP [.10]
- VP → Verb NP PP [.10]
- VP → Verb PP [.10]
- VP → VP PP [.10]
- VP → Aux VP [.10]
- VP → Verb VP [.10]
- VP → Inf Verb PP [.10]
- VP → Adv Verb NP [.20]
- PP → Preposition NP [1.00]

- NP^S → Pronoun [.30]
- NP^VP → Pronoun [.10]
- NP^PP → Pronoun [.10]
- NP^S → ProperNoun [.20]
- NP^VP → ProperNoun [.05]
- NP^PP → ProperNoun [.05]

…

- Nominal^NP → Noun [.20]
- Nominal^Nominal → Noun [.10]

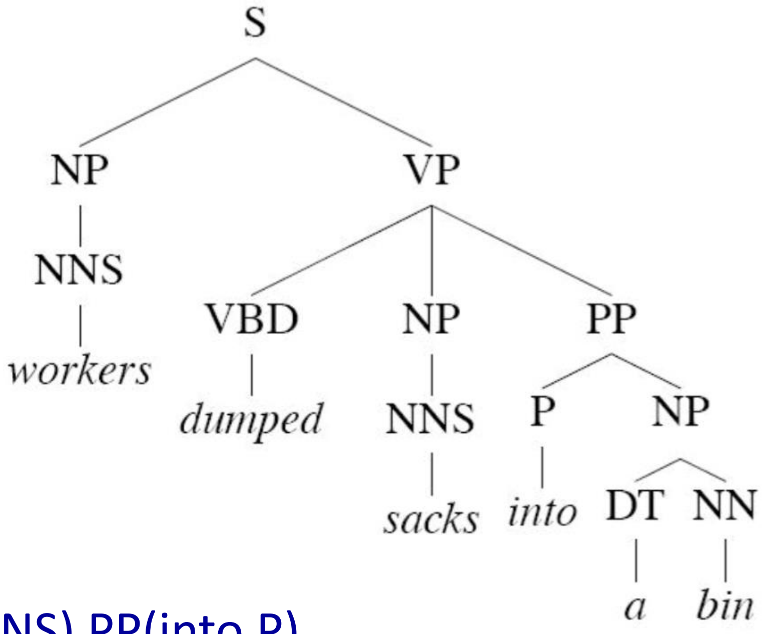
…

Review

Improving PCFG: Lexicalized PCFG

- (Review) Lexical head
  - E.g. N is the head of NP
  - E.g. V is the head of VP
  - The word in the phrase that is grammatically the most important
- VP → VBD NP PP

How to add lexical information to rules?



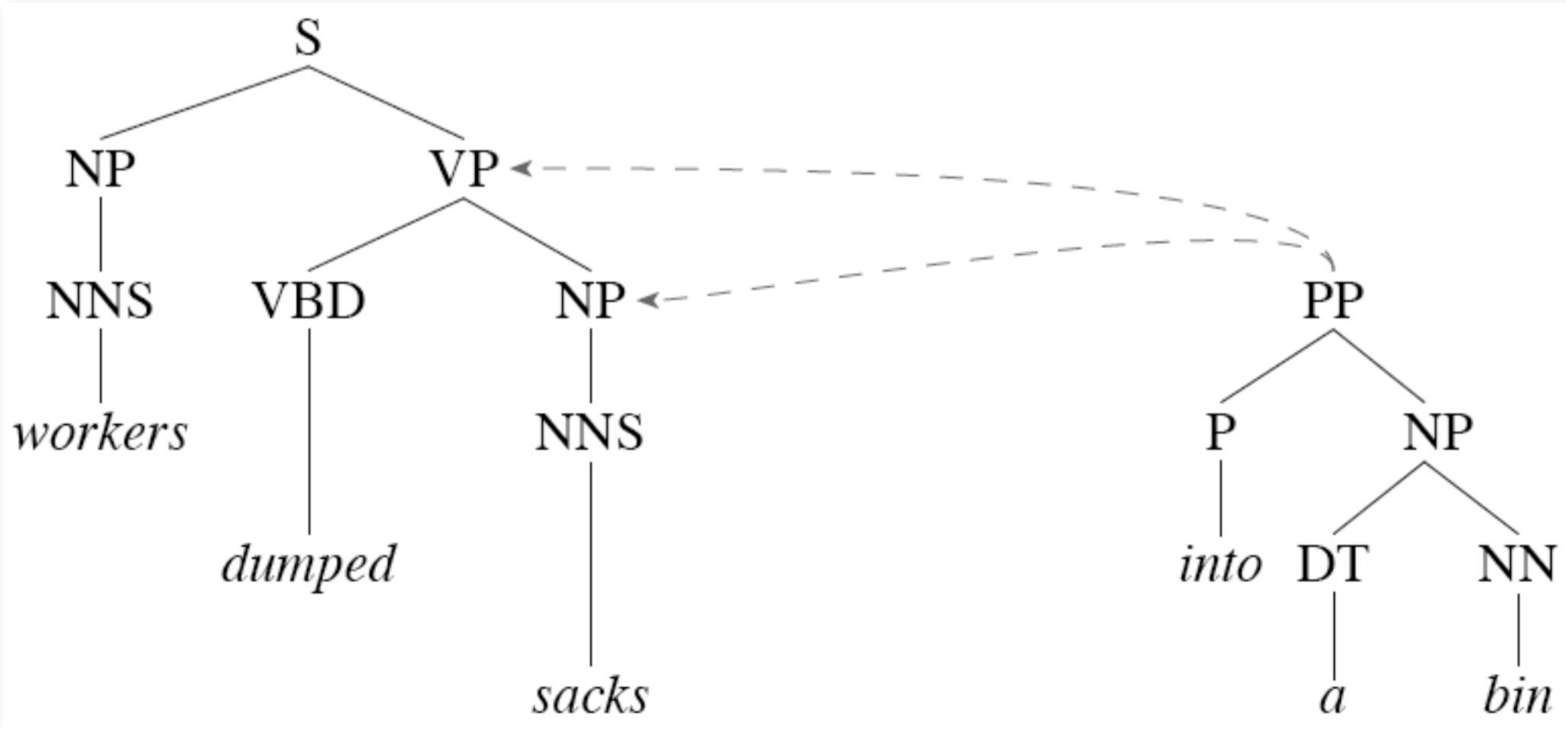
VP(**dumped**) → VBD(dumped) NP(sacks) PP(into)

VP(**dumped,VBD**) → VBD(dumped,VBD) NP(sacks,NNS) PP(into,P)

PP Attachment

Review

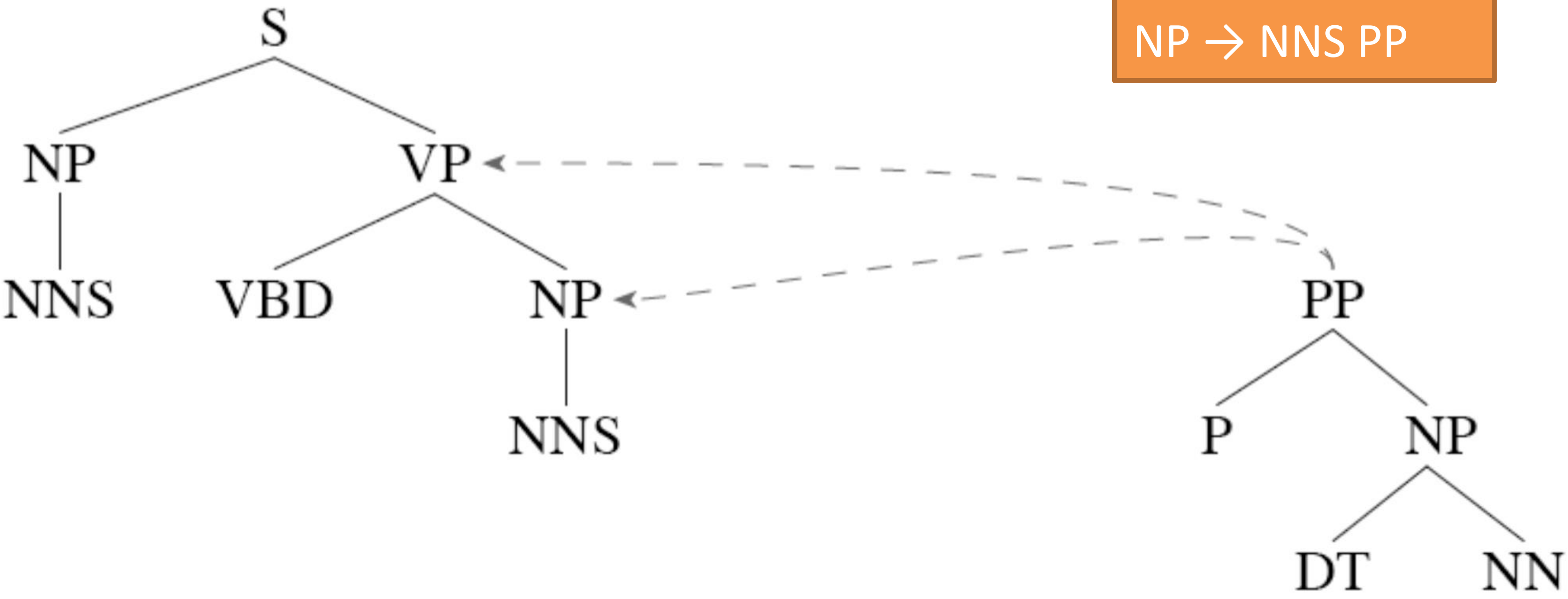
- E.g. Workers dumped sacks into a bin.



Review

PP Attachment

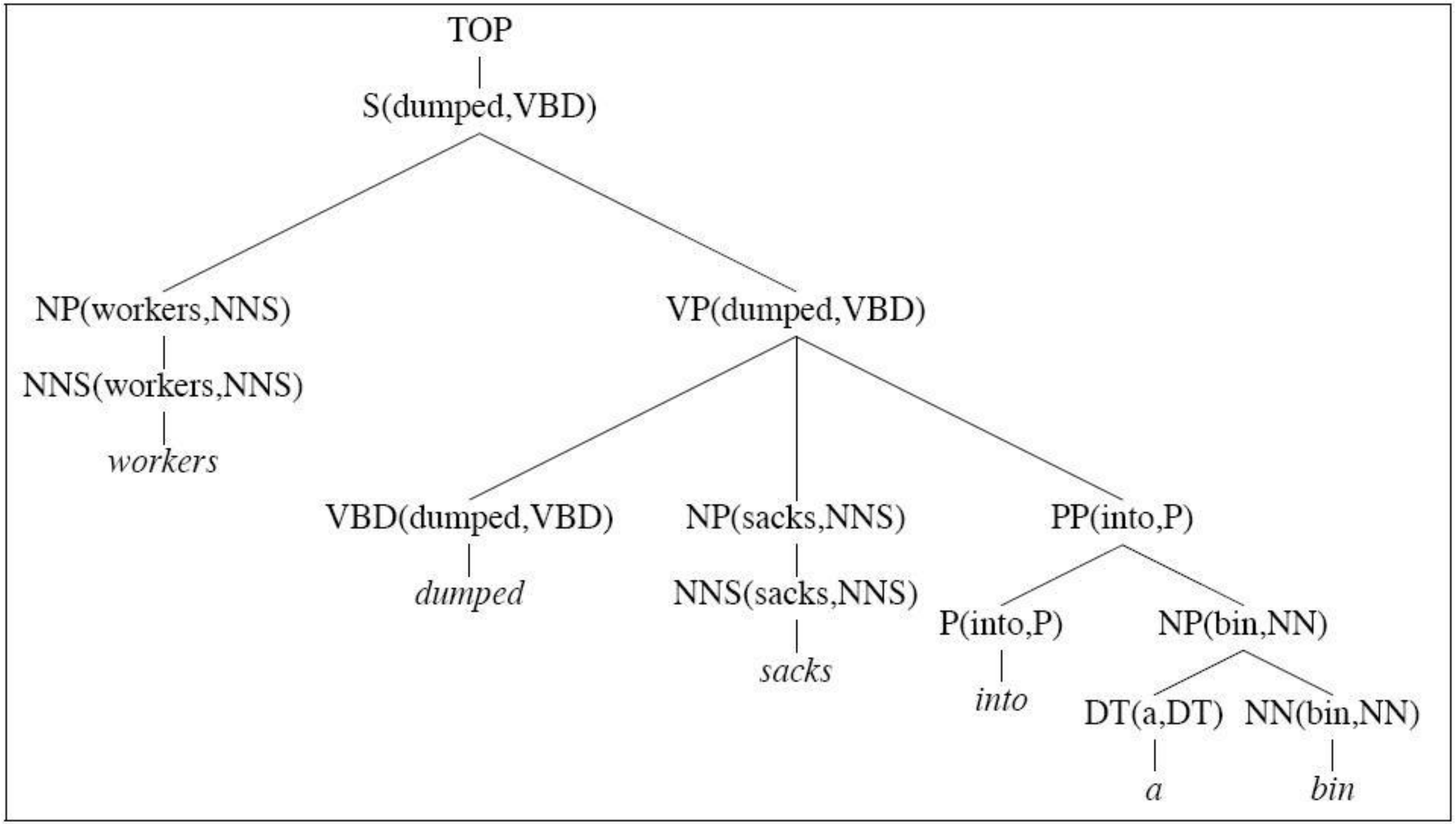
- NNS VBD NNS P DT NN

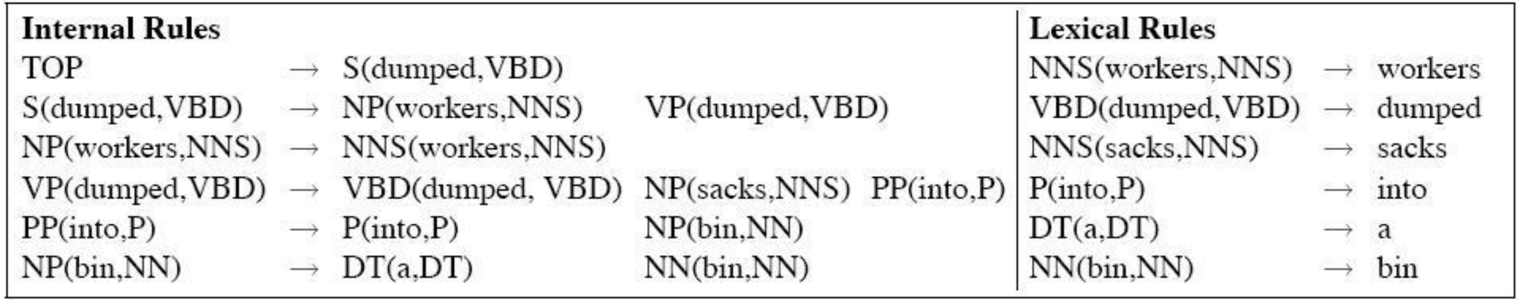


VP → VBD NP PP
NP → NNS PP

Improving PCFG: Lexicalized PCFG

Review



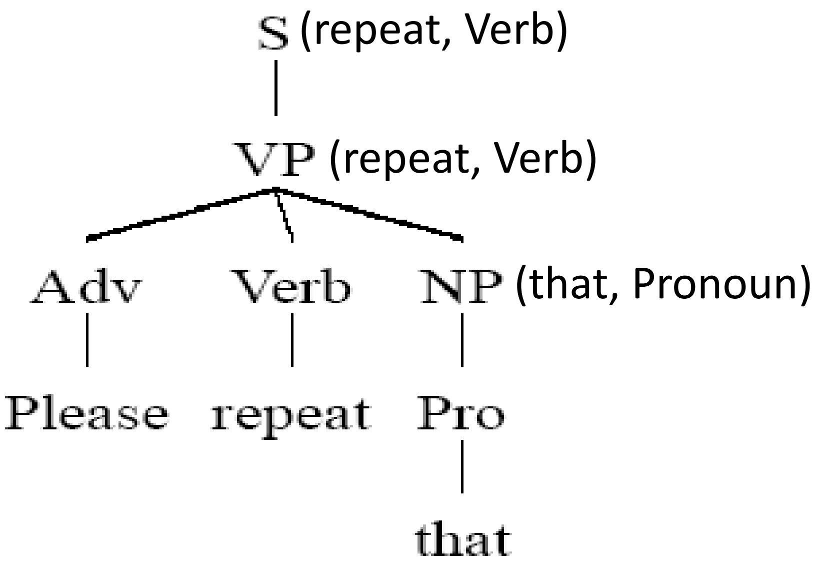


Internal Rules					**Lexical Rules**		
TOP	→	S(dumped,VBD)			NNS(workers,NNS)	→	workers
S(dumped,VBD)	→	NP(workers,NNS)	VP(dumped,VBD)		VBD(dumped,VBD)	→	dumped
NP(workers,NNS)	→	NNS(workers,NNS)			NNS(sacks,NNS)	→	sacks
VP(dumped,VBD)	→	VBD(dumped, VBD)	NP(sacks,NNS)	PP(into,P)	P(into,P)	→	into
PP(into,P)	→	P(into,P)	NP(bin,NN)		DT(a,DT)	→	a
NP(bin,NN)	→	DT(a,DT)	NN(bin,NN)		NN(bin,NN)	→	bin

A3-b.

VP(repeat) → Adv(please) Verb(repeat) NP(that)

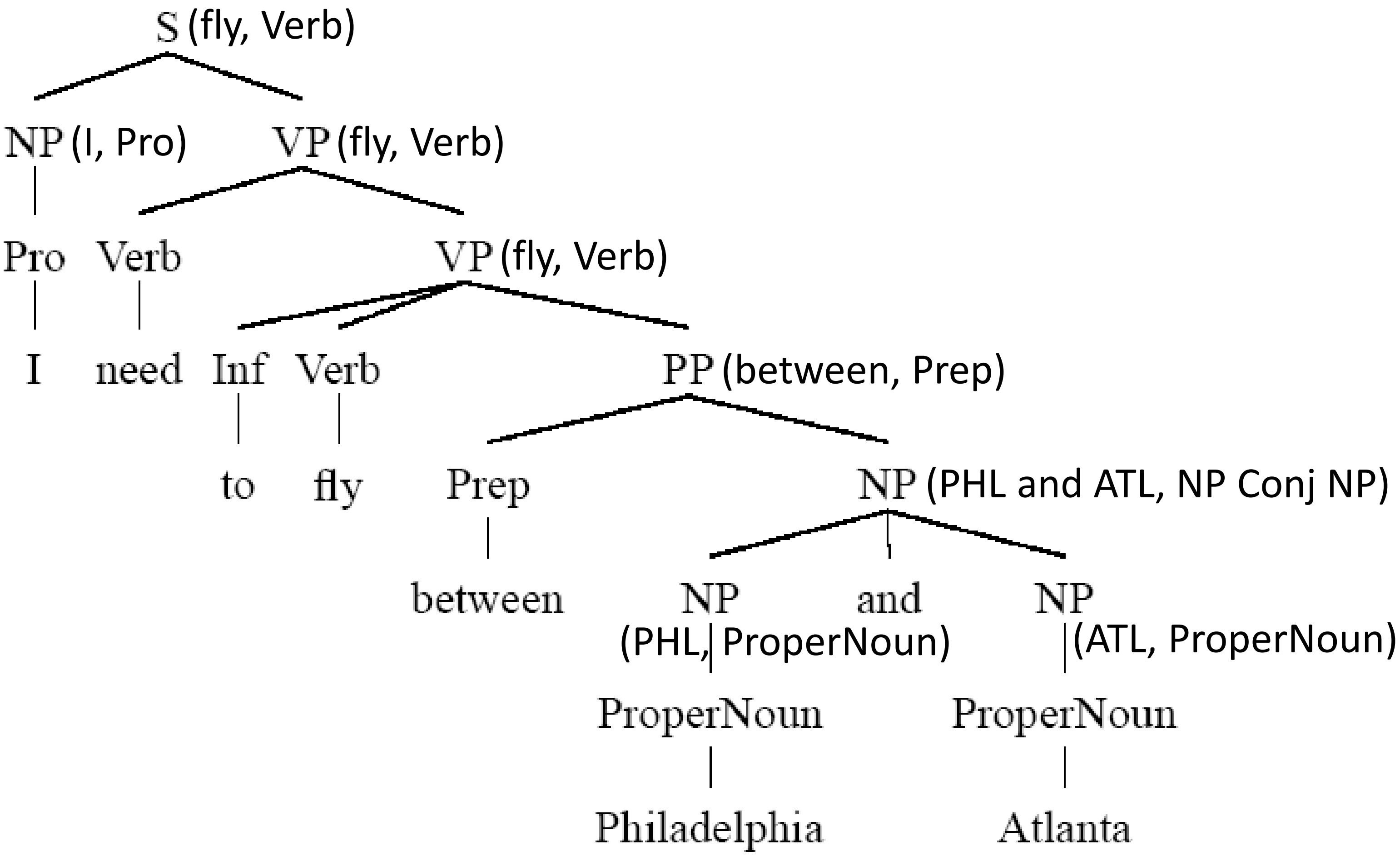
VP(repeat, Verb) → Adv(please, Adv) Verb(repeat, Verb) NP(that, Pronoun)



A3-b.

VP(fly) → Inf(to) Verb(fly) PP(between)

VP(fly, Verb) → Inf(to, Inf) Verb(fly, Verb) PP(between, prep)


S (fly, Verb)
NP (I, Pro)
VP (fly, Verb)
Pro
Verb
VP (fly, Verb)
I
need
Inf
Verb
PP (between, Prep)
to
fly
Prep
NP (PHL and ATL, NP Conj NP)
between
NP
and
NP
(PHL, ProperNoun)
(ATL, ProperNoun)
ProperNoun
ProperNoun
Philadelphia
Atlanta