

---

# Treebanks

LING 5200

---

Computational Corpus Linguistics

Nianwen Xue

# Outline

- Intuitions and tests for constituent structure
- Representing constituent structures
  - Continuous constituents
  - Discontinuous constituents
  - Types of traces
- Structural ambiguity resolution
  - PP attachment
  - Human and computer perspectives
- Building and accessing trees with NLTK

---

# Some words belong together

- Did [the dog] the children like chase the cat?
- Did the [dog the] children like chase the cat?

How do we decide what is a constituent?

# Substitution test

- Replace the constituent with a pro-form
  - The little boy fed the cat. →  
He fed her.
  - The little boy from next door fed the cat without a tail. →  
He fed her.
  - The little boy from next door fed the cat without a tail. →  
\* He from next door fed her without a tail.

# Substitution test

- ❑ Black cats detest green peas. ➡  
They detest them.
- ❑ These black cats detest those green peas. ➡  
They detest them.
- ❑ These black cats detest those green peas. ➡  
\* These they detest those them.

Assumption: only constituents can be substituted with proforms

# Pronouns are not the only proforms

- Put it **on the table**. → Put it **there**.
- Put it over **on the table**. → Put it over **there**.
  
- Put it **on the table** that's by the door. →
  - \* Put it **there** that's by the door.
- Put it over **on the table** that's by the door. →
  - \* Put it over **there** that's by the door.

# Pro-adjective

- I am **very happy**, ... .. and Linda is **so**, too.
- I am **very fond of Lukas**, ... .. and Linda is **so**, too.
- I am **very fond of** my nephew, ... \* ... and Linda is **so** of her niece.

# Moving NPs

- I fed **the cats**. →

**The cats**, I fed \_\_\_\_\_. (The dogs, I didn't.)

- I fed **the cats with long, fluffy tails**. →

**The cats with long, fluffy tails**, I fed \_\_\_\_\_. (The other cats, I didn't.)

\* **The cats**, I fed \_\_\_\_\_ **with long, fluffy tails**.

Assumption: only constituents can be moved

## Moving PP, ADJP,

- The cat strolled across the porch **with a confident air**.  
**With a confident air**, the cat strolled across the porch \_\_\_\_ →
- \* **With a**, the cat strolled across the porch \_\_\_\_  
**confident air**.
- Ali Baba returned from his travels **wiser than before**. →  
**Wiser than before**, Ali Baba returned from his travels \_\_\_\_.
- \* **Wiser than**, Ali Baba returned from his travels \_\_\_\_  
**before**.

# Moving ADVP

- They arrived at the concert hall **more quickly than they had expected.** →

**More quickly than they had expected,** they arrived at the concert hall \_\_\_\_.

- \* **More quickly than they,** they arrived at the concert hall \_\_\_\_ **had expected.**

# Can it be a sentence fragment in response to a question?

## ■ Noun phrase:

- What do you like?

The cats.

Cats with long, fluffy tails.

The cats with long, fluffy tails.

## ■ Prepositional phrase:

- How did the cat stroll across the porch?

With a confident air.

- Where did Ali Baba go?

On a long journey.

To New York.

# Can it be a sentence fragment in response to a question?

- Adjective phrase:
  - How did Ali Baba return?  
Wiser than before.  
Fairly jeg-lagged.
- Adverb phrase:
  - How did they do?  
Not badly.  
Surprisingly well.  
Much better than they had expected.

# Ungrammatical with non-constituents

- \* What did you feed \_\_\_\_ long, fluffy tails?
  - \* The cats with.
- \* How did the cat stroll across the porch \_\_\_\_ confident air?
  - \* With a.
- \* How did Ali Baba return from his travels \_\_\_\_ before?
  - \* Wiser than.
- \* How did they arrive at the concert hall \_\_\_\_ had expected?
  - \* More quickly than they.

# It cleft focus

## ■ Noun phrase

- Ordinary cats detest the smell of citrus fruits.

It is ordinary cats that detest the smell of citrus fruits.

## ■ Prepositional phrase

- The cat strolled across the porch with a confident air.

It was with a confident air that the cat strolled across the porch \_\_\_\_.

## ■ Adjective phrase

- Ali Baba returned from his travels wiser than before.

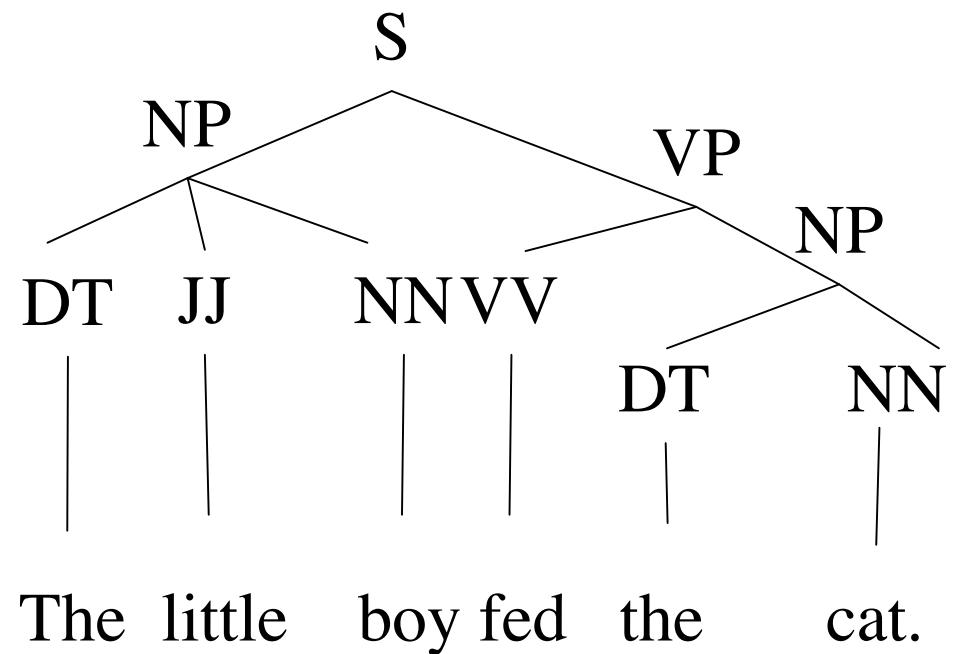
It was wiser than before that Ali Baba returned from his travels \_\_\_\_.

# Ungrammatical it cleft

- Ordinary cats detest **the smell of** citrus fruits.
  - \* It is **the smell of** that ordinary cats detest \_\_\_\_ citrus fruits.
- The cat strolled across the porch **with a confident** air.
  - \* It was **with a confident** that the cat strolled across the porch \_\_\_\_ air.
- Ali Baba returned from his travels **wiser than** before.
  - \* It was **wiser than** that Ali Baba returned from his travels \_\_\_\_ before.

# Representing constituent structure

(S (NP (DT the)  
(ADJ little)  
(NN boy))  
(VP (VV fed)  
(NP (DT the)  
(NN cat)))  
(PU .))



A constituent is exhaustively included in a pair of brackets.

A constituent is exhaustively dominated by a node.

# Bracket v.s. tree

- Labeled brackets
  - The label represents the category of the constituent
  - The text string within a bracket represents a constituent
  - Preferred representation scheme in corpus linguistics
- Tree diagram
  - Syntactic constituents are graphically represented as nodes in a tree
  - The nodes are labeled with the syntactic category of the constituent
  - Preferred illustration scheme in papers, textbooks.

---

# Penn Treebank representation scheme (Marcus 1993)

- Configurational representation with phrasal labels
- Non-configurational representation with functional tags
- Using empty categories to localize non-local dependencies

# PTB Phrasal labels

- S: Simple declarative sentence
- SBAR: Clause introduced by a subordinate conjunction
- SBARQ: Direction question introduced by a wh-word or wh-phrase
- SINV: Inverted declarative sentence
- SQ: Inverted yes-no question
- ADJP: adjective phrase
- ADVP: adverbial phrase
- CONJP: conjunction phrase
- FRAG: fragmentary phrase
- INTJ: interjection phrase
- LST: list
- **NAC**: not a constituent

# PTB Phrasal labels (cont'd)

- NP: noun phrase
- NX: used within certain complex NPs to mark the head of the NP.
- PP: prepositional phrase
- PRN: parenthetical
- PRT: particle
- QP: quantifier phrase
- RRC: reduced relative clause
- UCP: unlike coordinated phrase
- VP: Verb phrase
- WHADJP: Wh-adjective phrase
- WHADVP: Wh-adverb phrase
- WHNP: Wh-noun phrase
- WHPP: Wh-prepositional phrase
- X: unknown, uncertain, or unbracketable

# A PTB example

(S (NP-SBJ The Mortgage and equity real  
estate investment trust)

(ADVP last)

(VP (VBD paid)

(NP a dividend))

(PP-TMP (IN on)

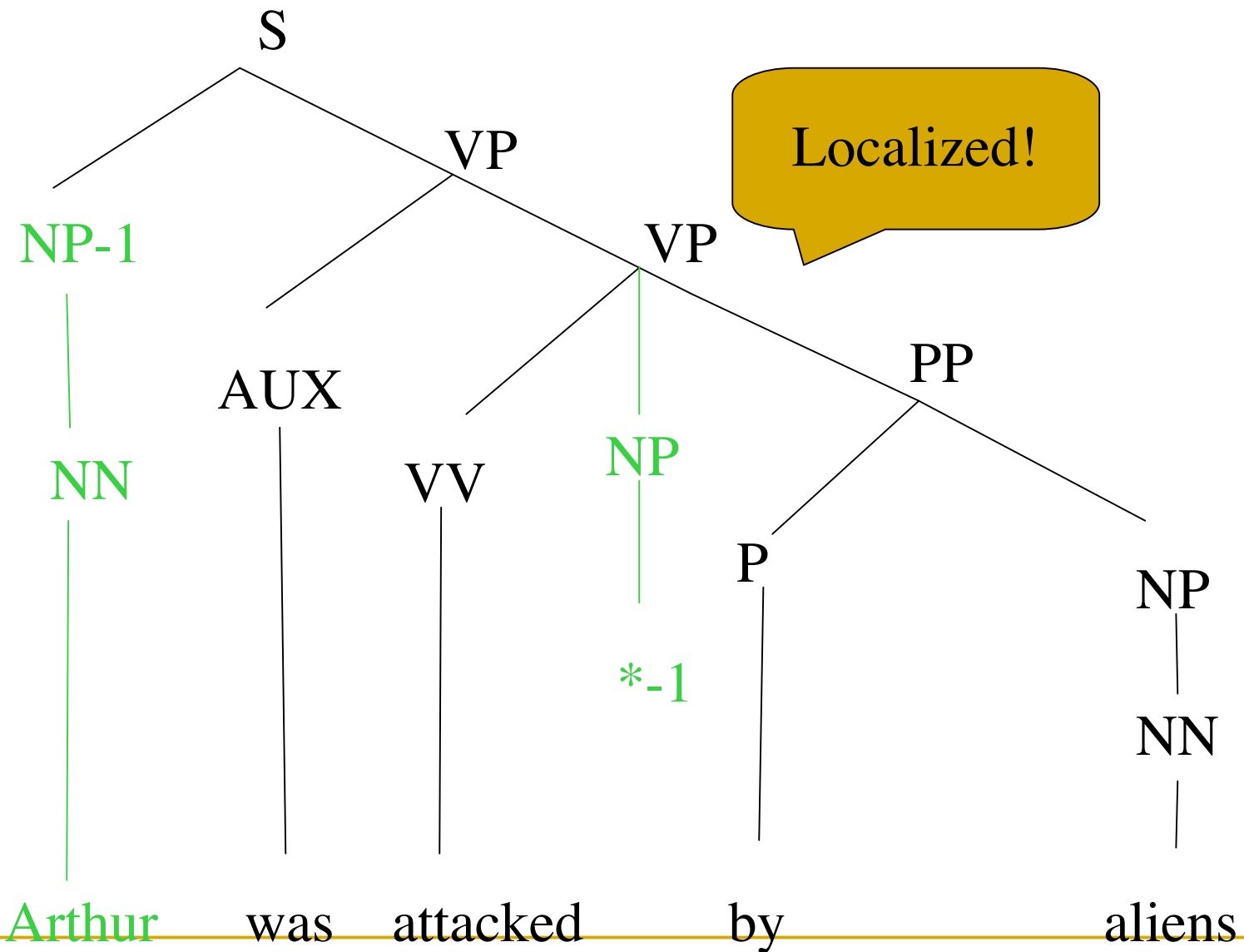
(NP August 1, 1988)))

---

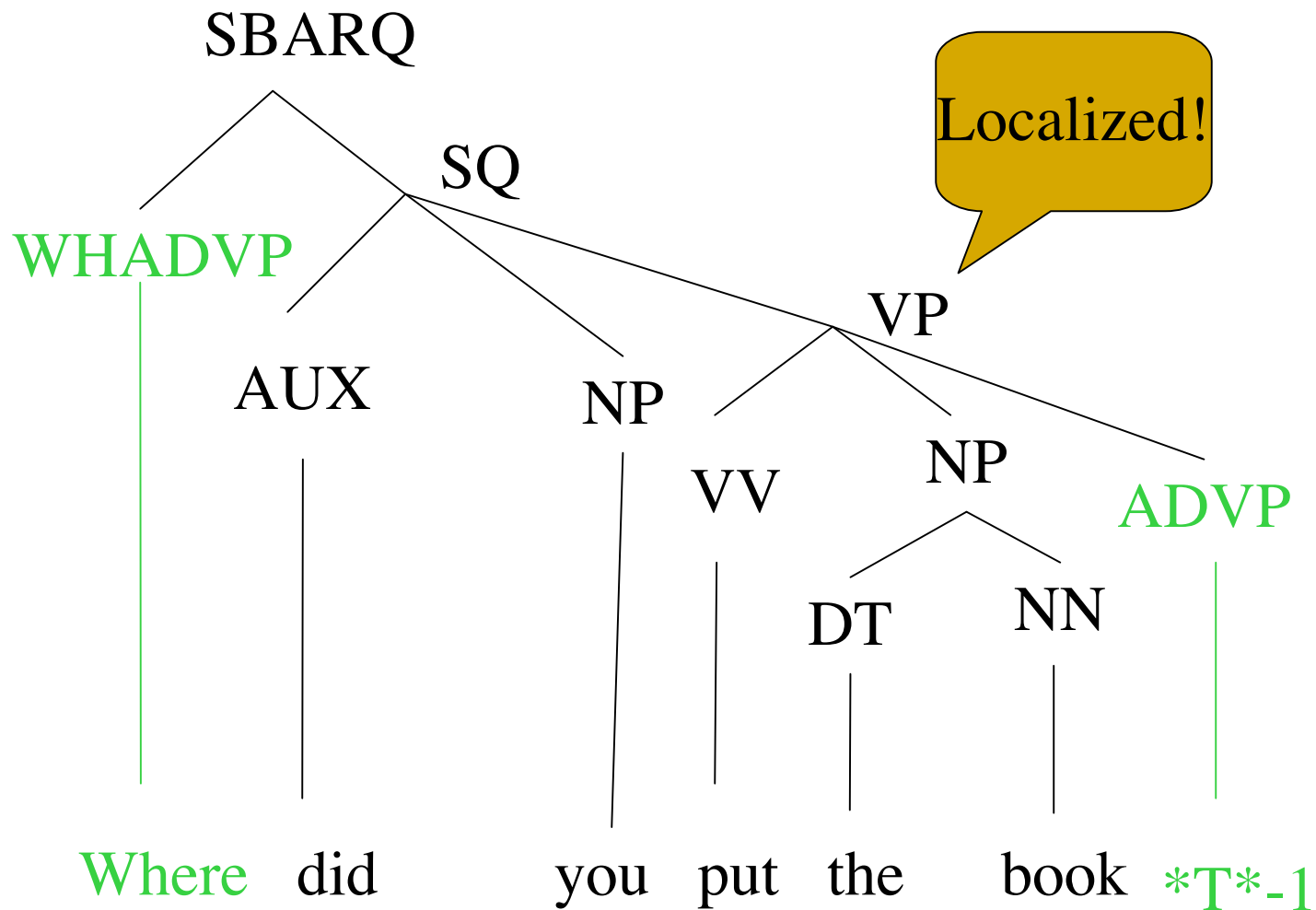
# Representing discontinuous constituents

- Is there a way of making discontinuous constituents continuous (or alternatively, making long-distance dependencies local)?
- The answer: using trace!
- Believers and doubters of trace

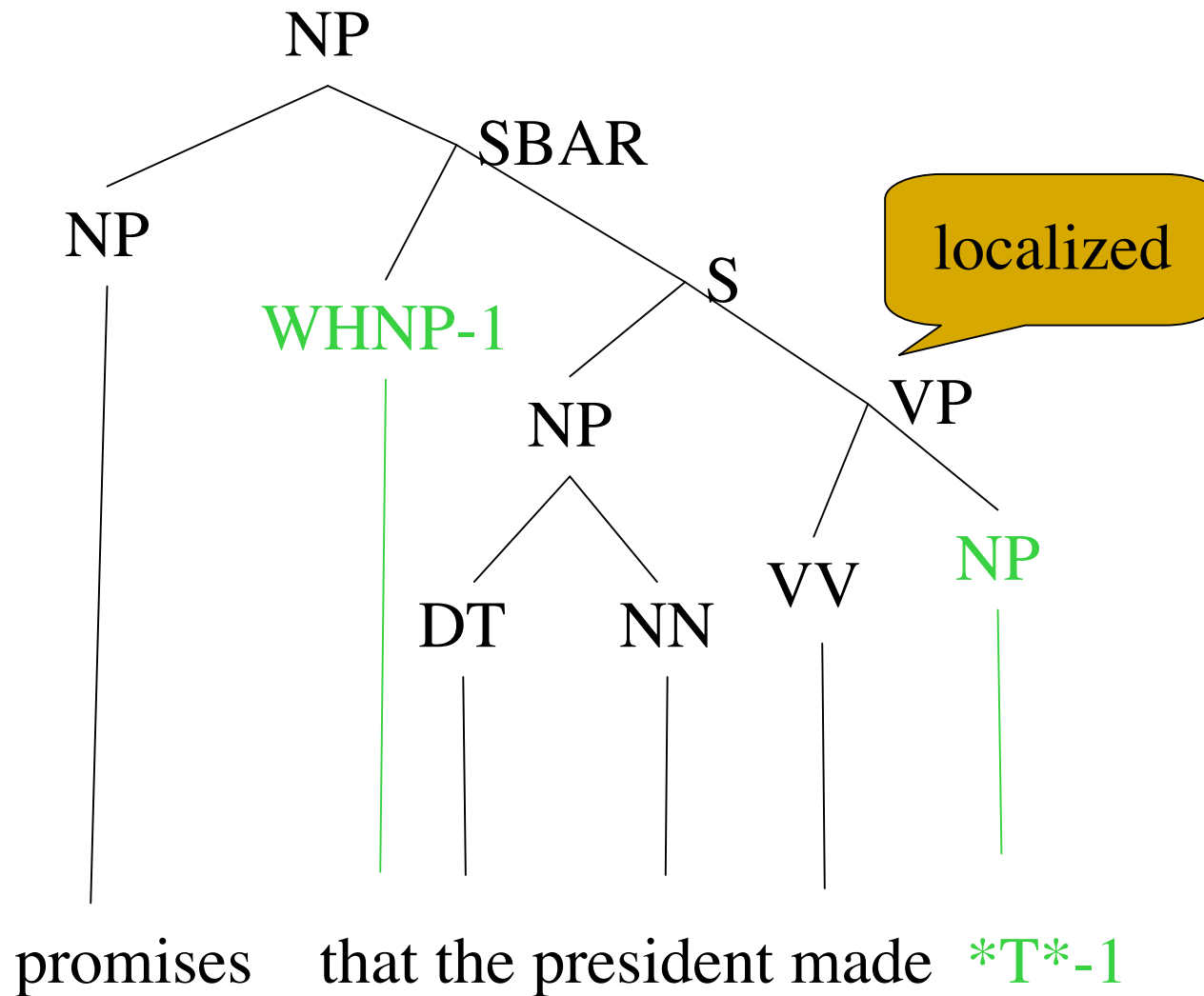
# Passivization



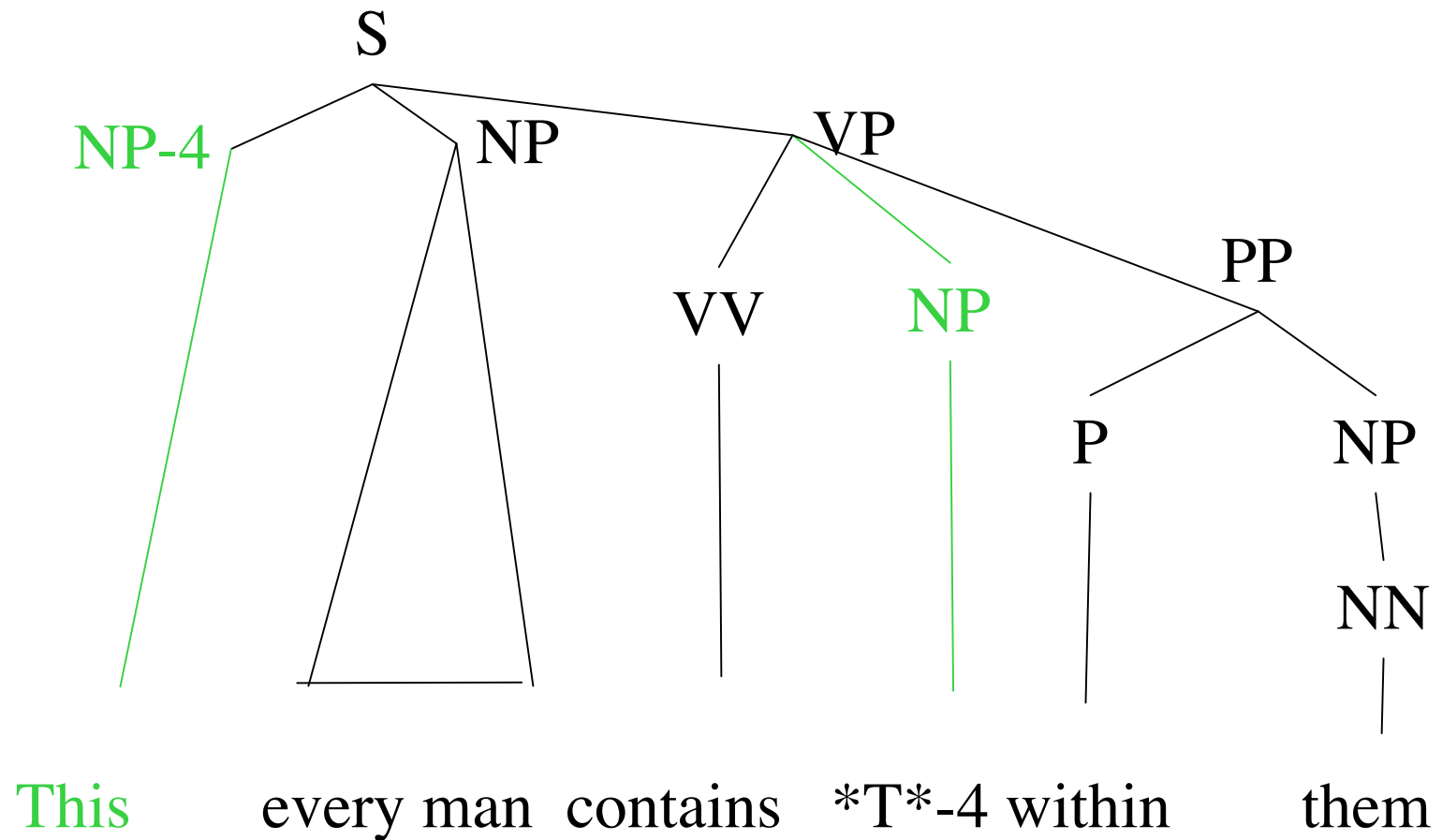
# WH-question



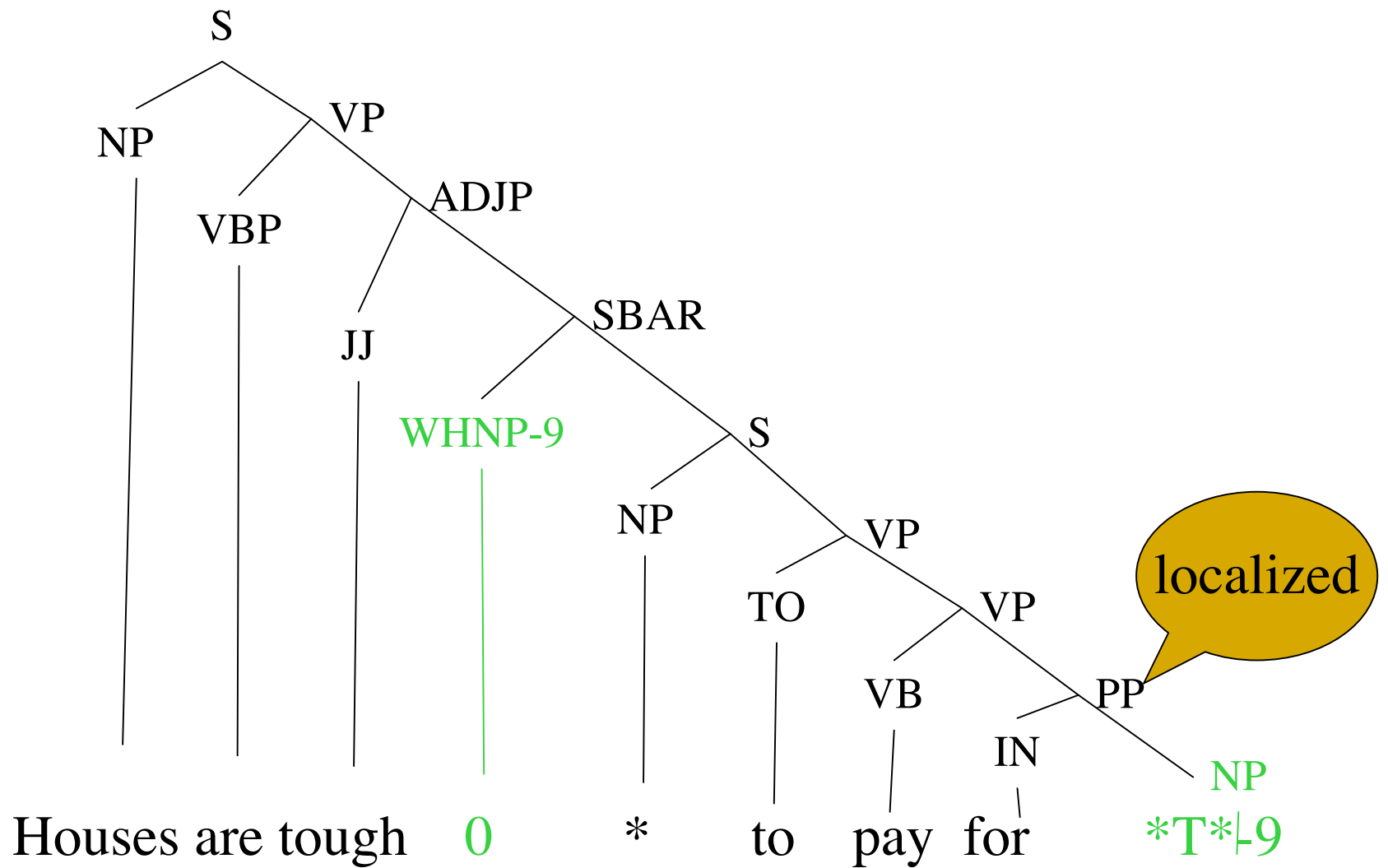
# Relative clause



# Topicalization



# Tough movement



---

# Structural ambiguity

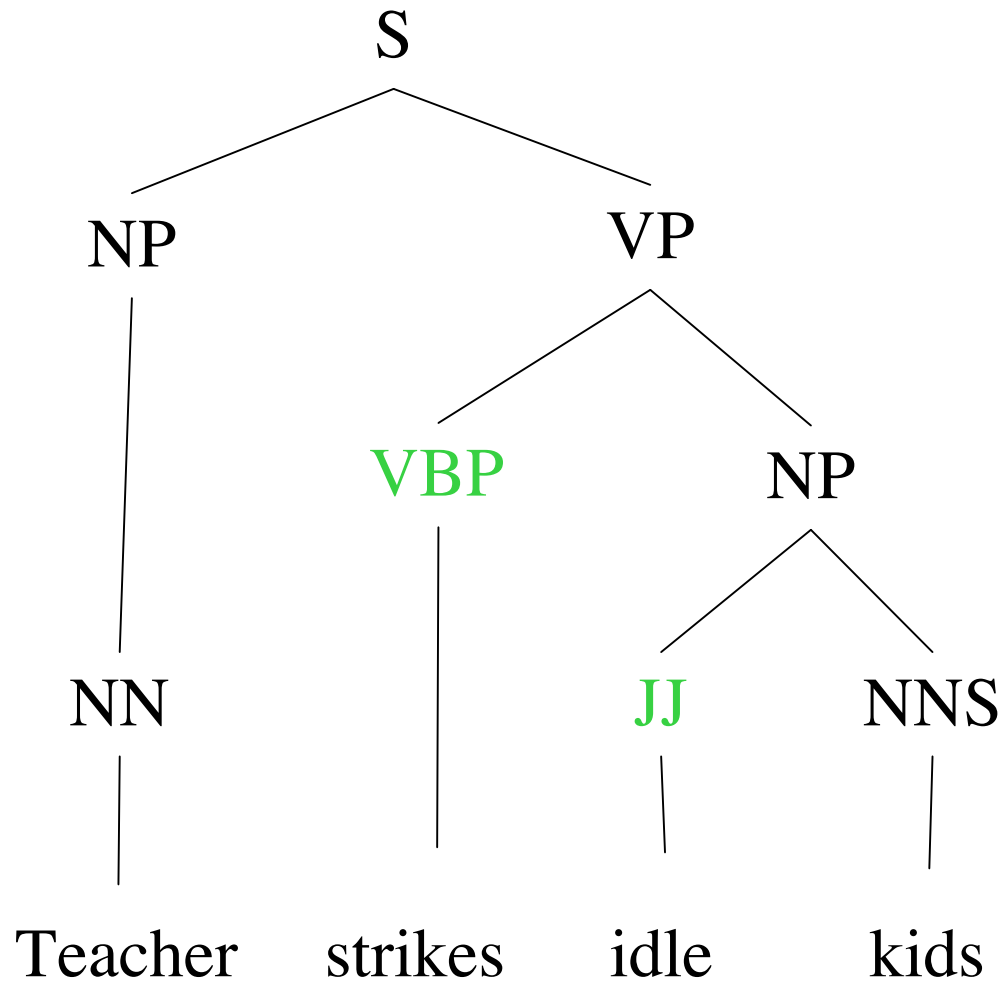
- Wanted: Man to take care of cow that does not smoke or drink.
- Question: how do we represent the interpretation that non-smoking and non-drinking man is sought to take care of cow?

---

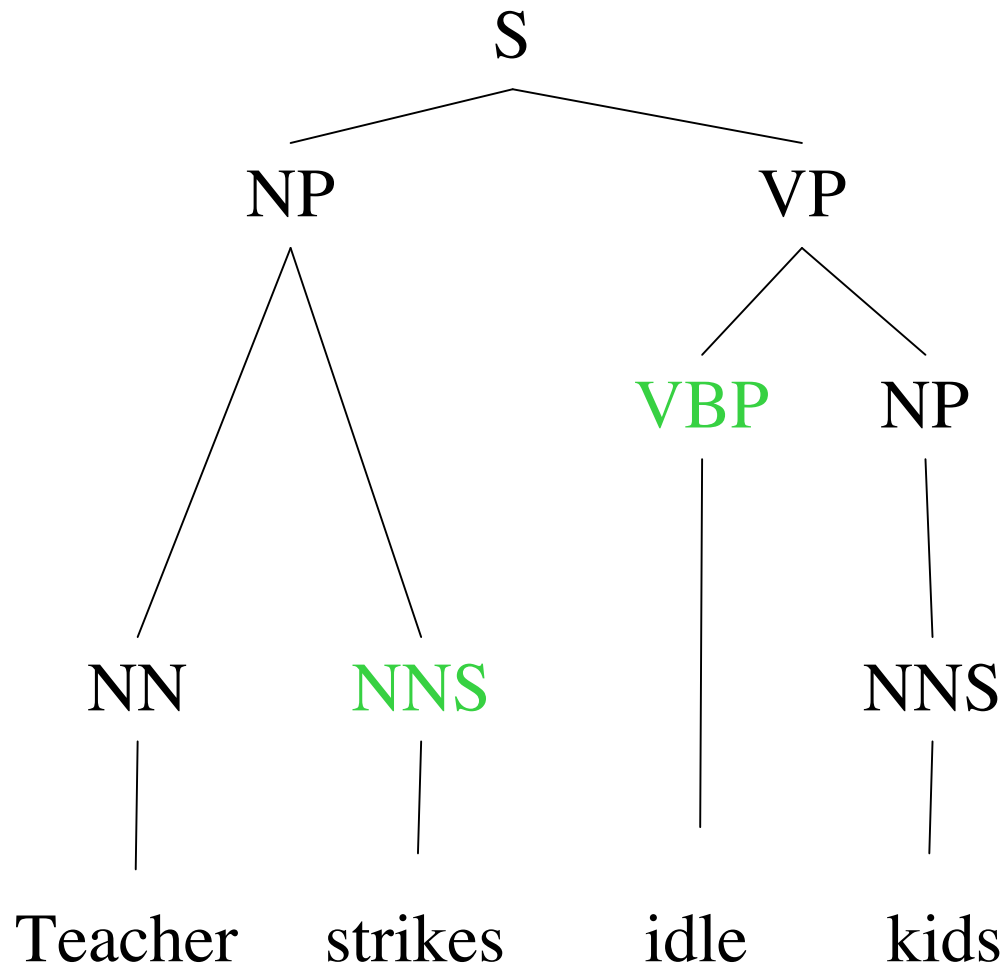
## Some examples

- Enraged cow injures farmer with ax
- Teacher Strikes Idle Kids
- Teller Stuns Man with Stolen Check

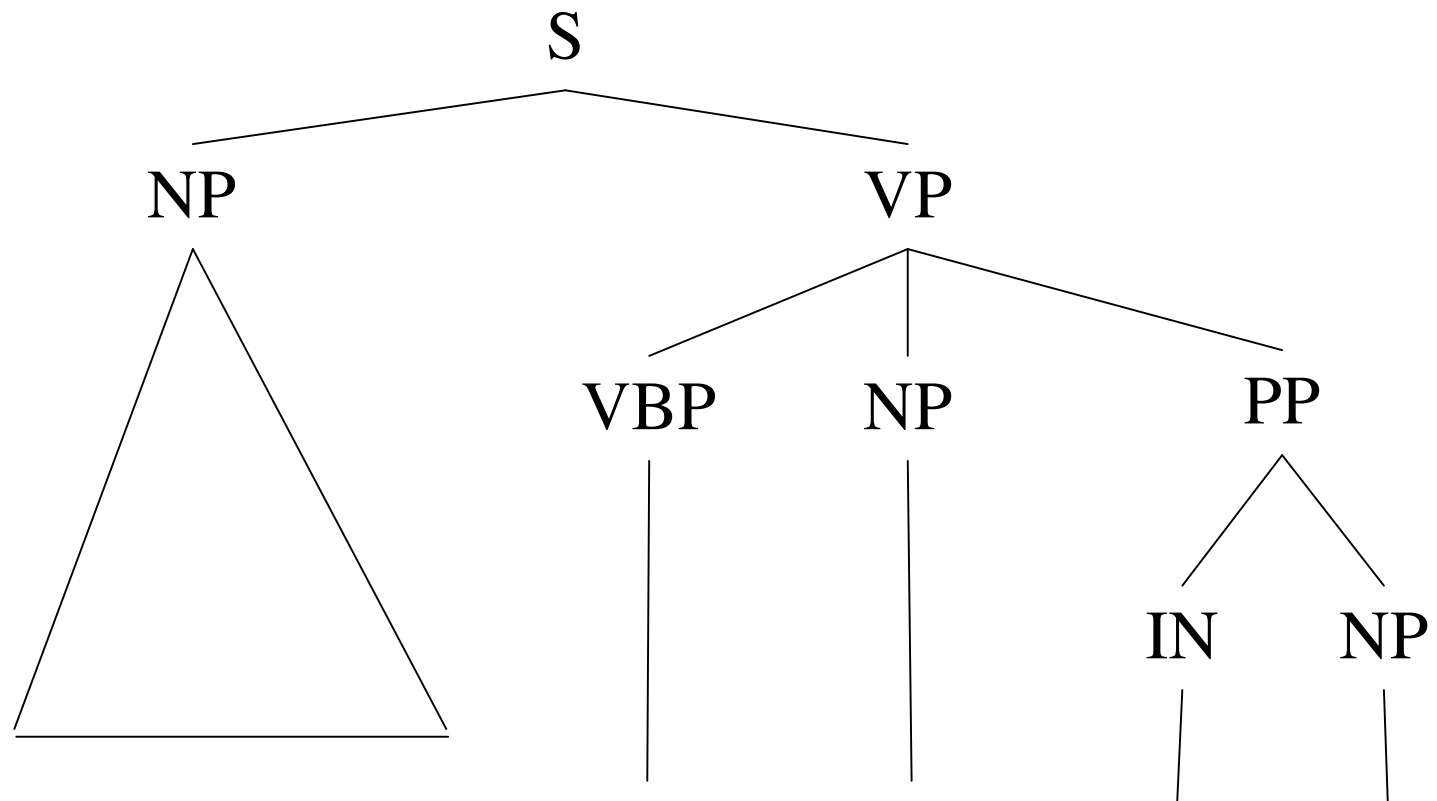
# Teacher strikes idle kids



# Teacher strikes idle kids

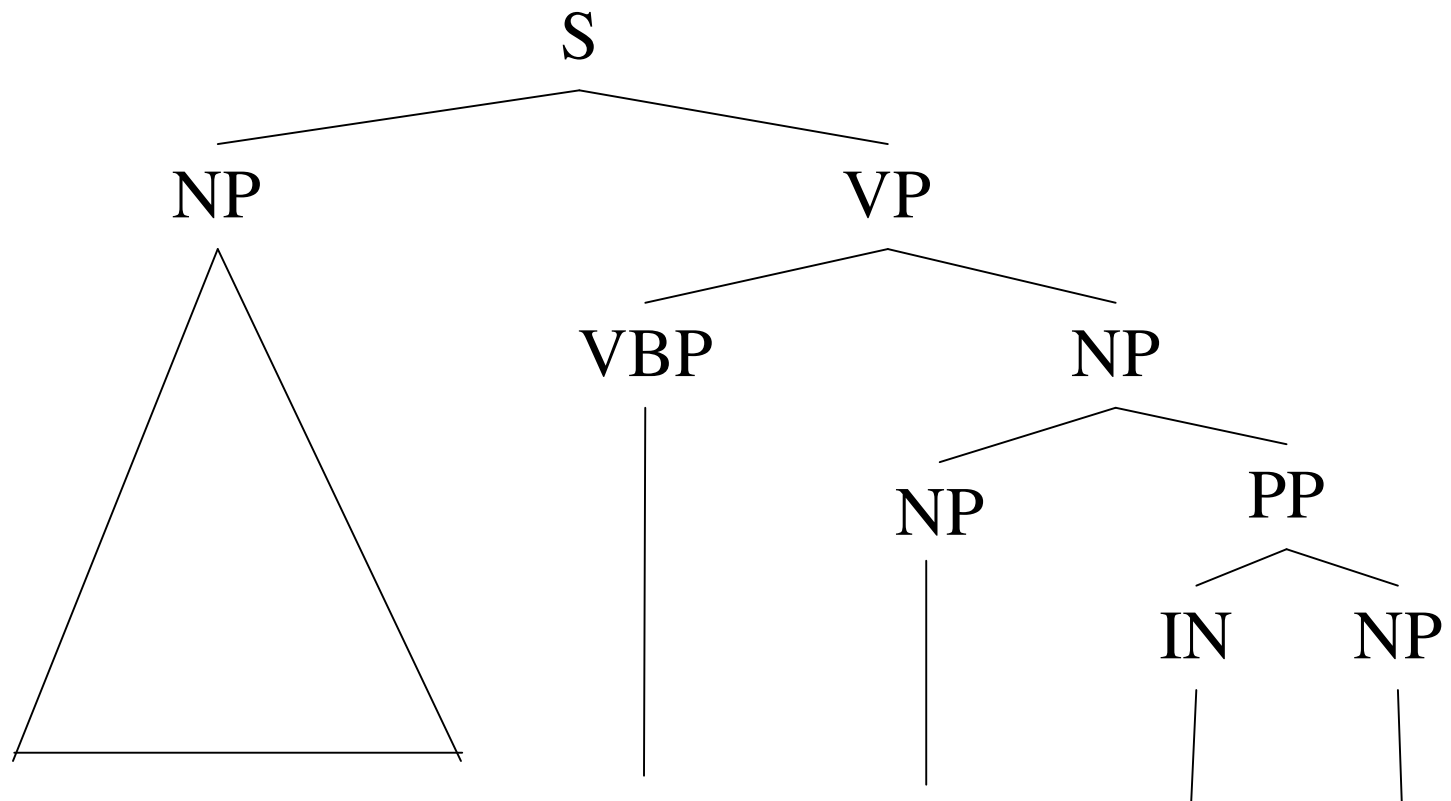


# Enraged cow injures farmer with ax



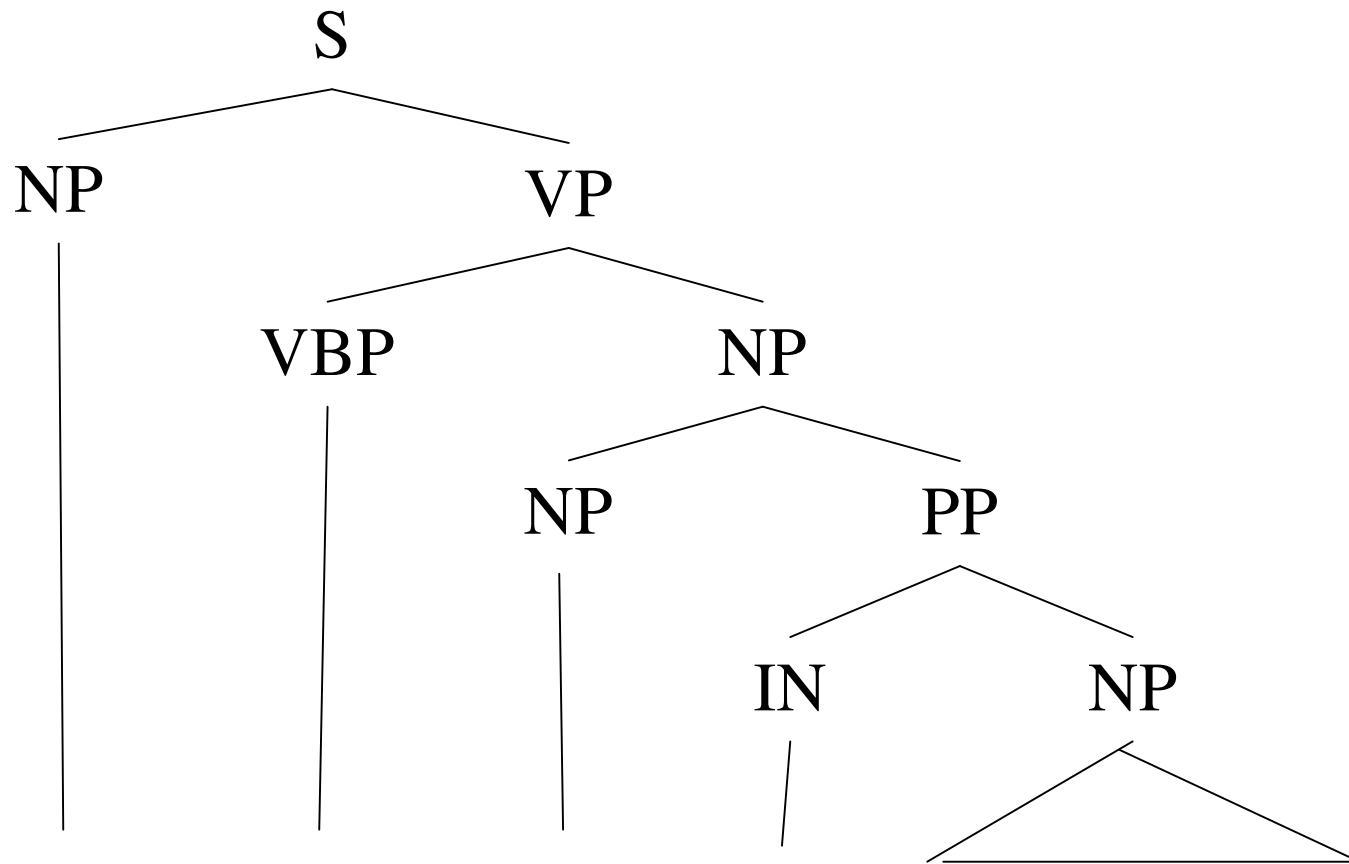
Enraged cow injures farmer with ax

# Enraged cow injures farmer with ax



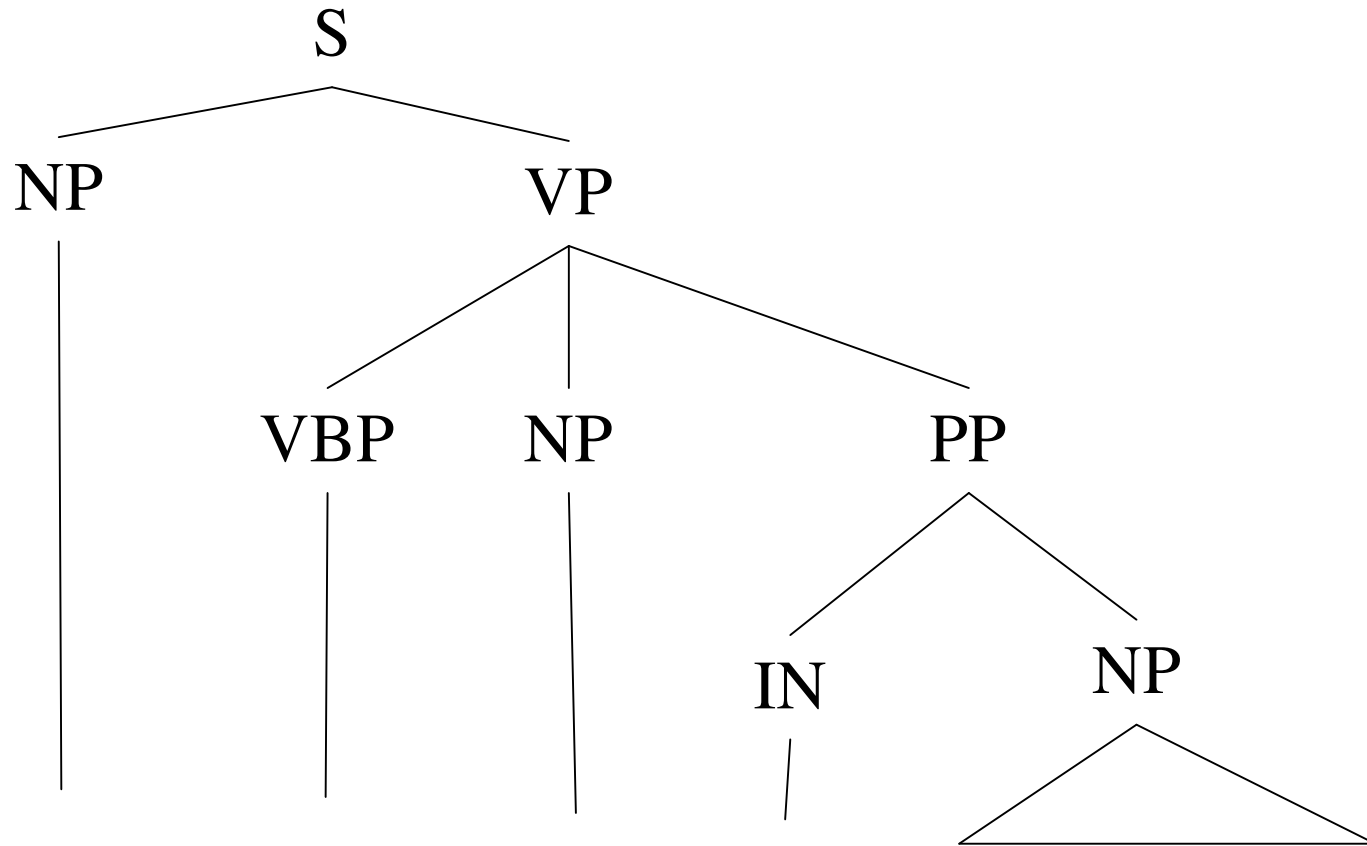
Enraged cow injures farmer with ax

# Teller Stuns Man with Stolen Check



Teller Stuns Man with Stolen Check

# Teller Stuns Man with Stolen Check



Teller Stuns Man with Stolen Check

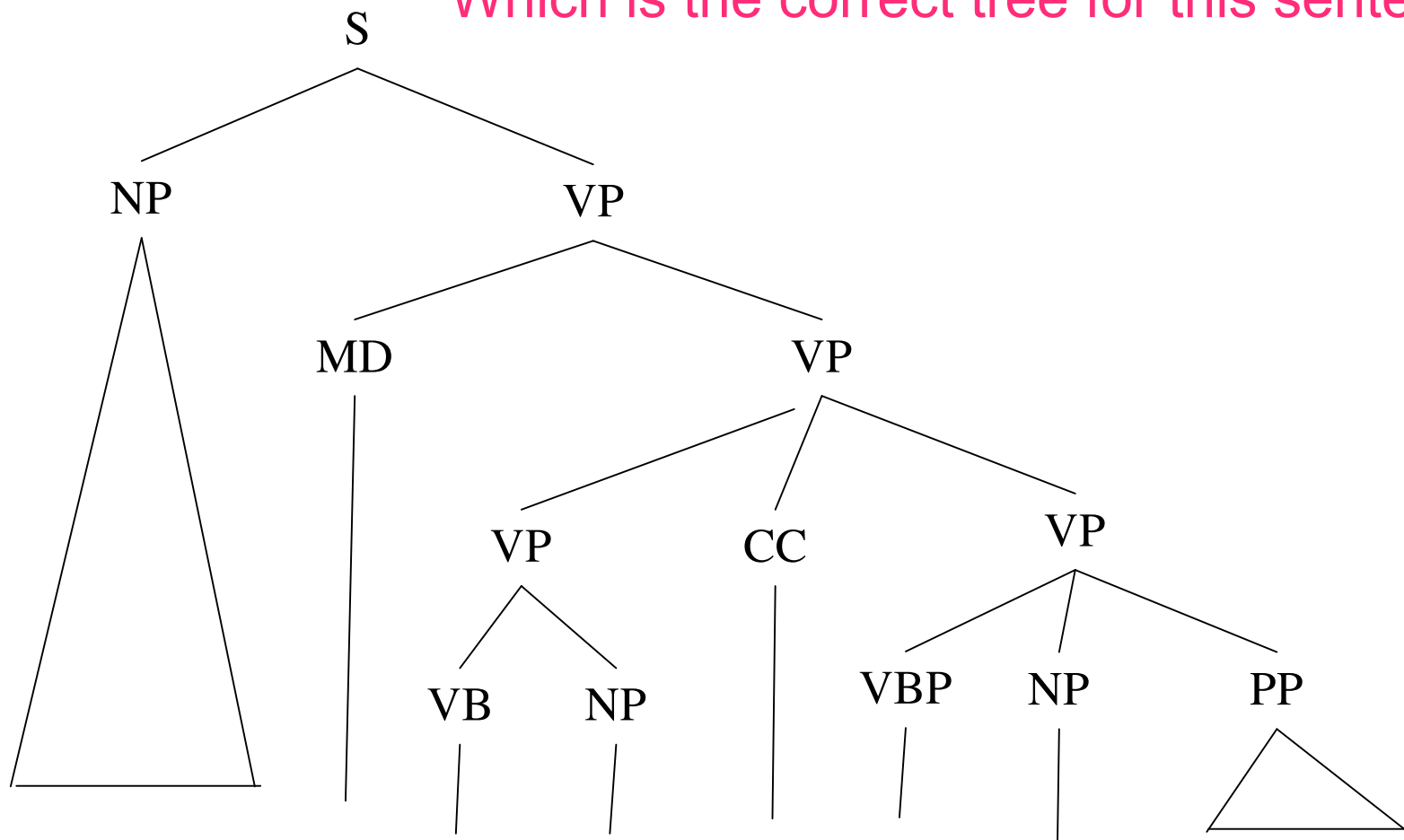
---

# Categorizing ambiguity

- Part of speech
- Conjunction
- Subordinate clause: higher or lower verbs?
- Prepositional phrase: verb or noun?

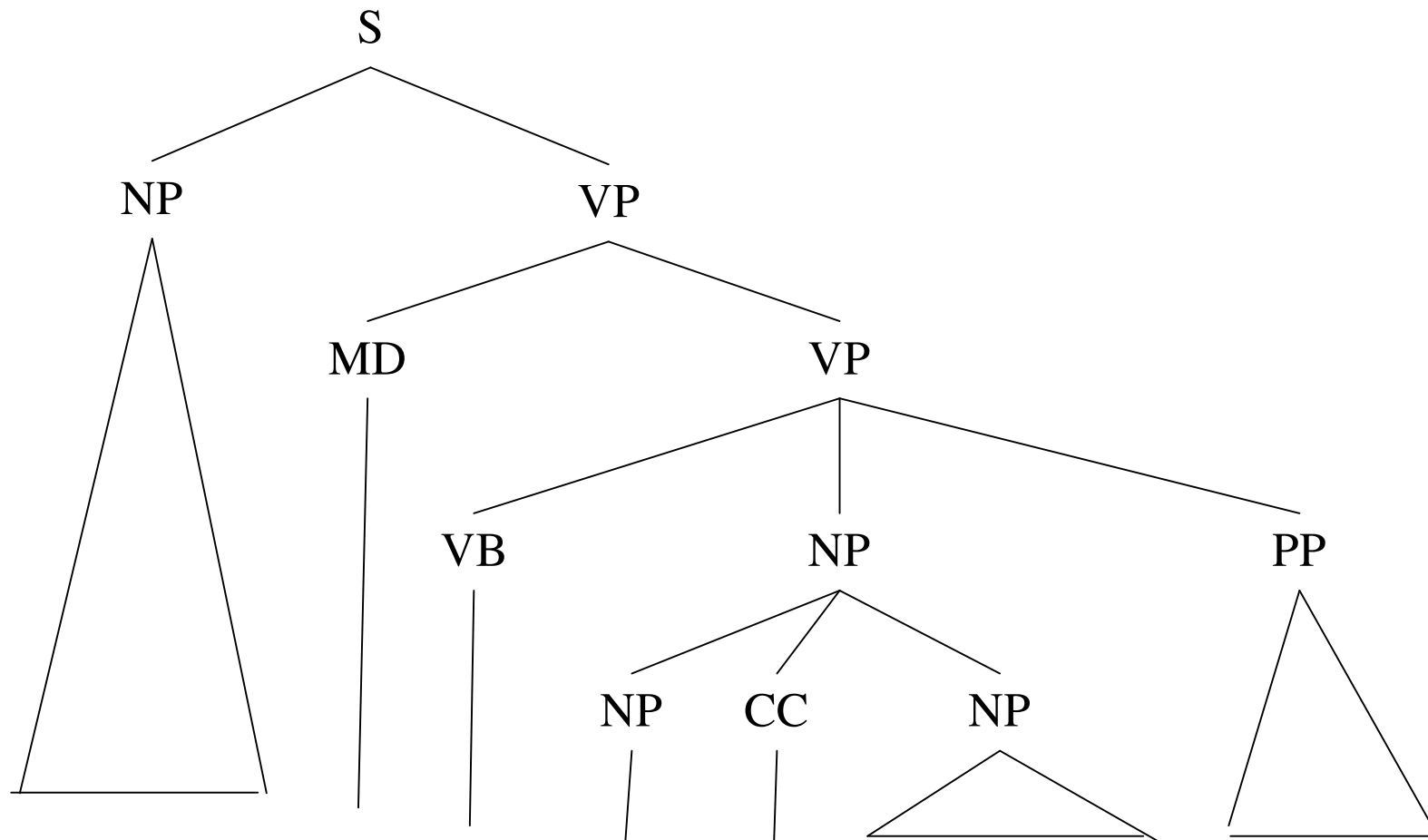
# Computer ambiguity v.s. human ambiguity

Which is the correct tree for this sentence?



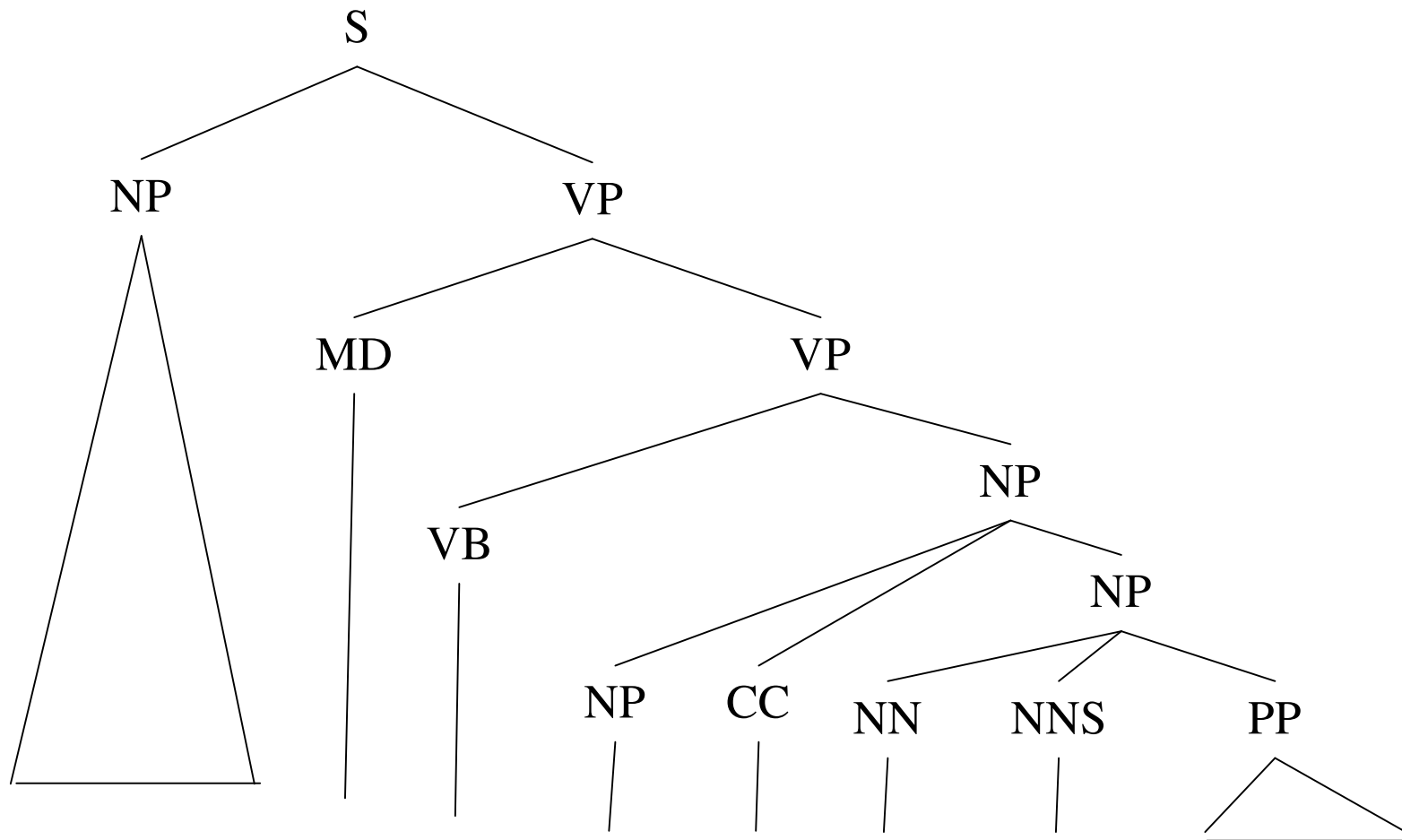
The post office will hold out discounts and service concessions as incentives

# Computer ambiguity v.s. human ambiguity



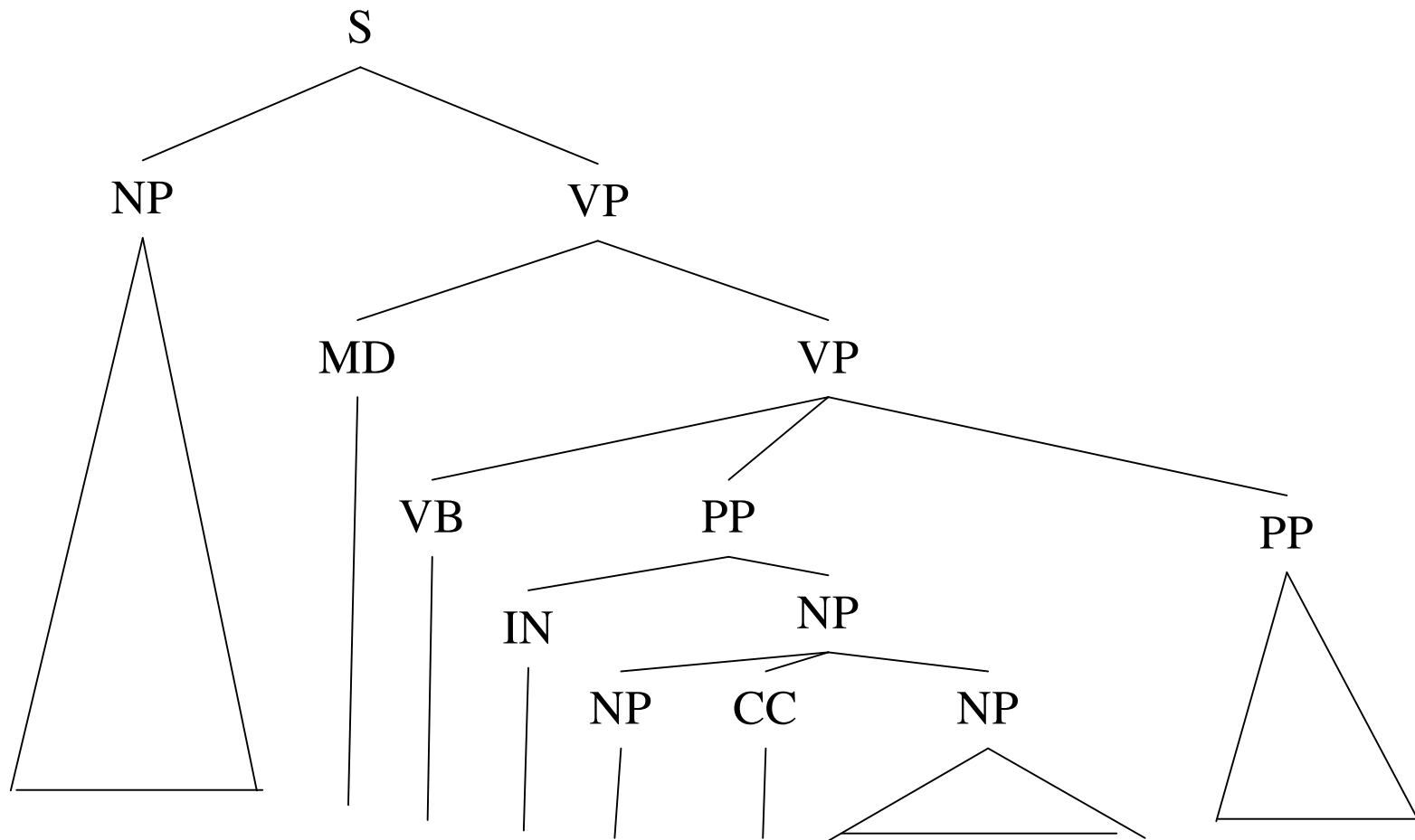
The post office will hold out discounts and service concessions as incentives

# Computer ambiguity v.s. human ambiguity



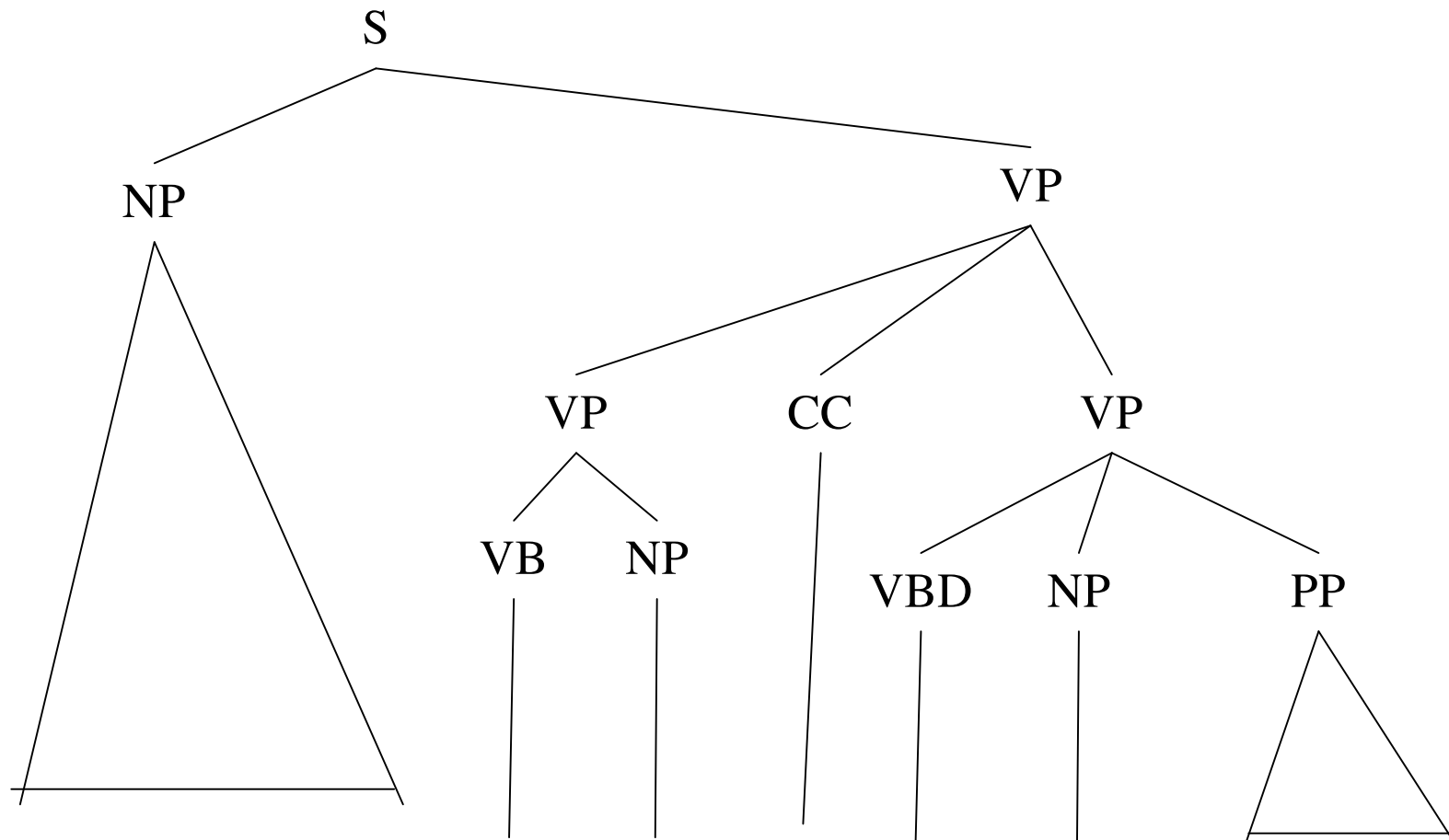
The post office will hold out discounts and service concessions as incentives

# Computer ambiguity v.s. human ambiguity



The post office will hold out discounts and service concessions as incentives

# Computer ambiguity v.s. human ambiguity



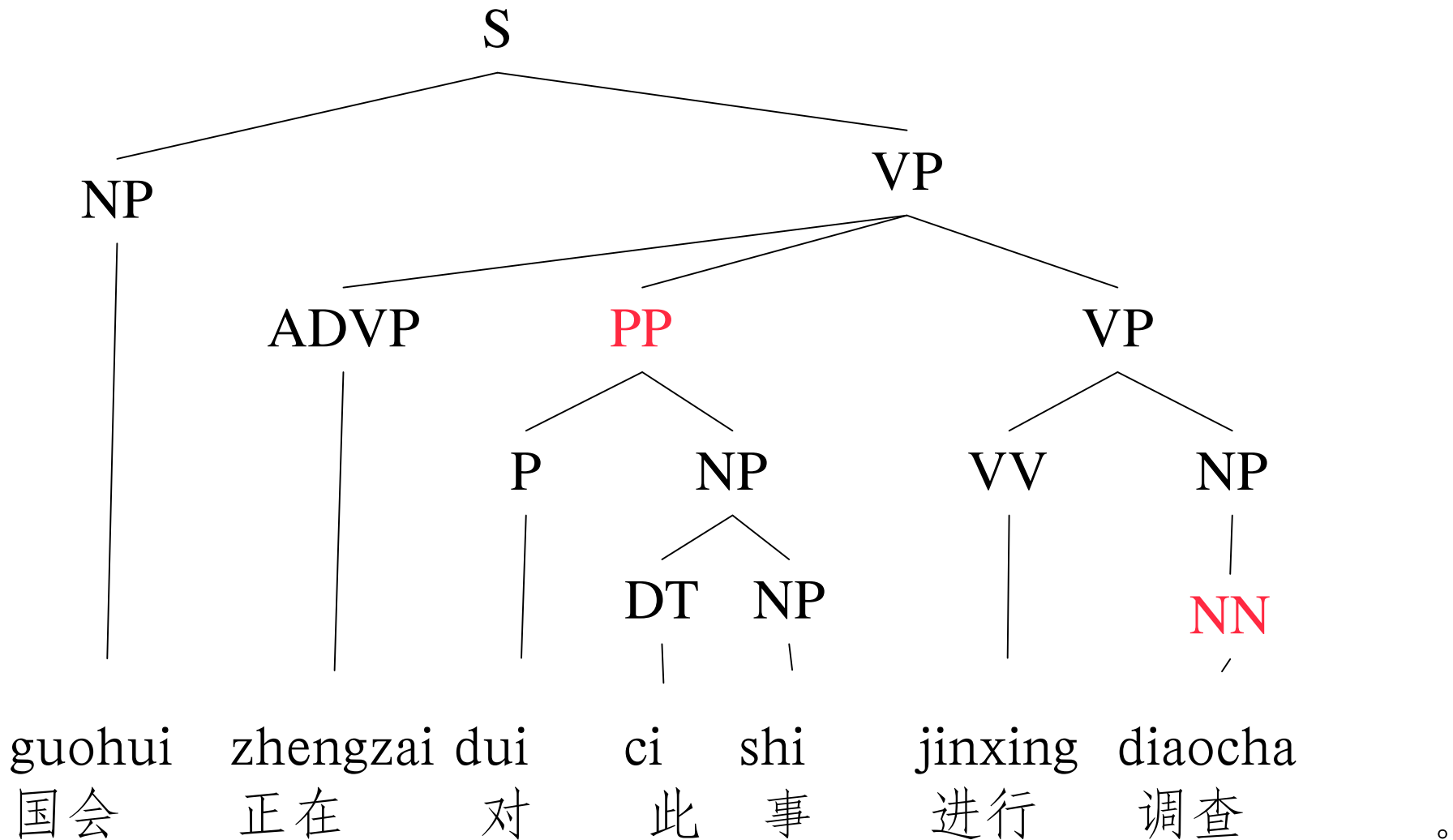
The post office will hold out discounts and service concessions as incentives

---

## Discussion question

- Can all semantic dependencies be represented with attachment at the appropriate level?

# Can all ambiguities be represented with a tree?



~~Congress presently toward this matter conduct investigation .~~

LING 51209, 2008  
“The Congress is investigating this matter.”

# Building a tree

```
>>> import nltk, re
>>> np1 = nltk.Tree('NP', ['teacher'])
>>> np2 = nltk.Tree('NP', ['idle', 'kids'])
>>> vp = nltk.Tree('VP', ['strikes', np2])
>>> s = nltk.Tree('S', [np1, vp])
```

# Using treebank data

```
>>> tree1 =  
    nltk.corpus.treebank.parsed_sents('wsj_00  
    01.mrg')[0]  
  
>>> print tree1[0]  
  
>>> print tree1[1]  
  
>>> tree1.node
```

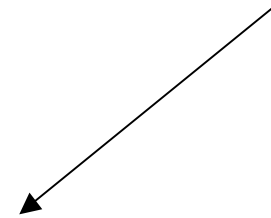
# Recursive function

```
>>> def traverse_tree(tree):  
    for child in tree:  
        if child.height() == 2:  
            print child.node  
        else:  
            print child.node  
            traverse_tree(child)
```

# Recursive function

```
>>> def traverse_tree(tree):  
    for child in tree:  
        if child.height() == 2:  
            print child.node  
        else:  
            print child.node  
            traverse_tree(child)
```

*Ending condition*



# Recursive function

```
>>> def get_subject(tree):  
    for child in tree:  
        if child.height() == 2:  
            return  
        elif child.node.endswith('SBJ'):  
            print child.leaves()  
        else:  
            get_subject(child)
```

# Hw4

```
>>> infile = open('hw4.txt', 'r')
>>> s_seg =
    nltk.data.load('tokenizers/punkt/english.pickle')
>>> lines = infile.readlines()
>>> tlines = ""
>>> for line in lines:
    tlines = tlines + ' ' + line
>>> newtlines = tlines.replace('\r\r', ' ')
>>> sents = s_seg.tokenize(newtlines)
```

## Hw 4 (cont'd)

```
>>> d = {}
```

```
>>> pat = r'\d+\.|\d+|\d+|\w+|\-\w+|\b\w+\b'
```

```
>>> for sent in sents:
```

```
    tokens = nltk.tokenize.regexp_tokenize(sent, pat)
```

```
    stems = [stemmer.stem(tok.lower()) for tok in tokens]
```

```
    for stem in stems:
```

```
        if stem in d:
```

```
            d[stem] += 1
```

```
        else:
```

```
            d[stem] = 1
```

---

## Hw4 (cont'd)

```
>>> items = [(v,k) for k, v in d.items()]
```

```
>>> items.sort()
```

```
>>> items.reverse()
```

```
>>> for v,k in items:  
    print k, v
```