

# Constituent structure and structural ambiguity

Feb. 16, 2006

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

# Some words belong together

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

# Do we have evidence for that?

- Did **the dog** chase **the cat**?
- Did **she** chase **him**?
  
- Did **the dog** **the children** like chase the cat?
- Did **the dog** **they** like chase the cat?
- Did **she** **they** like chase the cat?

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



# Pro-clause?

- I { know, suspect } **that they're invited.**  
I { know, suspect } **it.**
- I { imagine, think } **that they're invited.**  
I { imagine, think } **so.**

# 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  
\_\_\_\_\_.

- Adverb phrase

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

It was more quickly than they had expected that they arrived at



# Ungrammatical it cleft

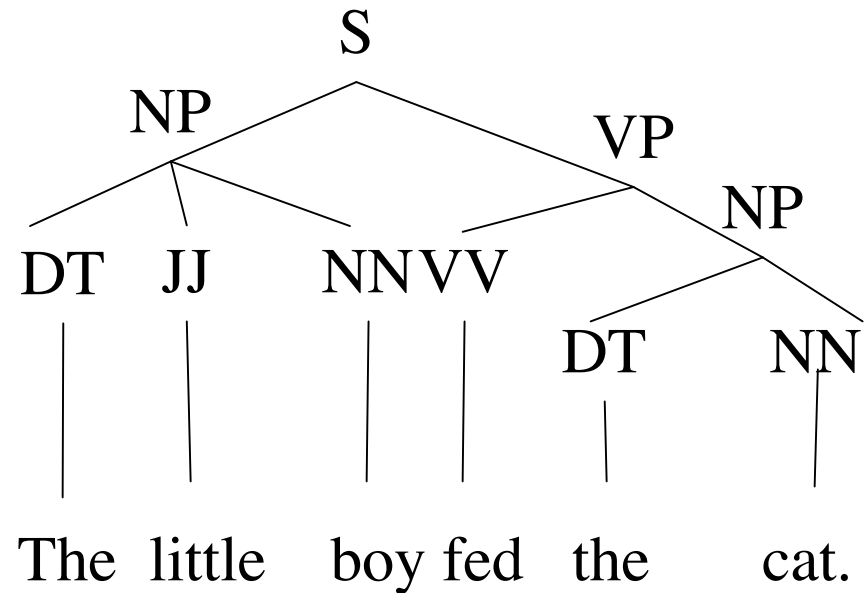
- Ordinary cats detest the smell of citrus fruits.  
It is the smell of that ordinary cats detest \_\_\_\_ citrus fruits.
- The cat strolled across the porch with a confident air.  
\* It was with a confident that the cat strolled across the porch \_\_\_\_ air.
- Ali Baba returned from his travels wiser than before.  
\* It was wiser than that Ali Baba returned from his travels \_\_\_\_ before.
- They arrived at the concert hall more quickly than they had expected.  
\* It was more quickly than that they arrived at the concert hall they has expected.

# Representing constituent structure

- 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: bracket bank?
- 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.

# Brackets v.s. tree

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



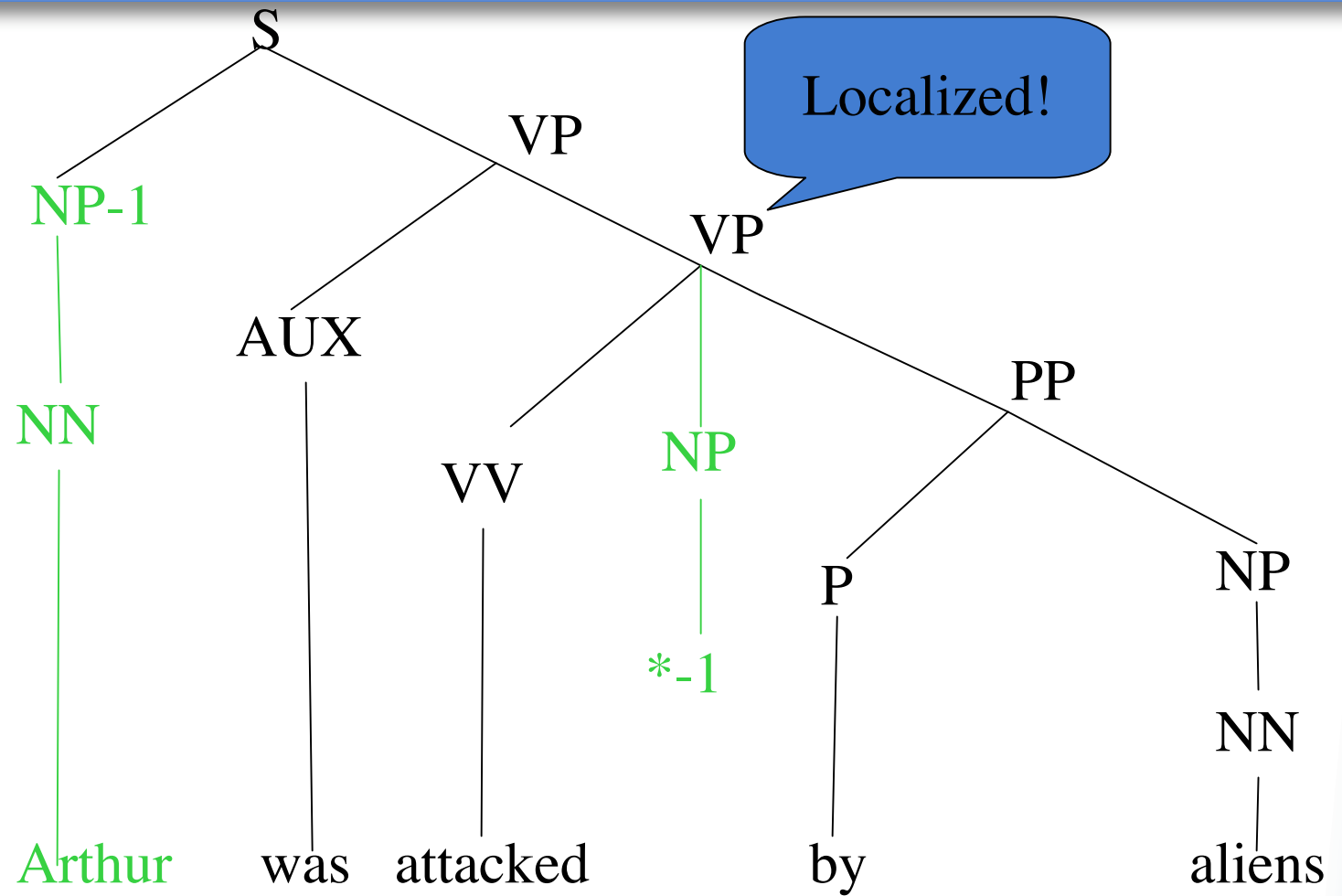
A constituent is exhaustively included in a pair of brackets.

A constituent is exhaustively dominated by a node.

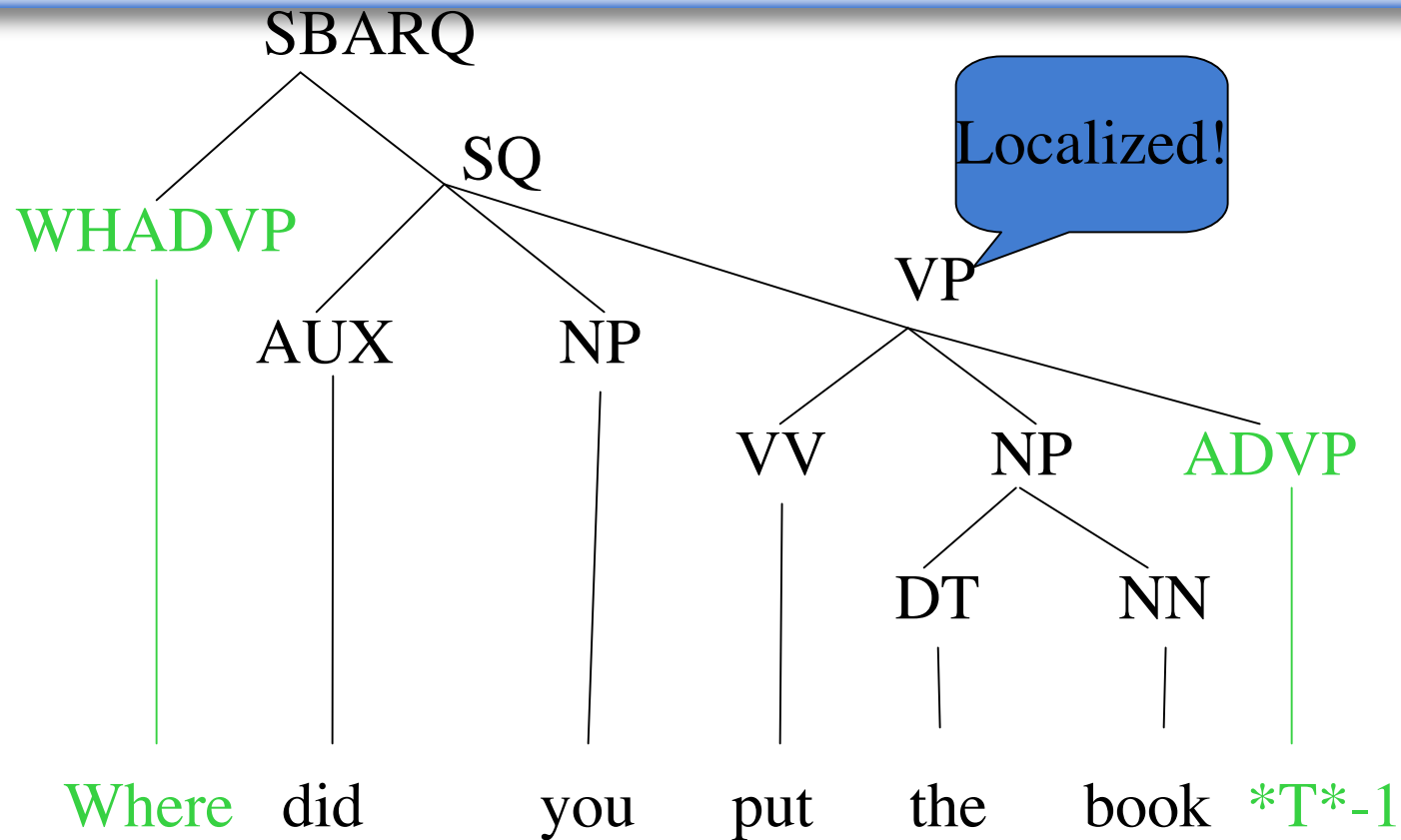
# Representing discontinuous constituents

- 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

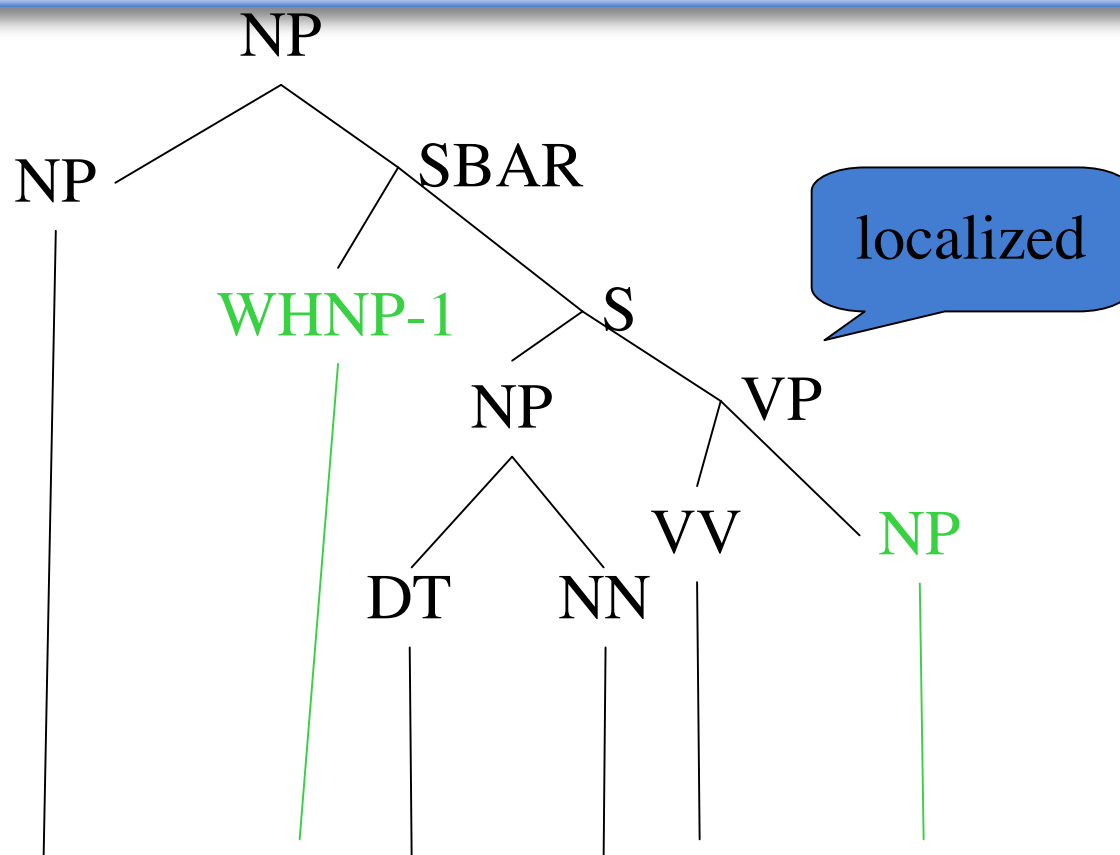
# Passivization



# WH-question

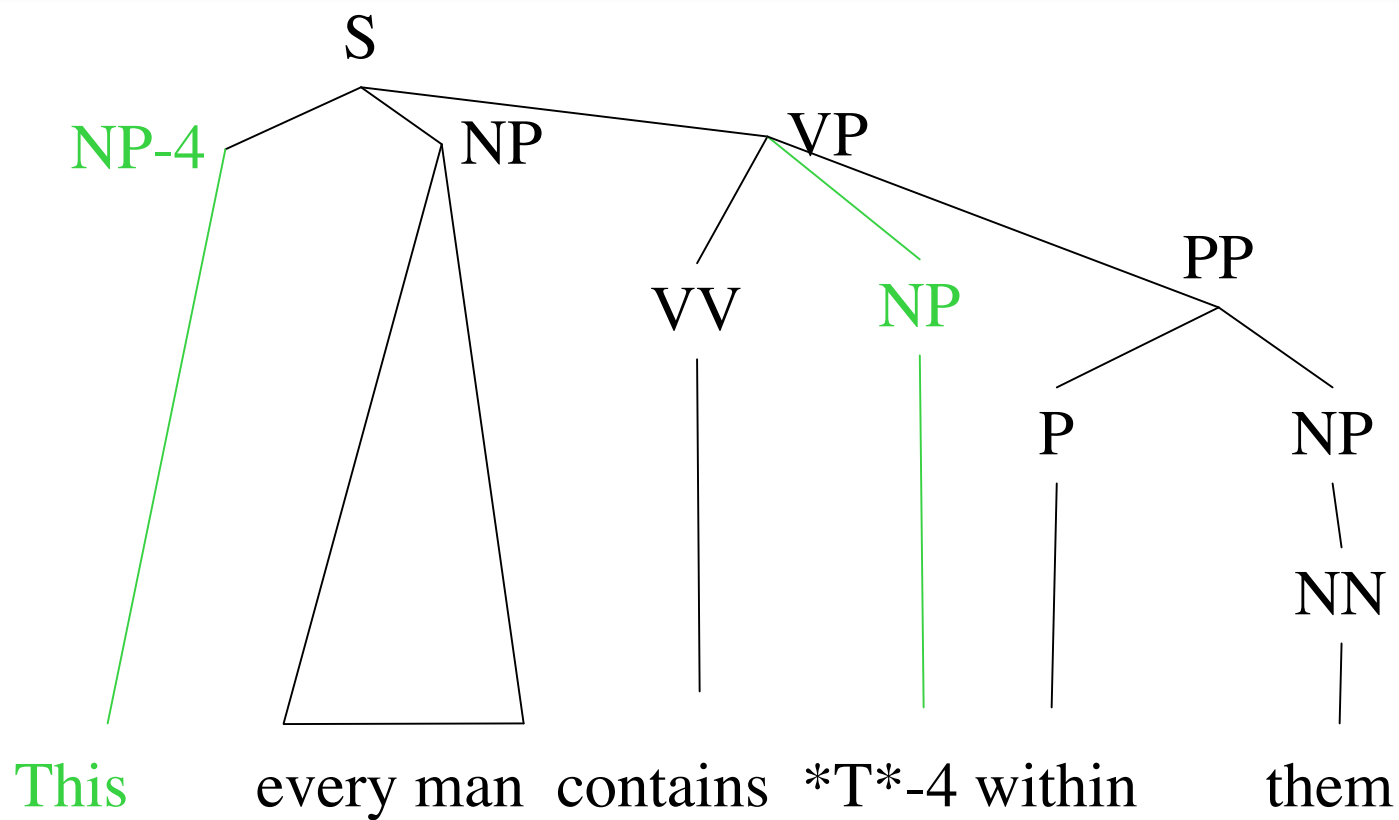


# Relative clause



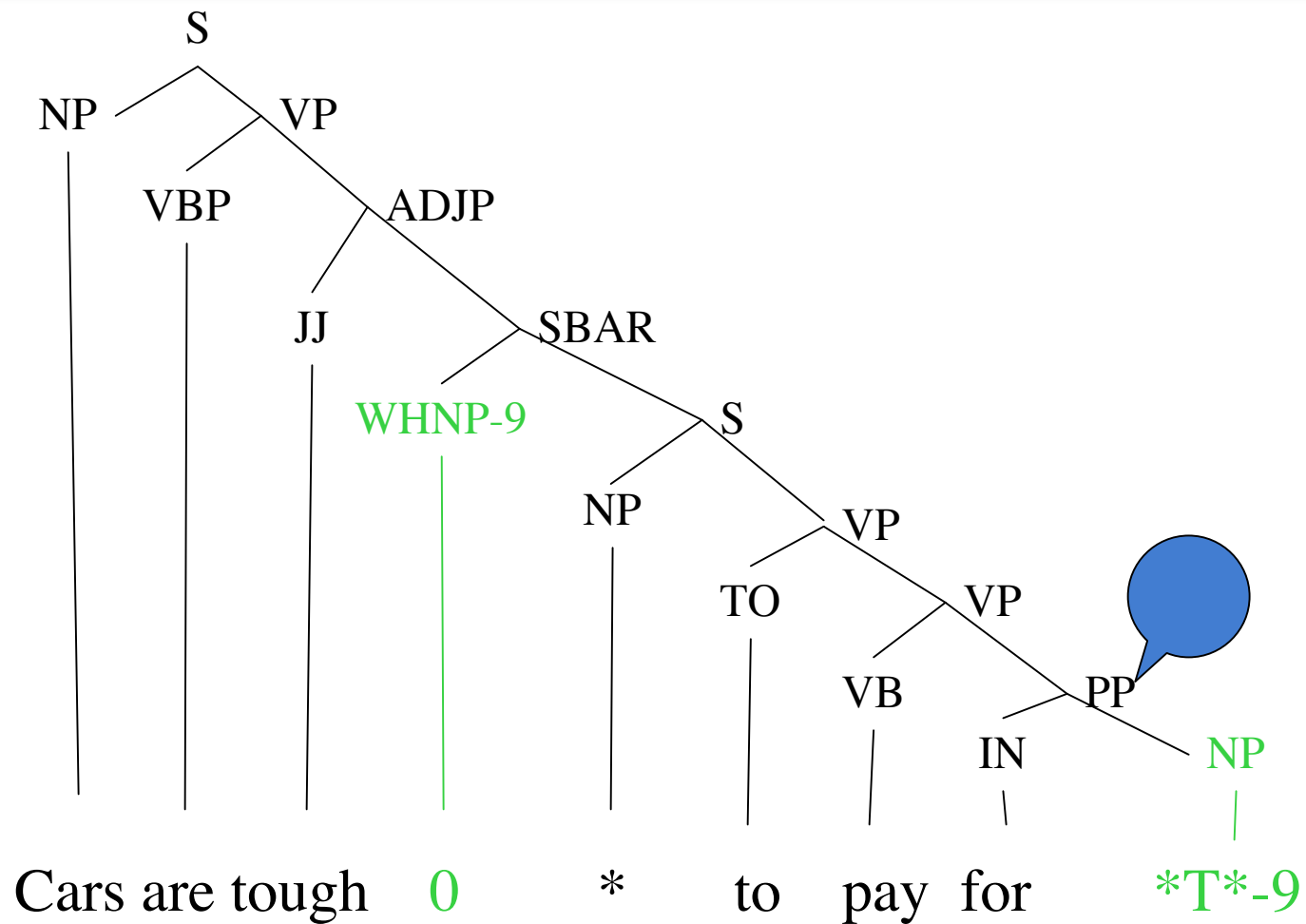
promises that the president made \*T\*-1

# Topicalization

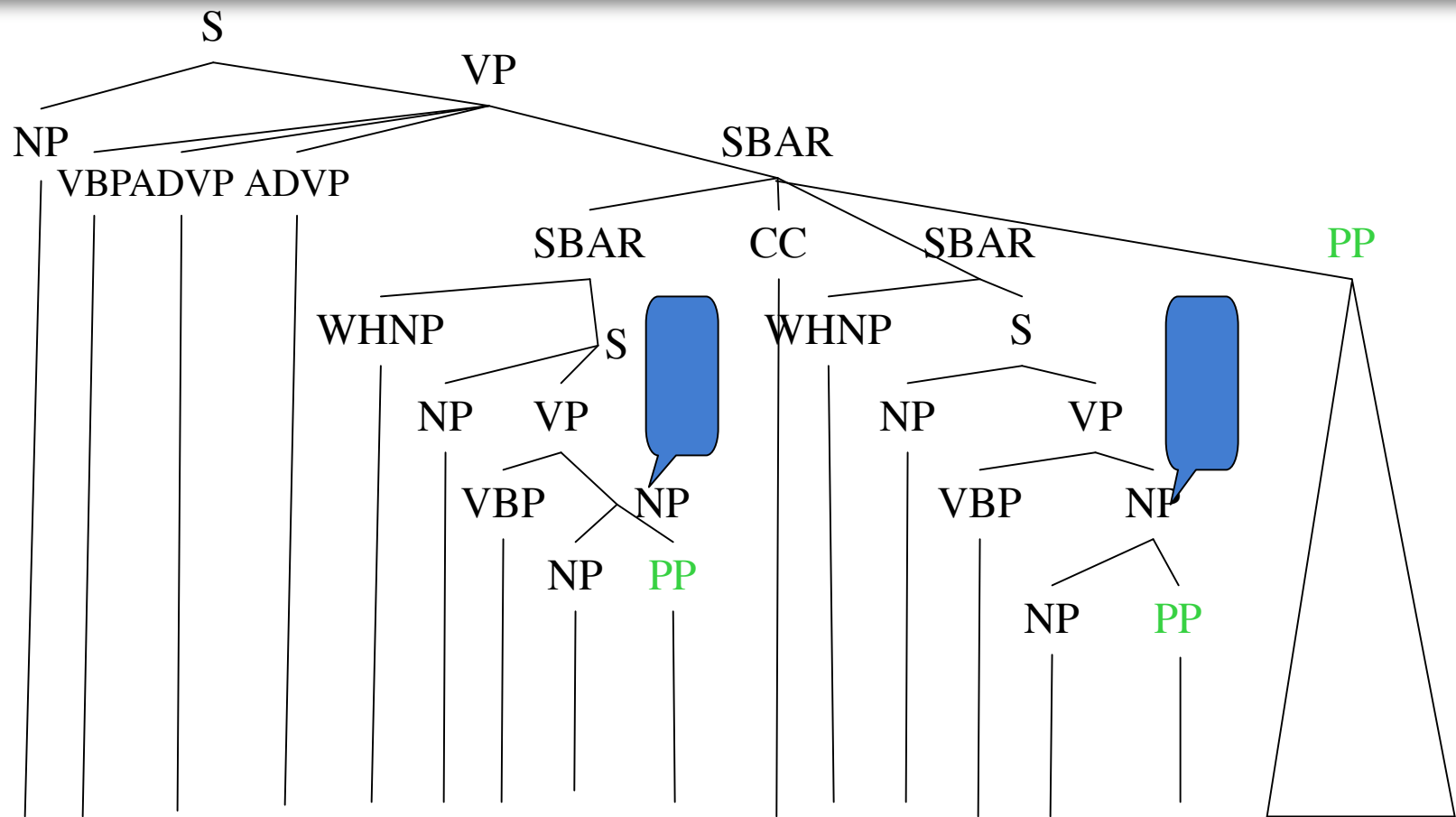




# Tough movement

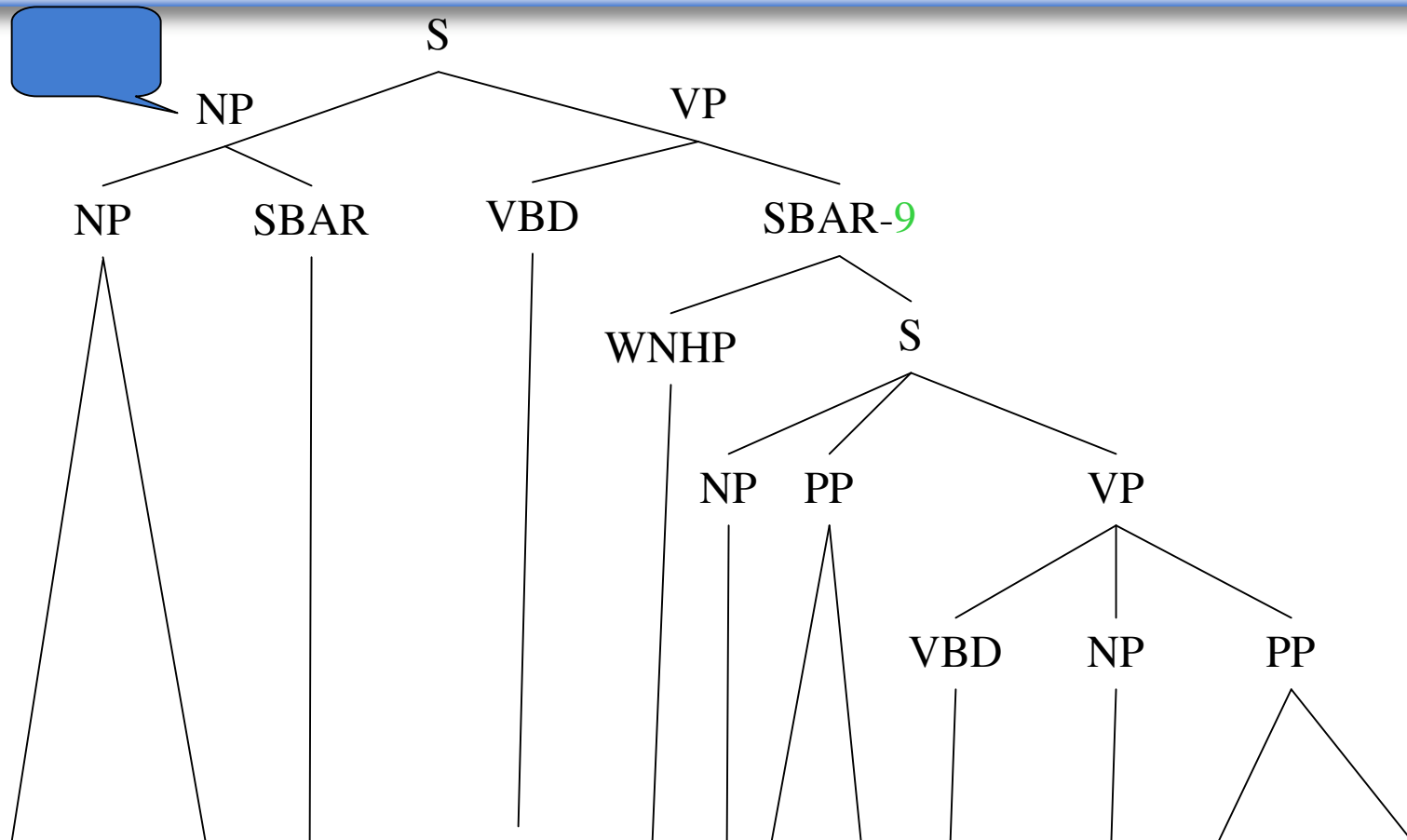


# Right-hand node raising



One knows better now who \*T\*-1 has bone \*RNR\*-5 and who \*T\*-2 has jelly \*RNR\*-5 in his spine.

# Interpret constituent here



A young woman \*ICH\*-9 entered whom she at once recognized \*T\*-2 as Jane Doe II.

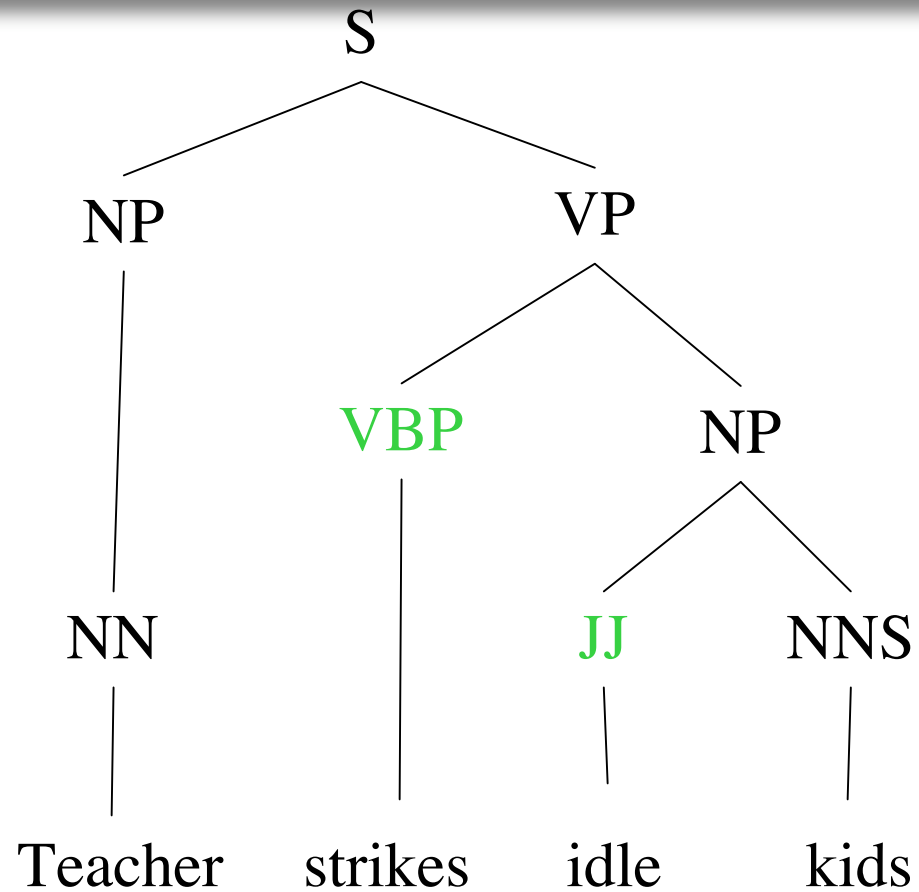
# Structural ambiguity

- 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?

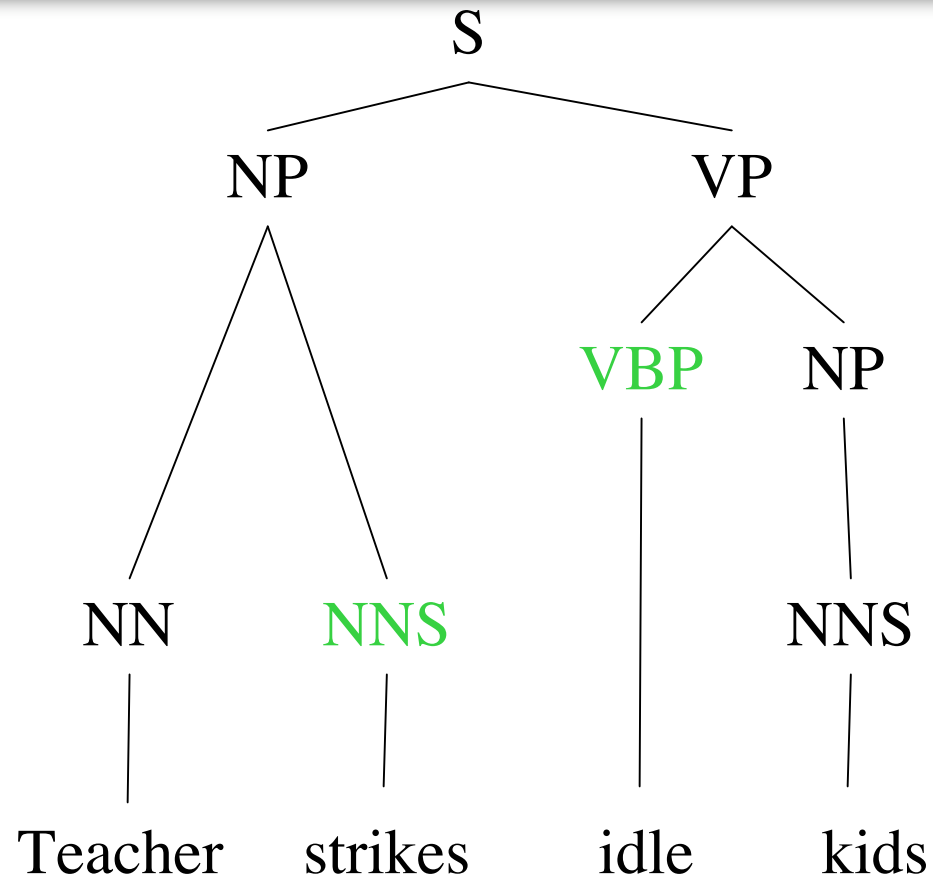
# Some examples

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

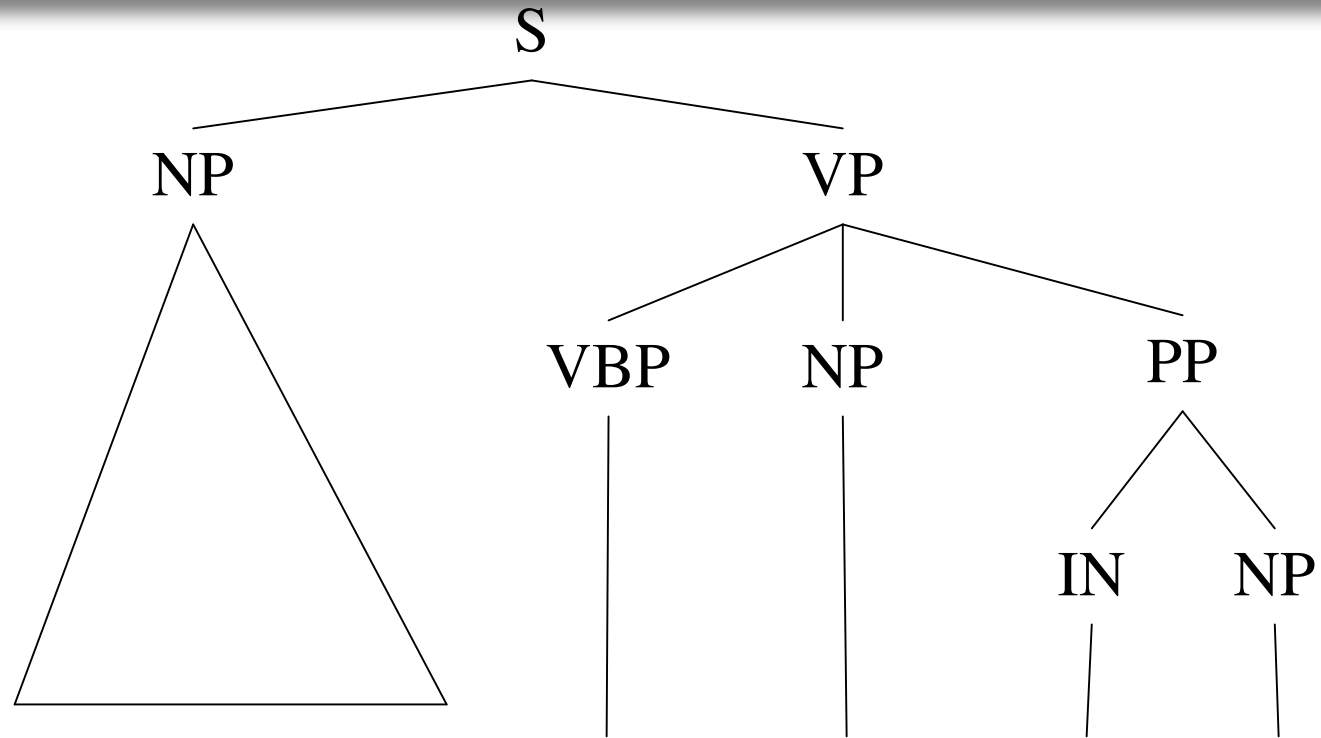
# Teacher strikes idle kids



# Teacher strikes idle kids



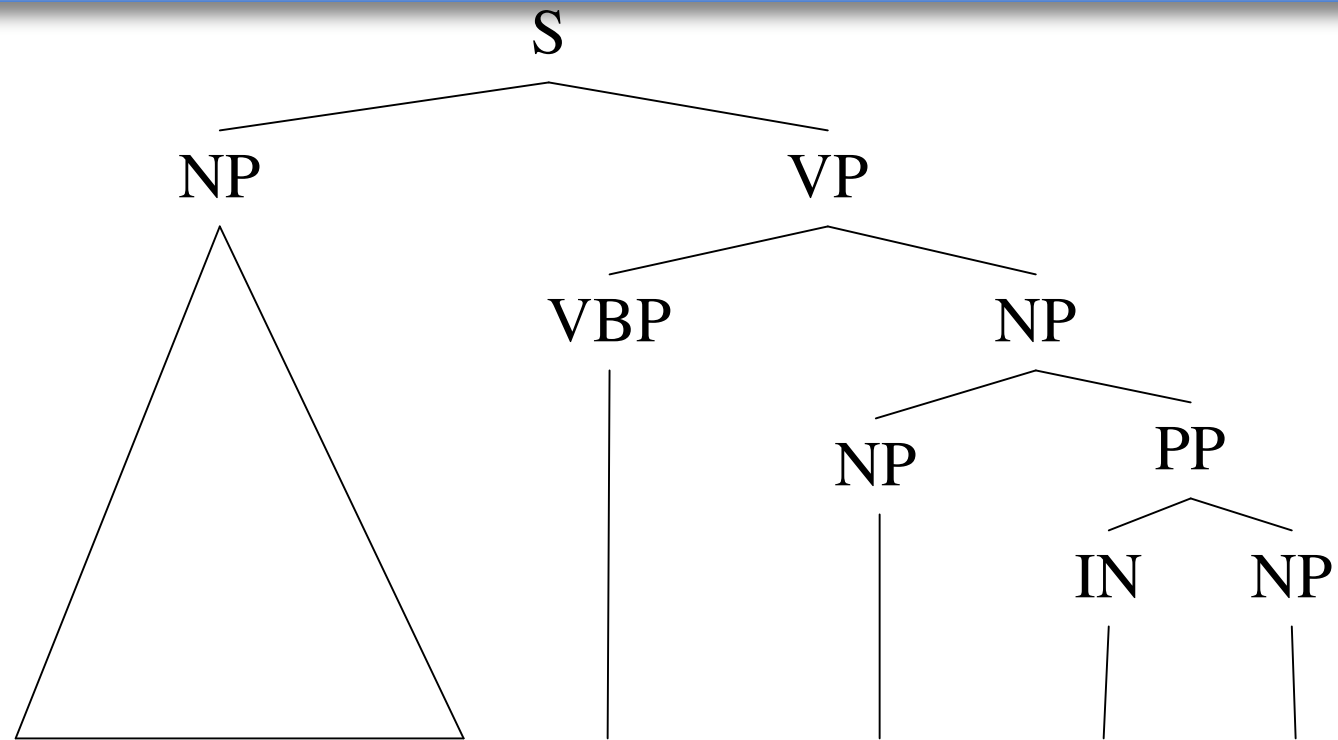
# Enraged cow injures farmer with ax



Enraged cow injures farmer with ax

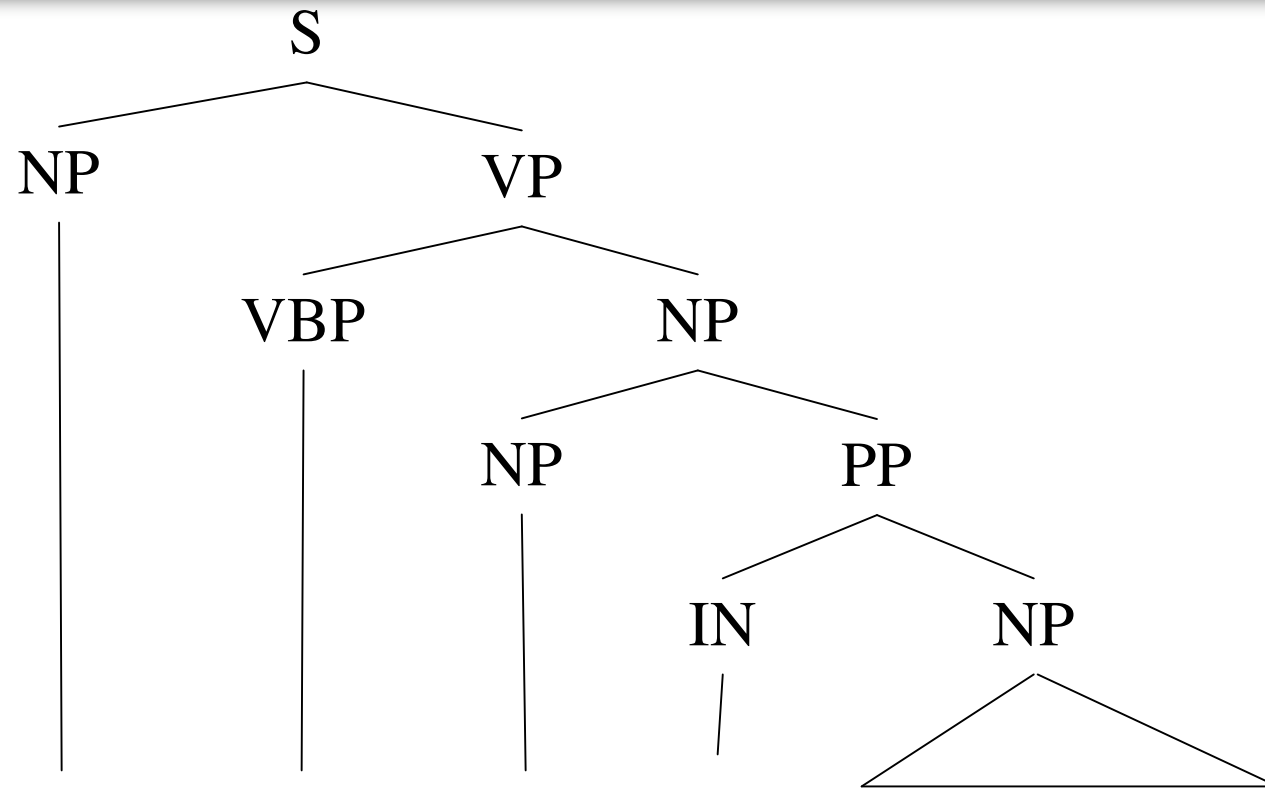


# Enraged cow injures farmer with ax



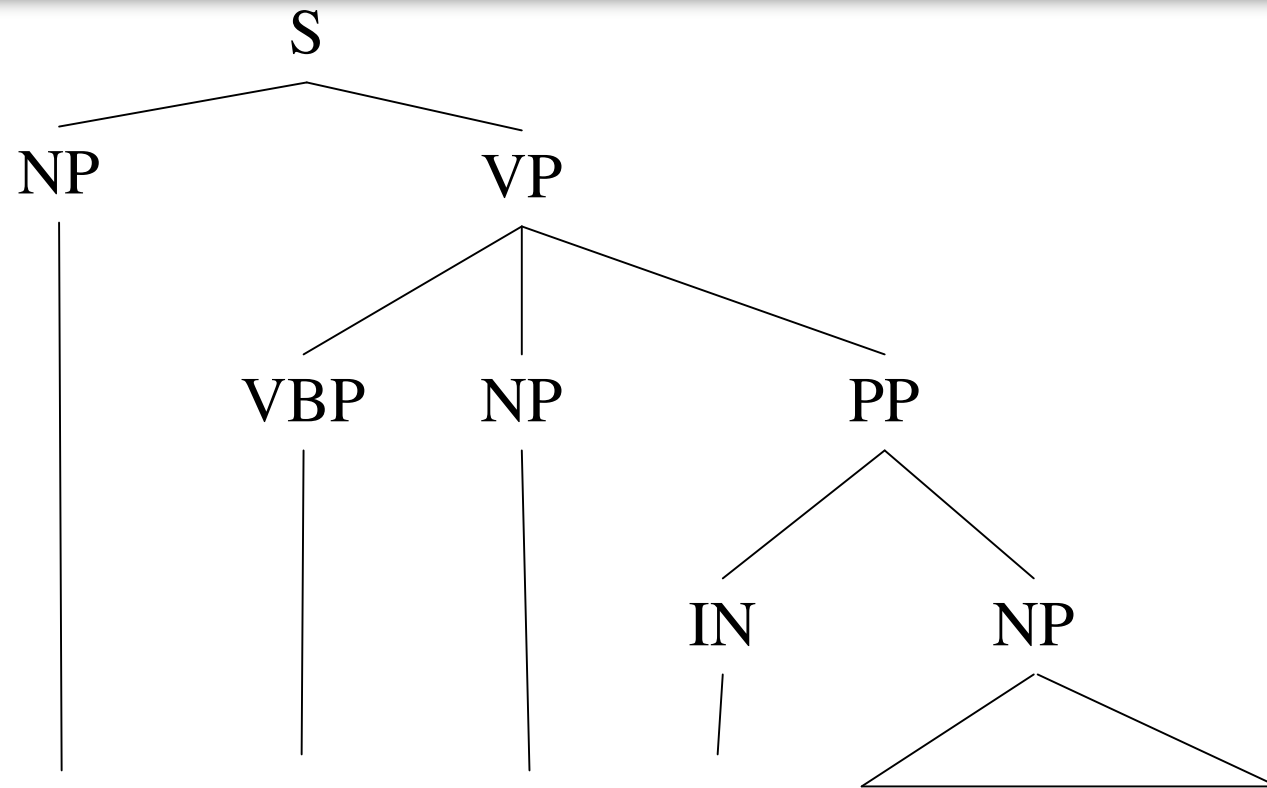
Enraged cow injures farmer with ax

# Teller Stuns Man with Stolen Check



Teller Stuns Man with Stolen Check

# Teller Stuns Man with Stolen Check

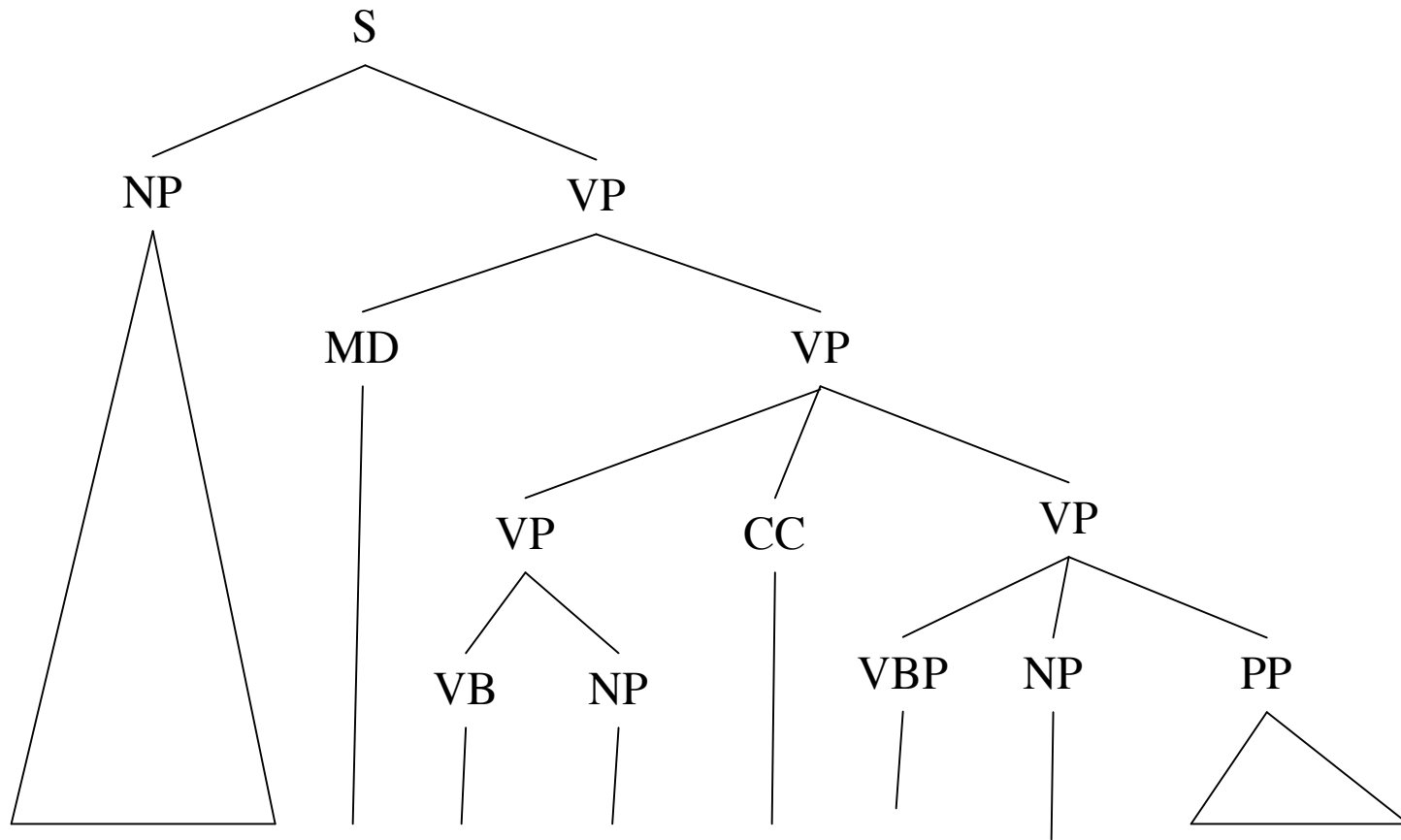


Teller Stuns Man with Stolen Check

# Categorizing ambiguity

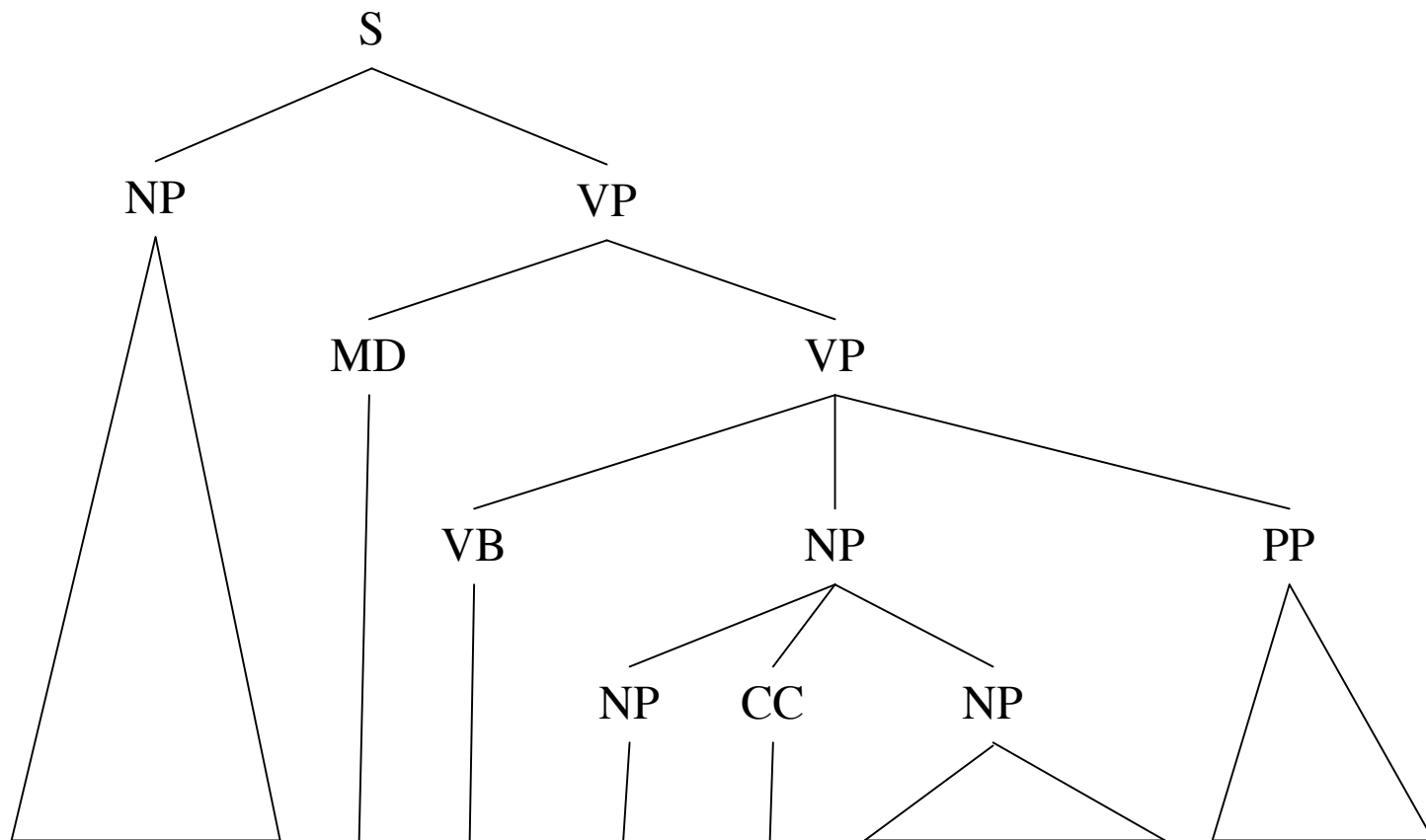
- Part of speech
- Conjunction
- Subordinate clause: higher or lower verbs?
- Prepositional phrase: verb or noun?

# Computer ambiguity v.s. human ambiguity



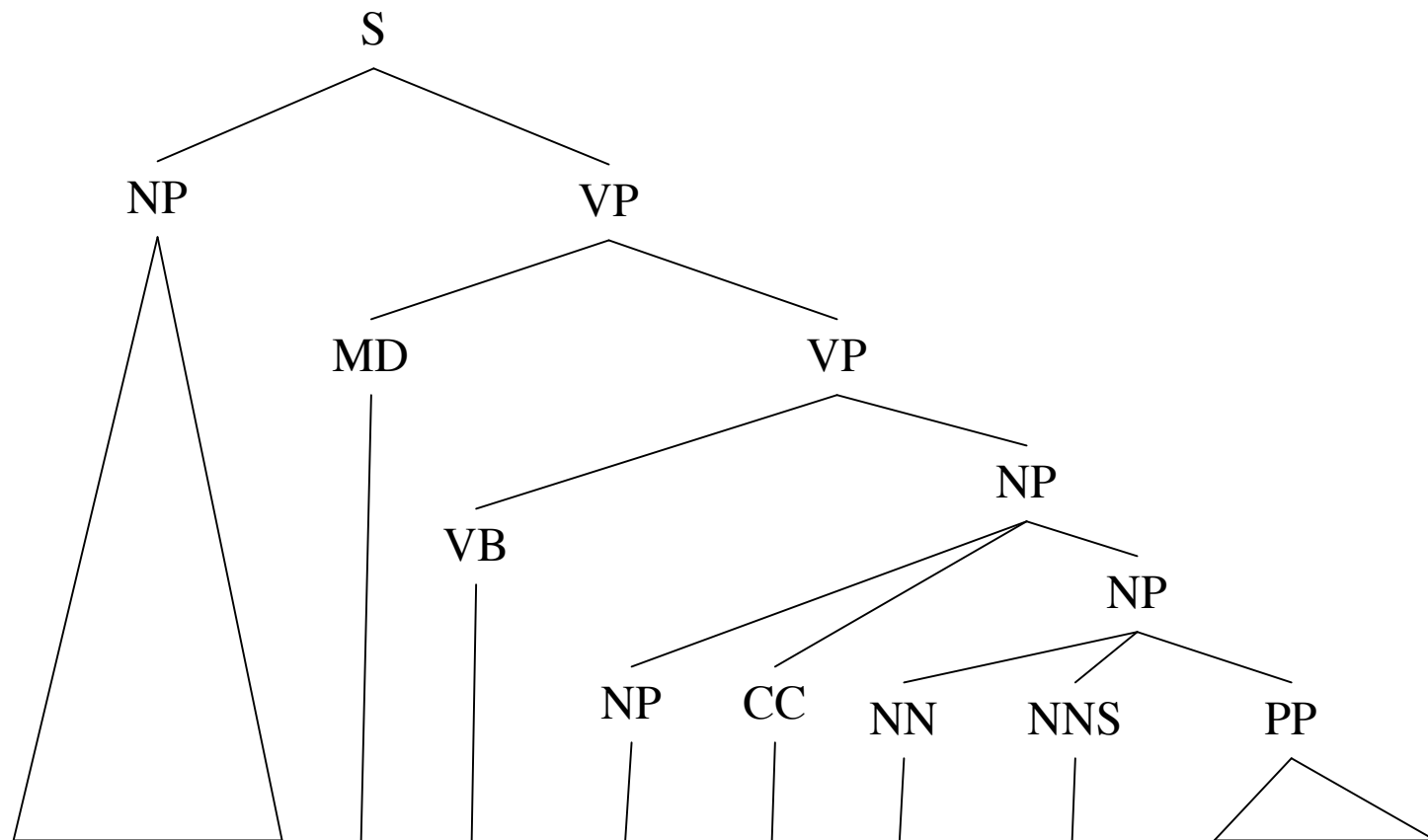
The post office will hold out discounts and service concessions as incentives

# Computer ambiguity v.s. human ambiguity



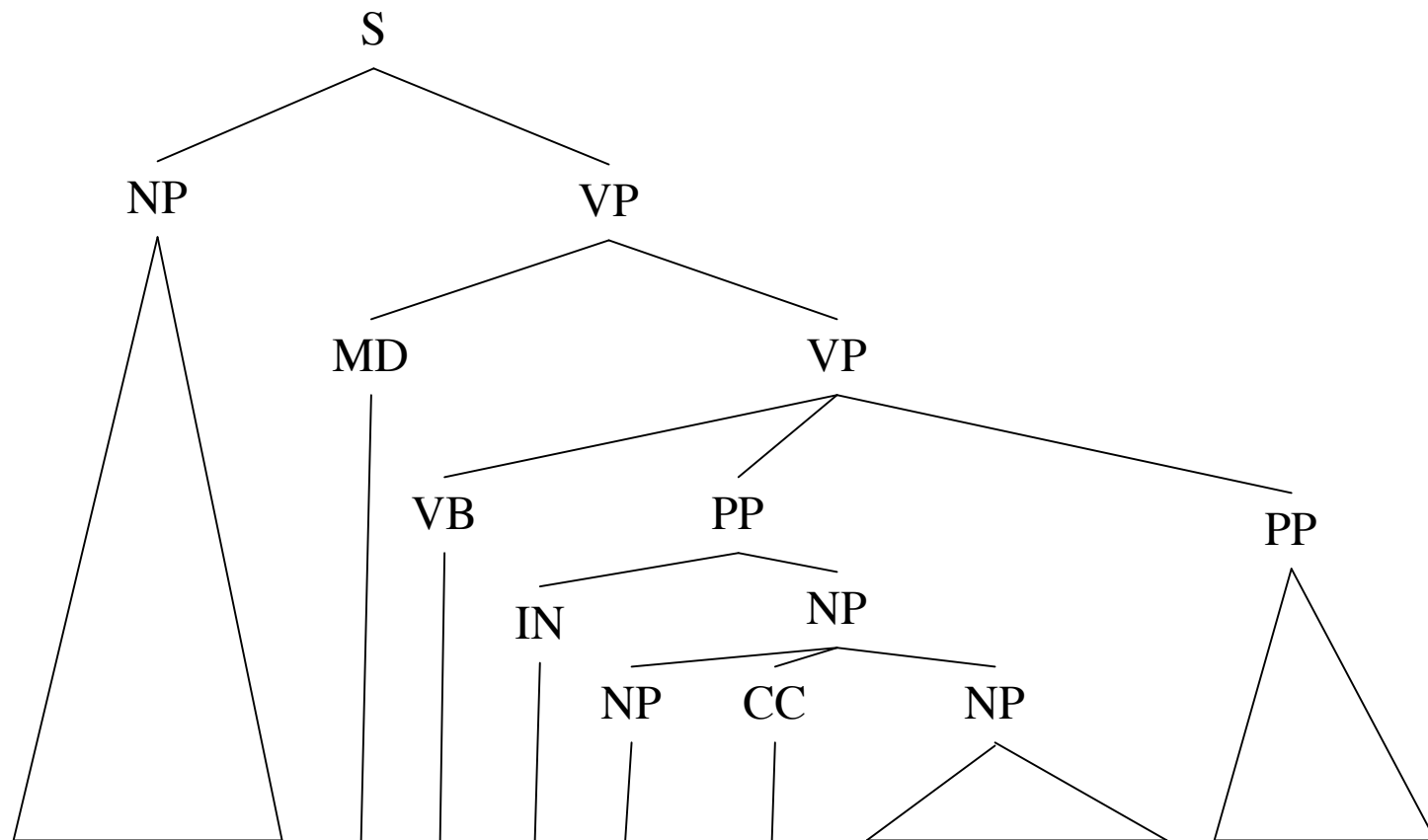
The post office will hold out discounts and service concessions as incentives

# Computer ambiguity v.s. human ambiguity



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

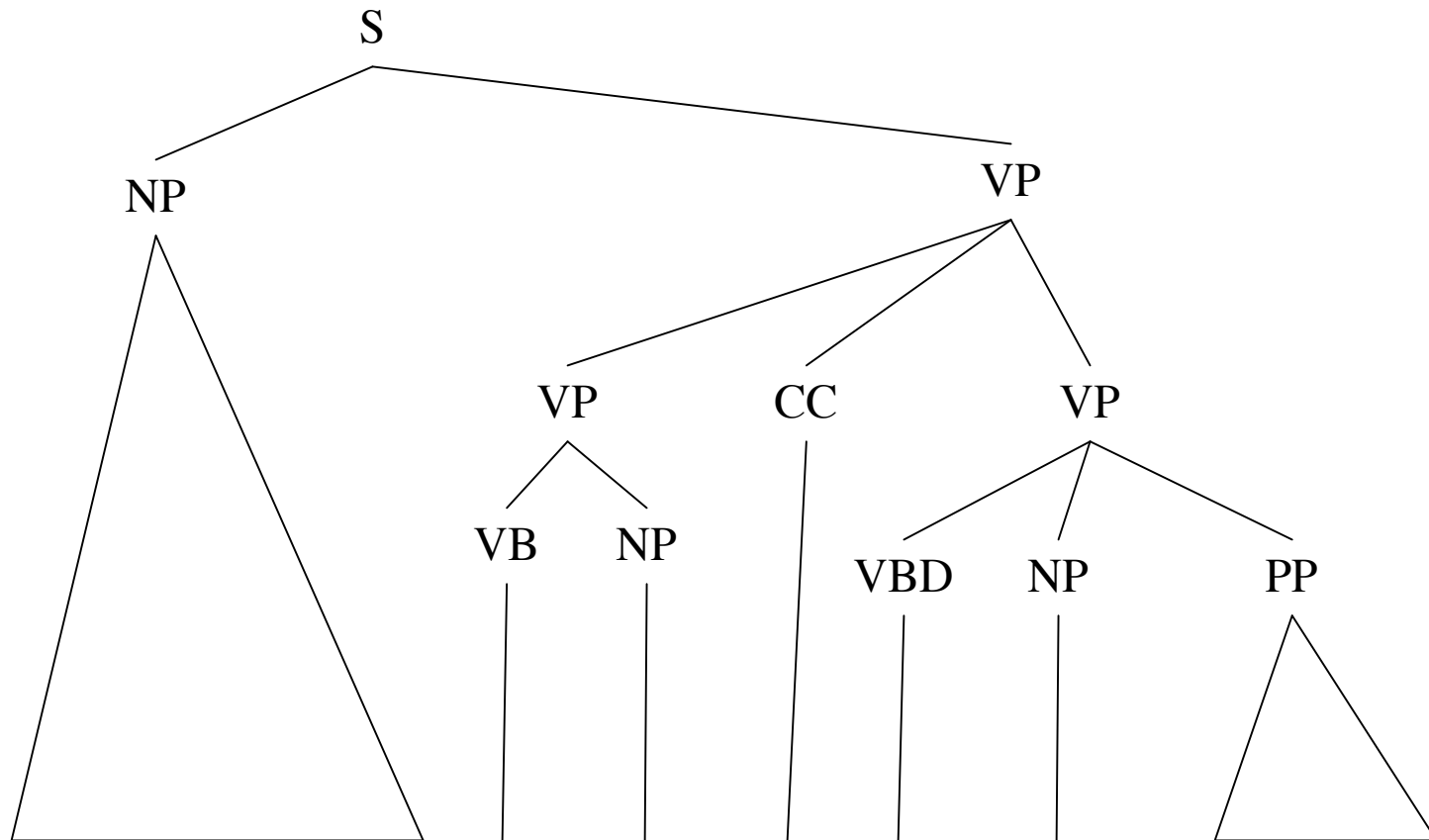
# Computer ambiguity v.s. human ambiguity



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

# Computer ambiguity vs human ambiguity

- Computer: I can't interpret this. There is so much ambiguity.
- Human: Ambiguity? What ambiguity?
- C: Well, first of all, "will" can be a noun or a modal verb
- H: No Silly, "will" can't be a noun here.
- C: how do I know that?
- H: You know because there is no such thing as "the post office will hold".
- C: huh???

# How many interpretations?

- The agency sees widespread use of the codes as a way of handling the rapidly growing mail volumes and controlling labor costs

# Discussion question

- Should a sentence be annotated with multiple interpretations in a treebank?

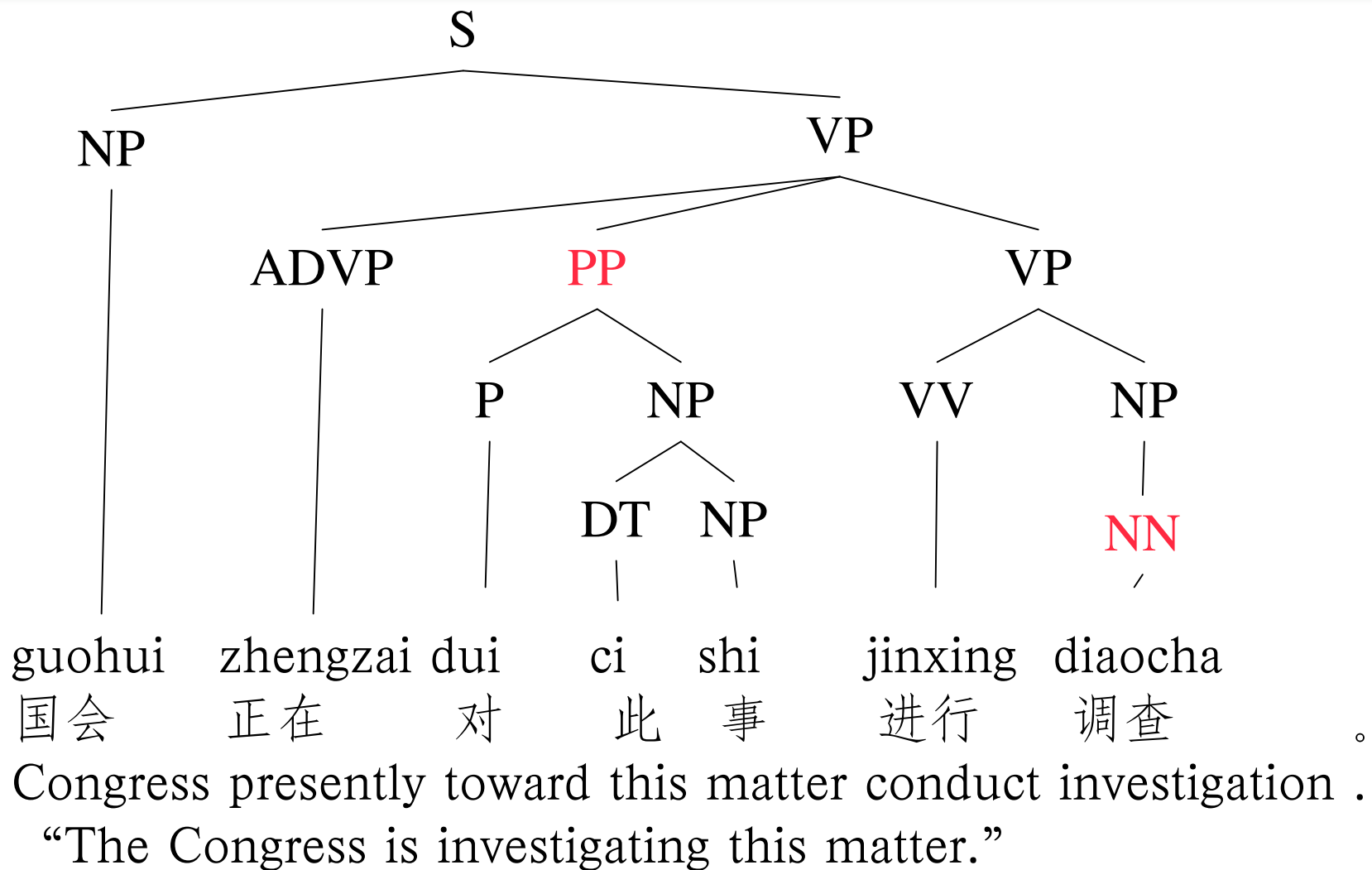
# Discussion question

- No, unless there is no way to resolve the ambiguity from the context provided in the text. Even then it may be a good idea to just choose one interpretation.
- Machines learn how ambiguities are resolved by examining a large amount of human annotated data and this approach is call **supervised machine learning**.

# Discussion question

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

# How do we represent the dependency between the PP and the Noun?



# Penn Treebank representation scheme

- 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)))

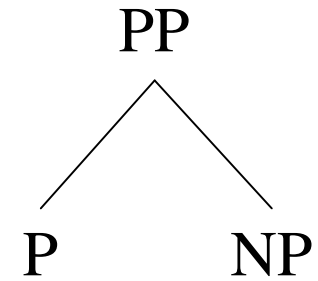
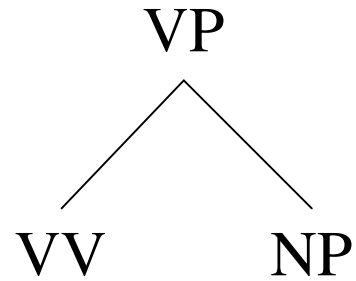
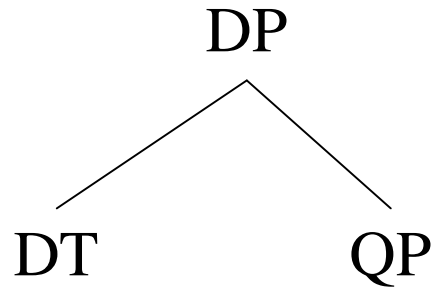
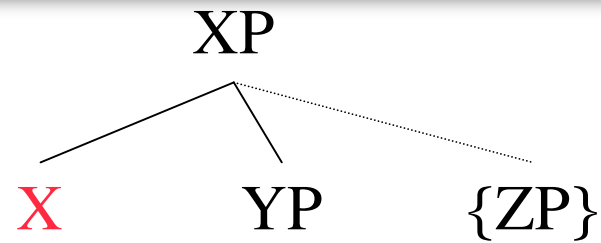
# CTB Phrasal labels

- ADJP: adjectival phrase
- ADVP: adverbial phrase
- **CLP**: classifier phrase, e.g., 一大帮人
- **CP**: complementizer phrase
- DNP: 的-phrase, e.g. 我的电脑
- DP: determiner phrase, e.g., 那三个人
- FRAG: fragmentary phrase
- **IP**: Inflectional phrase
- LCP: localizer phrase, e.g., 那三个人中
- LST: list, e.g., 一, 二, 三, 首先, 其次
- NP: noun phrase
- PP: prepositional phrase
- PRN: parenthetical
- QP: quantifier phrase
- UCP: unlike coordinate phrase
- VP: verb phrase

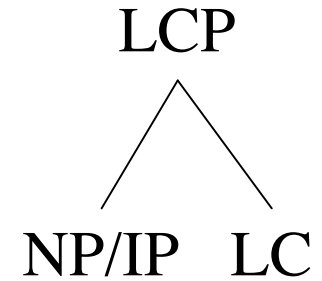
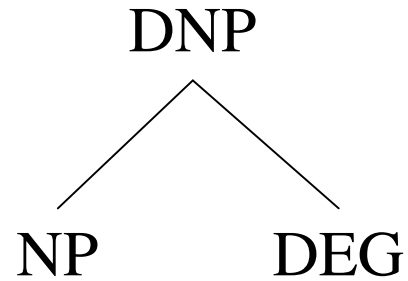
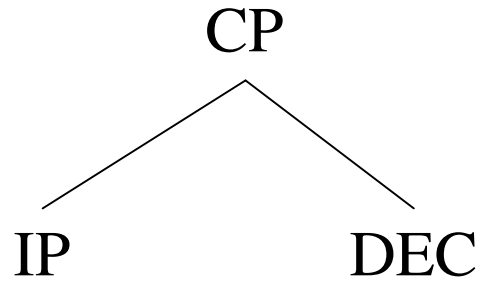
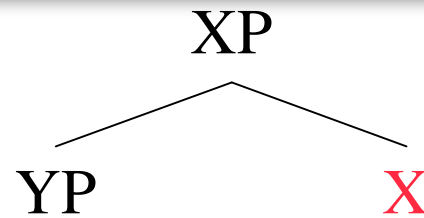
# An improvement (?) made in the Chinese Treebank

One grammatical relation per bracket

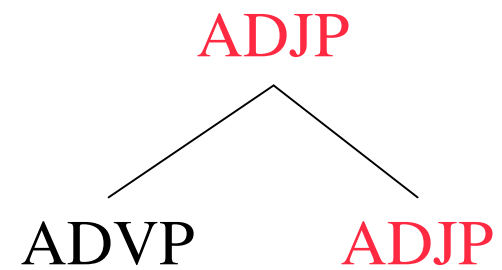
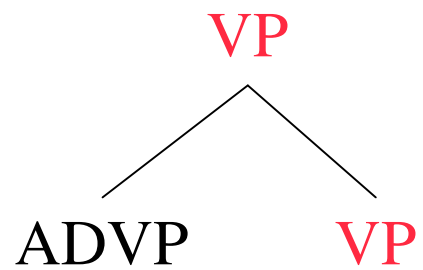
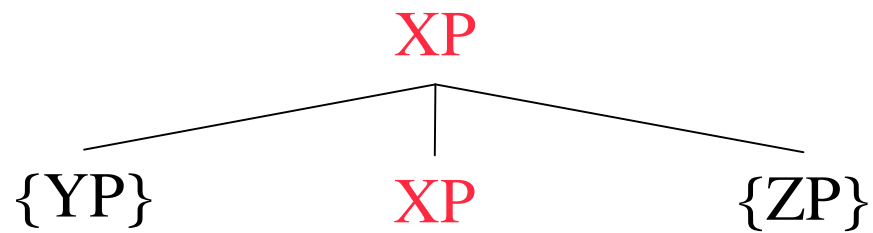
# Complementation (left-headed)



# Complementation (right-headed)

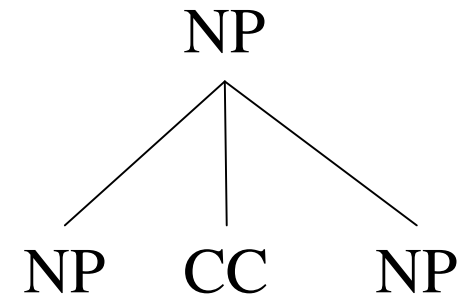
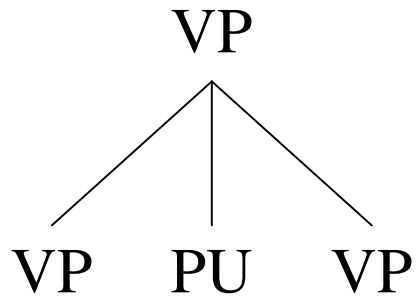
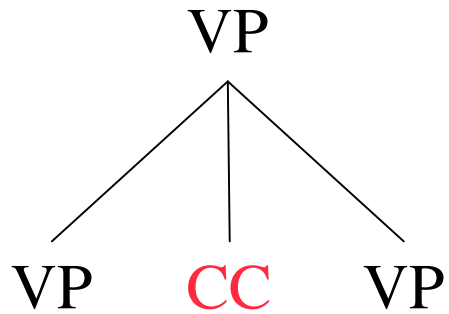
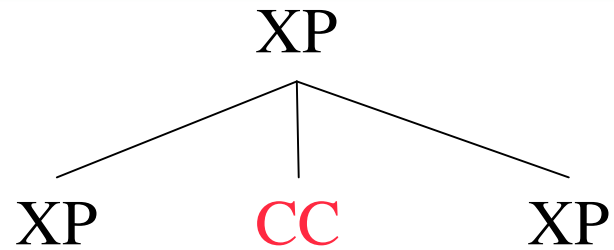


# Adjunction





# Coordination



# One grammatical relation per bracket

(IP (NP-SBJ (DNP (NP-PN (NR 美国)  
    (DEG 的)  
    (DP (DT 这)  
        (CLP (M 次))  
        (NP (NN 军事) (NN 行动))  
    (VP (PP-DIR (P 对)  
        (NP (DNP (NP (NP (NP-PN (NR 波斯湾))  
                    (NP (NN 地区))  
                    (CC 和)  
                    (NP (NN 世界))  
                    (DEG 的)  
                    (NP (NN 和平)  
                    (CC 与)  
                    (NN 安全)))  
        (VP (VV 造成)  
            (AS 了)  
            (NP-OBJ (ADJP (JJ 严重)  
                    (NP (NN 后果))))))

# Levy and Manning (2003)

- Is it harder to parse Chinese or Chinese Treebank?
- Should theoretical linguistics and corpus linguistics use the same criteria to decide what counts as a good parse and a good representation scheme?
- Should the quality of a corpus be validated by a parser?

# References

- Santorini, Beatrice and Anthony Kroch. 2000. The syntax of natural language: An online introduction using the Trees program.  
<http://www.ling.upenn.edu/~beatrice/syntax-textbook>
- Marcus, Mitchell P., Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational linguistics*, vol. 19
- Xue, Nianwen, Fei Xia, Fu-dong Chiou and Martha Palmer. 2005. Phrase structure annotation of a large Chinese corpus: the Chinese Treebank. *Natural Language Engineering*, 11(2)-207.