Brandeis University

Multilingual Treebanking

Nianwen Xue

7/12/2011

LSA Summer Institute

Multilingual Treebanking

- Treebanking is the process of mapping a sentence to its syntactic structure, usually in the form of a tree: a fully connected graph with a single root node
- Are trees sufficient to represent syntactic structures?

Treebanking by grammatical traditions

- Phrase structures
  - Penn Treebank (Generative Grammar: Extended Standard Theory)
  - LinGO Redwoods treebank (HPSG)
  - CCGBank (Combinatory Categorial Grammar)
- Dependency structures
  - The Prague Dependency Treebank
- Both
  - Tiger Treebank (German)

Penn Treebank

- Phrase structure annotation in the generative tradition
- The most influential treebank in NLP.
  - Google scholar citation: 3438 (Marcus et al 1993)

A little bit of history

- PTB I (Marcus et al 1993)
  - Context-free backbone
  - Skeletal structures
  - Limited empty elements
  - No argument/adjunct distinction
- PTB II (Marcus et al 1994)
  - Added function tags to mark up grammatical roles (thus argument/adjunct distinction, though not structurally)
  - Enriched the set of empty elements

A little bit of history

- Beyond PTB II
  - OntoNotes English Treebank annotation added more depth to the NP structure:
  - NML (“NoMinaL modifiers”)
    - (NP (NML human liver tumor) analysis)
  - *P* (place-holder):
    - (NP (NP K- (NML-1 *P*)) and (NP N- (NML-1 ras)))

http://papers.ldc.upenn.edu/Treebank_BioMedical_Addendum/TBguidelines-addendum.htm

PTB I Content

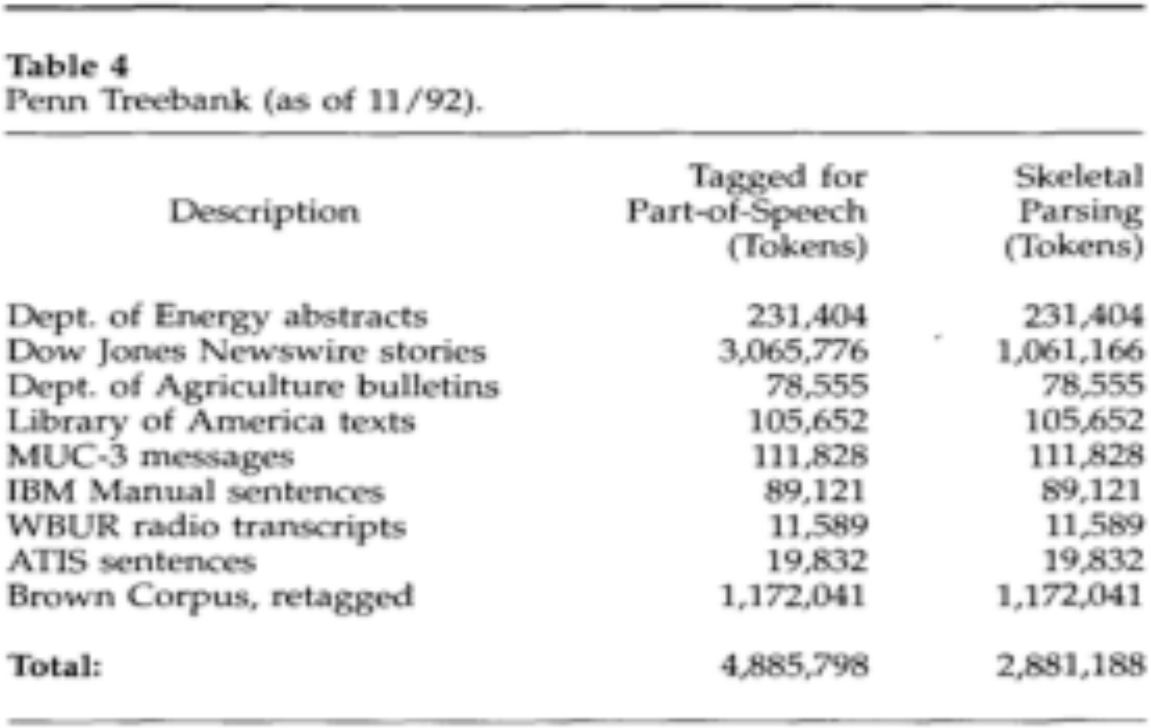


Table 4
Penn Treebank (as of 11/92).

Description	Tagged for Part-of-Speech (Tokens)	Skeletal Parsing (Tokens)
Dept. of Energy abstracts	231,404	231,404
Dow Jones Newswire stories	3,065,776	1,061,166
Dept. of Agriculture bulletins	78,555	78,555
Library of America texts	105,652	105,652
MUC-3 messages	111,828	111,828
IBM Manual sentences	89,121	89,121
WBUR radio transcripts	11,589	11,589
ATIS sentences	19,832	19,832
Brown Corpus, retagged	1,172,041	1,172,041
Total:	4,885,798	2,881,188

Most used

PTB II Content

- One million words of 1989 Wall Street Journal material annotated in Treebank-2 style.
- A small sample of ATIS-3 material annotated in Treebank-2 style.
- 300-page style manual for Treebank-2 bracketing, as well as the part-of-speech tagging guidelines.
- The contents of the previous Treebank CD-ROM (Version 0.5), with cleaner versions of the WSJ, Brown Corpus, and ATIS material (annotated in Treebank-1 style).
- Tools for processing Treebank data, including "tgrep," a tree-searching and manipulation package (note that usability of this release of tgrep is limited: users of Sun sparc systems should have no problem, but others may find the software to be difficult or impossible to port).

From the LDC website

PTB III Content

- This CD-ROM contains the following Treebank-2 Material:
  - One million words of 1989 Wall Street Journal material annotated in Treebank II style.
  - A small sample of ATIS-3 material annotated in Treebank II style.
  - A fully tagged version of the Brown Corpus.
- and the following new material:
  - Switchboard tagged, dysfluency-annotated, and parsed text
  - Brown parsed text

From the LDC website

Later Additions

- OntoNotes 4.0
  - 1.2M words of English Treebank
- Translations from other languages:
  - ECTB:
    - English Chinese Translation Treebank v 1.0
  - EATB:
    - English-Arabic Treebank v 1.0

PTB POS Tagset

1. CC Coordinating conjunction
2. CD Cardinal number
3. DT Determiner
4. EX Existential *there*
5. FW Foreign word
6. IN Preposition/subordinating participle conjunction
7. JJ Adjective
8. JJR Adjective, comparative
9. JJS Adjective, superlative
10. LS List item marker
11. MD Modal
12. NN Noun, singular or mass
13. NNS Noun, plural
14. NNP Proper noun, singular
15. NNPS Proper noun, plural
16. PDT Predeterminer
17. POS Possessive ending
18. PRP Personal pronoun
19. PP$ Possessive pronoun
20. RB Adverb
21. RBR Adverb, comparative
22. RBS Adverb, superlative
23. RP Particle
24. SYM Symbol (mathematical or scientific)
25. TO *to*
26. UH Interjection
27. VB Verb, base form
28. *VBD Verb, past tense*
29. VBG Verb, gerund/present
30. VBN Verb, past participle
31. VBP Verb, non-3rd ps. sing. present
32. VBZ Verb, 3rd ps. sing. present
33. WDT wh-determiner
34. WP wh-pronoun
35. WP$ Possessive wh-pronoun
36. WRB wh-adverb
37. # Pound sign
38. $ Dollar sign
39.. Sentence-final punctuation
40. , Comma
41. : Colon, semi-colon
42. (Left bracket character
43.) Right bracket character
44. " Straight double quote
45. ' Left open single quote
46. " Left open double quote
47. ' Right close single quote
48. " Right close double quote

PTB POS Tagging choices

- Based on the Brown Corpus, but simplified
- Merged lexically recoverable distinctions
  - No special tags for ‘be’,’do’, ‘have’, etc.
  - Still some residual tags: ‘to’
- Encode syntactic function where possible
  - The One/CD, the ones/NNS => the one/NN, the ones/NNS
- Allows multiple tags for a word in limited circumstances
  - JJ/NN, JJ/VBG, JJ/VBN, NN/VBG, RB/RP

Representation machinery

- Constituents with phrase labels
- Function tags
- Empty categories and co-indexation

Phrase labels

Label	Description
ADJP	Adjective phrase
ADVP	Adverbial phrase
NP	Noun phrase
PP	Prepositional phrase
S	Simple declarative sentence
SBAR	Clause introduced by subordination conjunction or 0
SBARQ	Direction question introduced with wh-word or wh-phrase
SINV	Declarative senence with subject-aux inversion
SQ	Subconstituent of SBARQ excluding wh-word or wh-phrase
VP	Verb phrase
WHADVP	Wh-adverb phrase
WHNP	Wh-noun phrase
WHPP	Wh-prepositional phrase
X	Constituent of unknow category

Dash tags (added in PTB II)

Text categories		Grammatical function	
-HLN	Headlines and datelines	-CLF	True clefts
-LST	List markers	-NOM	Non NPs that function as NPs
-TTL	titles	-ADV	Clausal and NP adverbials
Semantic Roles		-LGS	Logical subjects in passive
-VOC	vocatives	-PRD	Non-VP predicates
-DIR	Direction & trajectory	-SBJ	Surface subject
-LOC	location	-TPC	Topicalized and fronted constituents
-MNR	manner	-CLR	Closely related
-PRP	Purpose and reason	-DTV	Dative
-TMP	Temporal phrases	-PUT	Locative of PP of put
-BNF	Benefactive		
-EXT	extent		

Empty categories and coindexation (PTB II)

Empty categories		Pseudo attachment	
T	Trace of A'-movement (WH movement and topicalization)	*RNR*	Right node raising
(NP *)	Arbitrary PRO, controlled PRO, trace of passivization and raising	*ICH*	Interpret constituent here
0	Null complementizer and wh operator	*EXP*	expletive
U	Unit (of currency, etc.)	*PPA*	Permanent ambiguity
?	Ellipsed material of unknown category		
NOT	Anti-placeholder		

Raising

```
( (S
  (NP-SBJ-1
   (NP (NN Choice) )
   (PP (IN of)
    (NP (DT the) (NN volunteer) (NN military) ))
   (PP (IN in)
    (NP (DT the) (CD 1970s) )))
  (VP (VBD seemed)
   (S
    (NP-SBJ (-NONE- *-1) )
    (VP (TO to)
     (VP (VB doom)
      (NP (JJ national) (NN service) )
      (NP
       (NP (RB as) (RB much) )
       (PP (IN as)
        (NP (DT the) (NN draft) )))))))
  (. .) ))
```

(Subject) Control

```
( (S
    (NP-SBJ-5 (NNP Government) (NNS officials) )
    (VP (VBD tried)
      (PP-TMP (IN throughout)
        (NP (DT the) (NN weekend) ))
      (S
        (NP-SBJ (-NONE- *-5) )
        (VP (TO to)
          (VP (VB render)
            (NP (DT a) (JJ business-as-usual) (NN appearance) ))))
      (SBAR-PRP (IN in) (NN order)
        (S
          (NP-SBJ (-NONE- *-5) )
          (VP (TO to)
            (VP (VB avoid)
              (NP
                (NP (DT any) (NN sense) )
                (PP (IN of)
                  (NP (NN panic) ))))))))
    (. .) ))
```

Wh-movement (relative clause)

```
( (S (CC And)
    (NP-SBJ-1 (PRP we) )
    (VP (VBP hope)
      (S
        (NP-SBJ (-NONE- *-1) )
        (VP (TO to)
          (VP
            (VP (VB take)
              (NP (NN advantage) )
              (PP-CLR (IN of)
                (NP (NNS panics) )))
            (CC and)
            (VP (VB buy)
              (NP (NNS stocks) )
              (SBAR-TMP
                (WHADVP-2 (WRB when) )
                (S
                  (NP-SBJ (PRP they) )
                  (VP (VBP plunge)
                    (ADVP-TMP (-NONE- *T*-2) ))))))))))
    (. .) (" ") ))
```

Wh-movement (Question)

```
( (SBARQ
    (WHNP-1 (WP What) )
    (SQ (VBZ is)
      (NP-SBJ-2 (NN one) )
      (S
        (NP-SBJ (-NONE- *-2) )
        (VP (TO to)
          (VP (VB think)
            (NP (-NONE- *T*-1) )
            (PP-CLR (IN of)
              (NP (PDT all) (DT this) ))))))
    (. ?) ))
```

ICH (Extraposition)

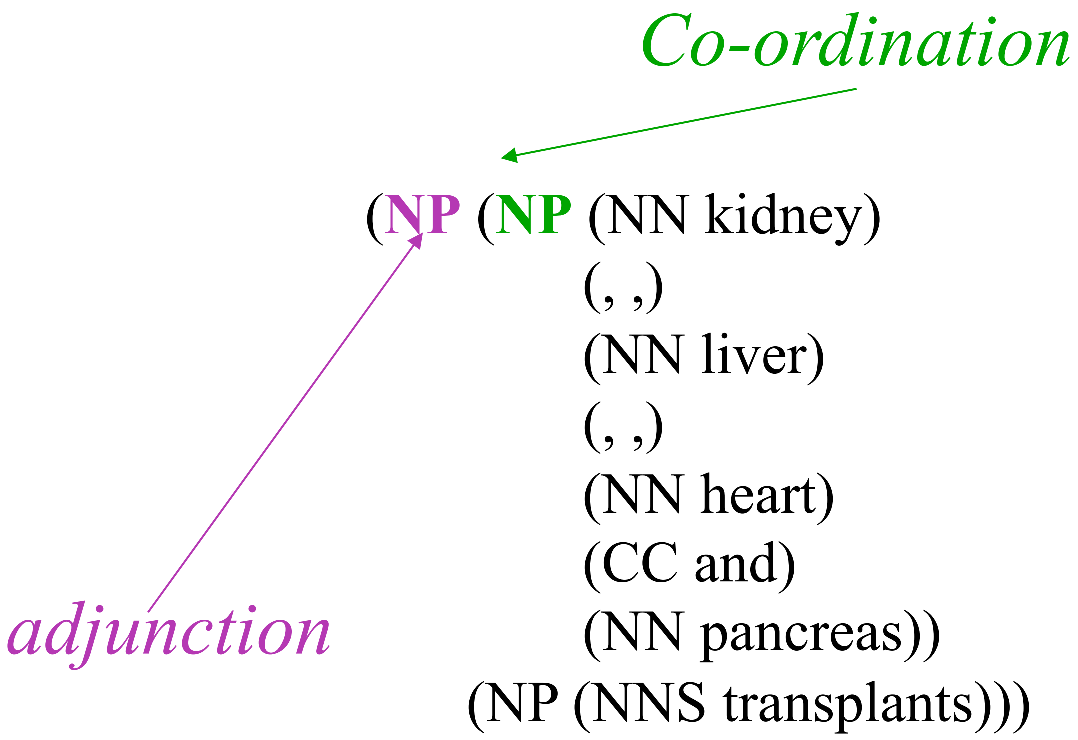
```
(S (NP-SBJ Plato)
   (VP knew
       (SBAR *ICH*-1)
       (NP-TMP yesterday)
       (SBAR-1 that
              (S (NP-SBJ Terry)
                 (VP would
                      (VP accept
                          (NP the honor)))))))
```

Linguistics and annotation

- Annotation is linguistics within a time frame
  - Analyzing a few sentences vs analyzing thousands of sentences consistently in a very short time
  - Data coverage and elegance of linguistic representation is good, but also need to ask:
    - Is my annotation reproducible by the machine
    - Is my annotation reproducible by other researchers?
    - Can my annotation be produced fast enough?
  - Tradeoffs may have to be made due to the time constraint
    - Plenty of evidence for that in the Penn Treebank II (but some have been fixed later)

Flat structures to save time

```
(NP (NN kidney)
        (, ,)
        (NN liver)
        (, ,)
        (NN heart)
        (CC and)
        (NN pancreas)
       (NNS transplants))
```

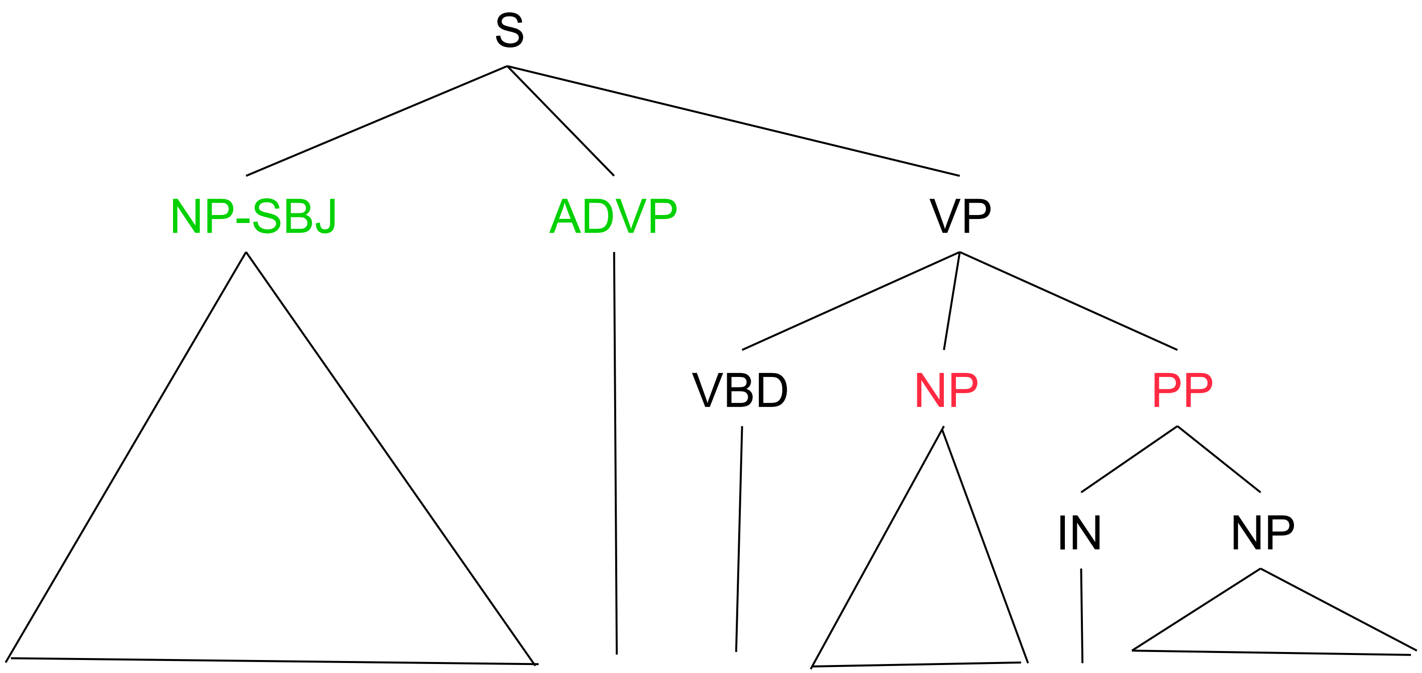


Co-ordination

adjunction

```
(NP (NP (NN kidney)
          (, ,)
          (NN liver)
          (, ,)
          (NN heart)
          (CC and)
          (NN pancreas))
    (NP (NNS transplants)))
```

No argument/adjunct distinction



The Mortgage and equity last paid a dividend on August 1, 1988
real estate investment trust

Quotes from Marcus et al 1993

“It proved to be very difficult for annotators to distinguish between a verb's arguments and adjuncts in all cases. Allowing annotators to ignore this distinction when it is unclear (attaching constituents high) increases productivity by approximately 150-200 words per hour. Informal examination of later annotation showed that forced distinctions cannot be made consistently.”

-CLR: Closely related

```
(SBAR (IN as)
    (S
      (NP-SBJ (DT the) (VBG graying) (NNS men) )
      (VP (VBD returned)
        (PP-CLR (TO to)
          (NP (PRP$ their) (NNS homes) )))))
```

18345 instances in the WSJ section of the PTB II

Can’t be properly addressed in syntax, but addressed in Propbank

CLR

```
(S
     (NP-SBJ (NNS well-wishers) )
     (VP (VBD stuck)
       (NP (JJ little) (NNP ANC) (NNS flags) )
       (PP-LOC-CLR (IN in)
         (NP (PRP$ their) (NN hair) ))))
   (CC and)
   (S
     (NP-SBJ (DT a) (NN man) )
     (VP (VBD tooted)
       (PP-LOC-CLR (IN on)
         (NP
           (NP (DT an) (NN antelope) (NN horn) )
           (VP (VBN wrapped)
             (NP (-NONE- *) )
             (PP-LOC-CLR (IN in)
               (NP (NNP ANC) (NNS ribbons) )))))))
```

Arg2?

PPA (maybe *RNR*?)

```
( (S
  (NP-SBJ (CD One) (JJ local) (NNP Phillips) (NN manager) )
  (VP (VBD said)
   (SBAR (-NONE- 0)
    (S
     (NP-SBJ
      (NP (DT a) (NN seal) )
      (PP-LOC (-NONE- *PPA*-1) ))
     (VP (VBD blew)
      (PP-LOC-1 (IN in)
       (NP
        (NP (CD one) )
        (PP (IN of)
         (NP
          (NP (DT the) (NN plant) (POS 's) )
          (NNS reactors) ))))))))
  (. .) ))
```

27 instances annotated in the entire WSJ Section of PTB II, most of which questionable

Evaluation

- Parseval, using the evalb software
  - http://nlp.cs.nyu.edu/evalb/
- Agreement among the annotators
  - There are easy exploits for the system, so you also want to calculate
- Agreement between an annotator and the benchmark

Chinese Treebank (1998 - ?)

邱福栋	Fu-Dong Chiou
蒋自新	Zixin Jiang
石美莎	Martha S. Palmer
夏　飞	Fei Xia
薛念文	Nianwen Xue
张美玉	Meiyu Chang
张修红	Xiuhong Zhang

(Xue, Xia, Chiou, Palmer 2005, JNLE)

CTB: overview

- Started in 1998 at Penn
- Supported by DOD, NSF, DARPA
- Latest version 7.0, 1.2M word Chinese corpus
  - Segmented, POS-tagged, syntactically bracketed
  - Phrase structure annotation
  - 94% ITA ***(Xue, Xia, Chiou, Palmer 2005)***
  - On-going expansion, another 1.2M words planned
- Additional layers of annotation
  - Propbank/Nombank

CTB: Milestones

Version	Year	Quantity (words)	Source	Propbank/ Nombank
CTB1.0	2001	100K	Xinhua	yes
CTB3.0	2003	250K	+HK News	yes
CTB4.0	2004	400K	+Sinorama	yes
CTB5.0	2005	500K	+Sinorama	yes
CTB6.0	2007	780K	+ BN	yes
CTB7.0	2010	1.2M	+BC,WB	yes

The Chinese Treebank: What’s the same?

- Same three layers:
  - Tokenization/word segmentation, part-of-speech tagging and syntactic parsing
- Same representation scheme
  - Phrase structure annotation, context-free grammar backbone
  - Function tags
  - Empty categories and their coindexation

The Chinese Treebank: What’s different?

- Word segmentation is a much more substantial task due to the orthographical conventions of Chinese
- Substantial difference in the POS tagset, reflecting the morphology-poor nature of the Chinese language
- Different choices at the syntactic parsing level, most notably the argument/adjunct distinction

Tagset comparison

Noun: Lack of number morphology
Verb: Lack of tense and aspect morphology
Adjectives/adverbs: no comparative and superlative forms
Preposition: prepositions and postpositions

<table>
<tr><th>Category</th><th>PTB</th><th>CTB</th><th>Category</th><th>PTB</th><th>CTB</th></tr>
<tr><td rowspan="6">verb</td><td rowspan="2">VBD,VBG,VBN,VBZ,VBP,VB</td><td rowspan="2">VV</td><td rowspan="4">noun</td><td>NN, NNS</td><td>NN</td></tr>
<tr><td rowspan="2">NP, NPS</td><td rowspan="2">NR</td></tr>
<tr><td rowspan="2">VBD,VBG,VBN,VBZ,VBP,VB</td><td rowspan="2">VA</td></tr>
<tr><td>NN</td><td>NT</td></tr>
<tr><td>VBD,VBG,VBN,VBZ,VBP,VB</td><td>VC</td><td>preposition</td><td>IN</td><td>P, LC,CS</td></tr>
<tr><td>VBD,VBG,VBN,VBZ,VBP,VB</td><td>VE</td><td rowspan="3">other</td><td rowspan="3">TO, MD,
POS, RP,
WDT, WP$
WP, WRB</td><td rowspan="3">BA, SB,
LB,DEC, DEG,
DEV,MSP, DER,
M, SP, DT</td></tr>
<tr><td>adjective</td><td>JJ, JJR, JJS</td><td>JJ</td></tr>
<tr><td>Adverb</td><td>RB, RBR, RBS</td><td>AD</td></tr>
</table>

Characteristics of Chinese

- No natural word boundary in text
- Pervasive pro-drop

  这 是 以前 *pro* 不 曾 遇到 的 新 问题 。

  this be before not already encounter DE new problem .

  “This is a problem we haven’t seen before.”
- Morphology-poor
  - No (explicit) tense, gender, person, number, agreement morphology

Word segmentation

日文章鱼怎么说 ?

日文　章鱼　怎么 说 ？

Japanese octopus how say

“How to say octopus in Japanese?”

日　文章　鱼　怎么 说 ？

Japan article fish how say

“???”

Word segmentation

日文章鱼怎么说？

日文 章鱼 怎么 说 ？
Japanese octopus how say
“How to say octopus in Japanese?”

日 文章 鱼 怎么 说 ？
Japan article fish how say
“???” Japanese octopus how?

Google Translate BETA

POS: verb or noun

美国 将 与 中国 讨论 贸易 赤字 。

U.S. will with China discuss trade deficit .

“The U.S. will discuss trade deficit with China.”

美国 将 与 中国 就 贸易 赤字 进行 讨论 。

U.S. will with China regarding trade deficit engage discussion .

“The U.S. will engage in a discussion on the trade deficit with china.”

POS: verb or noun

美国 将 与 中国 讨论 贸易 赤字 。

U.S. will with China discuss trade deficit .

“The U.S. will discuss trade deficit with China.”

“The United States will discuss trade deficit with China.”

美国 将 与 中国 就 贸易 赤字 进行 讨论 。

U.S. will with China regarding trade deficit en sion .

“The U.S. will engage in a discussion on the trad n china.”

“The United States trade deficit with China to discuss.”

Verb or preposition?

Google 用 33 亿 现金 收购 Double Click
Google use 33 billion cash buy Double Click

Google used 33 billion cash to buy Double Click
Google bought Double Click with 33 billion cash

Verb or preposition?

Google 用 33 亿 现金 收购 Double Click
Google use 33 billion cash buy Double Click

Google used 33 billion cash to buy Double Click
Google bought Double Click with 33 billion cash

Google spent 3.3 billion in cash Double Click

Google Translate BETA

Sentential complement or object control?

NP	V	NP	V	NP
他	希望	她	抢	银行
he	hope	she	rob	bank

“He hopes that she will rob the bank.”

他	逼	她	抢	银行
he	force	she	rob	bank

“He forced her to rob the bank.”

Sentential complement or object control?

NP	V	NP	V	NP
他	希望	她	抢	银行
he	hope	she	rob	bank

"He hopes that she will rob the bank."

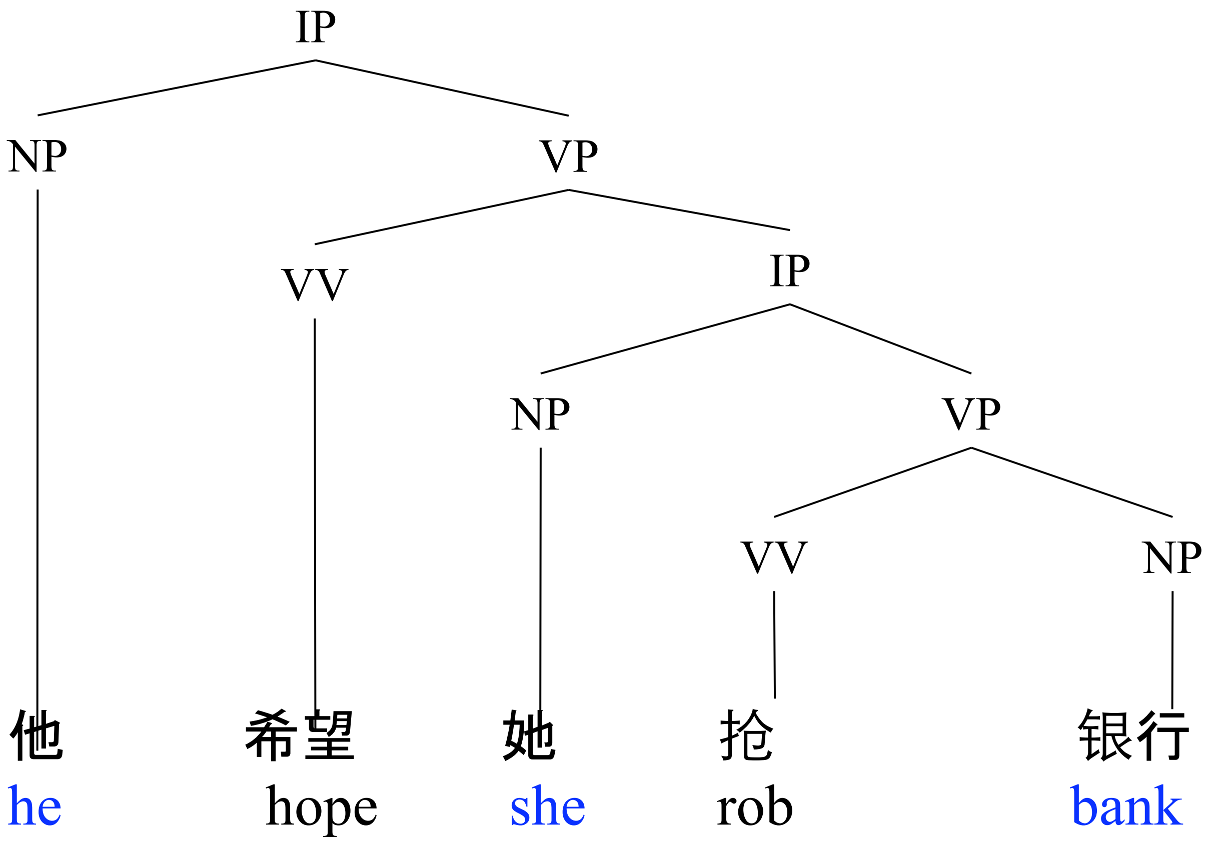
"He expressed the hope that her robbing the bank" Google Translate BETA

他	逼	她	抢	银行
he	force	she	rob	bank

"He forced her to rob the bank."

"He forced her robbing a bank." Google Translate BETA

Sentential complement



IP

NP VP

VV IP

NP VP

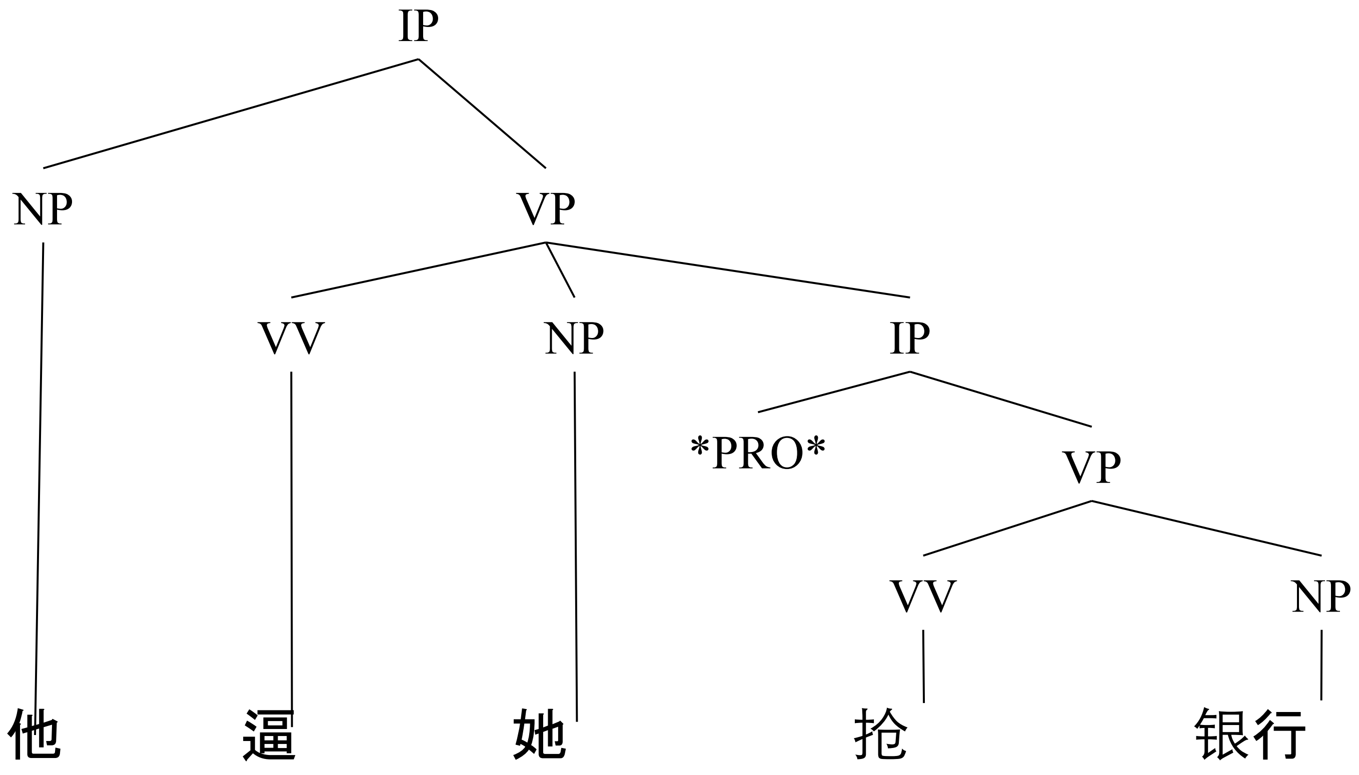
VV NP

他 希望 她 抢 银行

he hope she rob bank

“He hopes she will rob the bank.”

Object control



IP

NP VP

VV NP IP

PRO VP

VV NP

他 逼 她 抢 银行

he force she rob bank

“He forced her to rob the bank.”

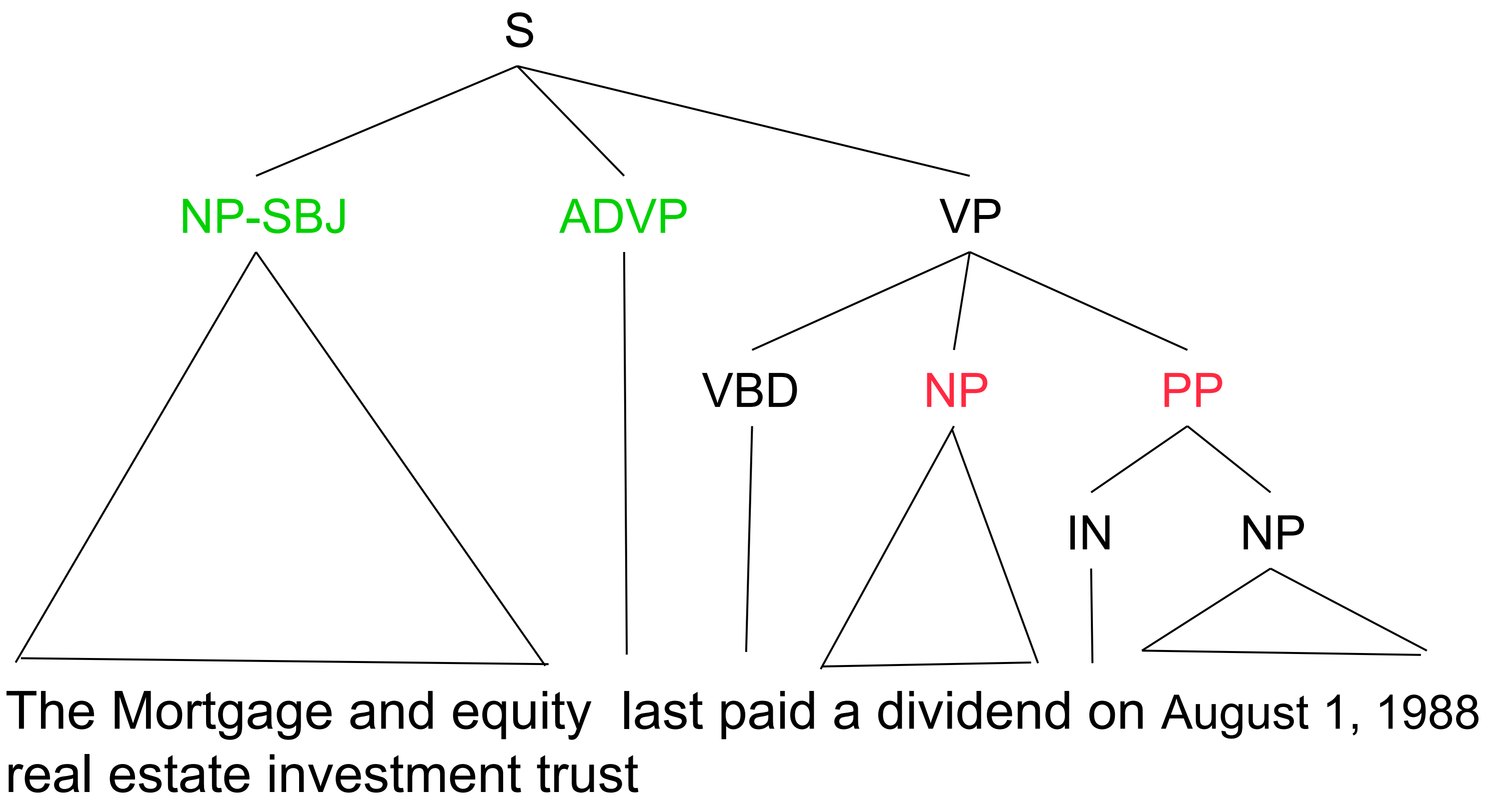
Sentential complement vs object control

- Can it take an existential construction as its complement?
- Can it take an idiom as its complement?
- Can it take a BEI construction as its complement?
- Can it take a topic construction as its complement?
- Can the complement clause have an aspectual marker?

Yes → Sentential complement

No → Object control

A Penn Treebank Example



S

NP-SBJ

ADVP

VP

VBD

NP

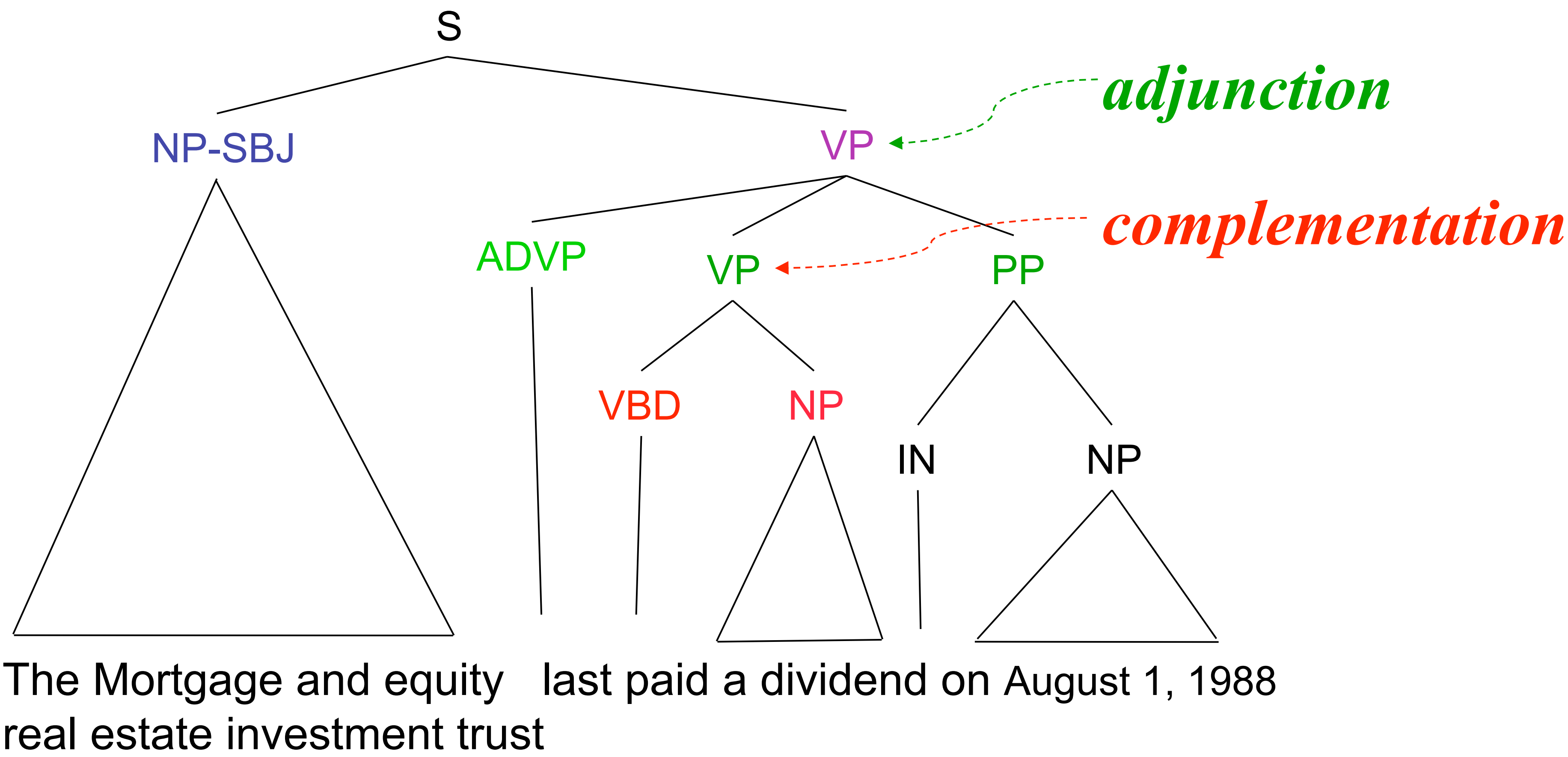
PP

IN

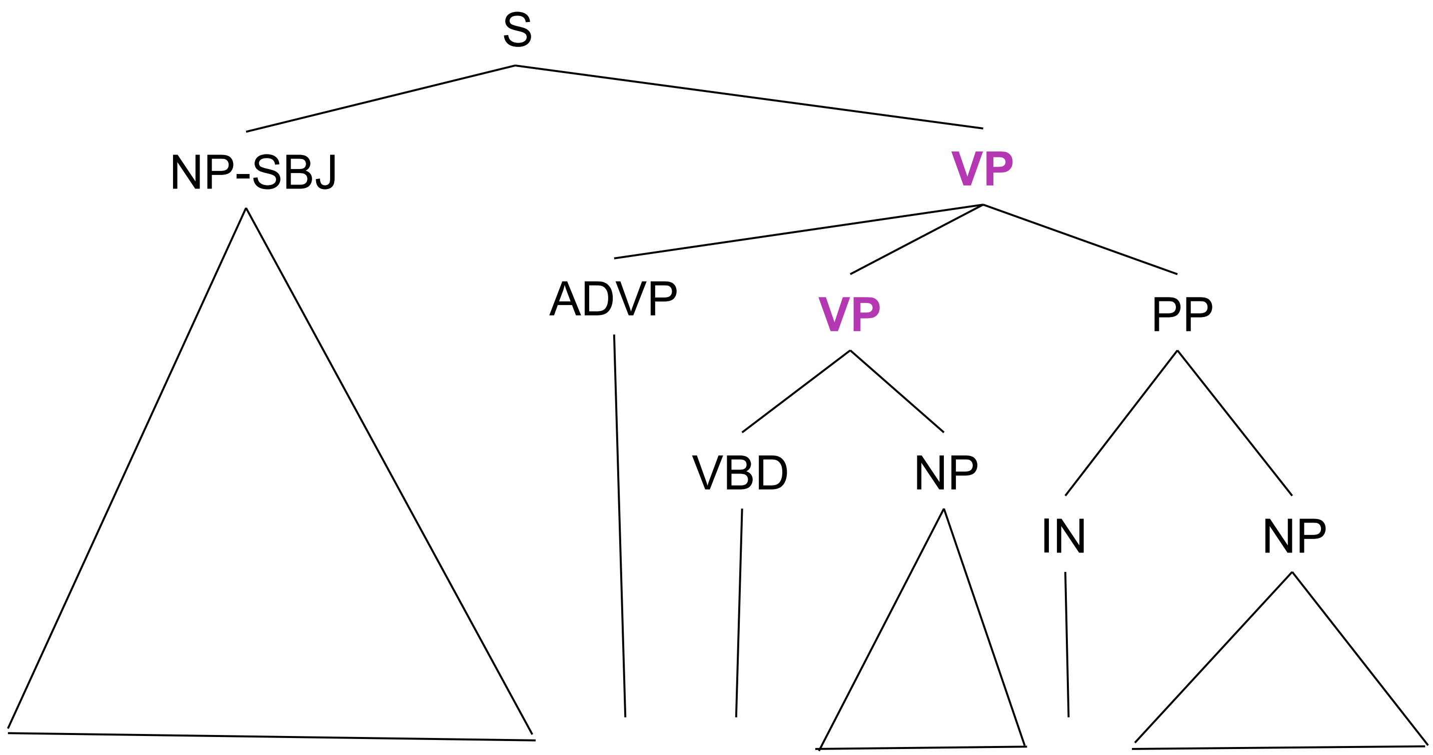
NP

The Mortgage and equity real estate investment trust last paid a dividend on August 1, 1988

Representing argument/adjunction distinction in (hypothetical) CTB annotation



Recursive structure!



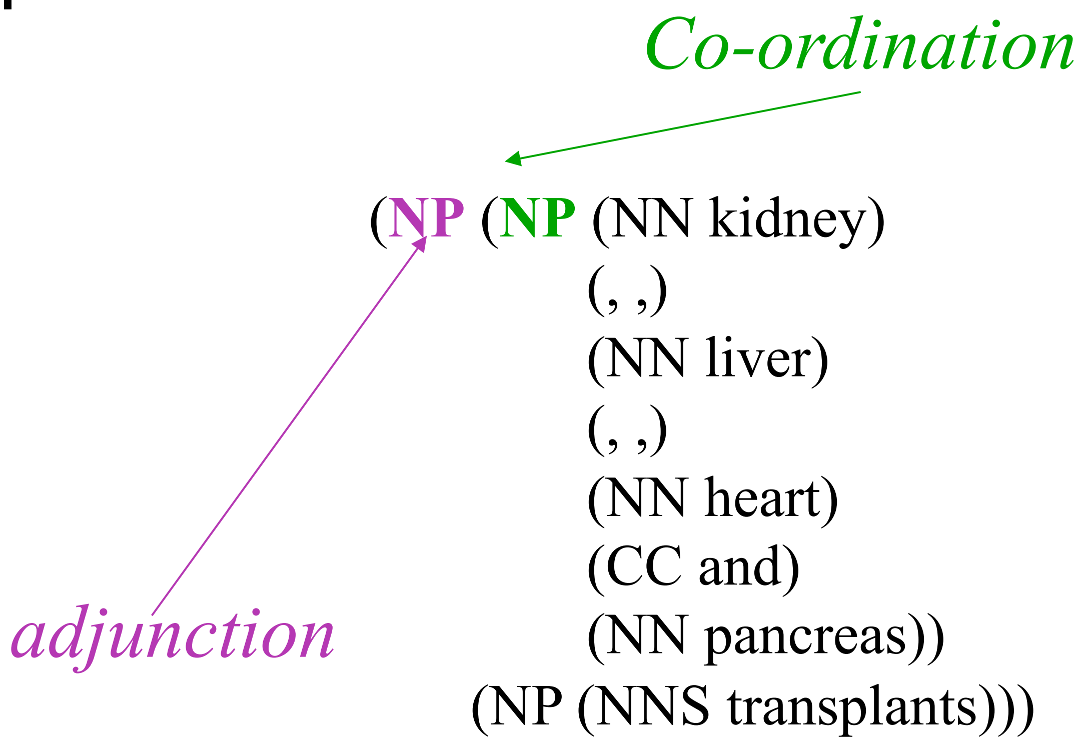
The Mortgage and equity last paid a dividend on August 1, 1988
real estate investment trust

A modification in the Chinese Treebank

One grammatical relation per bracket

Principles are hard to resist:

```
(NP (NN kidney)
        (, ,)
        (NN liver)
        (, ,)
        (NN heart)
        (CC and)
        (NN pancreas)
       (NNS transplants))
```

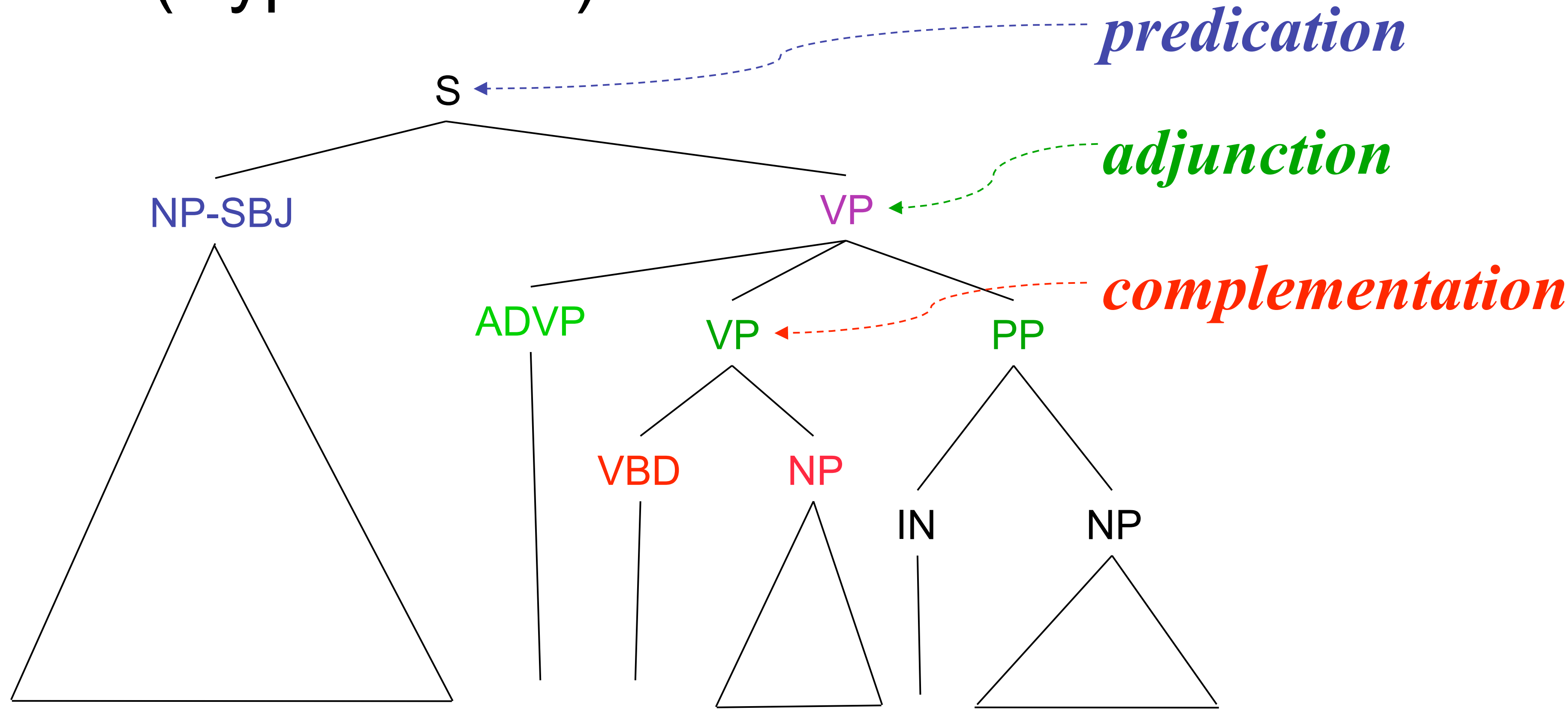


Co-ordination

adjunction

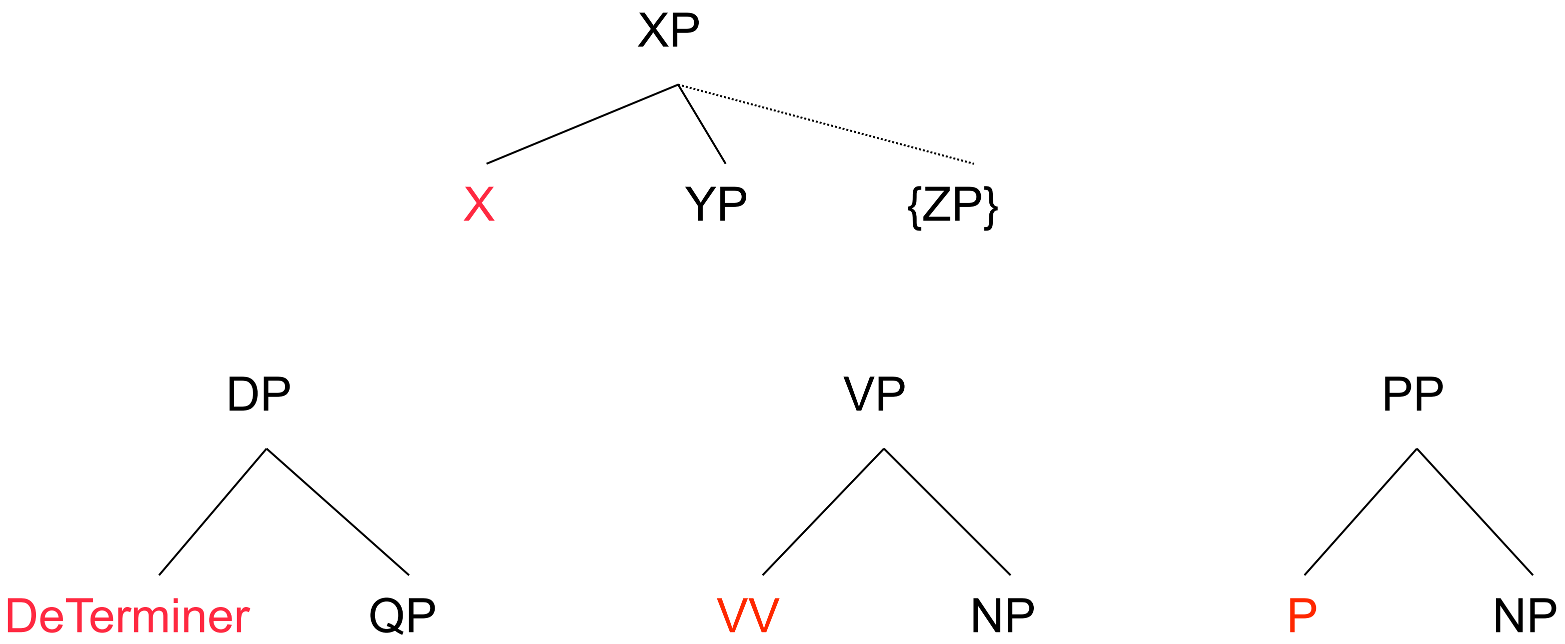
```
(NP (NP (NN kidney)
         (, ,)
         (NN liver)
         (, ,)
         (NN heart)
         (CC and)
         (NN pancreas))
     (NP (NNS transplants)))
```

(Hypothetical) CTB annotation



The Mortgage and equity real estate investment trust last paid a dividend on August 1, 1988

Complementation (left-headed)



XP

X YP {ZP}

DP

DeTerminer QP

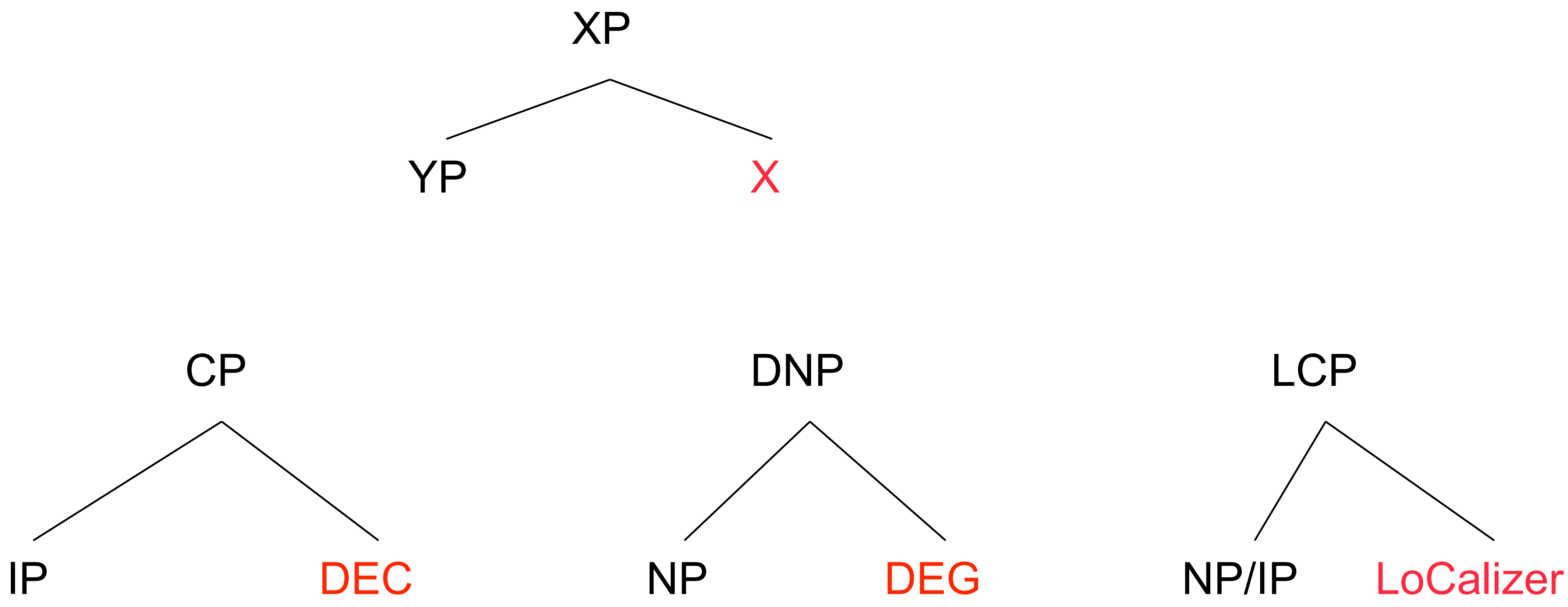
VP

VV NP

PP

P NP

Complementation (right-headed)



XP

YP X

CP

IP DEC

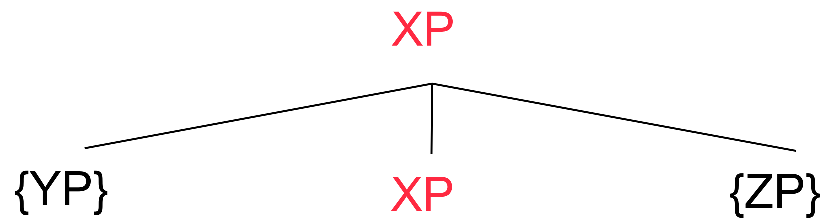
DNP

NP DEG

LCP

NP/IP LoCalizer

Adjunction

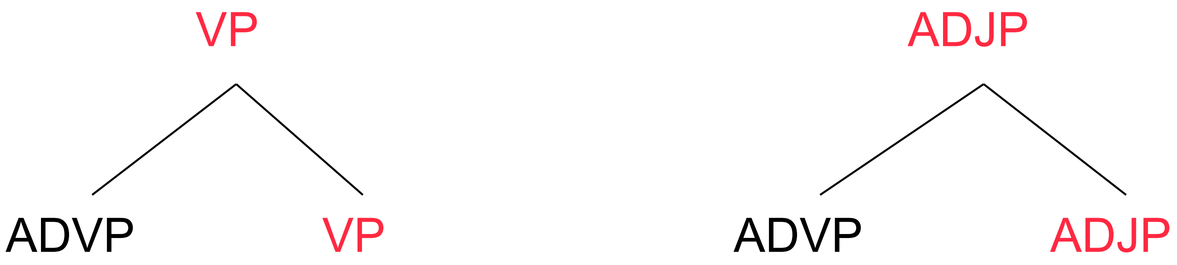


XP

{YP}

XP

{ZP}



VP

ADVP

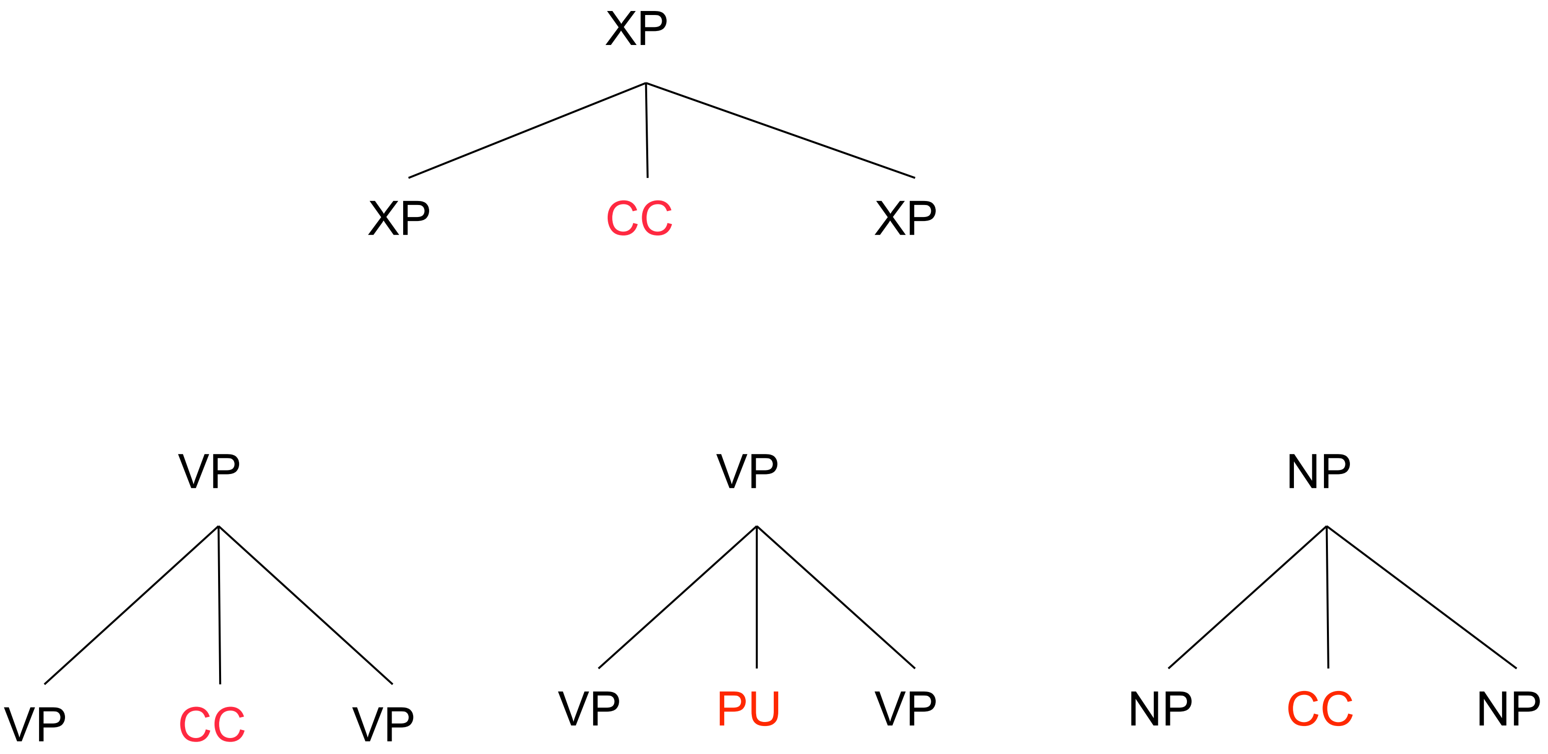
VP

ADJP

ADVP

ADJP

Coordination



XP
XP CC XP

VP
VP CC VP

VP
VP PU VP

NP
NP CC NP

Take-home points

- Treebanking has both linguistics and engineering aspects
- Treebanking has to be tailored to language-specific characteristics

References/Readings

- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Corpus of English: The Penn Treebank. Computational Linguistics. 19(2):313-330
- Mitchell Marcus , Grace Kim , Mary Ann Marcinkiewicz , Robert Macintyre , Ann Bies , Mark Ferguson , Karen Katz , Britta Schasberger. 1994. Penn Treebank: Annotating predicate argument structure. In ARPA Human Language Technology Workshop. Pages 114-119.
- Nianwen Xue, Fei Xia, Fu-Dong Chiou and Martha Palmer. 2005. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2):207-238.