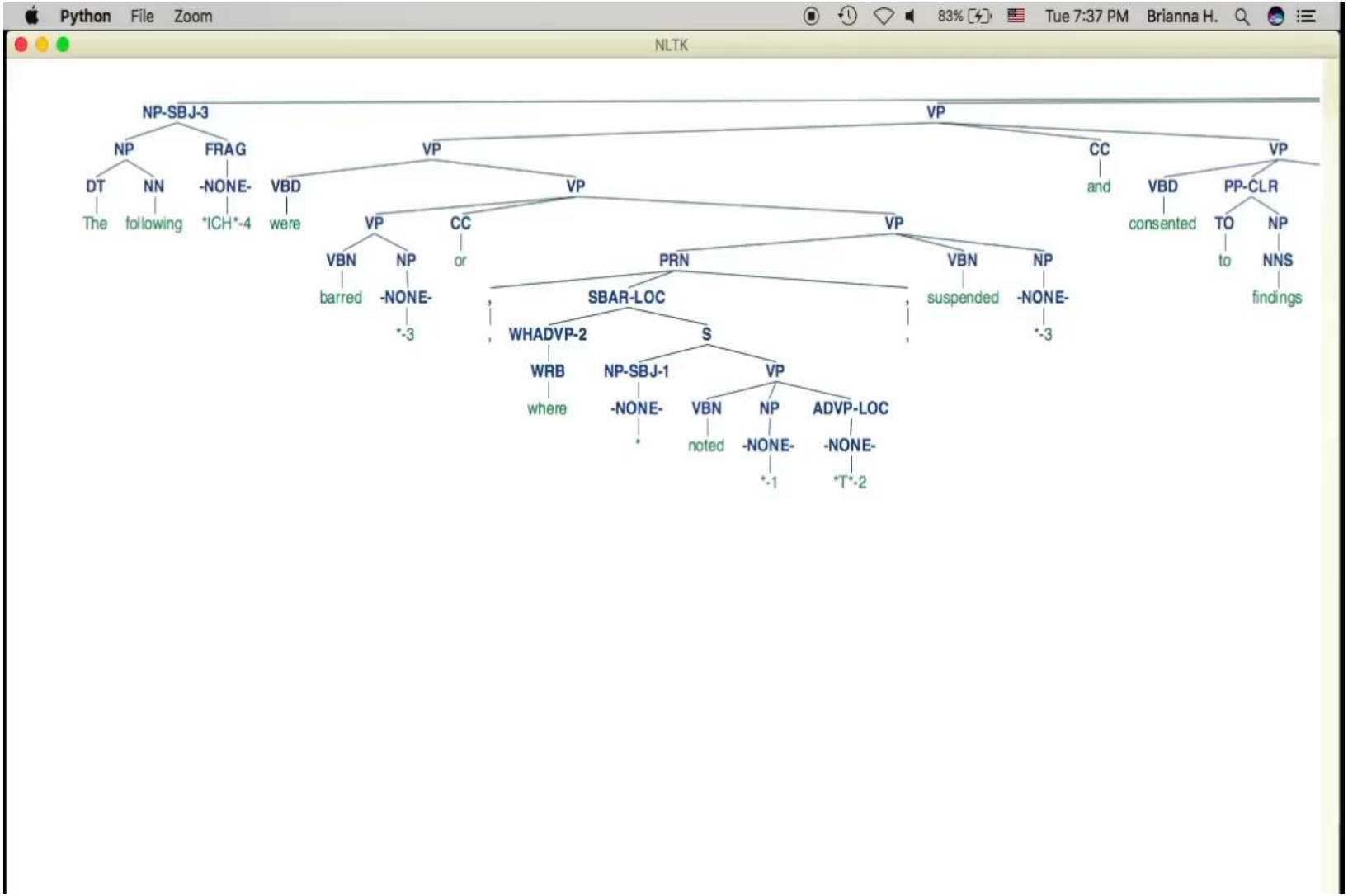
Introduction to treebanks


Python File Zoom
83% Tue 7:37 PM Brianna H.
NLTK
NP-SBJ-3
NP
FRAG
DT
NN
-NONE-
The
following
ICH-4
VP
VBD
were
VP
VP
CC
VP
VBN
NP
or
barred
-NONE-
*-3
PRN
VBN
NP
,
SBAR-LOC
,
suspended
-NONE-
*-3
,
WHADVP-2
S
WRB
NP-SBJ-1
VP
where
-NONE-
*
VBN
NP
ADVP-LOC
noted
-NONE-
-NONE-
*-1
T-2
VP
CC
VP
and
VBD
PP-CLR
consented
TO
NP
to
NNS
findings

Outline

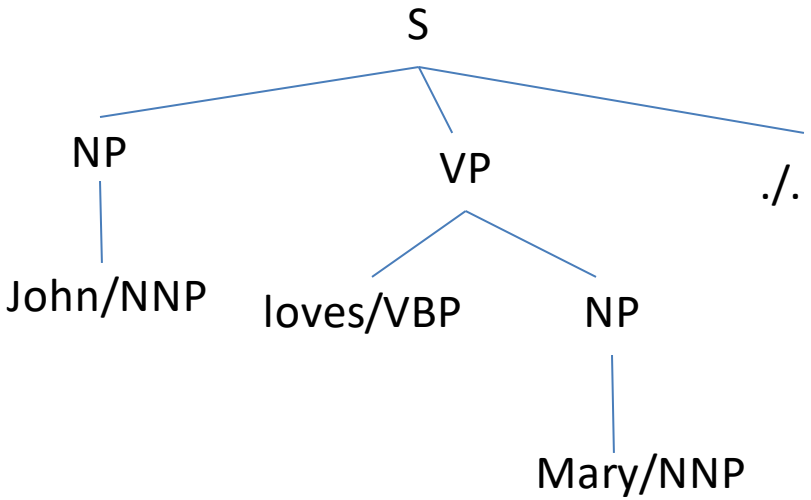
- Types of treebanks
  - (Syntactic) Treebank
  - PropBank
  - Discourse Treebank
- The English Penn Treebank
- Why do we need treebanks?

(Syntactic) Treebank

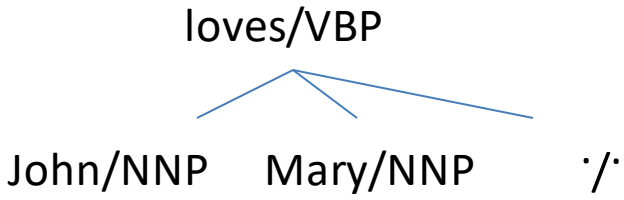
- Sentences annotated with syntactic structure (dependency structure or phrase structure)
- 1960s: Brown Corpus
- Early 1990s: The English Penn Treebank
- Late 1990s: Prague Dependency Treebank
- 1990s – now: Arabic, Chinese, Dutch, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Icelandic, Italian, Japanese, Korean, Latin, Norwegian, Polish, Spanish, Turkish, etc.

An example

- John loves Mary .



- (S (NP (NNP John))
  (VP (VBP loves)
  (NP (NNP Mary)))
  (. .))



PropBank

- Sentences annotated with predicate argument structure

- Ex: John loves Mary
  - "loves" is the predicate
  - "John" is Arg0 ("Agent")
  - "Mary" is Arg1 ("Theme")

- 2000s: The English PropBank, followed by the PropBanks for Chinese, Arabic, Hindi/Urdu, etc.

Discourse Treebank

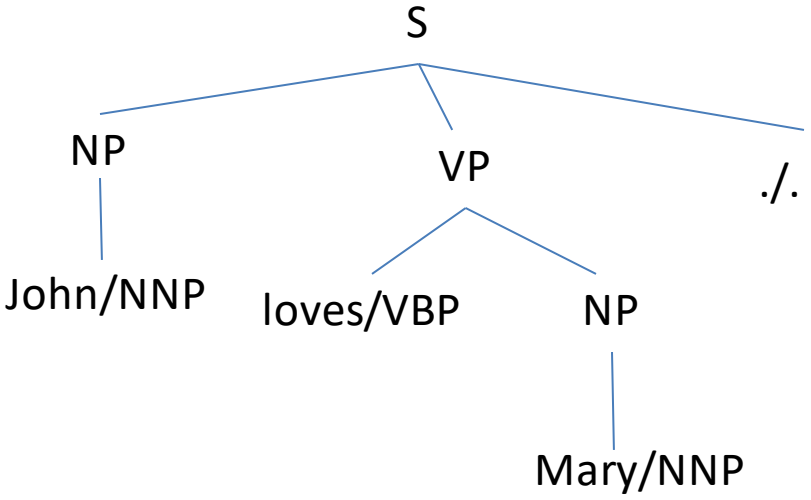
- 2006-2008: The English Discourse Treebank

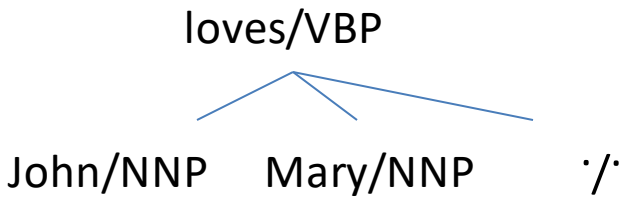
- The city's Campaign Finance Board has refused to pay Mr. Dinkins $95,142 in matching funds because his campaign records are incomplete.

Multi-representational, multi-layered treebank

- 2010-: Multi-representational, multi-layer Treebank for Hindi/Urdu

- The treebank includes both PS, DS, and PB.



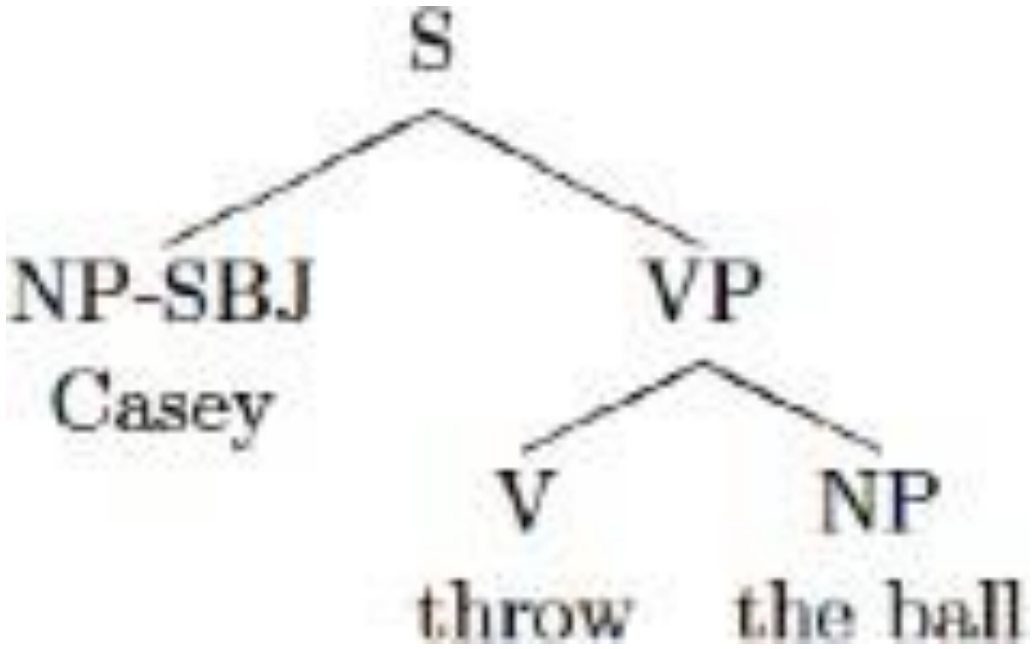


"loves" is predicate.
"John" is Arg0.
"Mary" is Arg1.

The English Penn Treebank (PTB)

- Developed at UPenn in early 1990s
- Most commonly used treebank in the CL field
- Data:
  - WSJ: 1-million words from 1987 to 1989
  - Others: Brown Corpus, ATIS, etc.
- Release:
  - 1992: version 1
  - 1995: version 2
  - 1999: version 3

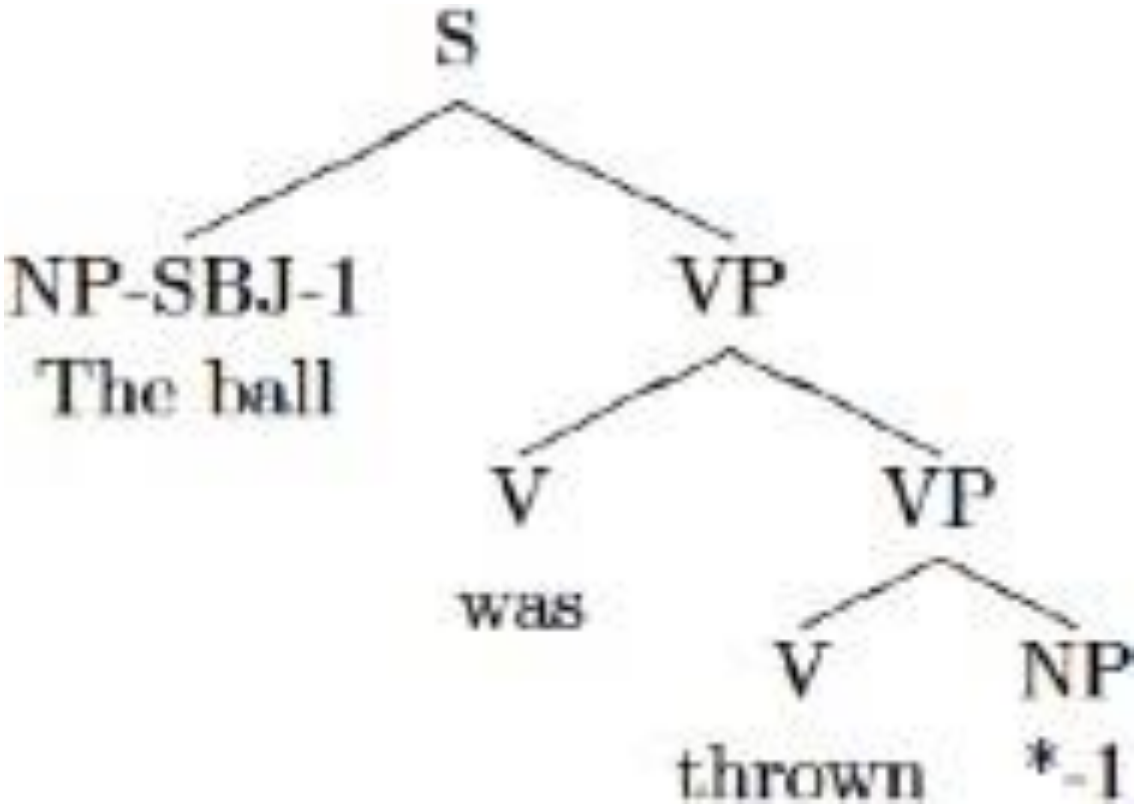
An example



The PTB Tagset

- Syntactic labels: e.g., NP, VP
- Function tags: e.g., -SBJ, -LOC
- Empty categories (ECs): e.g., *T* (for A-bar movement)
- Sub-categories for ECs: e.g., 0 (zero complementizers), NP* (PRO, A-movement)

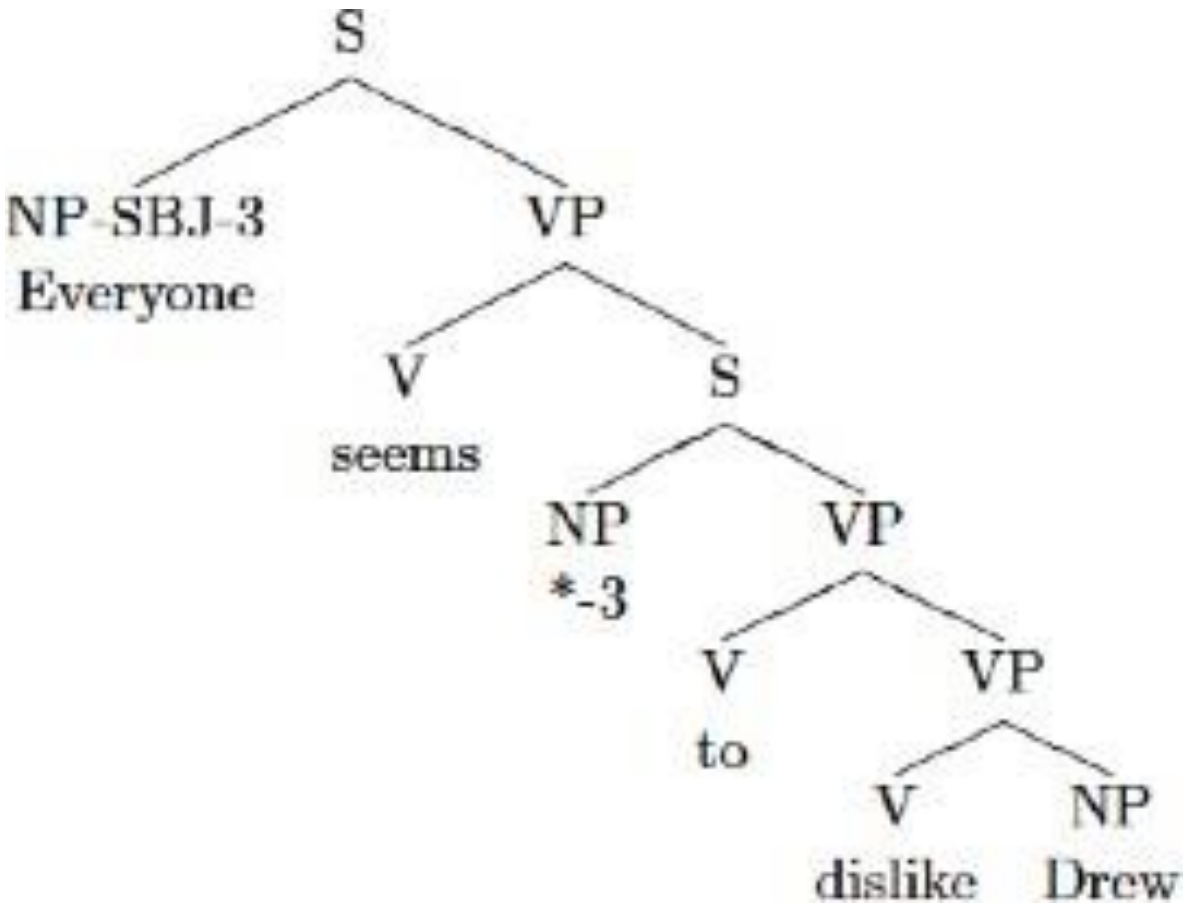
Passive



Clausal Complementation

```
(S (NP-SBJ he)
   (VP wrote
       (SBAR that
             (S (NP-SBJ he)
                (VP had
                    (VP given (PRT up)
                        (NP hope
                            (SBAR 0
                                  (S (NP-SBJ they)
                                     (VP would
                                         (ADVP-TMP ever)
                                         (VP agree
                                             (PP-CLR on
                                                     (NP anything)))))))))))))
```

Raising



Wh-Relative Clauses

```
(NP (NP answers)
    (SBAR (WHNP-6 that/which)
          (S (NP-SBJ-3 we)
             (VP 'd
                 (VP like
                     (S (NP-SBJ *-3)
                        (VP to
                            (VP have
                                (NP *T*-6)))))))))
```

Contact Relatives

```
(NP (NP answers)
    (SBAR (WHNP-3 0)
          (S (NP-SBJ-4 we)
             (VP 'd
                 (VP like
                     (S (NP-SBJ *-4)
                        (VP to
                            (VP have
                                (NP *T*-3)))))))))
```

Indirect Questions

```
( (S (NP-SBJ I)
     (VP forgot
         (SBAR (WHNP who)
               (S (NP-SBJ they)
                  (VP said
                      (SBAR (WHNP-2 0)
                            (S (NP-SBJ-1 they)
                               (VP wanted
                                   (S (NP-SBJ *-1)
                                      (VP to
                                          (VP hire
                                              (NP *T*-2)))))))))))
     .))
```

Punctuation

```
( (S (SBAR-ADV If
                (S (NP-SBJ-1 the judge)
                   (VP is
                       (VP impeached
                           (NP *-1)))))
     ,
     (SBAR-ADV as
                (S (NP-SBJ-2 *)
                   (VP is
                       (VP thought
                           (S (NP-SBJ *-2)
                              (ADJP-PRD likely))))))
     ,
     (NP-SBJ-3 he)
     (VP will
         (VP be
             (VP removed
                 (NP *-3)
                 (PP-DIR from
                         (NP office))
                 (ADVP-TMP immediately))))
     .))
```

FinancialSpeak

```
(S (NP-SBJ Copper)
   (VP finished
       (ADVP-CLR down
                 (NP 4.5 cents))
       ,
       (PP-CLR at
               (NP (NP $ 1.2345 *U*)
                   (NP-ADV a pound))))
   .)
```

Lists 1

```
( (S (NP-SBJ-1 It)
     (VP was
         (VP used
             (NP *-1)
             (S-CLR (NP-SBJ *)
                    (VP (VP (LST -LRB-
                                 1
                                 -RRB-)
                            to
                            (VP investigate
                                (NP wave behavior)))
                        ,
                        (VP (LST -LRB-
                                 2
                                 -RRB-)
                            to
                            (VP estimate
                                (NP the wave energy)))
                        ,
                        and
                        (VP (LST -LRB-
                                 3
                                 -RRB-)
                            forecast
                            (NP coastal changes))))))
     .))
```

Lists 2

```
( (S (NP-SBJ The aged care plan)
     (VP carries
         (NP these benefits)
         (PP for
             (NP (NP persons)
                 (PP over
                     (NP 65)))))
     :))
( (NP (LST 1)
      (NP Full payment)
      (PP of
          (NP (NP hospital bills)
              (PP for
                  (NP (NP stays)
                      (NP (QP up to 90) days)))))
      .))
( (NP (LST 2)
      (NP Full payment)
      (PP of
          (NP nursing home bills))
      (PP-TMP for
              (NP (NP (QP up to 180) days)
                  (PP-TMP following
                          (NP (NP discharge)
                              (PP from
                                  (NP a hospital))))))
      .))
( (NP (LST 3)
      (NP Hospital outpatient clinic diagnostic service)
      (PP for
          (NP (NP all costs)
              (PP in
                  (NP (NP excess)
                      (PP of
                          (NP (NP $ 20)
                              (NP-ADV a patient)))))))
      .))
```

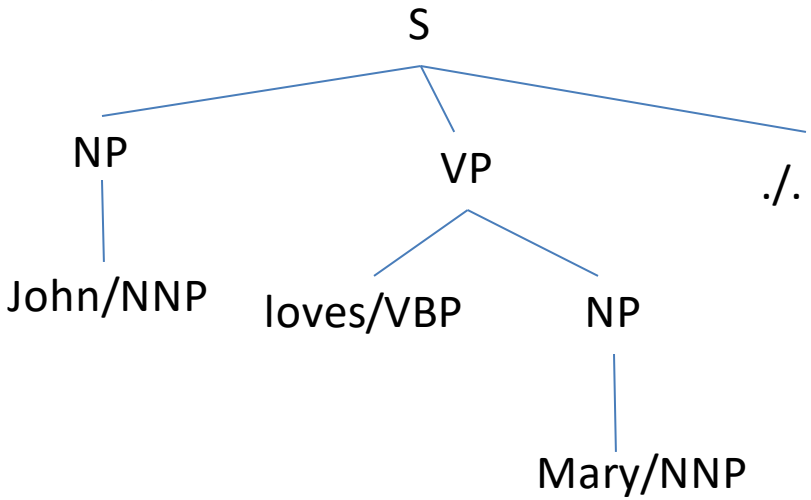
Why do we need treebanks?

- Computational Linguistics: (Session 6-7)
  - To build and evaluate NLP tools (e.g., word segmenters, part-of-speech taggers, parsers, semantic role labelers)
  - This leads to significant progress of the CL field
- Theoretic linguistics: (Session 2 and 5-6)
  - Annotation guidelines are like a grammar book, with more detail and coverage
  - As a discovery tool
  - One can test linguistic theories and collect statistics by searching treebanks.

CL example: Parsing

Input: John loves Mary .

Output:



S => NP VP .

NP => NNP

VP => VBP NP

NNP => John

NNP => Mary

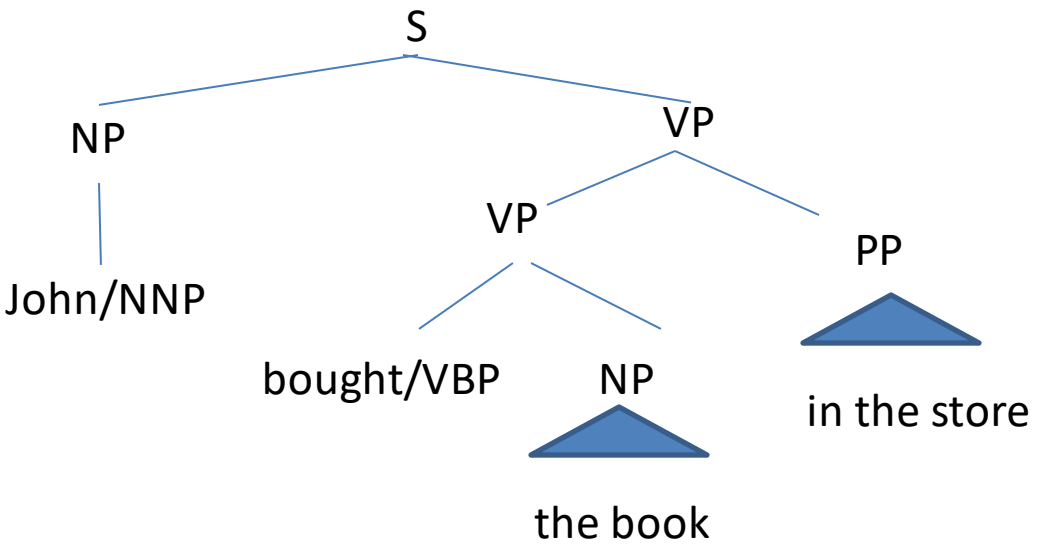
VBP => loves

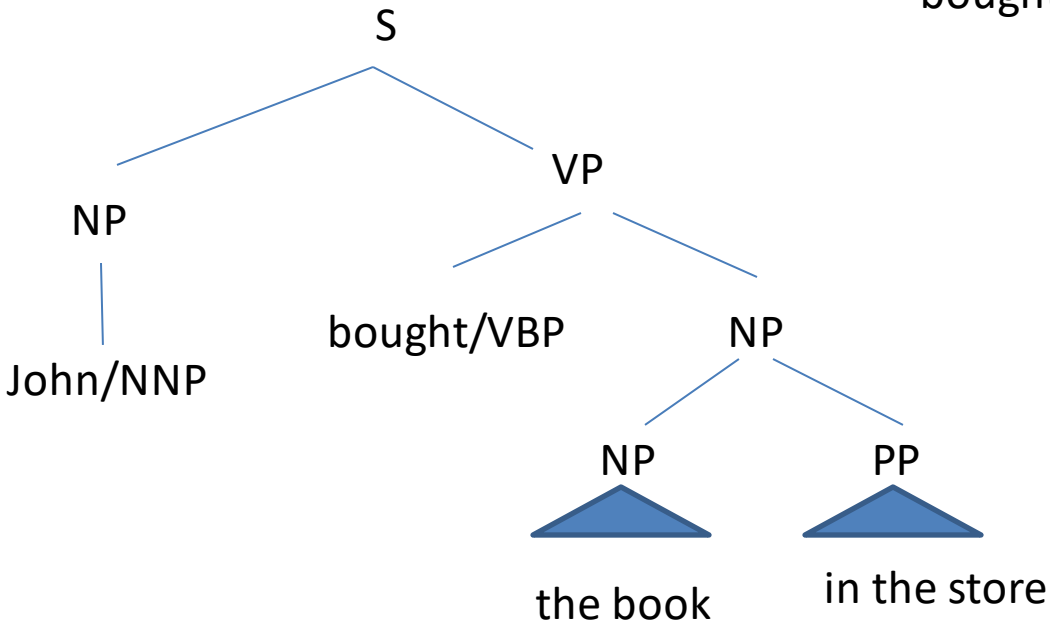
. => .

Ambiguity

PP attachment: John bought the book in the store

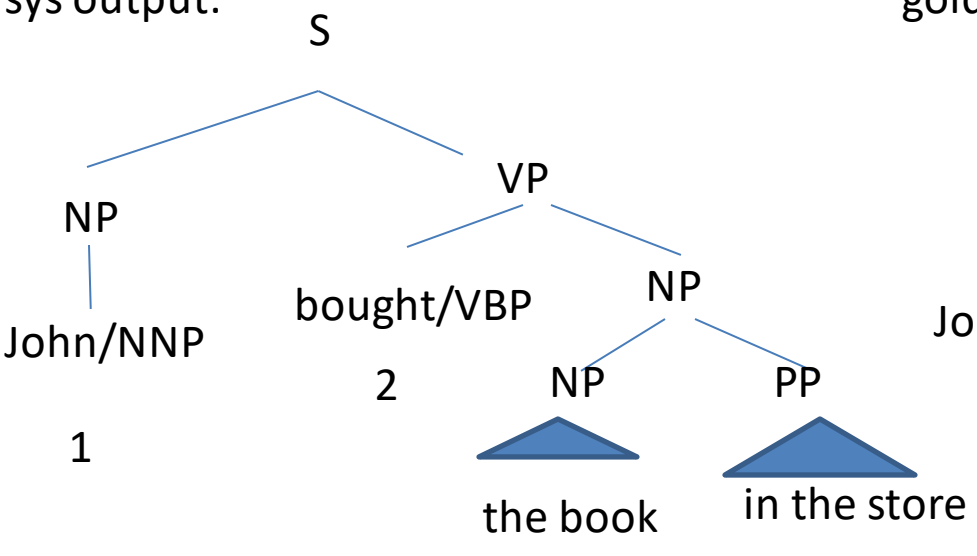
S => NP VP
NP => PN
VP => V NP
VP => VP PP
NP => NP PP
PP => P NP





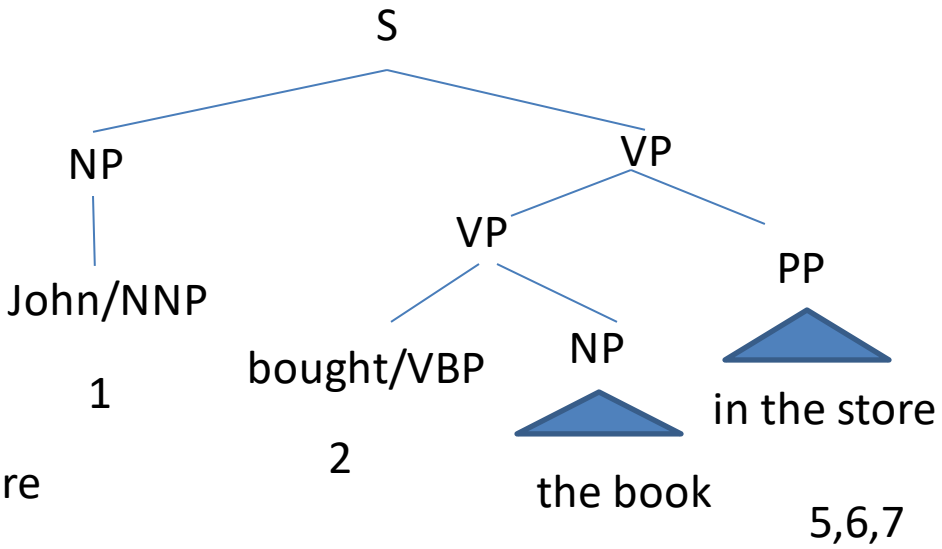
Labeled f-score

sys output:



(1, 7, S)
(1, 1, NP)
(2, 7, VP)
(3, 7, NP)
(3, 4, NP)
(5, 7, PP)
(6, 7, NP)

gold standard:



(1, 7, S)
(1, 1, NP)
(2, 7, VP)
(2, 4, VP)
(3, 4, NP)
(5, 7, PP)
(6, 7, NP)

Prec=6/7, recall=6/7, f-score=6/7

Parsing evaluation

- Evaluation:
  - precision, recall, f-score
  - Best f-score: around 91%

Thankyou