# The Penn Treebank


**Erhard W. Hinrichs**

**SfS-CL**

**Eberhard-Karls-Universität Tübingen**

# Penn treebank (as of 11/92)

| Description | POS Tagged | Parsed |
|---|---|---|
| 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 tokens:** | 4,885,798 | 2,881,188 |

# The Penn Treebank POS tagset (1)

- 36 POS tags
- 12 other tags (for punctuation and currency symbols)
- smaller than other tagset for English

| | |
|---|---|
| Brown Corpus | 87 |
| Lancaster-Oslo/Bergen (LOB) | 135 |
| Lancaster UCREL | 165 |
| London-Lund Corpus of Spoken English | 197 |

- motivation for smaller tagset: counteract data sparseness
- eliminate POS dstinctions that are syntactically recoverable (e.g., no separate tags for main verb inflections on have and do

# The Penn Treebank POS tagset (2)

| | | | | | |
|---|---|---|---|---|---|
| 1. | CC | Coord. conjunc. | 25. | TO | *to* |
| 2. | CD | Cardinal number | 26. | UH | Interjection |
| 3. | DT | Determiner | 27. | VB | V, base form |
| 4. | EX | Existential *there* | 28. | VBD | V, past tense |
| 5. | FW | Foreign word | 29. | VBG | V, gerund/pres. part. |
| 6. | IN | Prep./subord. conj. | 30. | VBN | V, past part. |
| 7. | JJ | Adject. | 31. | VBP | V, non-3rd ps. sing. pres. |
| 8. | JJR | Adject., comp. | 32. | VBZ | V, 3rd ps. sing. pres. |
| 9. | JJS | Adject., superl. | 33. | WDT | *wh*-det. |
| 10. | LS | List item marker | 34. | WP | *wh*-pronoun |
| 11. | MD | Modal | 35. | WP$ | Poss. *wh*-pronoun |

# The Penn Treebank POS tagset (3)

| | | | | | |
|---|---|---|---|---|---|
| 12. | NN | Noun, sing. or mass | 36. | WRB | *wh*-adverb |
| 13. | NNS | Noun, plural | 37. | # | Pound sign |
| 14. | NNP | Proper noun, sing. | 38. | $ | Dollar sign |
| 15. | NNPS | Proper noun, plural | 39. | . | Sent.-final punct. |
| 16. | PDT | Predeterminer | 40. | , | Comma |
| 17. | POS | Possessive ending | 41. | : | Colon, semi-colon |
| 18. | PRP | Personal pronoun | 42. | ( | L. bracket char. |
| 19. | PP$ | Poss. pronoun | 43. | ) | R. bracket char. |
| 20. | RB | Adverb | 44. | 〃 | Straight dbl. quote |
| 21. | RBR | Adverb, comp. | 45. | ‘ | L. open sngl. quote |
| 22. | RBS | Adverb, superl. | 46. | “ | L. open dbl. quote |
| 23. | RP | Particle | 47. | ’ | R. close sngl. quote |
| 24. | SYM | Symbol | 48. | ” | R. close dbl. quote |

# POS Annotation Strategy

Consistent syntactic function:

| Example: | Brown POS Tag | Penn Tag |
|---|---|---|
| *the one* | CD | NN |
| *the ones* | NNS | NNS |
| *both the boys* | ABX | PDT |
| *the boys both* | ABX | RB |
| *both of the boys* | ABX | NNS |
| *both boys and girls* | ABX | CC |

# Declarative Clauses

```
(S (NP-SBJ I)
    (VP consider
        (S (NP-SBJ Kris)
           (NP-PRD a fool))))
```

- SBJ, PRD are grammatical functions (GFs)
- small clauses are sentential

# Yes-No Questions

```
(SQ Was
    (NP-SBJ he)
    (ADVP-TMP ever)
    (ADJP-PRD successfull)
    ?)
```

- SQ mark inverted auxiliary structures
- TMP (for: temporal) a GF

# Wh-Questions

```
(SBARQ (WHNP-1 What
         (SQ is
             (NP-SBJ Tim)
             (VP eating
                 (NP *T*-1)))
         ?)
```

- SBARQ to mark WH-questions
- WH-prefixed labels:
  - WHNP, WHADVP, WHPP
  - co-indexed trace

# Passives

```
(S (NP-SBJ-1 The ball
   (VP was
       (VP thrown
           (NP *-1)
           (PP by
               (NP-LGS Chris)))))
```

- surface subject is tagged -SBJ,
- passive trace inserted after the verb,

# Control Structures

```
(S (NP-SBJ-1 Chris)
   (VP wants
       (S (NP-SBJ *-1)
          (VP to
              (VP throw
                  (NP the ball))))))
```

# Sentence Types

**S** — Simple declarative clause, i.e. one that is not introduced by a (possibly empty) subordinating conjunction or *wh*-word and that does not exhibit subject-verb inversion.

**SBAR** — Clause introduced by a (possibly empty) subordinating conjunction.

**SBARQ** — Direct question introduced by a *wh*-word or *wh*-phrase. should be bracketed as SBAR, not SBARQ.

**SINV** — Inverted declarative sentence, i.e. one in which the subject follows the tensed verb or modal.

**SQ** — Inverted yes/no question, or main clause of a *wh*-question, following the *wh*-phrase in SBARQ.

# Grammatical Roles (1)



**-BNF (benefactive)**

**-DTV (dative)**

**-LGS (logical subject)**

**-PRD (predicate)**

**-PUT** (locative PP of put)

**-SBJ (surface subject)**

**-TPC ("topicalized")**

**-VOC (vocative)**

**-DIR (direction)**

**-EXT (extent)**

# Grammatical Roles (2)





**-LOC (locative)**

**-MNR (manner)**

**-PRP (purpose or reason)**

**-TMP (temporal)**

**-CLR (closely related)**

**-CLF (cleft)**

**-HLN (headline)**

**-TTL (title)**

# Headline (HLN)

```
((NP-HLN (NP-LOC Chicago , IL)
                ,
                (NP-TMP May 8)
                --))
((S A fire broke out in an abandoned building .))
```

# DTV

marks the dative object in the unshifted form of the double object construction.

```
(S (NP-SBJ Aristotle)
   (VP gave
       (NP the book)
       (PP-DTV to
               (NP Plato))))
```

Compare with the shifted

```
(S (NP-SBJ Aristotle)
   (VP gave
       (NP Plato)
       (NP the book)))
```

# Empty Elements

***T*** (trace of A′-movement, including parasitic gaps)

**(NP *)** (arbitrary PRO, controlled PRO, and trace of A-movement)

**0** (null complementizer, including null *wh*-operator)

***U*** (unit)

***?*** (placeholder for ellipsed material)

***NOT*** (anti-placeholder in template gapping)

***RNR*** (pseudo-attach: right node raising)

***ICH*** (pseudo-attach: interpret constituent here)

***EXP*** (pseudo-attach: expletive)

***PPA*** (pseudo-attach: permanent predictable ambiguity)

# A′-movement

```
(SBARQ (WHADVP-439 Where)
       (SQ did
           (NP-SBJ you)
           (VP put
               (NP the book)
               (ADVP-PUT *T*-439)))
       ?)
```

# Extraposition

```
(S (NP-SBJ Chris)
   (VP knew
       (SBAR *ICH*-1)
       (NP-TMP yesterday)
       (SBAR-1 that
               (S (NP-SBJ Terry)
                  (VP would
                      (VP catch
                          (NP the ball)))))))
```

# Expletive *it*

```
(S (NP-SBJ (NP It)
           (S *EXP*-1))
   (VP is
       (NP a pleasure))
   (S-1 (NP-SBJ *)
        (VP to
            (VP teach
                (NP her)))))
```

# Unit (*U*)

```
(NP a
    (ADJP (QP $ 200 million) *U*)
    contract)
```

# Empty Comp. and Ellipsis

```
(S (NP-SBJ I)
   (VP believe
       (SBAR 0
             (S (NP-SBJ you)
                (VP are
                    (ADJP-PRD *?*))))))
```

# Right Node Raising



```
(S But
   (NP-SBJ-2 our outlook)
   (VP (VP has
           (VP been
               (ADJP *RNR*-1)))
       ,
       and
       (VP continues
           (S (NP-SBJ *-2)
              (VP to
                  (VP be
                      (ADJP *RNR*-1)))))
       ,
       (ADJP-1 defensive)))
```

# Attachment Ambiguity

```
(S (NP-SBJ I)
       (VP saw
           (NP (NP the man)
               (PP *PPA*-1))
           (PP-CLR-1 with
                     (NP the telescope))))
```

- PPA (for: *permanent predictable ambiguity*) marks an alternative attachment site for the PP
- CLR (for: *closely related*) GF for constituents whose argument/adjunct status is not clear.