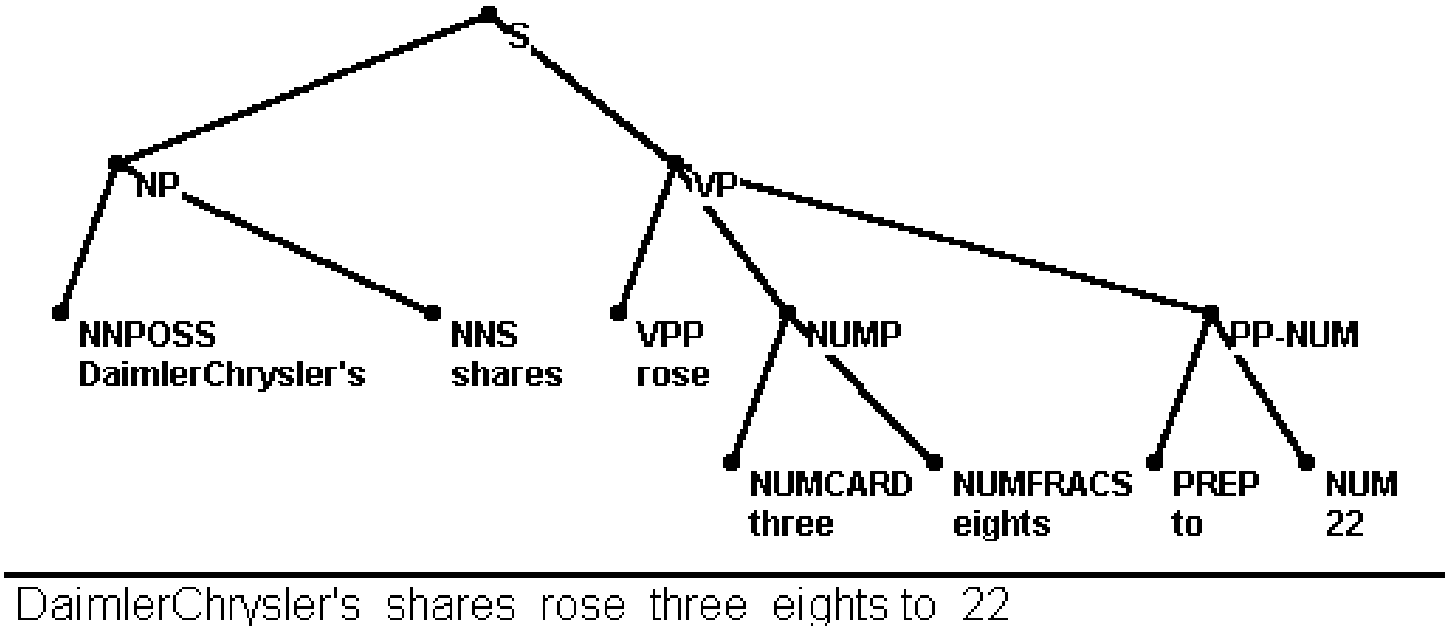
# Treebanks, Treebanking and Evaluation

# Phrase Structure Tree

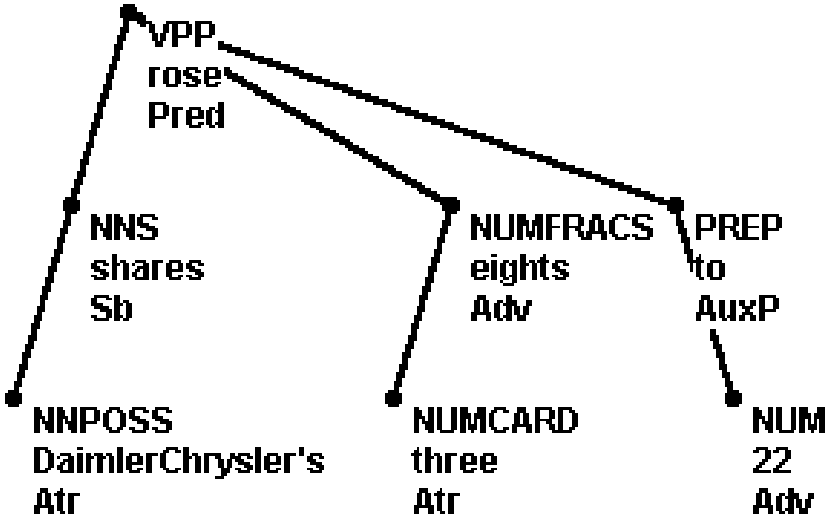
- Example:




((DaimlerChrysler's shares)$_{NP}$ (rose (three eights)$_{NUMP}$ (to 22)$_{PP\text{-}NUM}$ )$_{VP}$ )$_{S}$

# Dependency Tree

- Example:




DaimlerChrysler's shares rose three eights to 22

$rose_{Pred}(shares_{Sb}(DaimlerChrysler's_{Atr}),eights_{Adv}(three_{Atr}),to_{AuxP}(22_{Adv}))$

# Parser Development

- Use training data for learning phase
  - segment as needed (e.g., for heldout)
  - use all for
    - manually written rules (seldom today)
    - automatically learned rules/statistics
- Occasionally, test progress on Development Test Set
  - (simulates real-world data)
- When done, test on Evaluation Test Set
- ***Unbreakable Rule #1: Never look at Evaluation Test Data (not even indirectly, e.g. performance numbers)***

# Evaluation

- Evaluation of parsers (regardless of whether manual-rule-based or automatically learned)
- Repeat: Test against Evaluation Test Data
- Measures:
  - Dependency trees:
    - Dependency Accuracy, Precision, Recall
  - Parse trees:
    - Crossing brackets
    - Labeled precision, recall [F-measure]

# Dependency Parser Evaluation

- Dependency Recall:
  - $R_D$ = Correct(D) / |S|
    - Correct(D): number of correct dependencies
      - correct: word attached to its true head
      - Tree root is correct if marked as root
    - |S| - size of test data in words (since |dependencies| = |words|)
- Dependency precision (if output not a tree, partial):
  - $P_D$ = Correct(D) / Generated(D)
    - Generated(D) is the number of dependencies output
      - some words without a link to their head
      - some words with several links to (several different) heads

# Phrase Structure (Parse Tree) Evaluation

- Crossing Brackets measure
  - Example "truth" (evaluation test set):
    - ((the ((New York) - based company)) (announced (yesterday)))
  - Parser output - 0 crossing brackets:
    - ((the New York - based company) (announced yesterday))
  - Parser output - 2 crossing brackets:
    - (((the New York) - based) (company (announced (yesterday))))
- Labeled Precision/Recall:
  - Usual computation using bracket labels (phrase markers)

    T: $((\text{Computers})_{NP}\ (\text{are down})_{VP})_{S} \leftrightarrow$ P: $((\text{Computers})_{NP}\ (\text{are}\ (\text{down})_{NP})_{VP})_{S}$
    - Recall = 100%, Precision = 75%