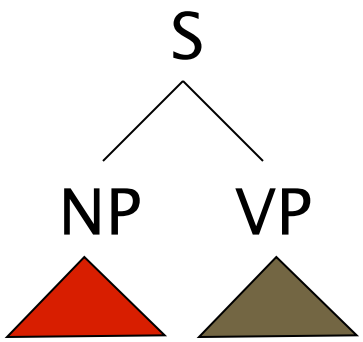
PCFG Independence Assumptions

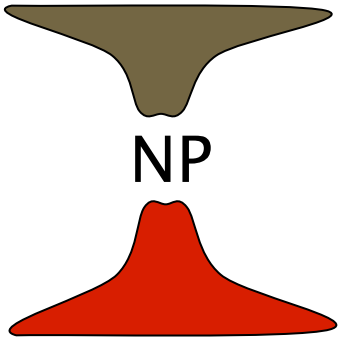
PCFGs and Independence

- The symbols in a PCFG define independence assumptions:

S → NP VP

NP → DT NN

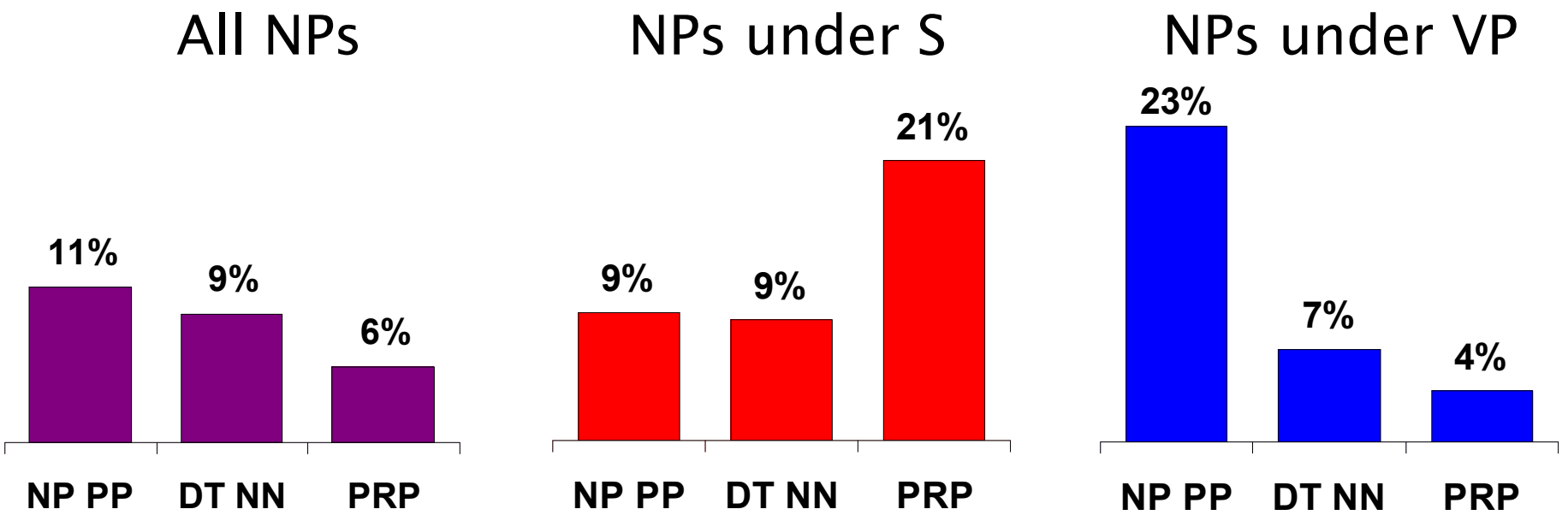




- At any node, the material inside that node is independent of the material outside that node, given the label of that node
- Any information that statistically connects behavior inside and outside a node must flow through that node's label

Non-Independence I

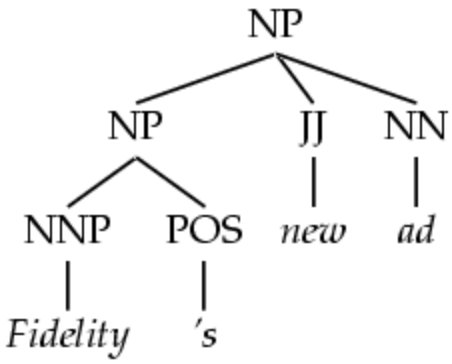
- The independence assumptions of a PCFG are often too strong

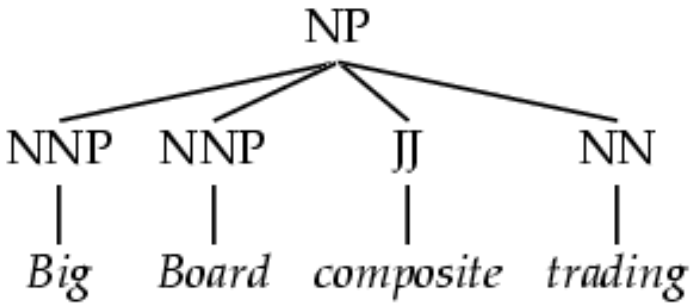


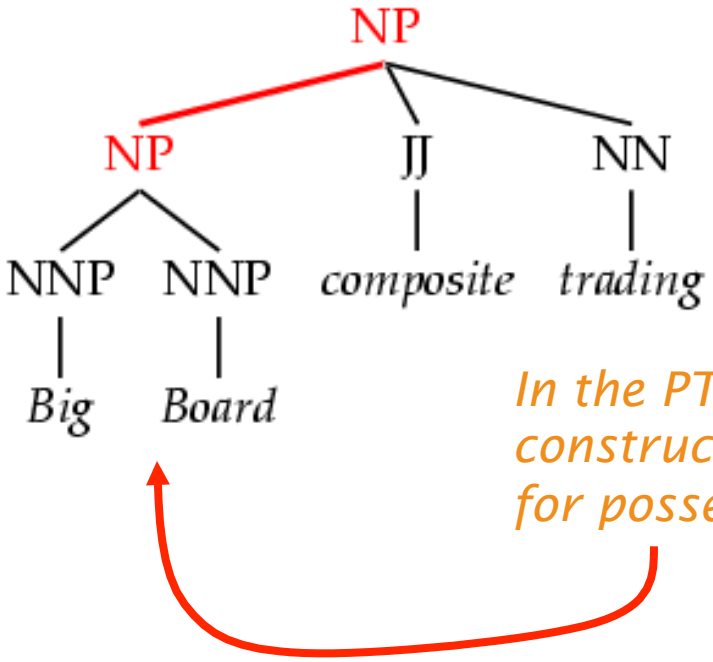
- Example: the expansion of an NP is highly dependent on the parent of the NP (i.e., subjects vs. objects)

Non-Independence II

- Symptoms of overly strong assumptions:
  - Rewrites get used where they don't belong



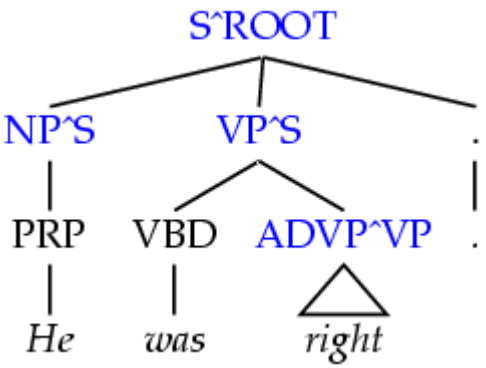




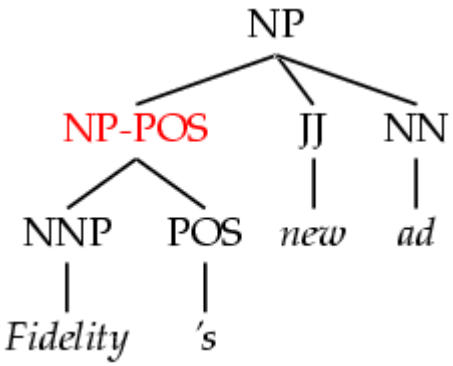
Refining the Grammar Symbols

- We can relax independence assumptions by encoding dependencies into the PCFG symbols, by state splitting:

Parent annotation
[Johnson 98]



Marking
possessive NPs



- Too much state-splitting ➔ sparseness (no smoothing used!)
- What are the most useful features to encode?

PCFG Independence Assumptions