CFGs and PCFGs

(Probabilistic)
Context-Free
Grammars

A phrase structure grammar

S → NP VP

VP → V NP

VP → V NP PP

NP → NP NP

NP → NP PP

NP → N

NP → *e*

PP → P NP

N → people

N → fish

N → tanks

N → rods

V → people

V → fish

V → tanks

P → with

people fish tanks

people fish with rods

Christopher Manning

Phrase structure grammars = context-free grammars (CFGs)

- G = (T, N, S, R)
  - T is a set of terminal symbols
  - N is a set of nonterminal symbols
  - S is the start symbol ($S \in N$)
  - R is a set of rules/productions of the form $X \rightarrow \gamma$
    - $X \in N$ and $\gamma \in (N \cup T)^*$

- A grammar G generates a language L.

Phrase structure grammars in NLP

- $G = (T, C, N, S, L, R)$
  - T is a set of terminal symbols
  - C is a set of preterminal symbols
  - N is a set of nonterminal symbols
  - S is the start symbol ($S \in N$)
  - L is the lexicon, a set of items of the form $X \rightarrow x$
    - $X \in P$ and $x \in T$
  - R is the grammar, a set of items of the form $X \rightarrow \gamma$
    - $X \in N$ and $\gamma \in (N \cup C)^*$
- By usual convention, S is the start symbol, but in statistical NLP, we usually have an extra node at the top (ROOT, TOP)
- We usually write *e* for an empty sequence, rather than nothing

A phrase structure grammar

S → NP VP
VP → V NP
VP → V NP PP
NP → NP NP
NP → NP PP
NP → N
NP → *e*
PP → P NP

N → *people*
N → *fish*
N → *tanks*
N → *rods*
V → *people*
V → *fish*
V → *tanks*
P → *with*

people fish tanks

people fish with rods

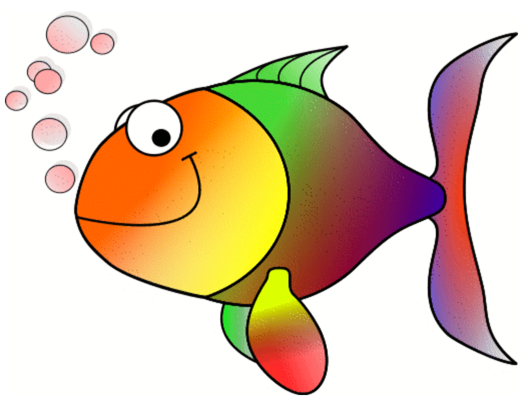
Probabilistic – or stochastic – context-free grammars (PCFGs)

- G = (T, N, S, R, P)
  - T is a set of terminal symbols
  - N is a set of nonterminal symbols
  - S is the start symbol (S ∈ N)
  - R is a set of rules/productions of the form $X \to \gamma$
  - P is a probability function
    - P: R $\to$ [0,1]
    - $\forall X \in N, \sum_{X \to \gamma \in R} P(X \to \gamma) = 1$
- A grammar G generates a language model L.

$$\sum_{\gamma \in T*} P(\gamma) = 1$$

A PCFG

S → NP VP	1.0	N → *people*	0.5
VP → V NP	0.6	N → *fish*	0.2
VP → V NP PP	0.4	N → *tanks*	0.2
NP → NP NP	0.1	N → *rods*	0.1
NP → NP PP	0.2	V → *people*	0.1
NP → N	0.7	V → *fish*	0.6
PP → P NP	1.0	V → *tanks*	0.3
		P → *with*	1.0



[With empty NP removed so less ambiguous]

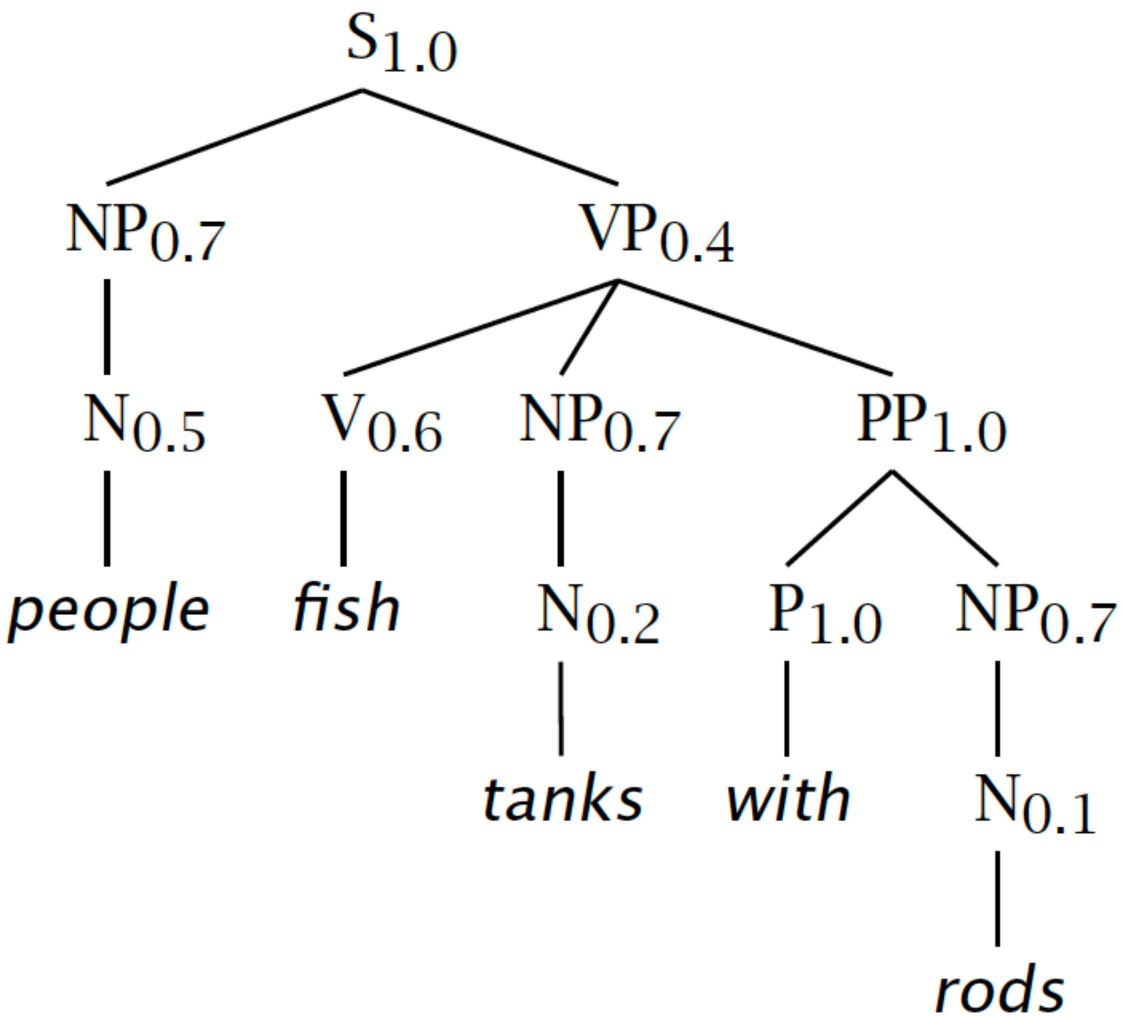
The probability of trees and strings

- P(*t*) – The probability of a tree *t* is the product of the probabilities of the rules used to generate it.
- P(*s*) – The probability of the string *s* is the sum of the probabilities of the trees which have that string as their yield

$$P(s) = \Sigma_j P(s, t) \text{ where } t \text{ is a parse of } s$$

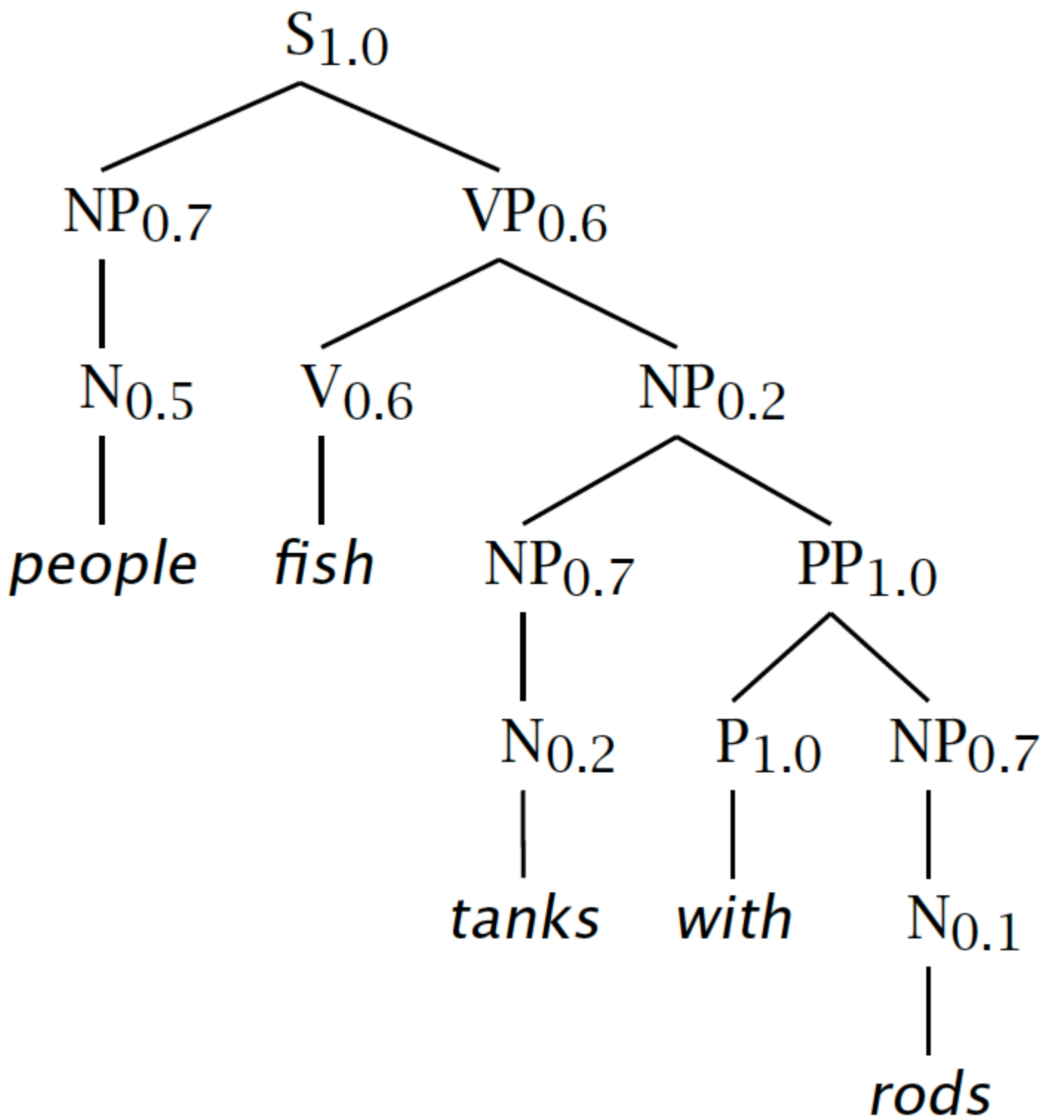
$$= \Sigma_j P(t)$$

t_1:



$S_{1.0}$
$NP_{0.7}$
$VP_{0.4}$
$N_{0.5}$
$V_{0.6}$
$NP_{0.7}$
$PP_{1.0}$
people
fish
$N_{0.2}$
$P_{1.0}$
$NP_{0.7}$
tanks
with
$N_{0.1}$
rods

t_2:


S1.0
NP0.7
VP0.6
N0.5
V0.6
NP0.2
people
fish
NP0.7
PP1.0
N0.2
P1.0
NP0.7
tanks
with
N0.1
rods

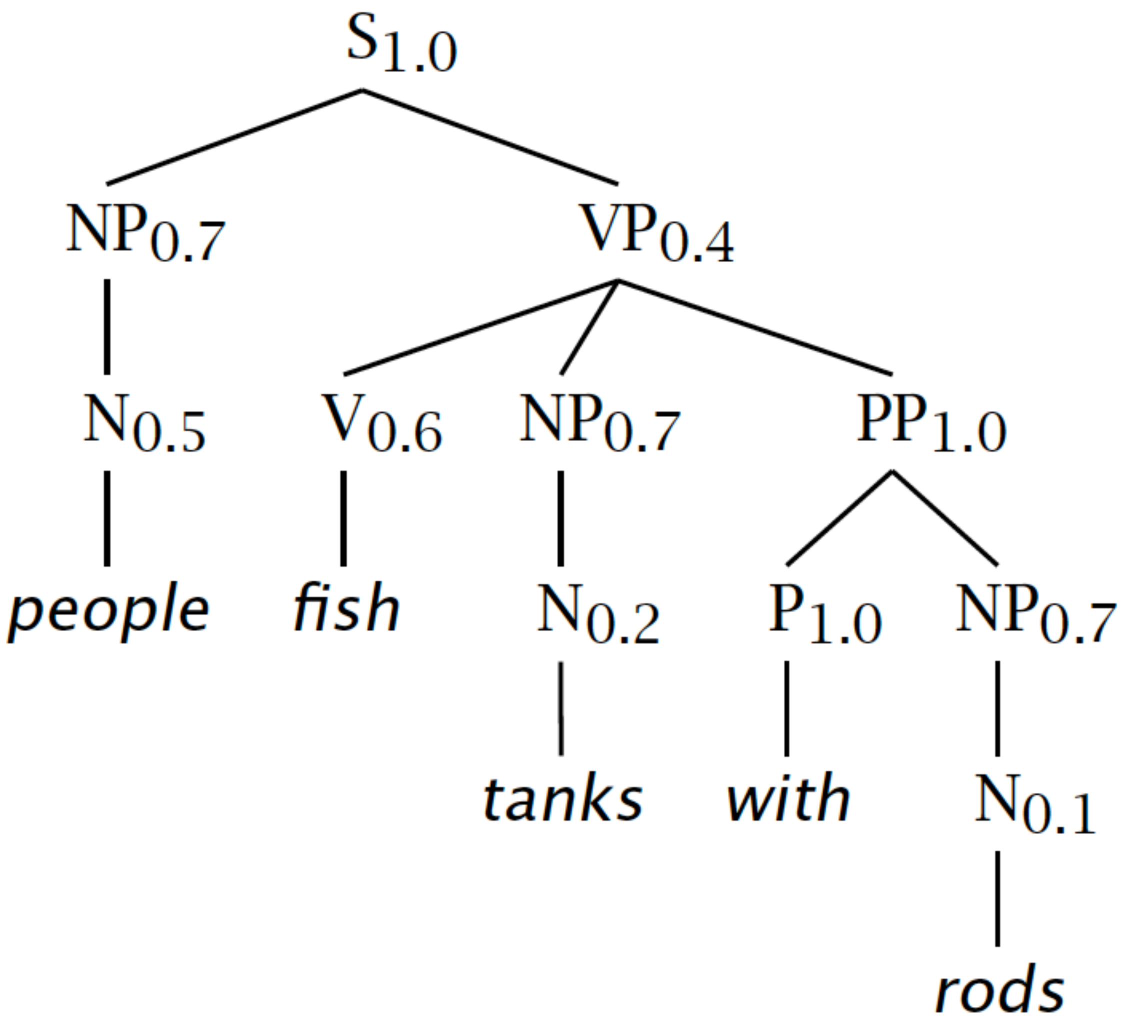
Tree and String Probabilities

- s = *people fish tanks with rods*
- $P(t_1) = 1.0 \times 0.7 \times 0.4 \times 0.5 \times 0.6 \times 0.7$
  $\times 1.0 \times 0.2 \times 1.0 \times 0.7 \times 0.1$
  $= 0.0008232$

  Verb attach
- $P(t_2) = 1.0 \times 0.7 \times 0.6 \times 0.5 \times 0.6 \times 0.2$
  $\times 0.7 \times 1.0 \times 0.2 \times 1.0 \times 0.7 \times 0.1$
  $= 0.00024696$

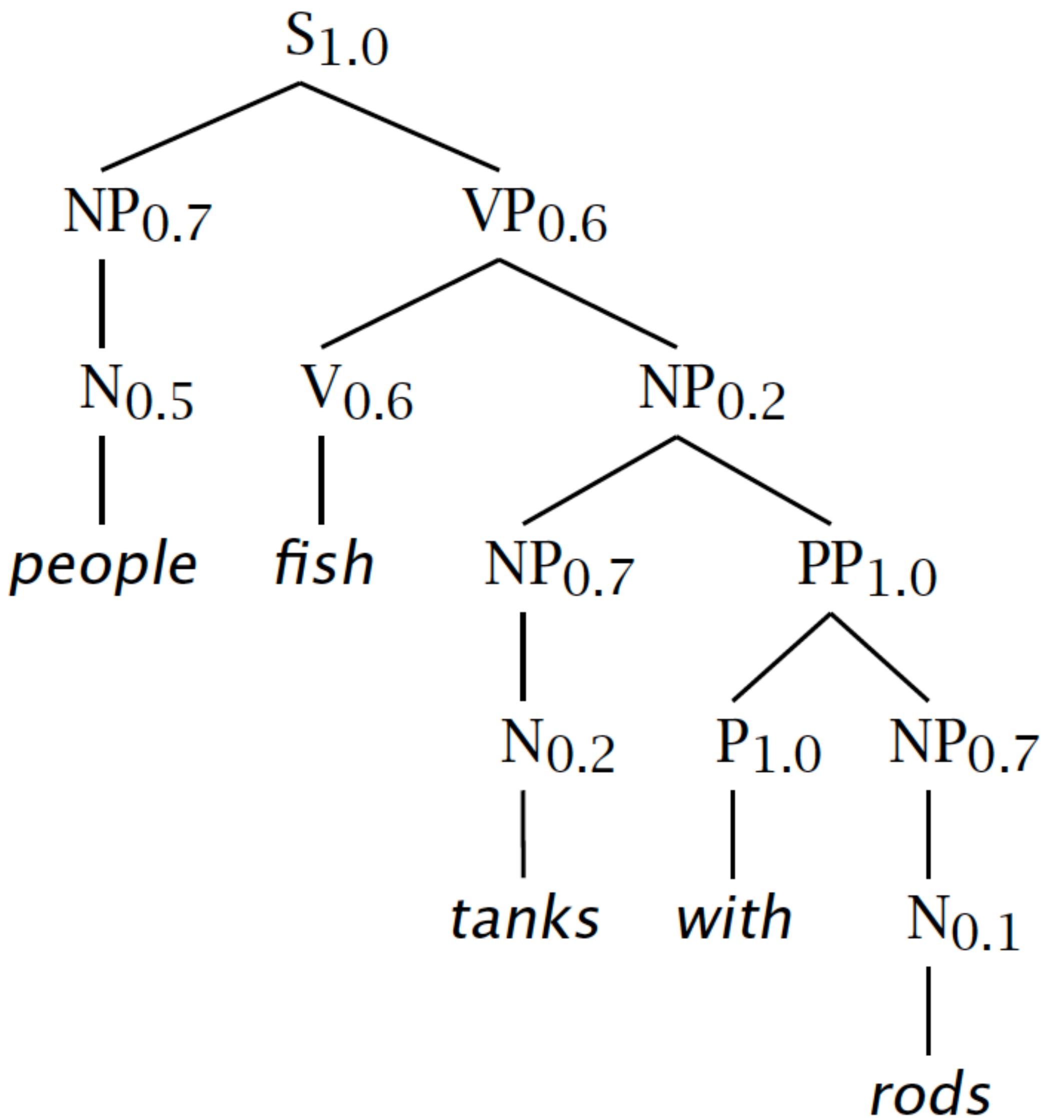
  Noun attach
- $P(s) = P(t_1) + P(t_2)$
  $= 0.0008232 + 0.00024696$
  $= 0.00107016$

t_1:



S1.0
NP0.7
VP0.4
N0.5
V0.6
NP0.7
PP1.0
people
fish
N0.2
P1.0
NP0.7
tanks
with
N0.1
rods

t_2:



S1.0
NP0.7
VP0.6
N0.5
V0.6
NP0.2
people
fish
NP0.7
PP1.0
N0.2
P1.0
NP0.7
tanks
with
N0.1
rods

CFGs and PCFGs

(Probabilistic) Context-Free Grammars