Computational Psycholinguistics

Lecture 7: Probabilistic Models of Human Sentence Processing

Afra Alishahi

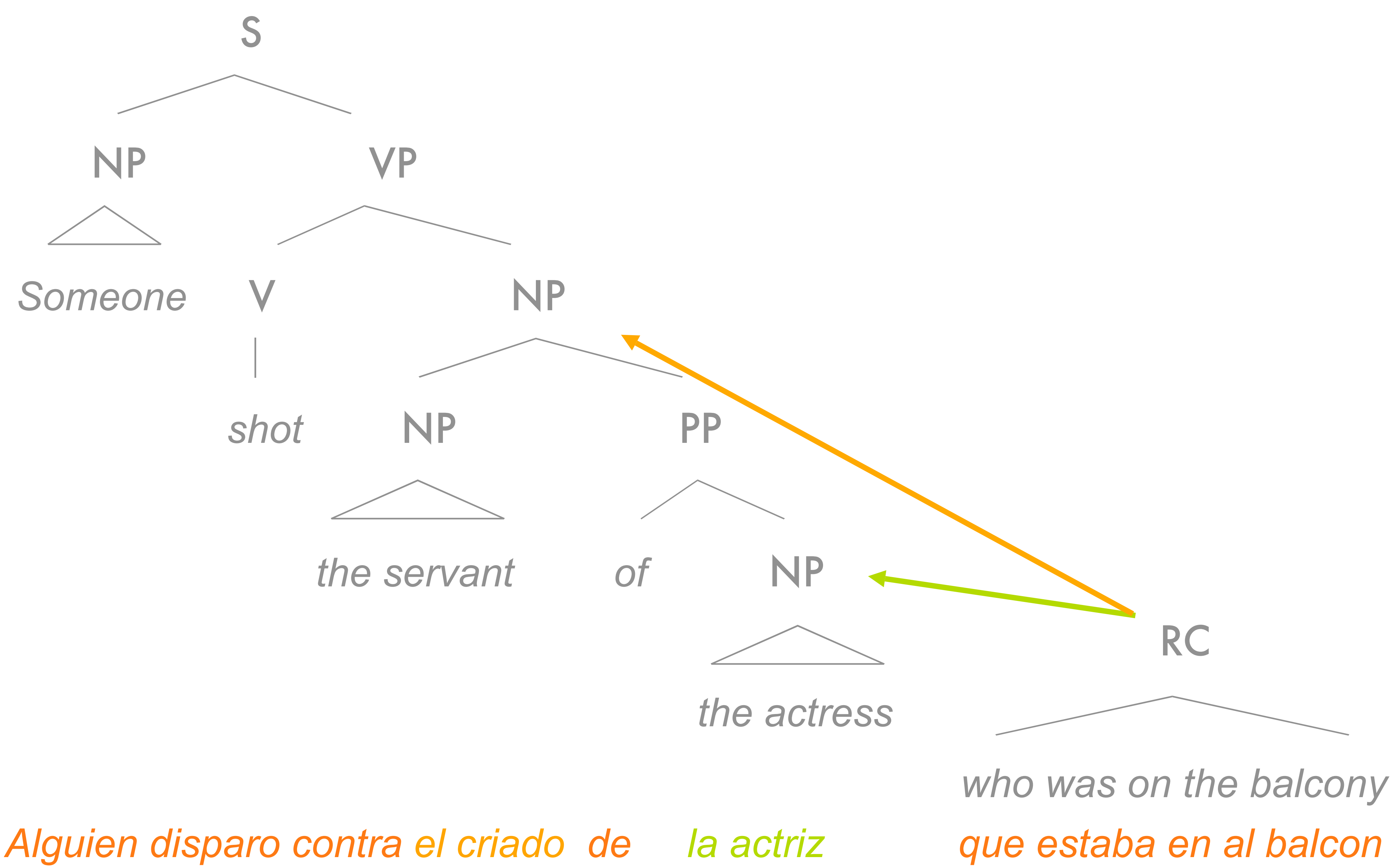
December 15, 2008

Probabilistic Syntax Processing

- Lexical frequencies can contribute to resolving many ambiguities, but not all.
- Does human parser keep track of structural as well as lexical frequencies?
  - Sometimes in contrast with previously suggested principles, such as Late Closure (Frazier)

Someone shot the servant of the actress who was on the balcony.

Relative Clause Attachment



Cross-linguistic RC Preferences

Language	Off-line	On-line
Spanish	high	low
French	high	low
Italian	high	low
Dutch	high	
German	high	low(early), high(late)
English	low	low
Arabic	low	
Norwegian	low	
Swedish	low	
Romanian	low	

- Experienced-based treatment of structural ambiguity?

Tuning Hypothesis

- Tuning Hypothesis (Mitchell et al., 1995):
  - human parser deals with ambiguity by initially selecting the syntactic analysis that has worked most frequently in the past.
  - Further evidence: school children's preferences before and after a period of two weeks in which exposure to high/low examples was increased (Cuetos et al., 1996)
- How to formalize this hypothesis?

The Competition Model

- The Competition Model (MacWhinney et al. 1984)
  - Goal: map from the formal level (surface forms, syntactic constructions, etc) to functional level (meaning, intention)
  - Approach: probabilistically combine various surface cues for choosing the correct functional interpretation
- Focus on the combination of cues, and how the probabilities vary from language to language
  - E.g., assigning thematic roles to grammatical positions (English: word order; German: morphological cues)

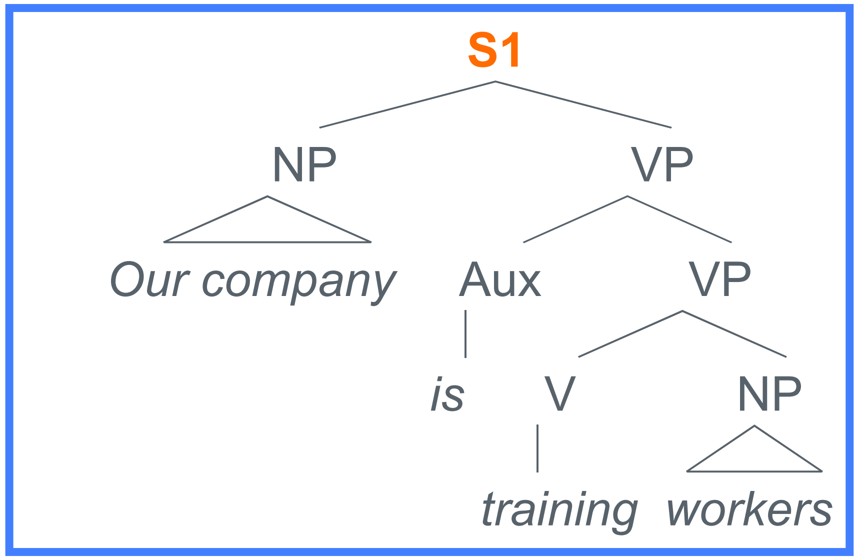
Cue Validity

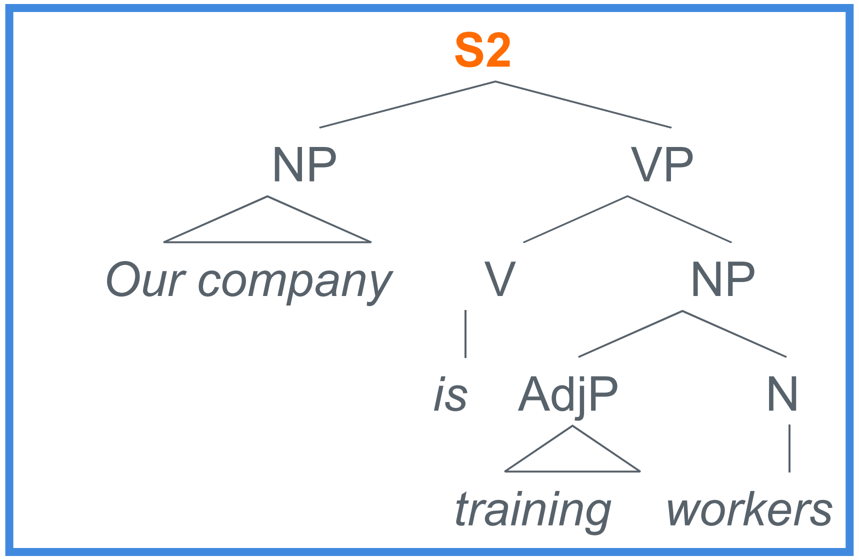
- Cue validity $v(c,i)$: contribution of a cue c to an interpretation i
  - $v(c,i) = \text{availability}(c) \times \text{reliability}(c,i)$
  - $P(c) \times P(i|c) = P(c,i)$
- Combining various cues: $\prod_i P(A|c_i)$
- Comparing two interpretations A and B:

$$P(A|C) = \frac{\prod_i P(A|c_i)}{\prod_i P(A|c_i)+\prod_i P(B|c_i)}$$

Probabilistic Parsing

- Considering the N sentences seen in the past, choose the structure with the highest probability





- How to calculate the probability of a sentence?
  - Maximum likelihood estimation: P(S) = C(S) / N
  - Grain problem: C(S1) = C(S2) = 0; better use probabilities of the smaller chunks, but how small?

Stochastic CFGs

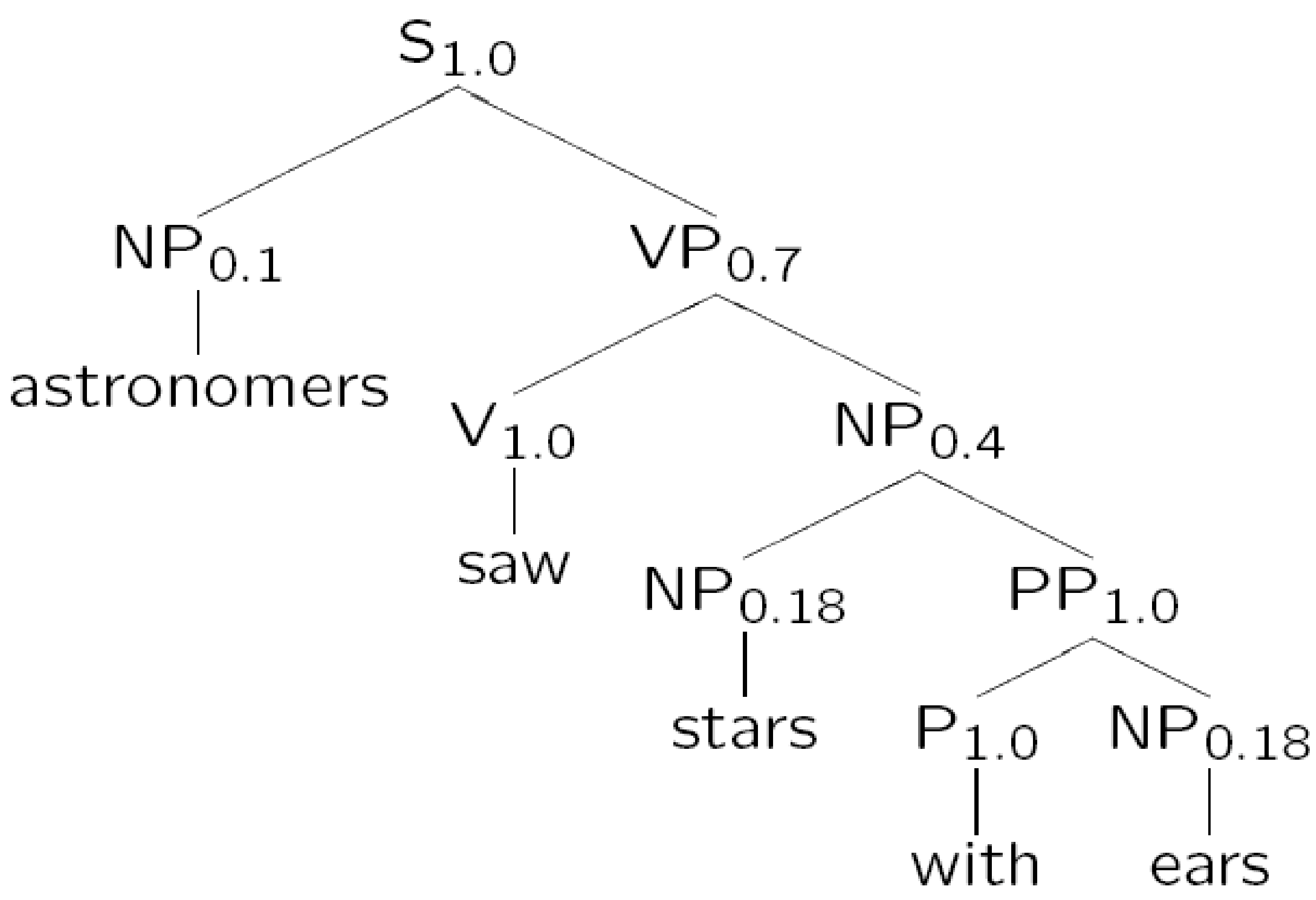
- Augment standard context free grammars by annotating grammar rules with probabilities.

S → NP VP	1.0	NP → NP PP	0.4
PP → P NP	1.0	NP → astronomers	0.1
VP → VP NP	0.7	NP → ears	0.18
VP → VP NP	0.3	NP → saw	0.04
P → with	1.0	NP → stars	0.18
V → saw	1.0	NP → telescopes	0.1

- Probabilities of all rules with the same LHS sum to one
- Probability of a parse is the product of the probabilities of all rules applied

Parse Ranking

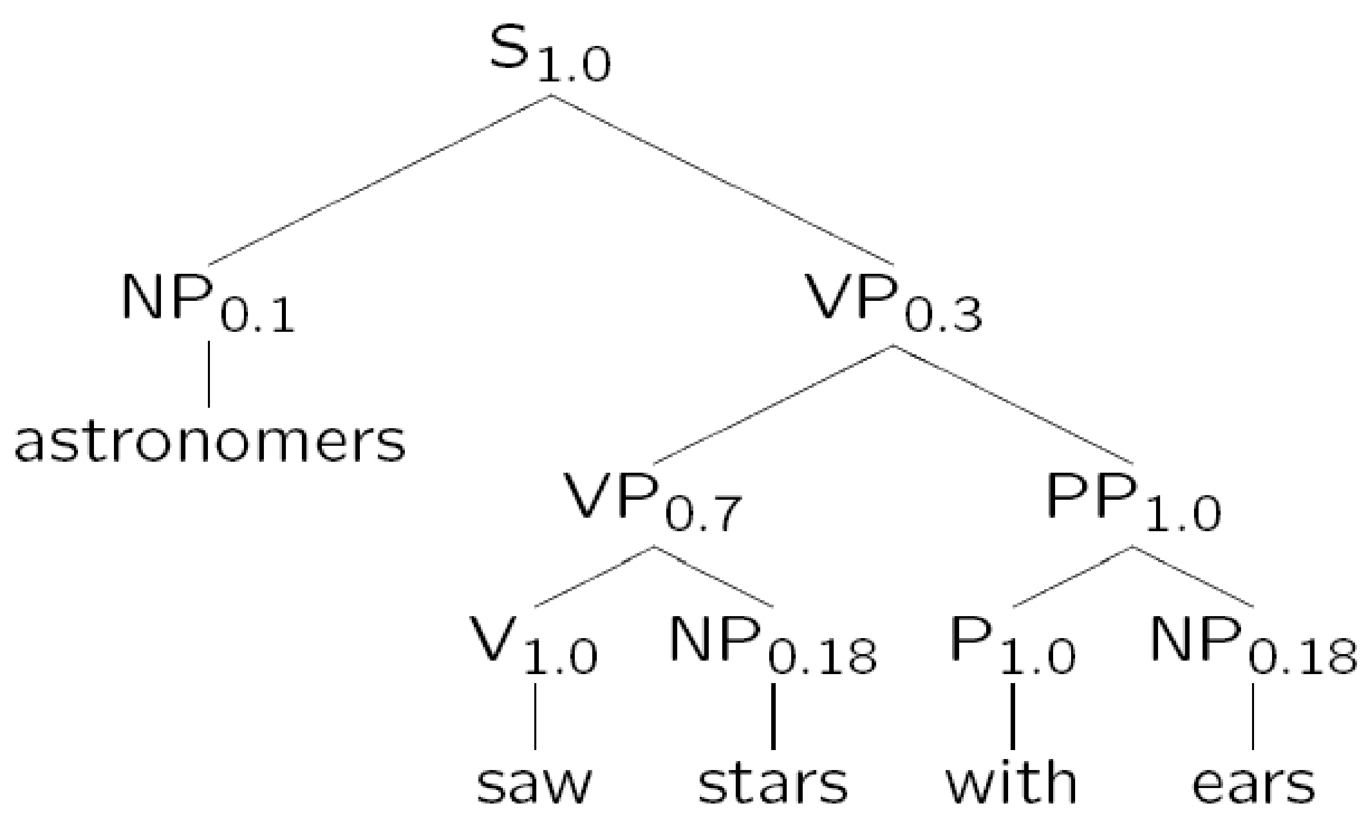
t_1:



$$P(t_1) = 1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0009072$$

Parse Ranking

t_2:



$$P(t_1) = 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0006804$$

Jurafsky (1996)

- Psycholinguistic model of lexical and syntactic access and disambiguation
- Probability of a parse is a combination of
  - Stochastic CFGs
  - Frame probabilities of individual items
- Architecture: incremental, bounded parallel
  - Computation of parse probabilities is incremental
  - Least probable parses are pruned

Frame Preferences

The women discussed the dogs on the beach.

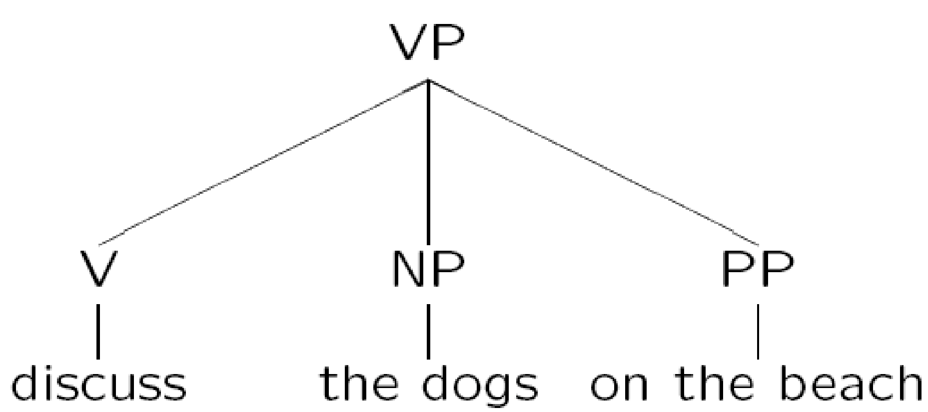
t_1: The women discussed them (the dogs) while on the beach.

✓ t_2: The women discussed the dogs which were on the beach.

$p(\text{discuss}, \langle \text{NP PP} \rangle) = 0.24$

VP → V NP XP   0.15

t_1:

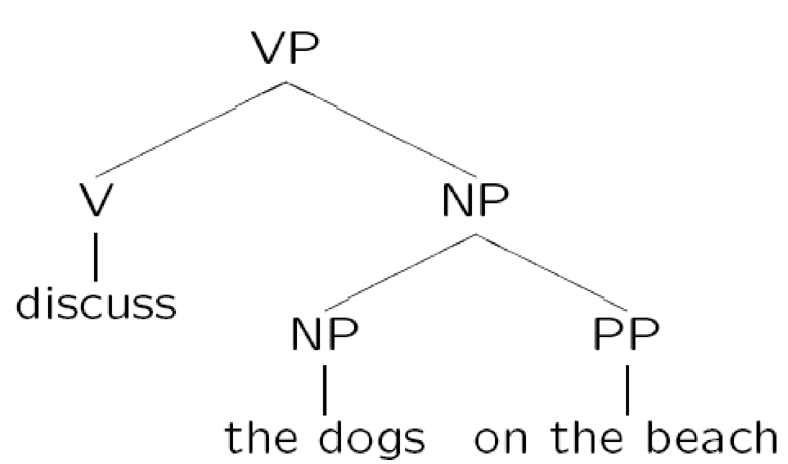


$p(t_1) = 0.15 \times 0.24 = 0.036$ (dispreferred)

$p(\text{discuss}, \langle \text{NP} \rangle) = 0.76$

VP → V NP   0.39
NP → NP XP   0.14

t_2:



$p(t_2) = 0.76 \times 0.39 \times 0.14 = 0.041$ (preferred)

Frame Preferences

The women kept the dogs on the beach.

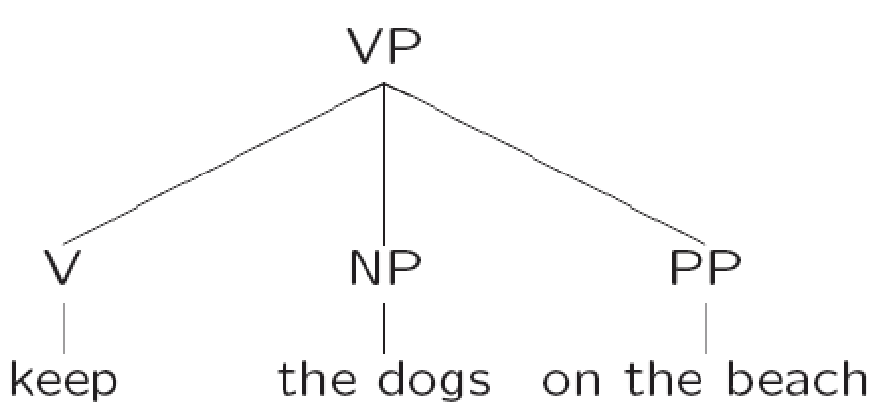
✓ t_1: The women kept them (the dogs) on the beach.

t_2: The women kept the dogs which were on the beach.

$p(\text{keep}, \langle \text{NP XP[pred +]} \rangle) = 0.81$

VP → V NP XP 0.15

t_1:

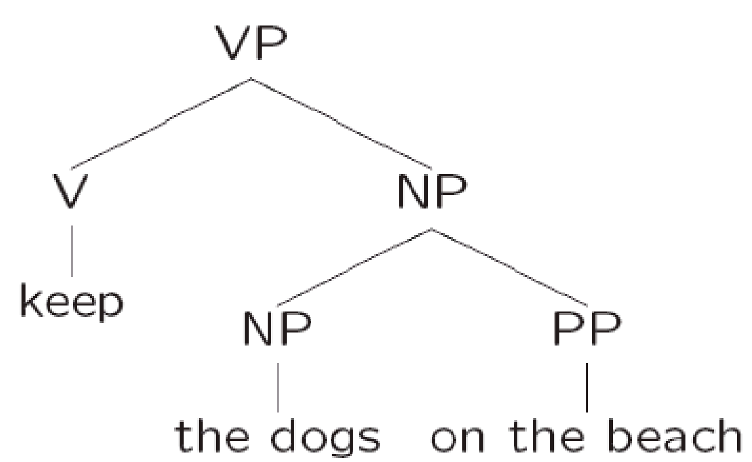


$p(t_1) = 0.15 \times 0.81 = 0.12$ (preferred)

$p(\text{keep}, \langle \text{NP} \rangle) = 0.19$

VP → V NP 0.39
NP → NP XP 0.14

t_2:

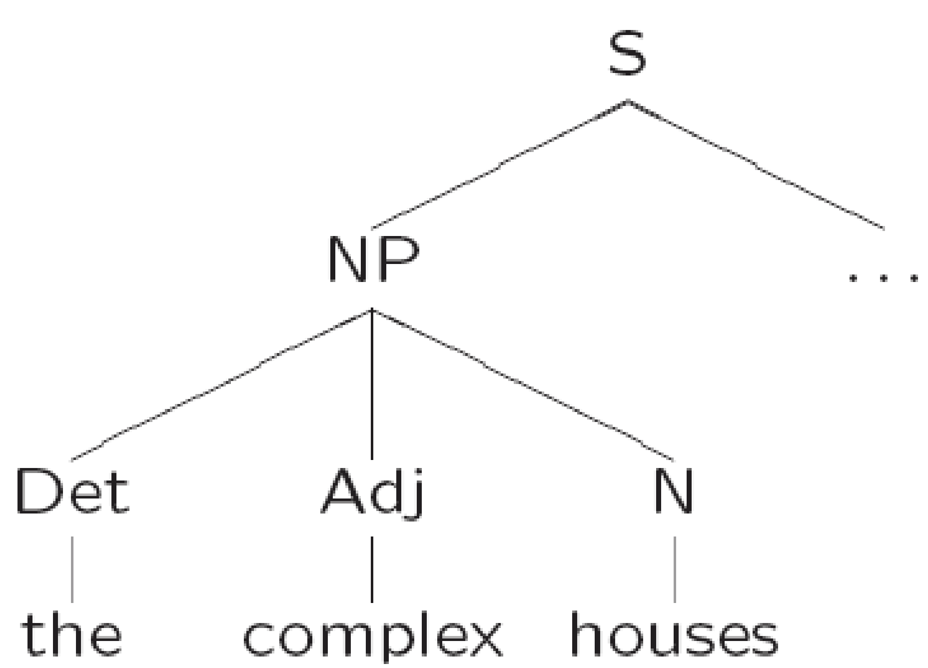


$p(t_2) = 0.19 \times 0.39 \times 0.14 = 0.01$ (dispreferred)

Construction Preferences

$\text{S} \rightarrow \text{NP} \ldots$	0.92
$\text{NP} \rightarrow \text{Det Adj N}$	0.28
$\text{N} \rightarrow \text{ROOT s}$	0.23
$\text{N} \rightarrow \text{house}$	0.0024
$\text{Adj} \rightarrow \text{complex}$	0.00086

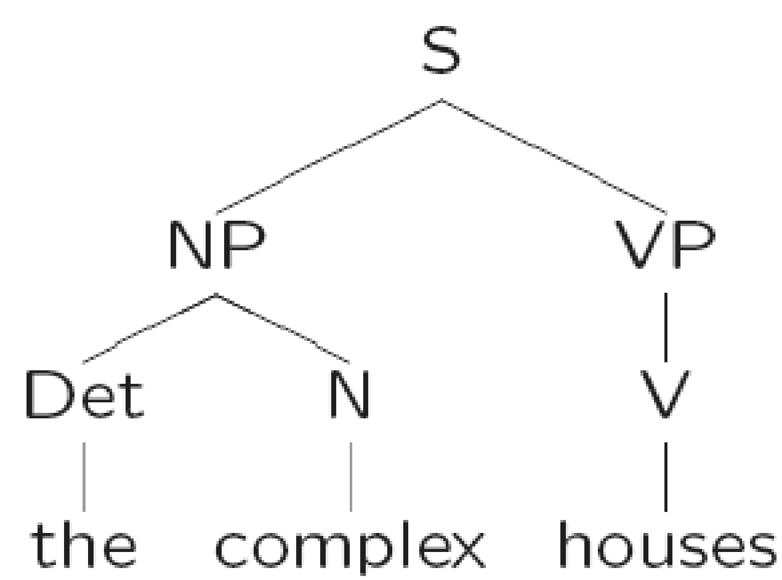
t_1:



$p(t_1) = 1.2 \times 10^{-7}$ (preferred)

$\text{NP} \rightarrow \text{Det N}$	0.63
$\text{S} \rightarrow [\text{NP}\ _{\text{VP}}[\text{V} \ldots$	0.48
$\text{N} \rightarrow \text{complex}$	0.000029
$\text{V} \rightarrow \text{house}$	0.0006
$\text{V} \rightarrow \text{ROOT s}$	0.086

t_1:

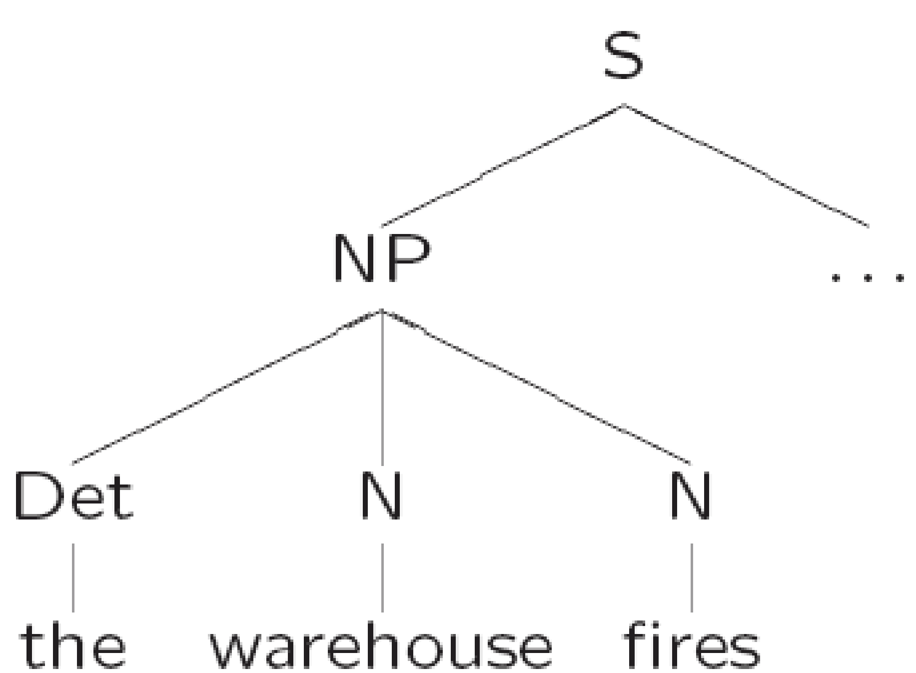


$p(t_1) = 4.5 \times 10^{-10}$ (dispreferred)

Construction Preferences

S → NP ...	0.92
NP → Det N N	0.28
N → fire	0.00072
N → ROOT s	0.23

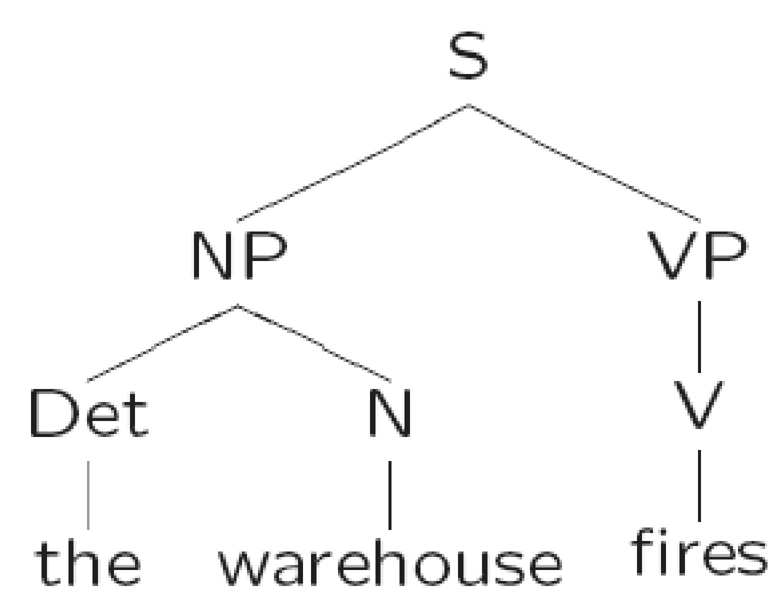
t_1:



$p(t_1) = 4.2 \times 10^{-5}$ (preferred)

NP → Det N	0.63
S → [NP $_{VP}$[V ...	0.48
V → fire	0.00042
V → ROOT s	0.086

t_1:



$p(t_1) = 1.1 \times 10^{-5}$ (dispreferred)

Beam Search and Garden Path

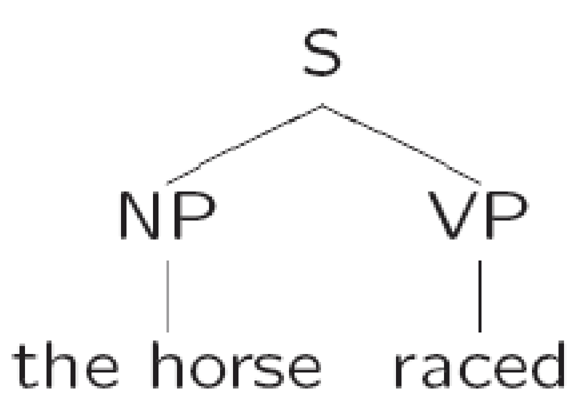
- Prune low probability parses via beam search
    - Assumption: if the relative probability of a parse with respect to the best parse drops below a certain threshold, it will be pruned
- Pruned parses are predicted to reflect garden-path sentences

Frame and Construction Probs

The horse raced past the barn fell.

$p(\text{race}, \langle \text{NP} \rangle) = 0.92$

t_1:

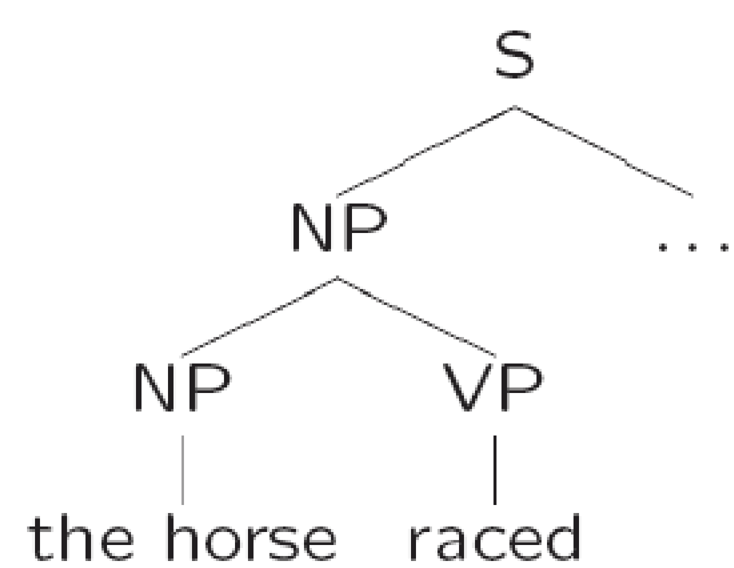


$p(t_1) = 0.92$ (preferred)

$p(\text{race}, \langle \text{NP NP} \rangle) = 0.08$

$\text{NP} \rightarrow \text{NP XP} \quad 0.14$

t_2:



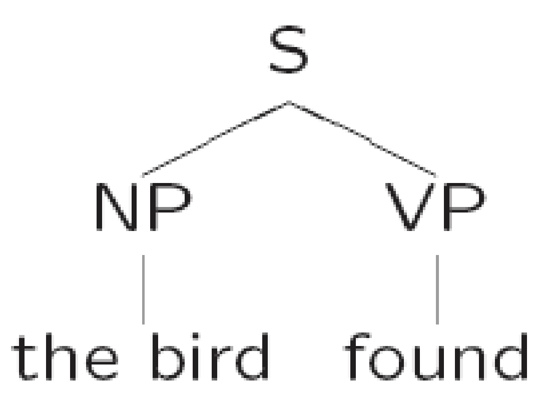
$p(t_1) = 0.0112$ (dispreferred)

Frame and Construction Probs

The bird found in the room died.

$p(\text{find}, \langle \text{NP} \rangle) = 0.38$

t_1:

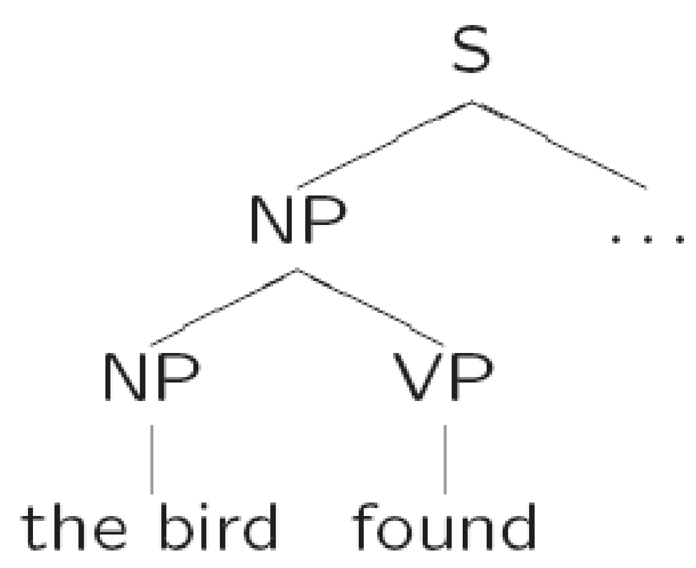


$p(t_1) = 0.38$ (preferred)

$p(\text{find}, \langle \text{NP NP} \rangle) = 0.62$

NP $\rightarrow$ NP XP   0.14

t_2:



$p(t_1) = 0.0868$ (dispreferred)

Setting Beam Width

sentence	probability ratio
the complex houses ...	267:1
the horse raced ...	82:1
the warehouse fires ...	3.8:1
the bird found ...	3.7:1

Claim: a tree is pruned, and therefore a garden-path, if the probability ration is greater than **5:1**

Open Issues

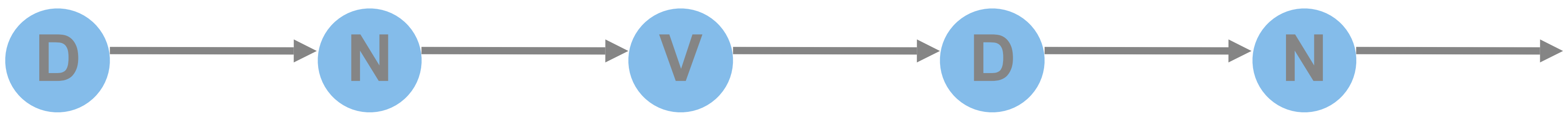
- Incrementality: can we make more fine grained predictions about the time course of ambiguity
- Relative difficulty: Jurafsky doesn't distinguish the relative difficulty of parses/interpretations that remain in the beam
- Memory: no account for memory load within a sentence (e.g. centre embeddings)
- Coverage: small, manually designed lexicon and grammar; tested on a handful of examples

A wide-coverage model: ICMM

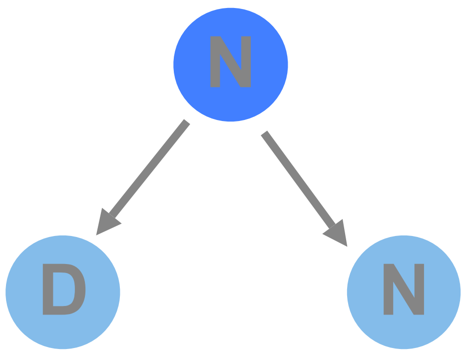
- ICMM: Incremental Cascaded Markov Model (Crocker & Brants, 2000)
  - Standard HMM POS tagger for lexical categories, similar to SLCM
  - Structural probabilities computed as in a SCFG
- Wide coverage:
  - A fully implemented parser, trained on parsed corpora (Brown, WSJ, NEGRA)
  - Adapted to operate incrementally

Probabilistic Tagging & Parsing

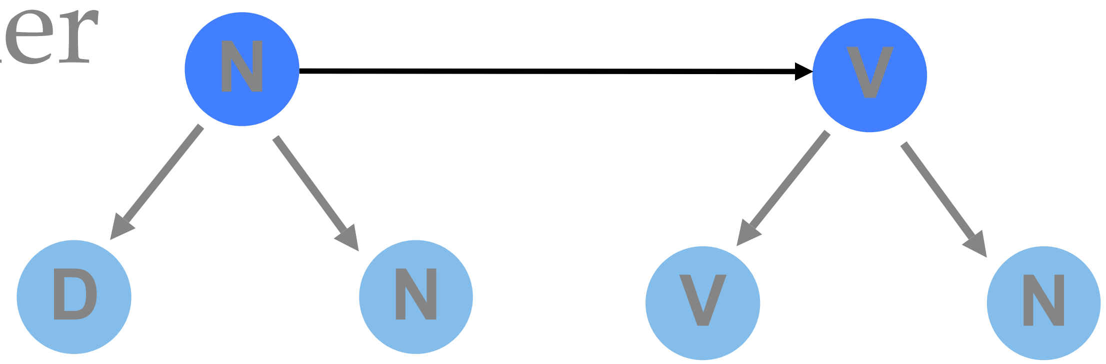
- Markov Models for part-of-speech tagging use `horizontal' probabilities (e.g., SLCM)



- Stochastic CFGs use `vertical' probabilities (e.g., Jurafsky)



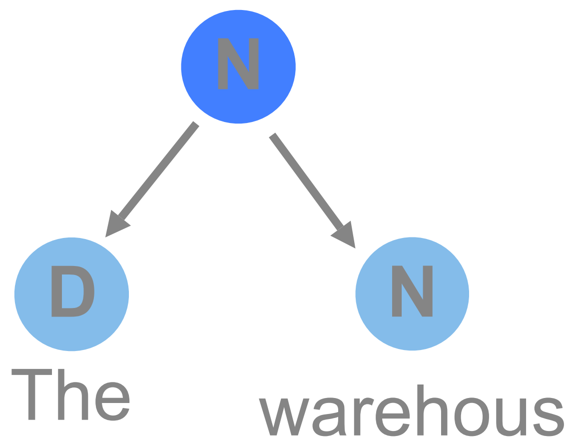
- Cascaded Markov Models apply `horizontal' probabilities to levels higher than parts-of-speech



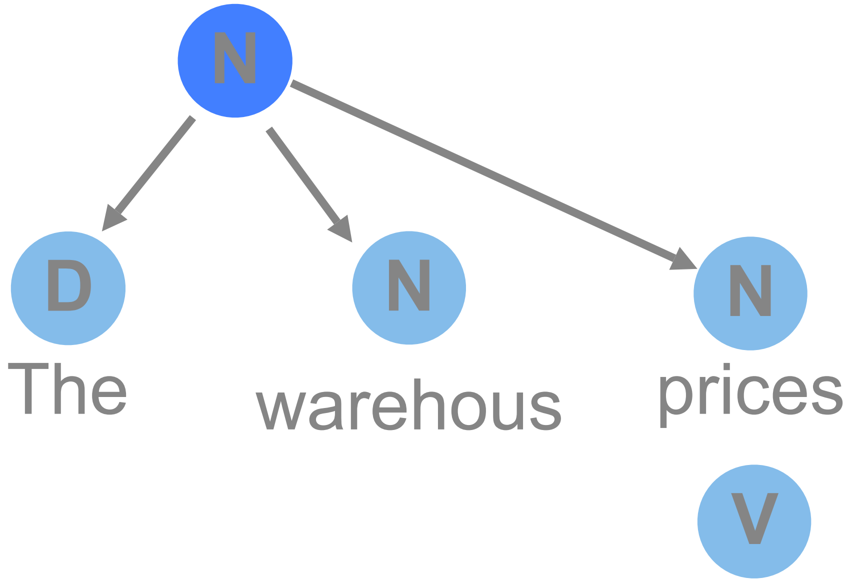
Incremental Cascaded Markov Models

- A parse consists of different layers of nodes
  - Each Markov model layer consists of a series of nodes corresponding to phrasal (syntactic) categories
  - Transitions correspond to trigram category probabilities
- Incremental (word-by-word) processing
  - Build hypotheses for all layers as soon as a word is read
  - Use each Markov model layer as a probabilistic filter, where only highest probability sequences are passed to the next layer

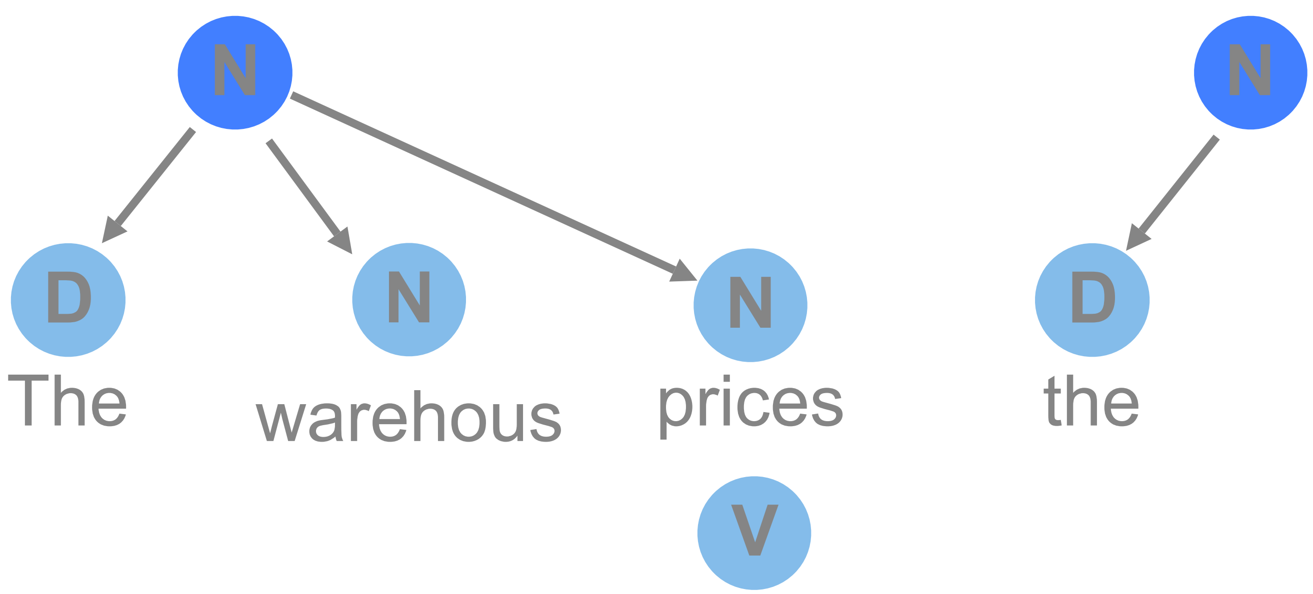
ICMM



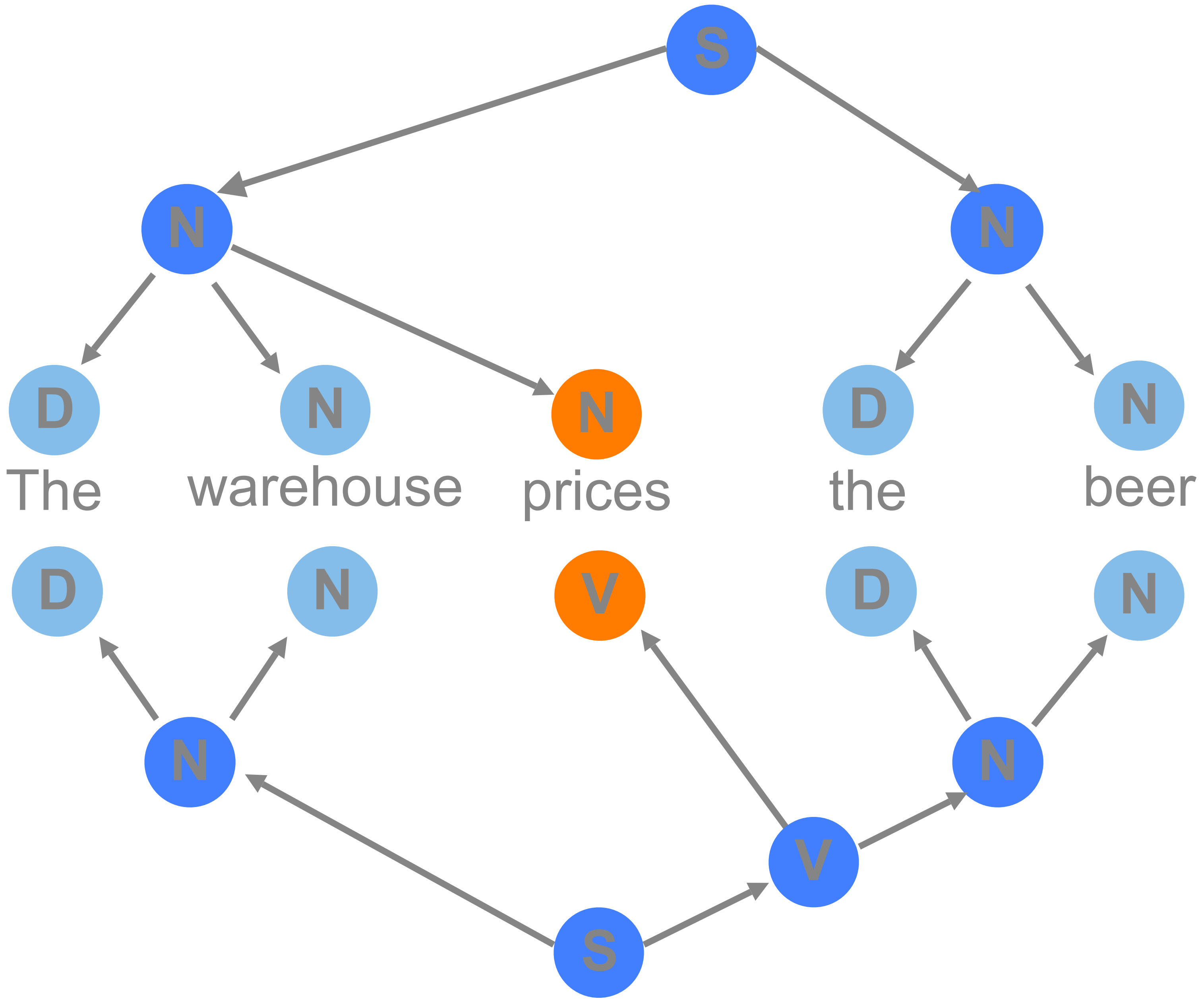
ICMM



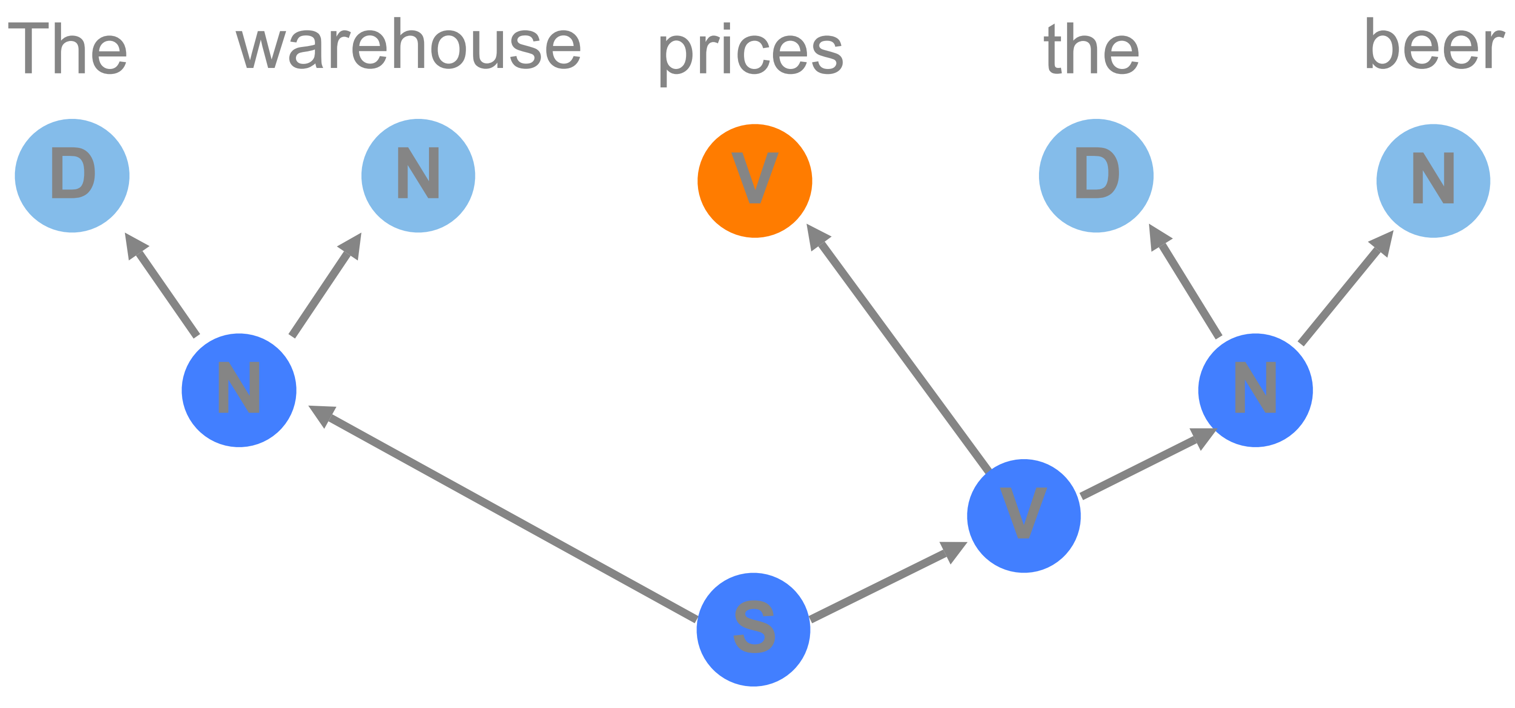
ICMM



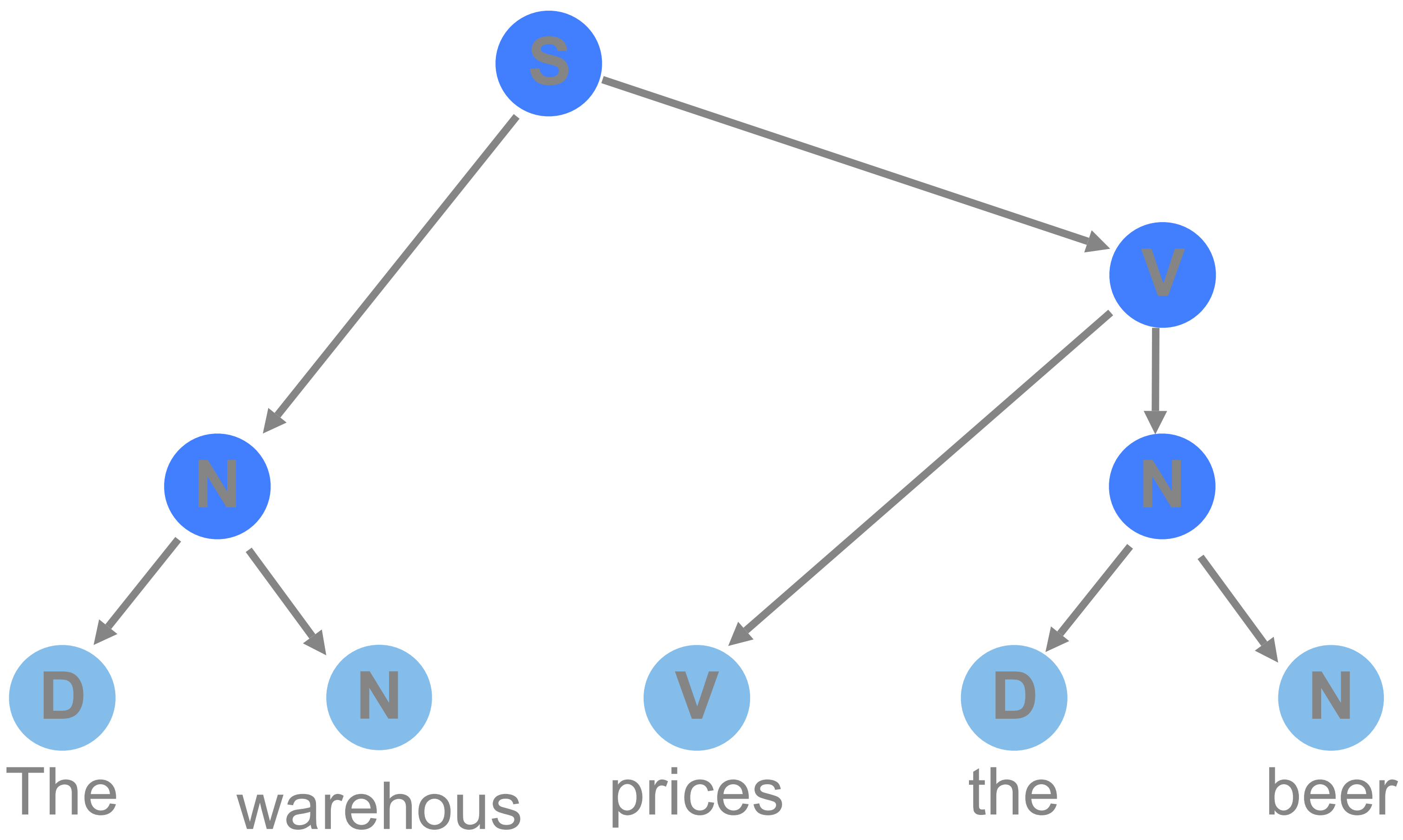
ICMM



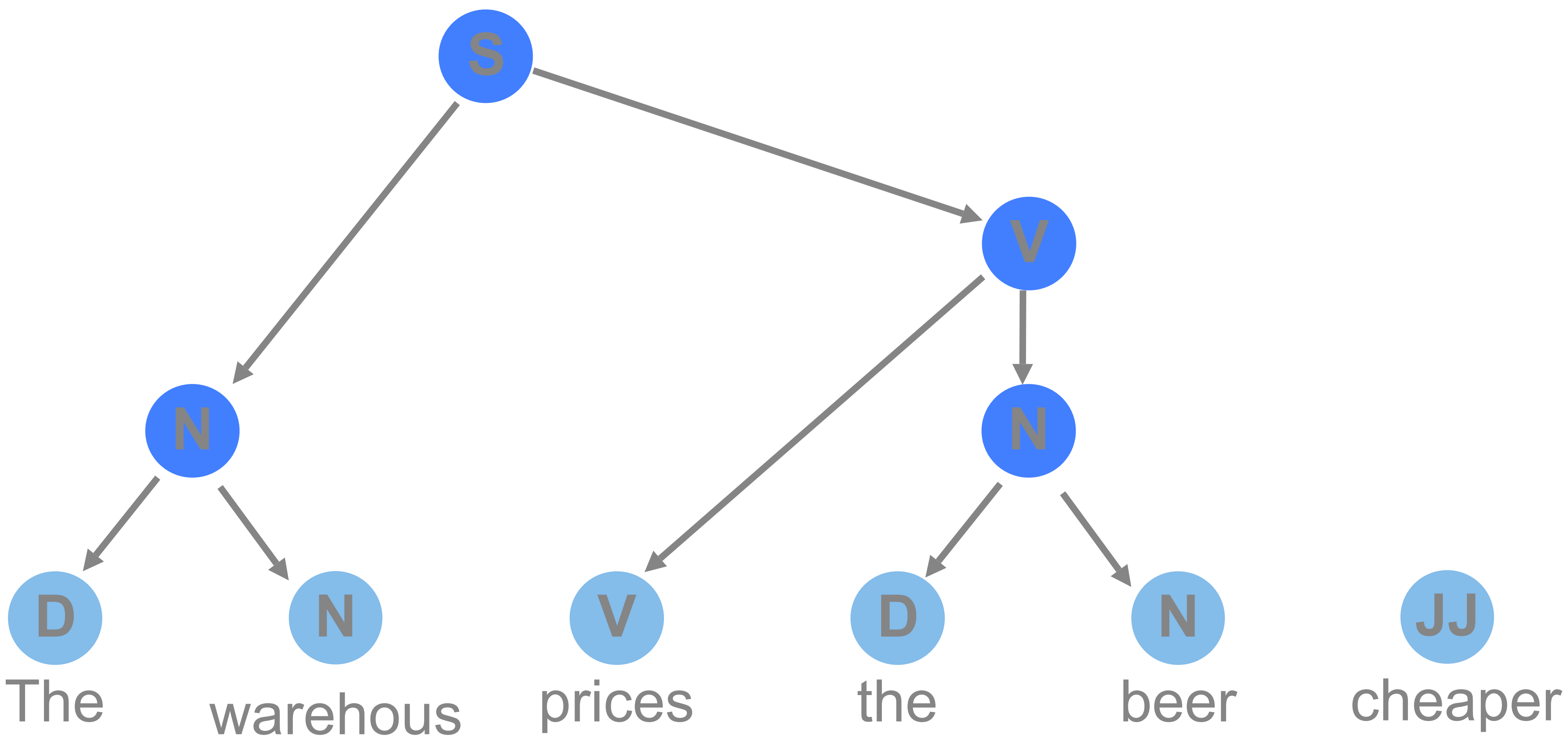
ICMM



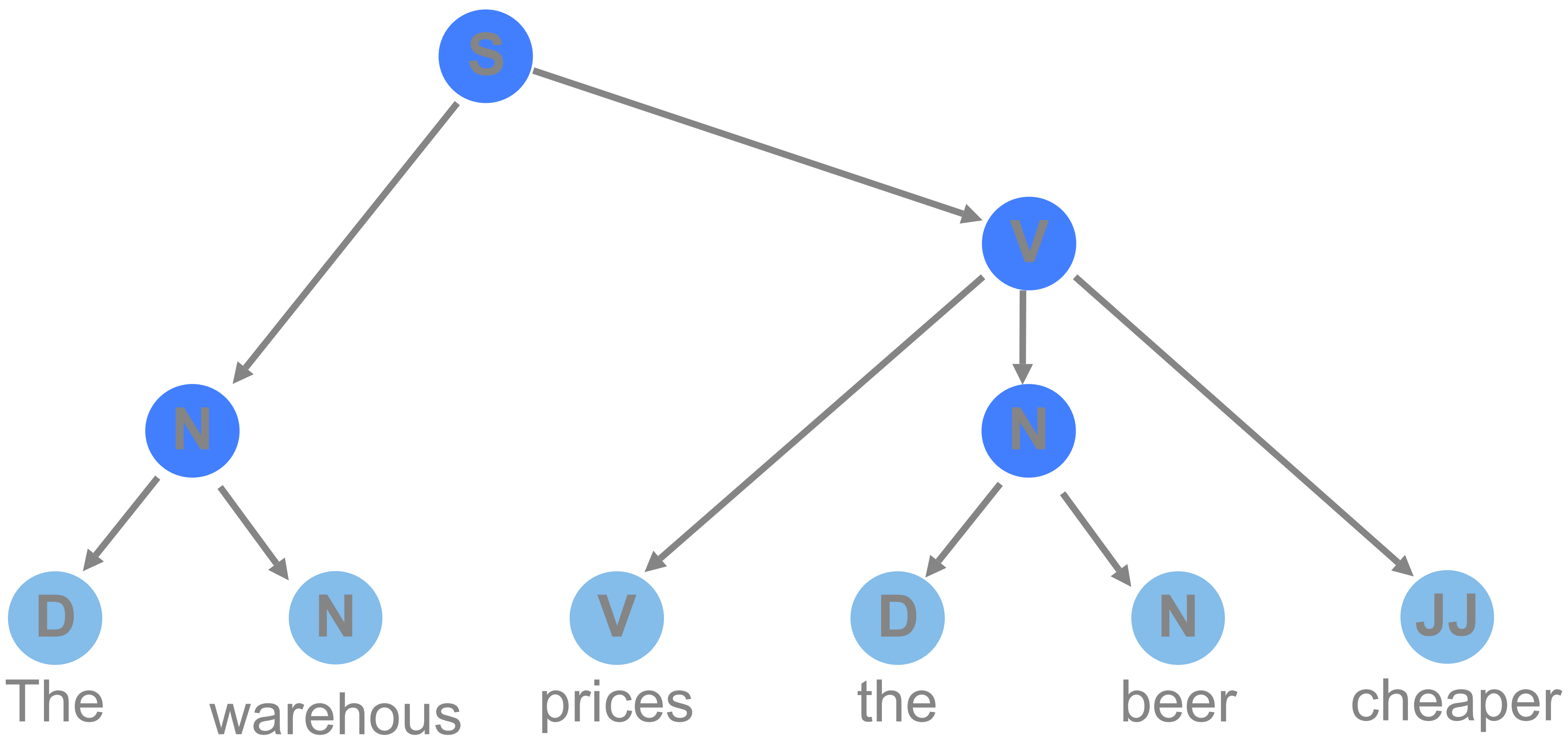
ICMM



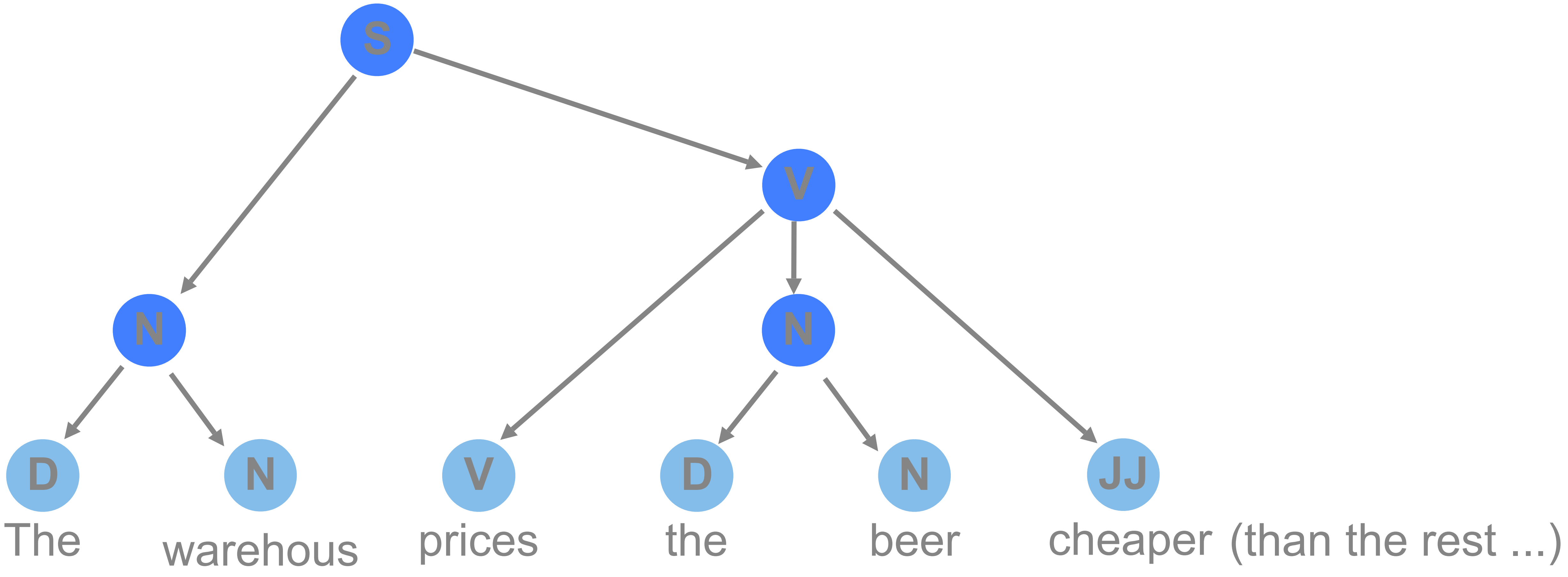
ICMM



ICMM



ICMM



S
N
V
D
N
V
N
JJ
D
N
The
warehous
prices
the
beer
cheaper
(than the rest ...)

ICMM: Summary

- Advantages:
  - Wide coverage: accounts for a range of experimental findings concerning lexical and syntactic ambiguities
  - Cognitive plausibility: the model is incremental and uses limited memory
- Limitations:
  - Makes predictions about time course, but only at a coarse-grained level
  - Does not include verb subcategorization preferences

Summary & Conclusions

- Motivation: People process language: rapidly, robustly, and accurately
  - Experimental evidence for probabilistic mechanisms
- SLCM: Simple, robust account of lexical category disambiguation
- Jurafsky: Probabilistic parser that models a range of local ambiguities
- ICMM: Incremental, broad coverage parser, combines SLCM & Jurafsky

Remaining Problems

- Integrating plausible parsing mechanisms:
  - Either bounded parallel, or serial (momentary parallel) with reanalysis
- Investigating more plausible `optimal functions’
  - More linguistically informed probabilistic models (lexical, semantic ...)
  - Integration with non-probabilistic decision strategies (e.g., recency)
  - More sophisticated integration of memory load constraints