# Computational Psycholinguistics
# Lecture 6: Probabilistic Parsing

Matthew W. Crocker

Marty Mayberry

# Probabilistic Language Processing

- Task of comprehension: recover the correct interpretation
- Goal: Determine the most likely analysis for a given input:

$$\arg\max_i P(s_i) \text{ for all } s_i \in S$$

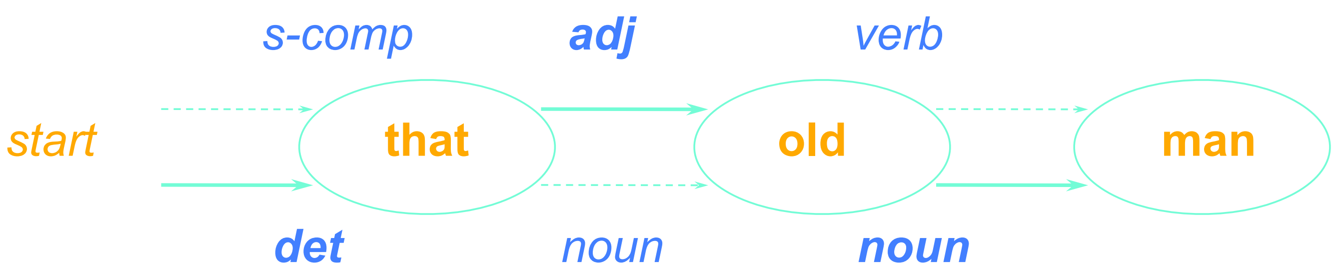
- *P* **hides a multitude of sins:**
  - P corresponds to the degree of belief in a particular interpretation
  - Influenced by recent utterances, experience, non-linguistic context
- P is usually determined by frequencies in corpora or completions
- To compare probabilities (of the Si), we assume parallelism. How much?

# The Model: A Simple POS Tagger

- Find the best category path ($t_1 \ldots t_n$) for an input sequence of words ($w_1 \ldots w_n$):

$$P(t_0,\ldots t_n, w_0,\ldots w_n) \approx \prod_{i=1}^{n} P(w_i \mid t_i) P(t_i \mid t_{i-1})$$

- Initially preferred category depends on two parameters:
  - Lexical bias: $P(w_i|t_i)$
  - Category context: $P(t_i|t_{i-1})$
- Categories are assigned incrementally: Best path may require revision



# 2 Predictions

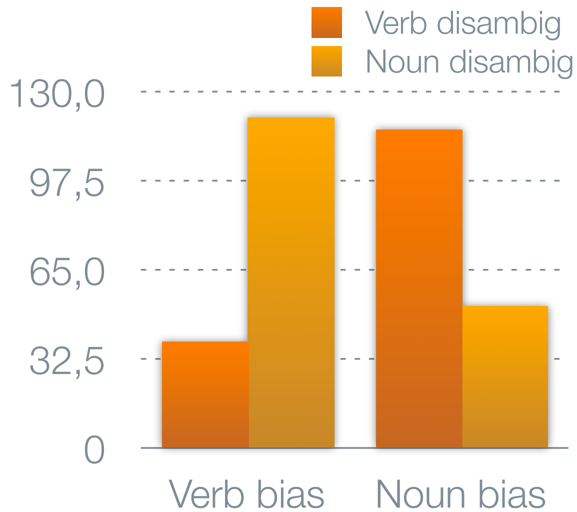
- The Statistical Hypothesis:
  - Lexical word-category frequencies are used for initial category resolution
- The Modularity Hypothesis:
  - Initial category disambiguation is modular, and not determined by (e.g. syntactic) context
- Two experiments investigate
  - The use word-category statistics
  - Autonomy from syntactic context

# Statistical Lexical Category

- Initially preferred category depends on: $P(t_0,\ldots t_n, w_0,\ldots w_n) \approx \prod_{i=1}^{n} P(w_i \mid t_i) P(t_i \mid t_{i-1})$

- Categories are assigned incrementally
  - the warehouse ***prices*** the beer very modestly
  - DET N N / V **V!**
  - the warehouse ***prices*** are cheaper than the rest
  - DET N N / V **N** ...
  - the warehouse ***makes*** the beer very carefully
  - DET N N / V **V**
  - the warehouse ***makes*** are cheaper than the rest
  - DET N N / V **N!** ...

- Lexical bias: P(wi|ti)
- Category context: P(ti|ti-1)
- Trained on the Susanne corpus

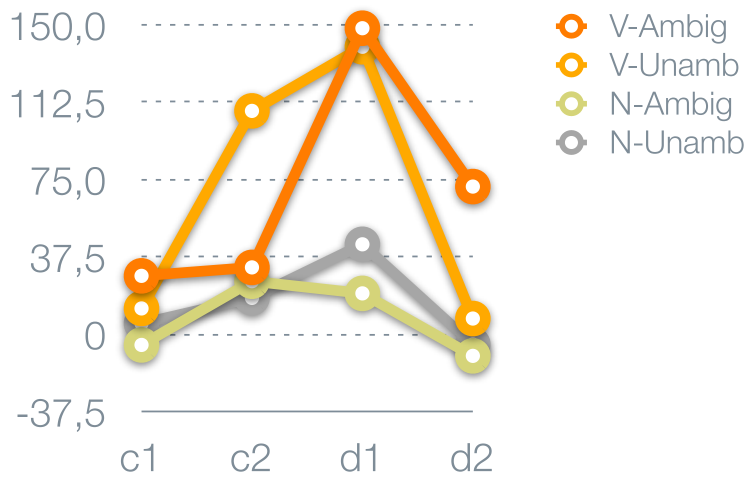



- Interaction between bias and disambiguation
- Lexical category frequency determines initial category decisions

# Modular Disambiguation?

- Do initial decisions reflect integrated use both lexical and syntactic constraints/biases or just (modular) lexical category biases?
  - N/V bias with immediate/late syntactic disambiguation as noun

- Main effect of bias at disambiguation:
  - Initial decisions ignore syntactic context.
  - Problematic for lexicalist syntactic theories
  - At c2, VA/VU difference is significant
  - Implies lexical category doesn't include number (?!)




a) **[V-bias, N-disamb]** The warehouse **makes are** cheaper than the rest.
b) **[V-bias, N-unamb]** The warehouse **make is** cheaper than the rest.
c) **[N-bias, N-disamb]** The warehouse **prices are** cheaper than the rest.
d) **[N-bias, N-unamb]** The warehouse **price is** cheaper than the rest.

# ‘That’ Ambiguity (Juliano & Tanenhaus)

***That* experienced** diplomat(s) would be very helpful ... [DET]

The lawyer insisted ***that* experienced** diplomat(s) would be very helpful [Comp]

- Initially: det=.35 comp=.11 Post-verbally: comp=.93 det=.06
- Found increased RT when dispreferred (according to context) is forced
- Advocates bigram over unigram:

P(that|comp)= 1, P(that|det)=.171

P(comp|verb)=.0234, P(det|verb)=.0296

P(comp|start)=.0003, P(det|start)=.0652

| $t_i$ | *Comp* | *Det* |
|---|---|---|
| $t_{i-1}$ = verb | .0234 | .0051 |
| $t_{i-1}$ = start | .0003 | .0111 |

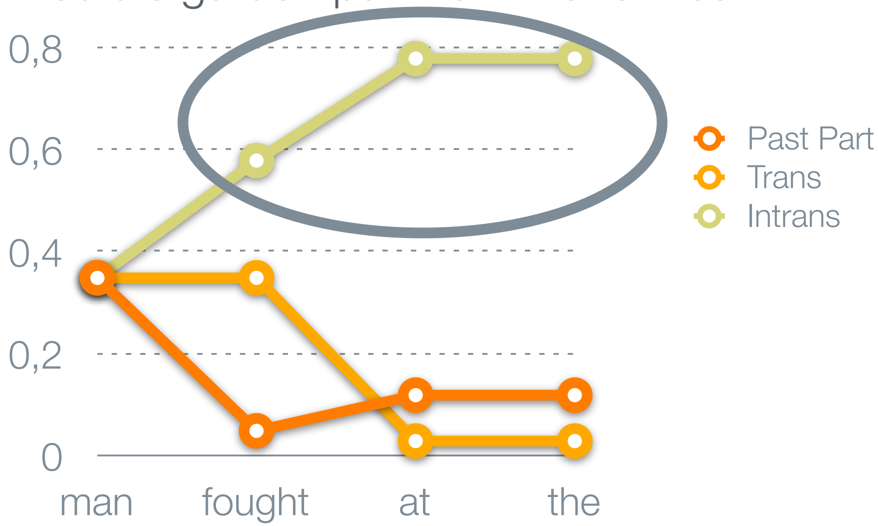
# Internal Reanalysis

- The tagger model predicts internal reanalysis for some sequences.
- Viterbi: revise most likely category sequence based on new evidence
- Right context in RR/MV ambiguities: [MacDonald 1994]
  - The sleek greyhound *raced at the track* won the event
  - The sleek greyhound *admired at the track* won the event
- *raced* = intrans bias, *admired* = trans bias
- Increased RT (blue) indicate transitivity bias is used

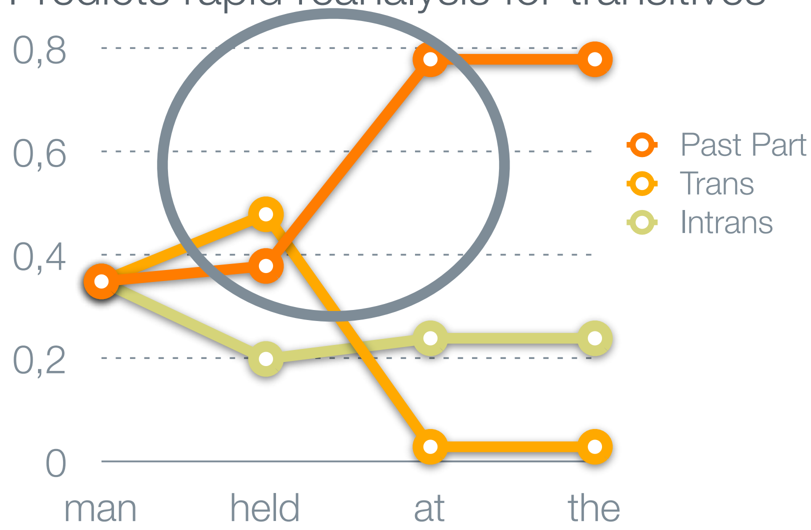
# An SLCM Account

- Assume transitive/intransitive POS categories, extract frequencies from the Susanne corpus: *The man fought at the police station fainted* [intransitive]
  *The man held at the police station fainted* [transitive]

Predicts garden path for intransitives




Predicts rapid reanalysis for transitives



# SLCM Summary

- Psychologically plausible: lower statistical complexity than other models
- High accuracy in general: explains why people perform well overall
- Explains where people have difficulty
  - Statistical: category frequency **drives** initial category decisions
  - Modular: syntax structure **doesn't determine** initial category decisions
  - Bigram evidence: "that" ambiguity [Juliano and Tanenhaus]
  - Reanalysis of verb transitivity for 'reduced relatives' [MacDonald]

# Comments on the SLCM

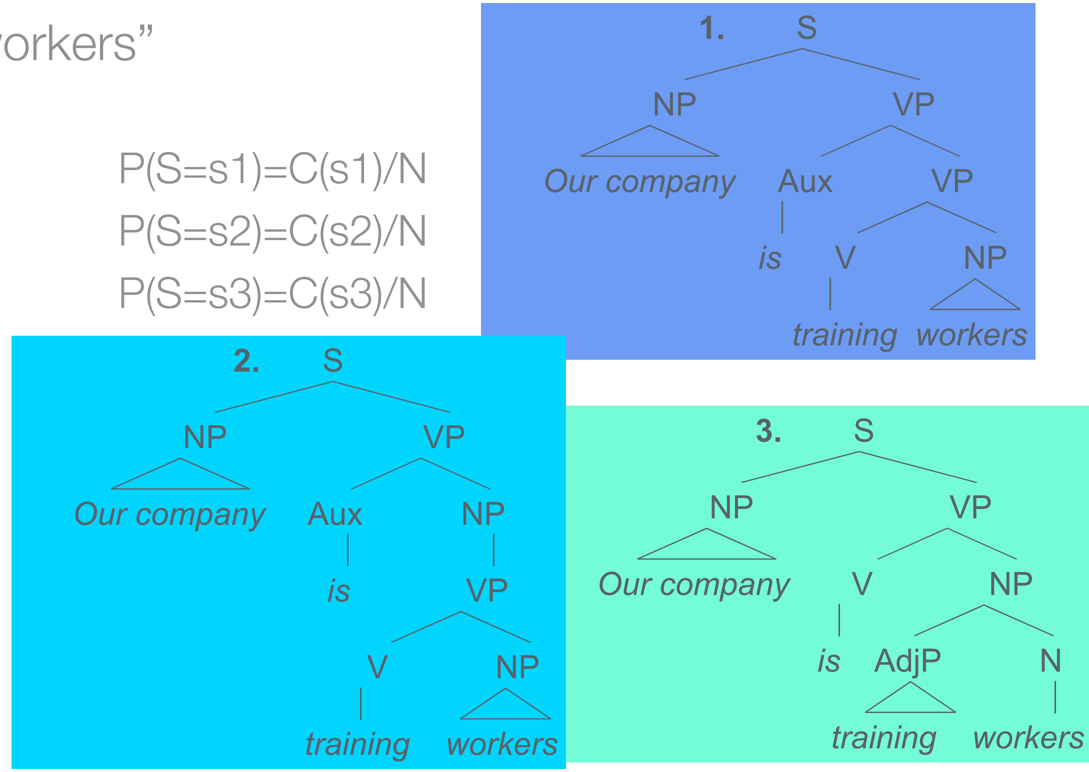
- combines optimality with psychological plausibility
- category preference appears truly frequency-based
- indication of which features are exploited [e.g. transitivity, not number]
- Implications for the Grain Problem?
  - Bigrams used, but not structure ?
  - Transitivity but not number ?

# Estimating P: The Grain Problem

- Suppose you have been exposed to N sentences in your lifetime
- “Our company is training workers”

P(S=s1)=C(s1)/N
P(S=s2)=C(s2)/N
P(S=s3)=C(s3)/N

- Problem: P=0, often
- Solution:
  - Estimate P, by combining probabilities of smaller chunks



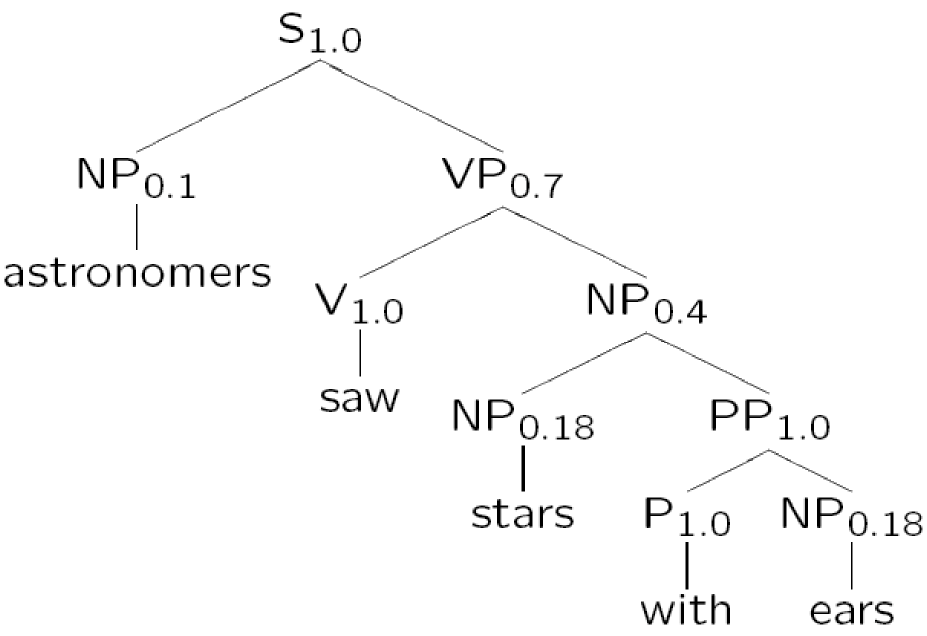
# Probabilistic Grammars

- Context-free rules annotated with probabilities
  - 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.
- Example (Manning and Schütze 1999)

| | | | |
|---|---|---|---|
| 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 |

# Parse Ranking

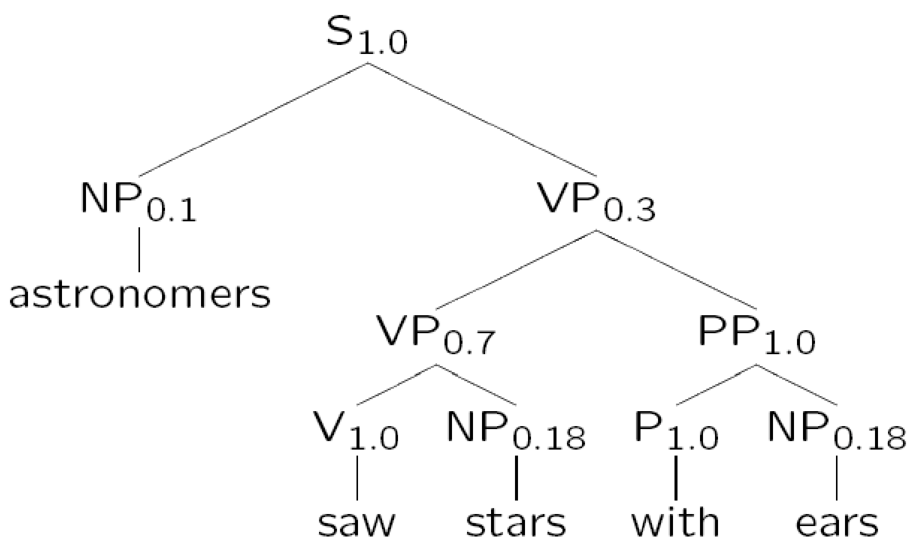
$t_1$:




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

# 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
- Exploits concepts from statistical parsing
  - Probabilistic CFGs
  - Bayesian modeling frame probabilities
- Architecture: Probabilistic, bounded, parallel parser
  - Parses are "pruned" (removed from memory) if they fall outside the "beam"
    - E.g. if they are too improbable with respect to the best parse
  - Pruned parses are predicted to reflect garden-path sentences

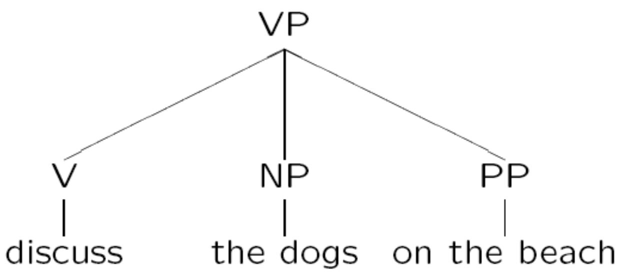
# Frame Preferences

- The women discussed the dogs on the beach.
  - t1. The women discussed them (the dogs) while on the beach. (10%)
  - t2. The women discussed the dogs which were on the beach. (90%)

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

VP → V NP XP  0.15

$t_1$:

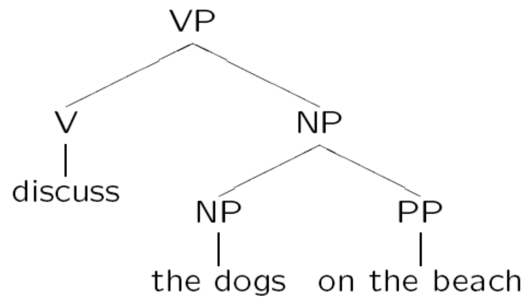



$p(t_1) = 0.15 \times 0.24 = 0.036$ (dispreferred)

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

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)

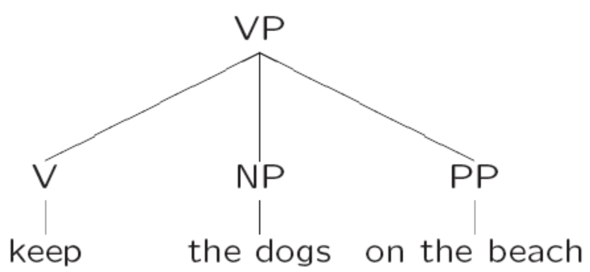
# Frame Preferences

- The women kept the dogs on the beach.
  - t1. The women kept the dogs which were on the beach. (10%)
  - t2. The women kept them (the dogs) while on the beach. (90%)

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

VP → V NP XP   0.15

$t_1$:

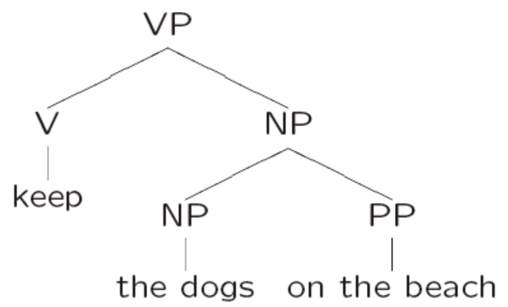



$p(t_1) = 0.15 \times 0.81 = 0.12$ (preferred)

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

VP → V NP   0.39
NP → NP XP   0.14

$t_2$:

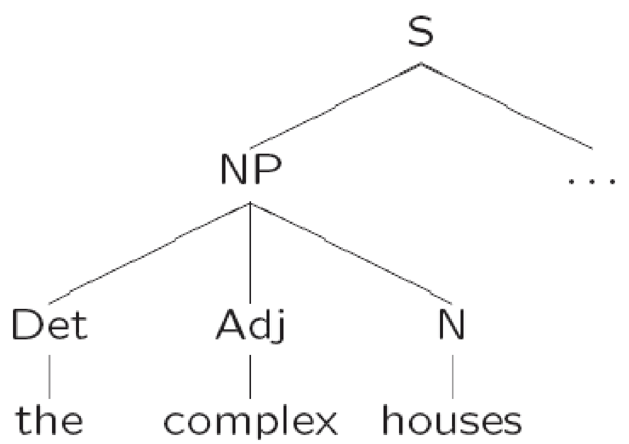



$p(t_2) = 0.19 \times 0.39 \times 0.14 = 0.01$ (dispreferred)

# Construction Preferences

| | |
|---|---|
| S → NP ... | 0.92 |
| NP → Det Adj N | 0.28 |
| N → ROOT s | 0.23 |
| N → house | 0.0024 |
| Adj → complex | 0.00086 |

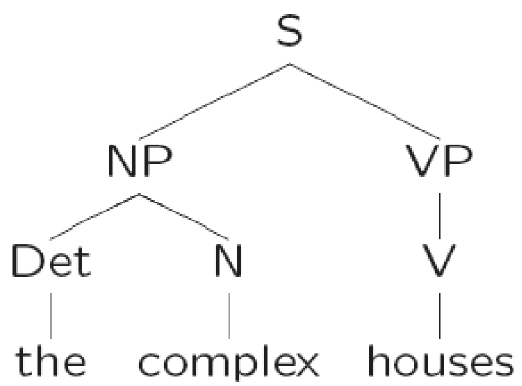
$t_1$:




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

| | |
|---|---|
| NP → Det N | 0.63 |
| S → [NP $_{VP}$[V ... | 0.48 |
| N → complex | 0.000029 |
| V → house | 0.0006 |
| V → ROOT s | 0.086 |

$t_1$:

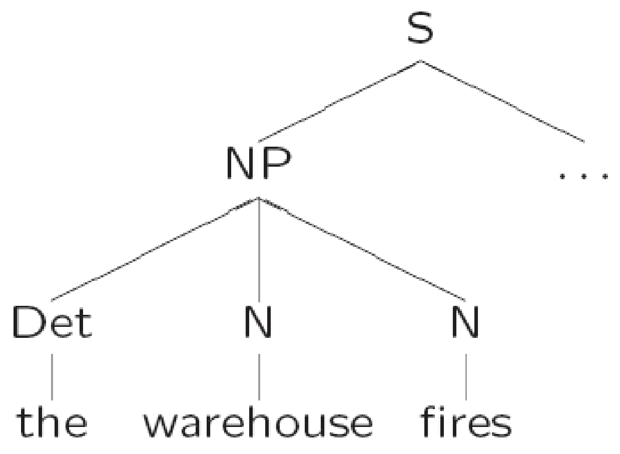



$p(t_1) = 4.5 \times 10^{-10}$ (dispreferred)

# Construction Preferences

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

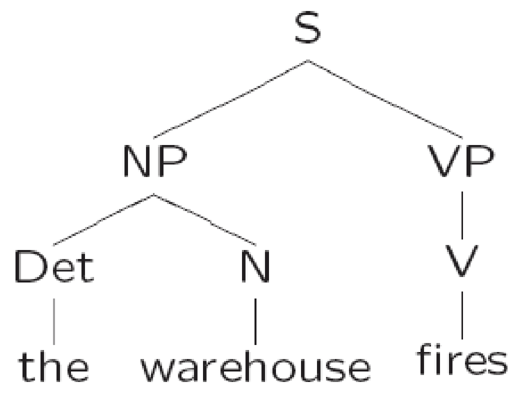
$t_1$:




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

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

$t_1$:



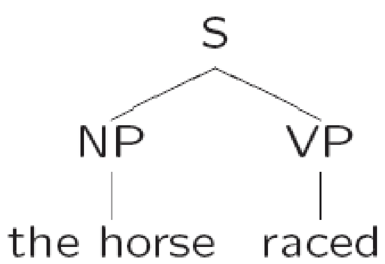

$p(t_1) = 1.1 \times 10^{-5}$ (dispreferred)

# Frame and Construction Probs

*"The horse raced past the barn fell."*

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

$t_1$:

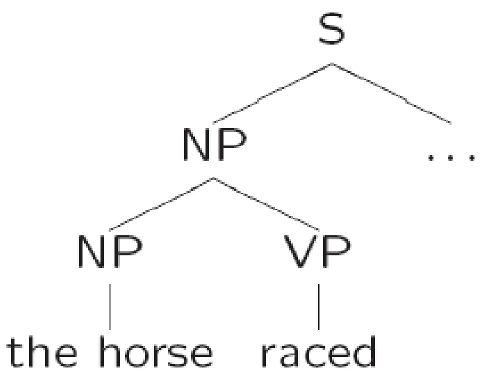



$p(t_1) = 0.92$ (preferred)

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

NP → NP XP 0.14

$t_2$:



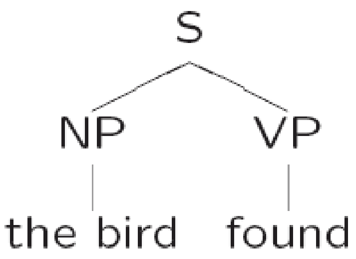

$p(t_1) = 0.0112$ (dispreferred)

# Frame and Construction Probs

*"The bird found died"*

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

$t_1$:

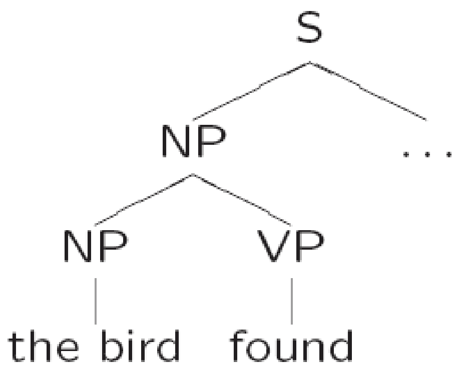



$p(t_1) = 0.38$ (preferred)

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

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

$t_2$:




$p(t_1) = 0.0868$ (dispreferred)

# Setting Beam Width

- Assumption: if the relative probability of a parse with respect to the best parse drops below a certain threshold, it will be pruned

| 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)
- Cross-linguistics: Does the model work well for languages other than English?