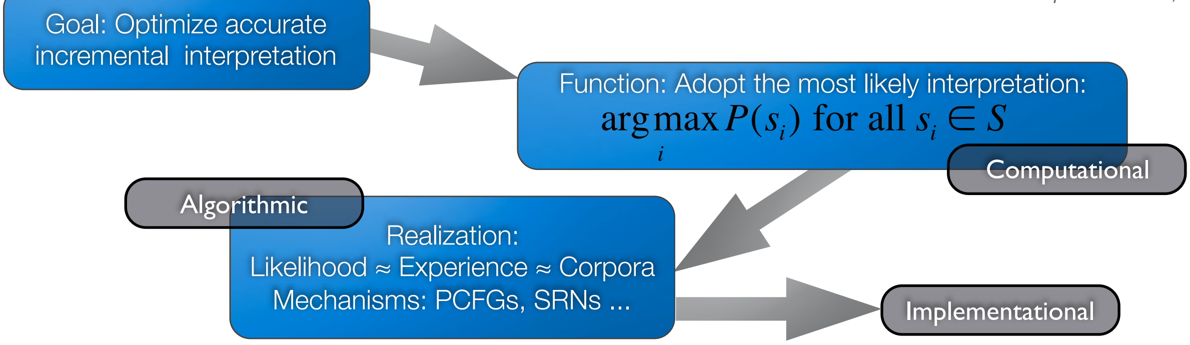
Computational Psycholinguistics

Lecture 7: **Probabilistic Parsing**

Matthew W. Crocker
crocker@coli.uni-sb.de

Crocker. *Rational Models of Comprehension*, 2005.

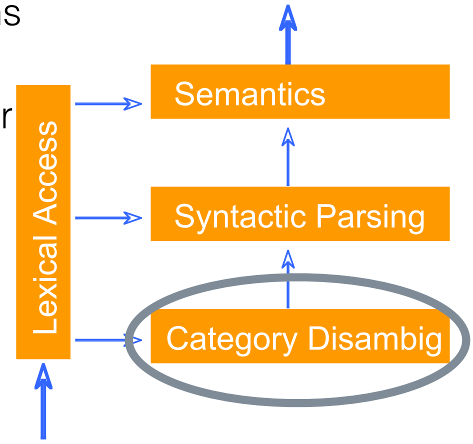


$$\arg\max_i P(s_i) \text{ for all } s_i \in S$$

- Empirical: lexical access, word category/sense, subcategorization
- Rational: accurate, robust, broad coverage
- Rational Models:
  - explain accurate performance in general: i.e. rational behaviour
  - explain specific observed human behavior: e.g. for specific phenomena

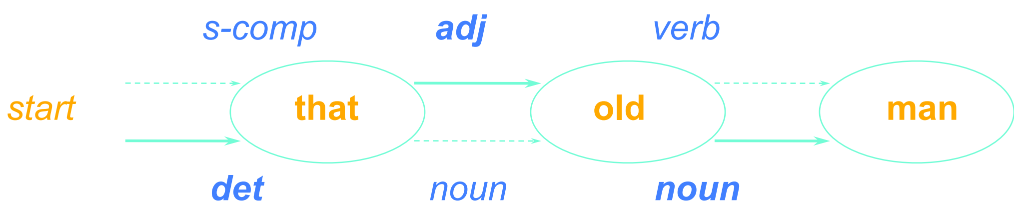
Lexical Category Disambiguation

- Sentence processing involves the resolution of lexical, syntactic, and semantic ambiguity.
  - Solution 1: These are not distinct problems
  - Solution 2: Modularity, divide and conquer
- Category ambiguity:
  - *Time flies like an arrow.*
- Extent of ambiguity:
  - **10.9%** (types) **65.8%** (tokens) (Brown Corpus)



The Model: A Simple POS Tagger

- Find the best category path ($t_1 \ldots t_n$) for an input sequence of words ($w_1 \ldots w_n$): $P(t_0, \ldots t_n, w_0, \ldots w_n)$
- Initially preferred category depends on two parameters:
  - Lexical bias: $\boldsymbol{P(w_i|t_i)}$
  - Category context: $\boldsymbol{P(t_i|t_{i-1})}$
- Categories are assigned incrementally: Best path may require revision



SLCM Summary

- High accuracy in general & psychologically plausible
- 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]
  - Explains "local coherence" effects:
    "The coach smiled at the player *tossed* the frisbee …"

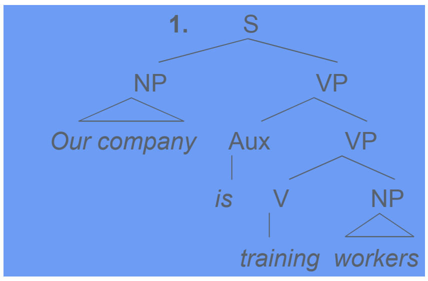
Estimating P: The Grain Problem

- Suppose you have been exposed to N sentences in your lifetime

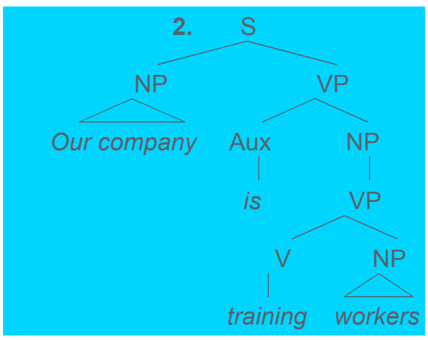
$$\arg\max_i P(s_i) \text{ for all } s_i \in S$$

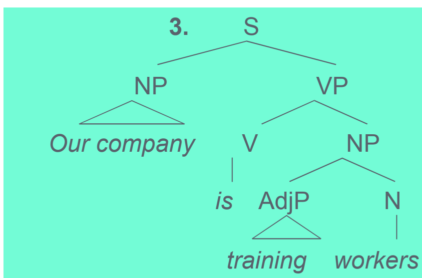
- "Our company is training workers"

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





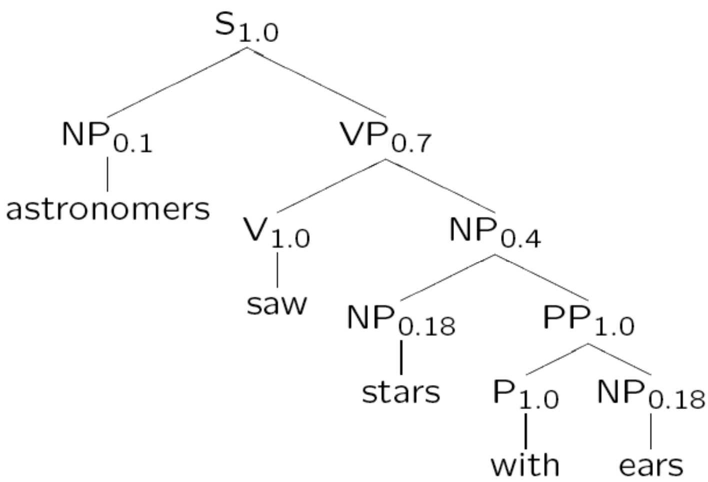
PCFGs: a quick reminder

- 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 PP	0.3	NP → saw	0.04
P → with	1.0	NP → stars	0.18
V → saw	1.0	NP → telescopes	0.1

Parse Ranking

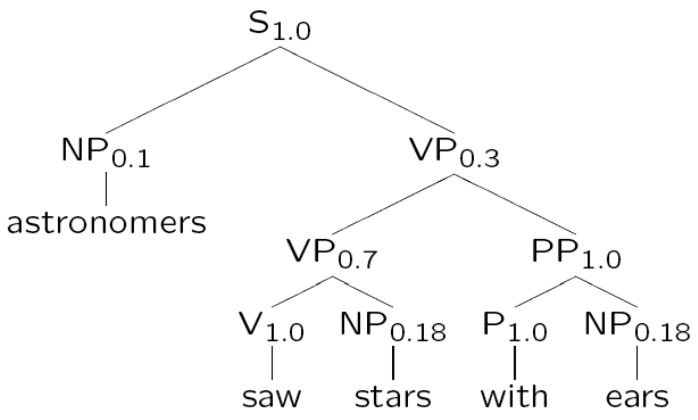
t_1:



$$P(t_1) = 1.0\times0.1\times0.7\times1.0\times0.4\times0.18\times1.0\times1.0\times0.18 = 0.0009072$$

Parse Ranking

t_2:



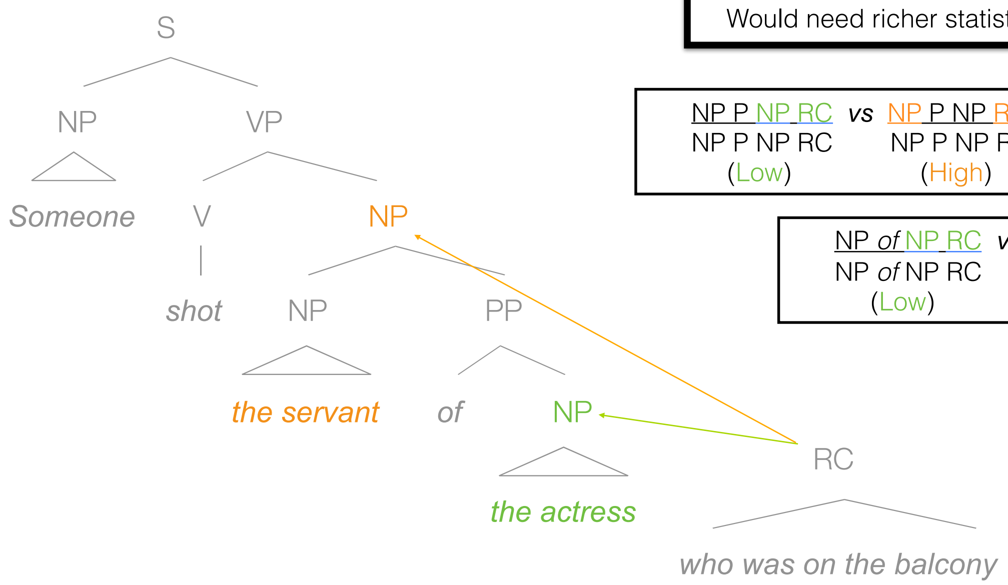
$$P(t_1) = 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0006804$$

Recall the Grain Problem

Note: PCFG-derived probabilities will be the same for both structures.
Would need richer statistics to capture!

NP P NP RC	*vs*	NP P NP RC
NP P NP RC		NP P NP RC
(Low)		(High)

NP *of* NP RC	*vs*	NP *of* NP RC
NP *of* NP RC		NP *of* NP RC
(Low)		(High)



Alguien disparo contra el criado de la actriz que estaba en al balcon

Methodological advantages

- Transparently combine symbolic and stochastic mechanisms
  - Associate probabilities with rules and representation
- Scaleable, predictive models
  - Supervised training is well understood
  - Independent empirical basis for establishing the parameters
- Blurring the boundary between rational and empirical
  - Combines existing theories with mechanisms that learn from experience
  - Do probabilities encode “hidden” knowledge/representations?

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

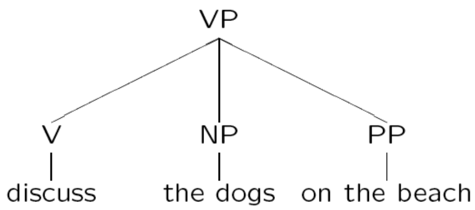
Frame Preferences

- "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:

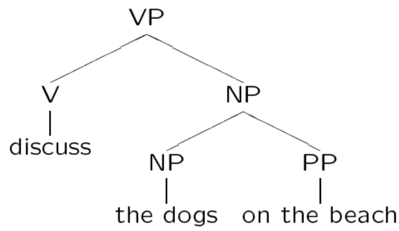


$p(t_1) = 0.15 \times 0.24 = 0.036$ (dispreferred)

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

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)

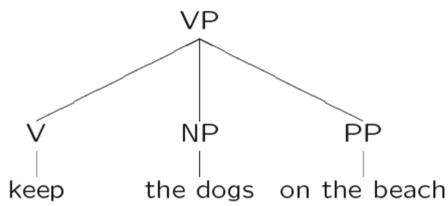
Frame Preferences

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

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

VP → V NP XP 0.15

t_1:

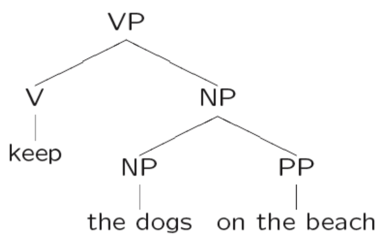


$p(t_1) = 0.15 \times 0.81 = 0.12$ (preferred)

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

VP → V NP 0.39
NP → NP XP 0.14

t_2:

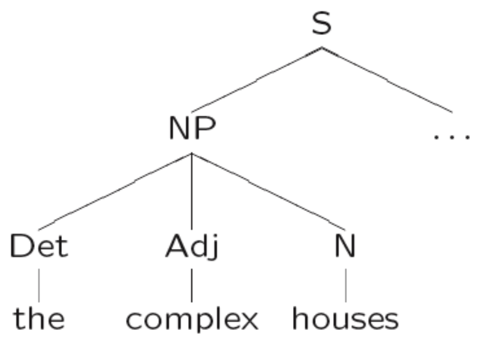


$p(t_2) = 0.19 \times 0.39 \times 0.14 = 0.01$ (dispreferred)

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

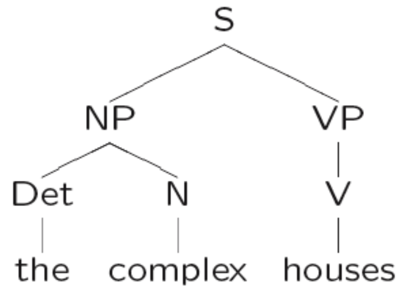
t_1:



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

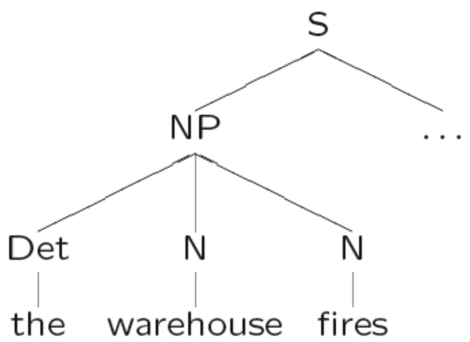


$p(t_1) = 4.5 \times 10^{-10}$ (dispreferred)

Construction Preferences

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

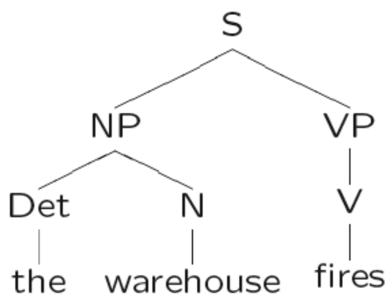
t_1:



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

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

t_1:



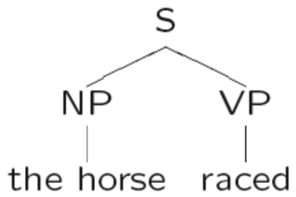
$p(t_1) = 1.1 \times 10^{-5}$ (dispreferred)

Frames *and* Constructions

"The horse raced past the barn fell."

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

t_1:

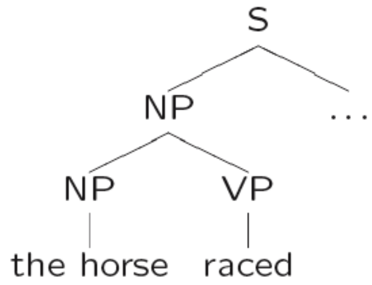


$p(t_1) = 0.92$ (preferred)

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

NP $\rightarrow$ NP XP   0.14

t_2:



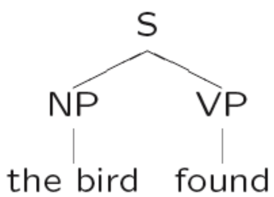
$p(t_1) = 0.0112$ (dispreferred)

Frame and Construction Probs

"The bird found died"

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

t_1:

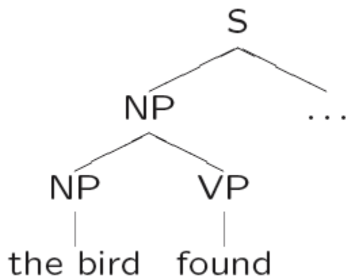


$p(t_1) = 0.38$ (preferred)

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

NP $\rightarrow$ NP XP   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 resolution:
    - What about when category preferences go against syntactic possibilities
- 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), as there is no ambiguity
    - Gibson (1992) used a similar "beam" approach with a memory load heuristic
- Does the model make the right predictions when scaled up?

Psychological Plausibility

- Are wide-coverage, probabilistic models cognitively plausible?
- Broad coverage probabilistic parsers:
  - High accuracy: 86% precision/recall
  - Robust: Analyse all and ill-formed input
  - But: Non-incremental & massively parallel
- What is the general performance of probabilistic parser that:
  - Has restricted memory resources
  - Strictly incremental parsing (and pruning)

Design of the Experiment

- Adapted a standard Stochastic Context Free Grammar:
  - Incremental Processing: full processing on each word, no lookahead
    - Immediate pruning: reduces memory requirements
    - Pruning: active/inactive/both
      - Variable Beam: edges close to best are kept (like Jurafsky)
      - Fixed Beam: fixed number of best edges are kept
- Training: Wall street journal sections 2-21
- Testing: From section 22 (1578 sentences of length 40 or less)

Results for Incremental SCFG

- Baseline performance:
  - Recall: 68.82%
  - Precision: 73.77%
  - Chart size: 141,650
  - Avg # of analysis per span: 18.7
  - Speed: 1.8 Tokens/Sec

F-Score: 71.21

- Restricted model:
  - Recall: 68.82%
  - Precision: 73.66%
  - Chart size: 1.15%
  - Avg # of analysis per span: 2
  - Speed: 301 Tokens/Sec
  - Fixed beam (inactive: 2 active: 4)

F-Score: 71.16

Interim Summary

- Wide coverage grammar, good overall performance
  - Accounts for specific lexical/syntactic local ambiguities
  - Sacrifices linguistic fidelity/richness
- Cognitive plausibility? Brants & Crocker (2000)
  - Psychological Plausibility: Incrementality & Restricted Memory
  - No degradation in accuracy
  - Memory: 100 x less
  - Speed: 100 x faster

Summary of Jurafsky

- Probabilistic grammars offer rational account of lexical and syntactic disambiguation in parsing
- Can be easily scaled, and also restricted to meet considerations of cognitive plausibility
- Jurafsky's model, however, does not explain behaviour (i.e. reading times) beyond POS tag models (but does yield a syntactic analysis).
- Also, coarse-grained linking hypothesis to processing difficulty.