Computational Psycholinguistics

Lecture 7: Probabilistic Parsing

Matthew W Crocker

Computerlinguistik
Universität des Saarlandes

Rational Analysis

- Hypothesis: People approach optimal adaptation to the task of language understanding.

Rational Analysis: *when a cognitive system is optimally adapted*

- Goals: *Obtain the most likely interpretation*
- Environment: *Input is incremental and ambiguous*
- Computational: *Finiteness, 'foregrounded' interpretation*

Constructing a Rational Analysis:

1. Derive the Optimal Function
2. Test against the empirical data
3. Revise the Optimal Function

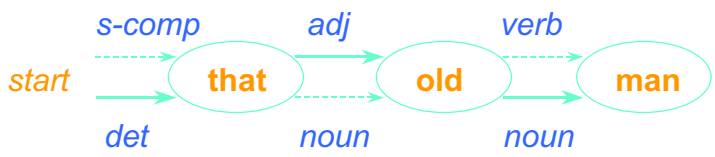
- Use probabilistic frameworks to reason about rational choice
  - Initial hypothesis: The optimal function is one which maximises the likelihood of obtaining the correct intepretation of an utterance

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:
  - Lexical bias: $P(w_i|t_i)$
  - Category context: $P(t_i|_{ti-1})$
- Categories are assigned incrementally
- Best category path may require revision

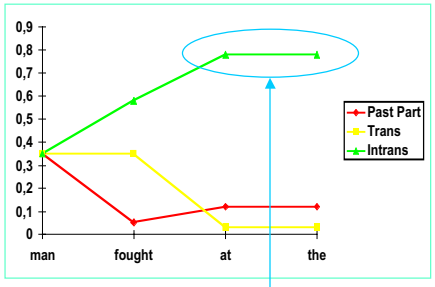


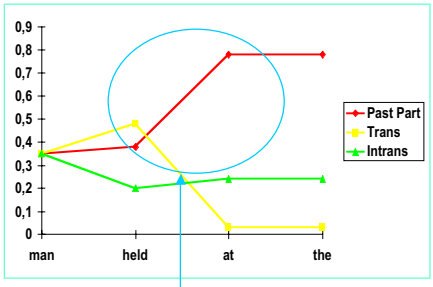
Internal Reanalysis

- 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 bias is used

An SLCM Account

- Assume transitive/intransitive subcategories
  - Extracted transitivity from the Susanne corpus
  - Simulation with (similar) examples:
    - The man *fought at the police* station fainted [intransitive]
    - The man *held at the police* station fainted [transitive]





Correctly predicts the garden path effect

Correctly predicts immediate reanalysis

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 → initial category decisions ✓
  - Modular: syntax → initial category decisions ✗
  - Bigram effect: "that" ambiguity [Juliano and Tanenhaus]
  - Reanalysis of verb transitivity for 'reduced relatives' [MacDonald]

➔ Comments:
  - combines optimality with psychological plausibility
  - category preference appears truly frequency-based
  - indication of which features are exploited [e.g. transitivity, not number]

Probabilistic Parsing

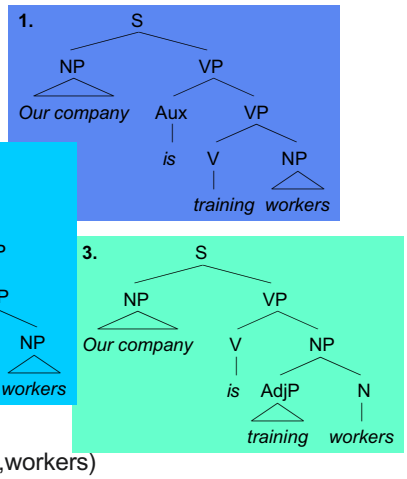
- Determine the most likely analysis for a given input:

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

- Use estimates based on frequencies in prior experience
  - I.e. estimate P(Si) based on frequencies
- Not that simple: what about context?
  - Should recent utterances affect the probability more than old utterances?
  - Does the current situation (non-linguistic) affect probabilities?
- Mechanisms:
  - To compare probabilities (of the S_i), we assume parallelism. How much?
  - How does the human parsing mechanism affect estimation and exploitation of probabilities in both parsing and reanalysis

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
- E.G.
  - P(training|Adj) vs. P(training|V)
  - P(NP|training,workers) vs P(VP|training,workers)

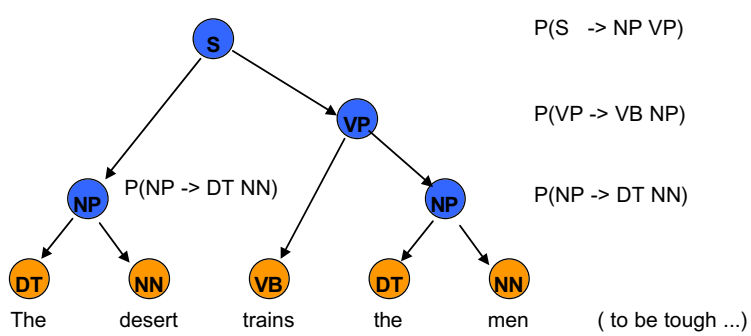


Towards a wide-coverage model: ICMM

- Crocker and Brants (2000): A Wide-Coverage Probabilistic Model of Human Sentence Processing
  - ICMM: Incremental Cascaded Markov Model
    - Standard HMM POS tagger for lexical categories, similar to SLCM
    - Structural probabilities computed as in a SCFG
    - Cascaded Markov Models are also used to help filter out structures
  - Wide coverage:
    - A fully implemented, wide coverage parser
    - Trained on parsed corpora: Brown, WSJ, NEGRA
    - Not hand coded
    - Adapted to operate incrementally
- Issues:
  - Linguistic representations (particularly lexical) are not very rich:
    - I.e. no full estimation of Frame probabilities, only some sub-categories
  - Psychological plausibility: what about memory requirements?

Stochastic Context-Free Parsing

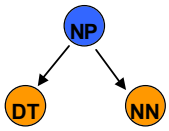
- Context-free grammar rules are associated with probabilities
- Probability of a parse is the product of the rules' probabilities
- Best parse: $\boxed{\arg\max_i P(s_i) \text{ for all } s_i \in S}$



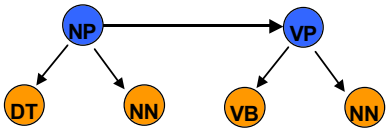
Markov Models for Tagging & Parsing

- Markov Models for part-of-speech tagging use "horizontal" probabilities: SLCM (Corley & Crocker) $P = P(W_i|T_i)xP(T_i|T_{i-1})$

- Stochastic context-free grammars use "vertical" probabilities

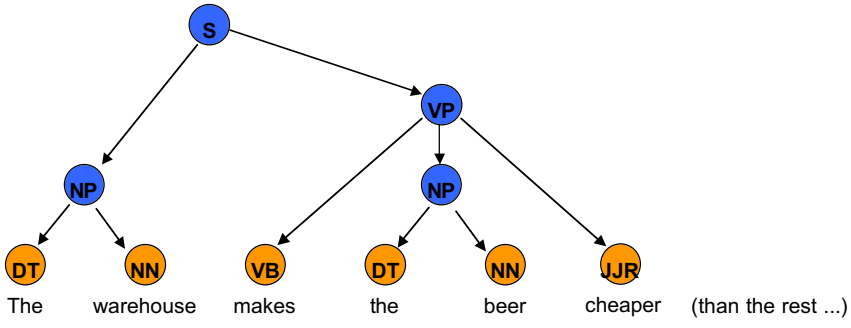


- Cascaded Markov Models apply "horizontal" probabilities to levels higher than parts-of-speech



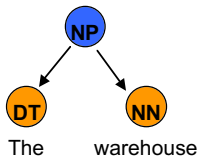
Incremental Cascaded Markov Models

- Incremental (word-by-word) processing
- Build hypotheses for all layers as soon as a word is read
- Filter hypotheses with Markov Models



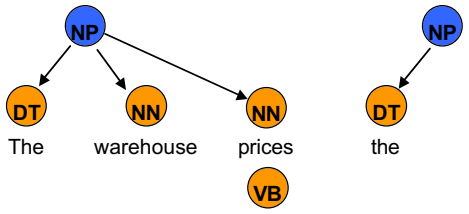
Incremental Cascaded Markov Models

- Incremental (word-by-word) processing
- Build hypotheses for all layers as soon as a word is read
- Filter hypotheses with Markov Models



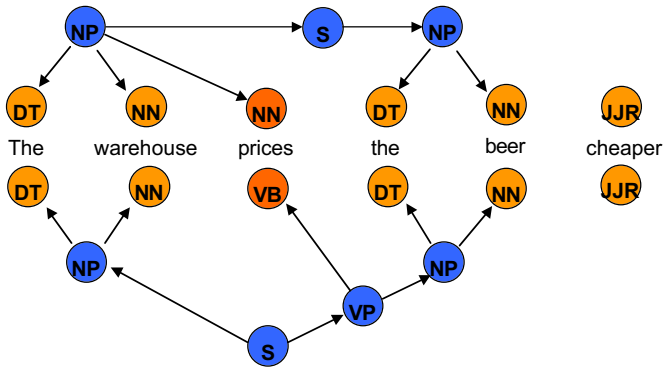
Incremental Cascaded Markov Models

- Incremental (word-by-word) processing
- Build hypotheses for all layers as soon as a word is read
- Filter hypotheses with Markov Models



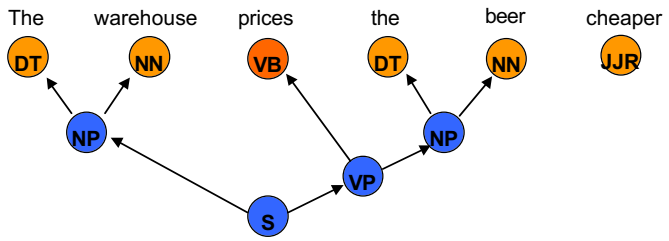
Incremental Cascaded Markov Models

- Incremental (word-by-word) processing
- Build hypotheses for all layers as soon as a word is read
- Filter hypotheses with Markov Models



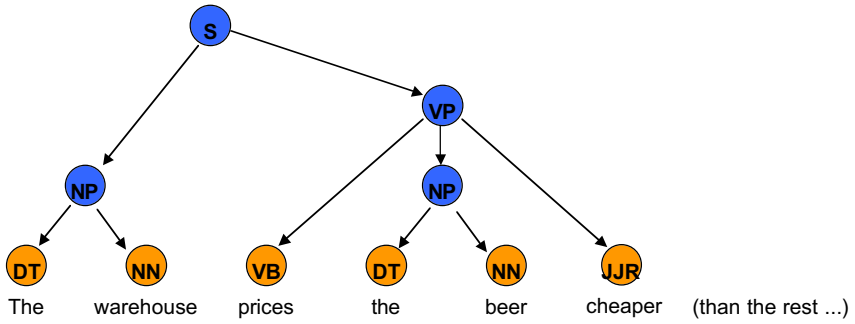
Incremental Cascaded Markov Models

- Incremental (word-by-word) processing
- Build hypotheses for all layers as soon as a word is read
- Filter hypotheses with Markov Models



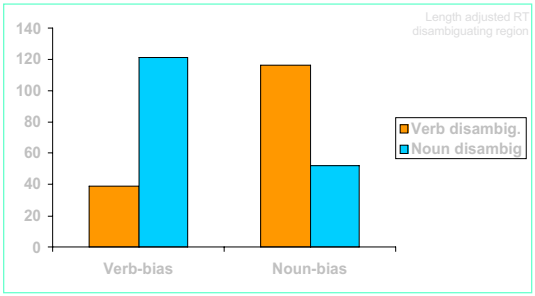
Incremental Cascaded Markov Models

- Incremental (word-by-word) processing
- Build hypotheses for all layers as soon as a word is read
- Filter hypotheses with Markov Models



Noun-Verb Ambiguity: Crocker & Corley (1998)

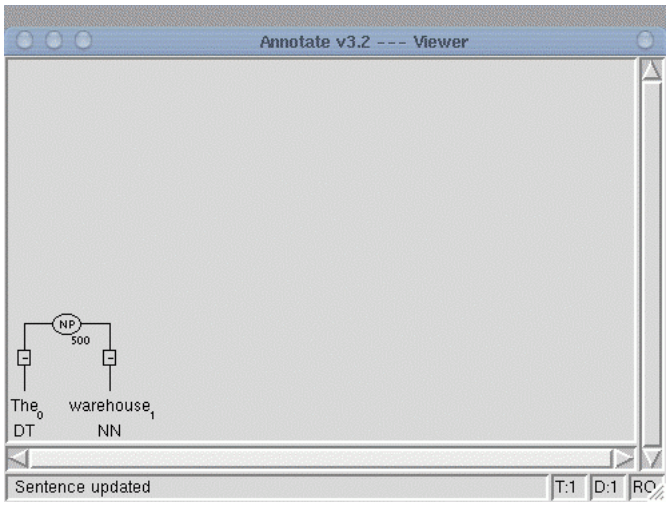
- Categories are assigned incrementally
  - the warehouse *prices* the beer very modestly
  - DET N N / V V!
  - the warehouse *makes* are cheaper than the rest
  - DET N V / N N! ...



- Lexical category frequency determine initial decisions

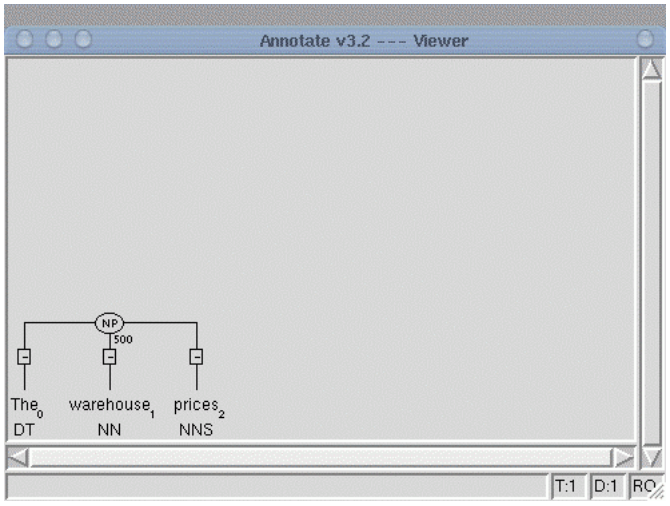
Modelling Noun-Verb ambiguity: 1

- *The warehouse*



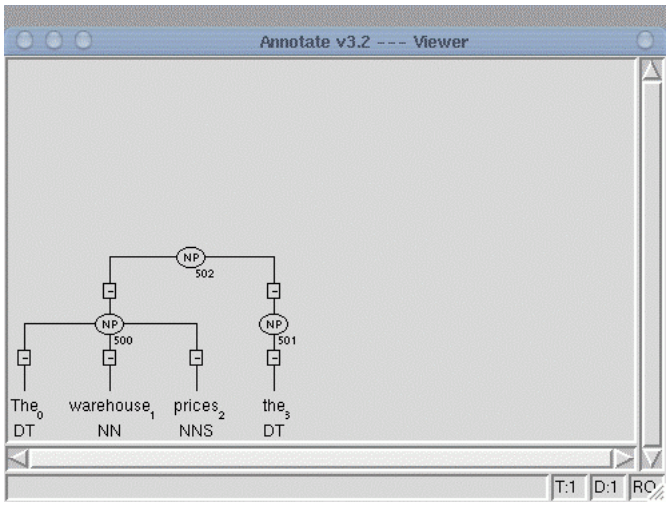
Modelling Noun-Verb ambiguity: 2

- *The warehouse prices*



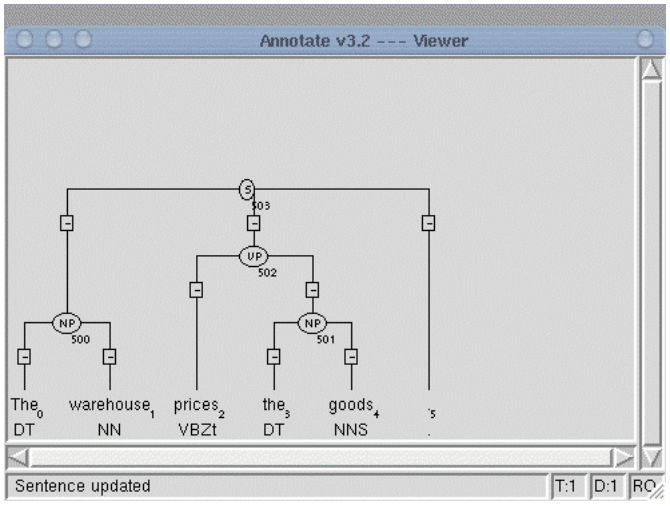
Modelling Noun-Verb ambiguity: 3

- *The warehouse prices the*



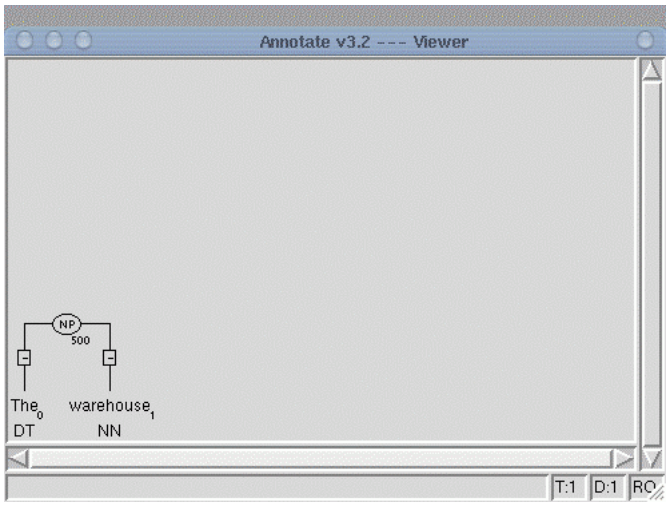
Modelling Noun-Verb ambiguity: 4

- *The warehouse prices the goods*



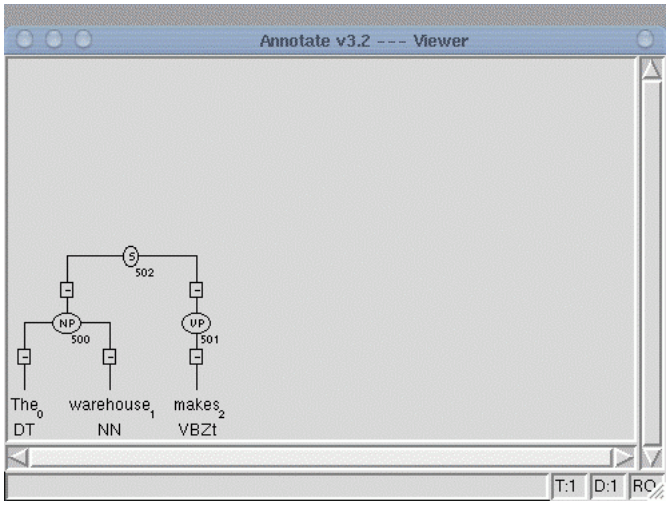
Modelling Noun-Verb ambiguity: 5

- *The warehouse*



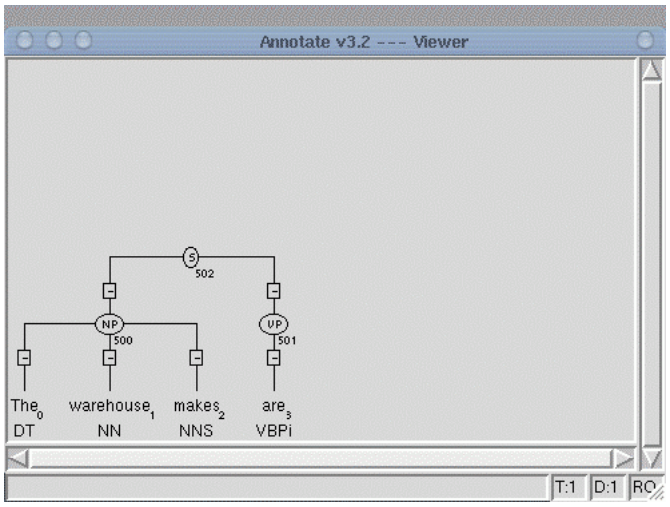
Modelling Noun-Verb ambiguity: 6

- *The warehouse makes*



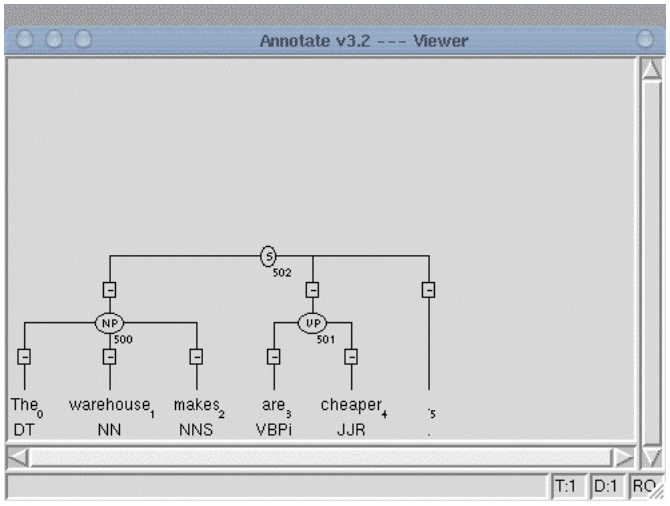
Modelling Noun-Verb ambiguity: 7

- *The warehouse makes* *are*



Modelling Noun-Verb ambiguity: 8

- *The warehouse makes are* *cheaper*

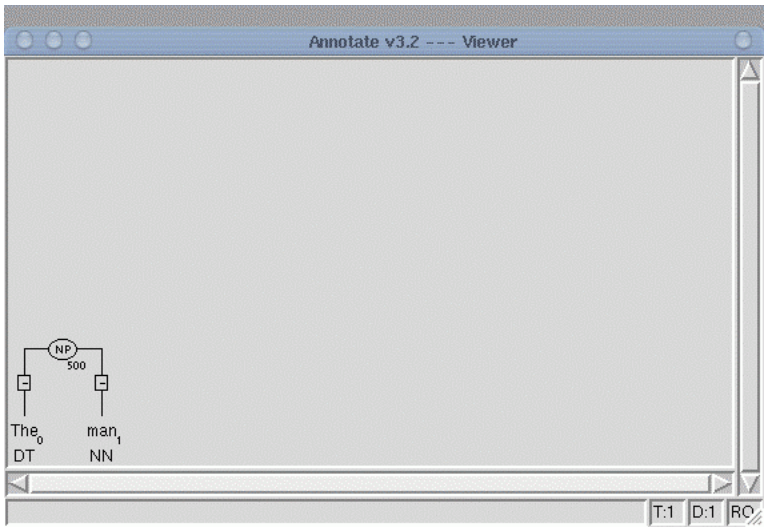


Reduced-Relatives & Post Ambiguity

- Reduced-relative: verb-bias/post-ambiguity (MacDonald)
  - *The man held at the station was innocent.*
    - *Transitive bias: good disambiguation:* ***minimal GP effect***
  - *The man raced to the station was innocent*
    - *Intransitive bias: poor disambiguation:* ***garden-path effect***
  - *The dog shown at the track won four trophies*
    - ***Unambiguous***

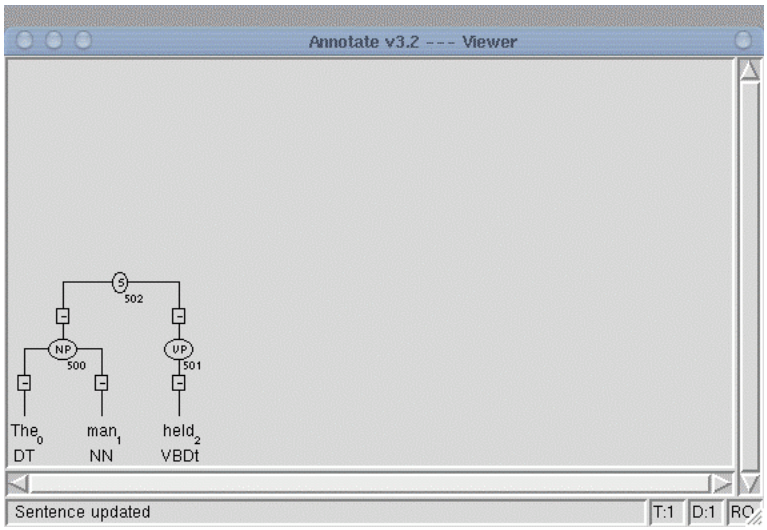
Reduced Relative Ambiguity: 1

- *The man*



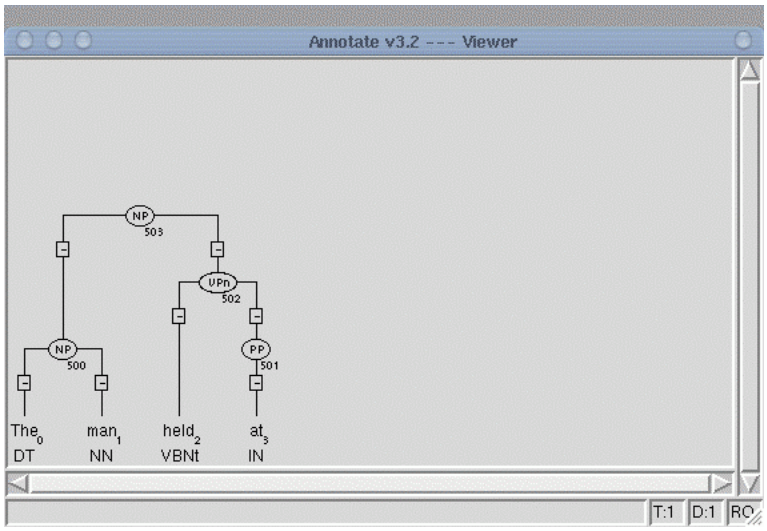
Reduced Relative Ambiguity: 2

- *The man held*



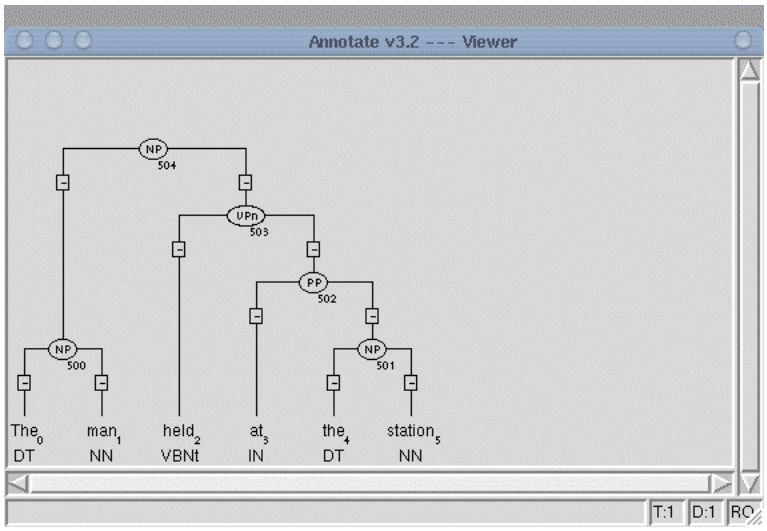
Reduced Relative Ambiguity: 3

- *The man held at*



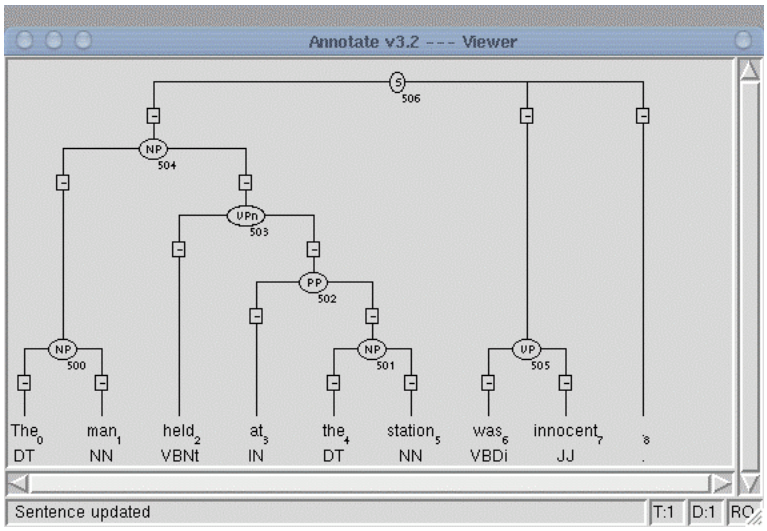
Reduced Relative Ambiguity: 4

- *The man held at the station*



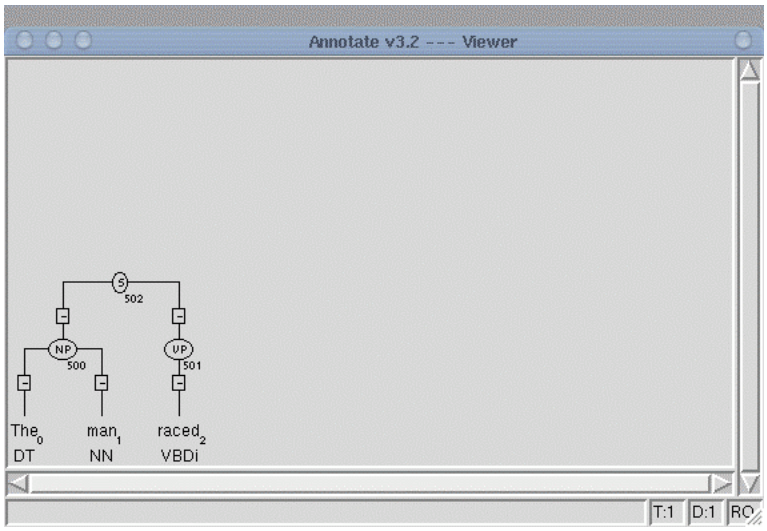
Reduced Relative Ambiguity: 5

- *The man held at the station was innocent*



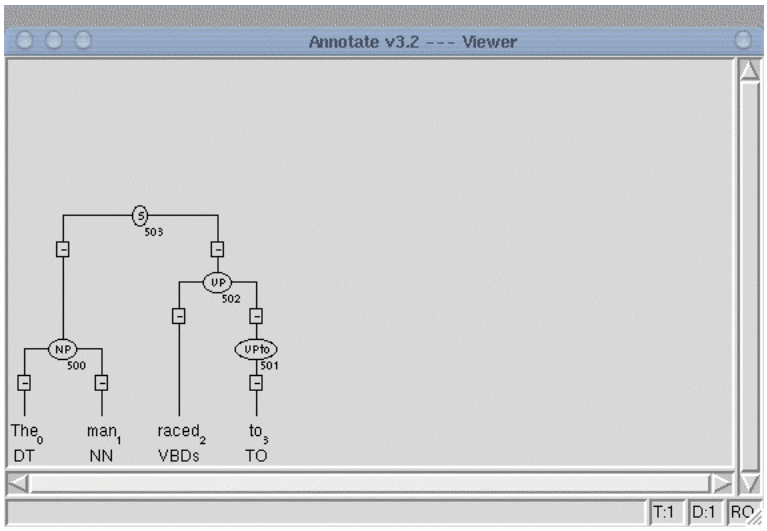
Reduced Relative Ambiguity: 6

- *The man raced*



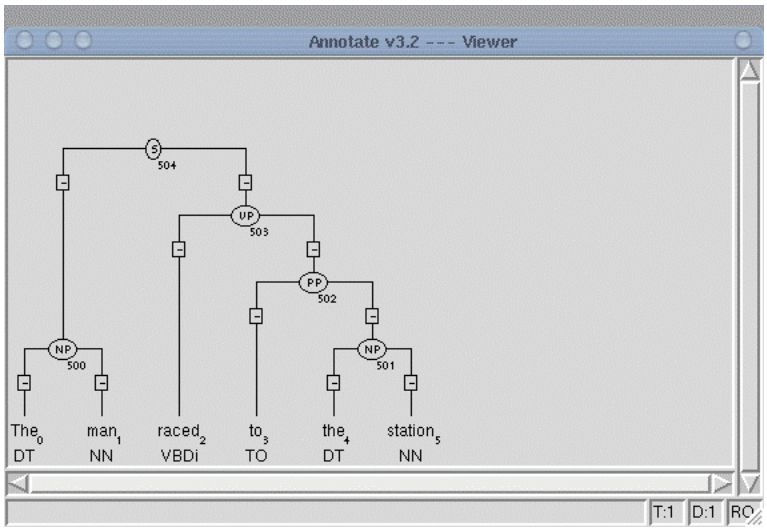
Reduced Relative Ambiguity: 7

- *The man raced to*



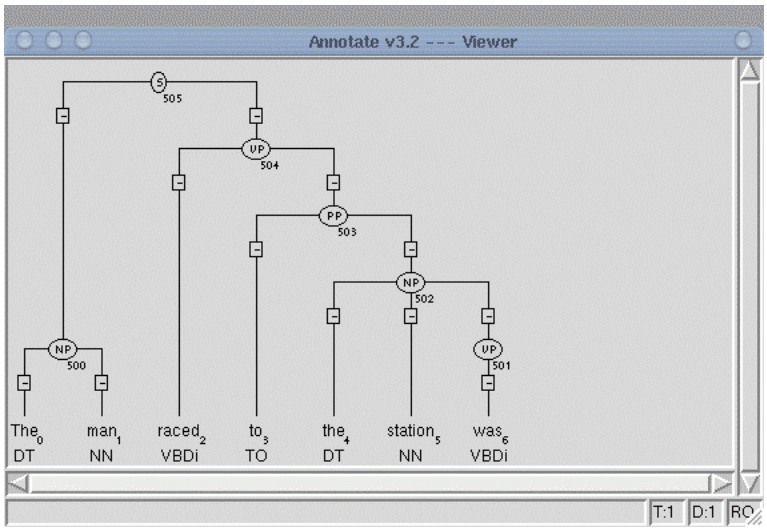
Reduced Relative Ambiguity: 8

- *The man raced to the station*



Reduced Relative Ambiguity: 9

- *The man raced to the station was innocent*



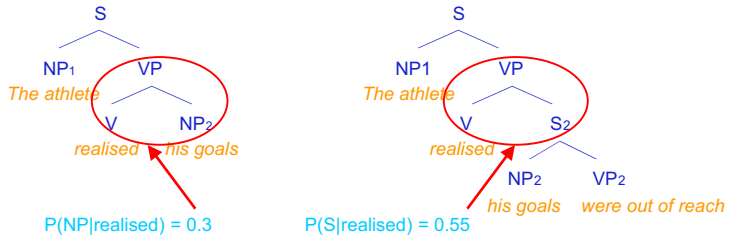
Summary of Data

- Noun/Verb category ambiguity: (Corley & Crocker, MacDonald)
  - The warehouse makes/prices …
- Reduced-relative: verb-bias/post-ambiguity (MacDonald)
  - The dog raced at the track won four trophies
  - The dog admired at the track won four trophies
- 'That' = det or comp: (Julian & Tanenhaus, also Gibson)
  - That experienced diplomat(s) would be ...
  - The lawyer insisted that experiences diplomat(s) is/are
- Low attachment preference

➔ NP/S & NP/Z attachment preference, against likelihood
  - The athlete realized his shoes were out of reach
  - While the man was sailing the ship sank

NP/S Ambiguity: A Problem for Likelihood?

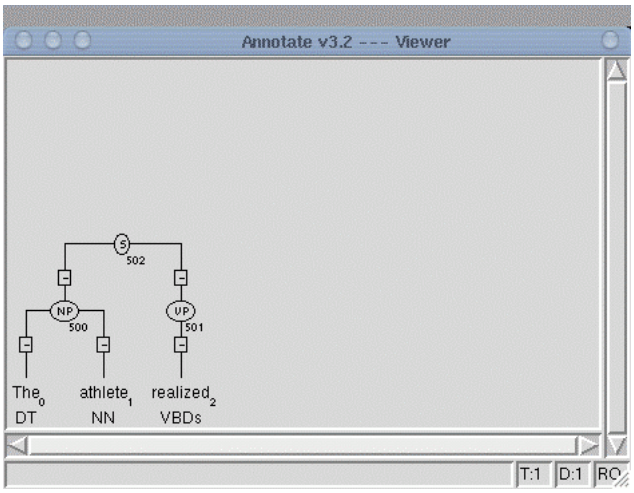
- NP/S Complement Ambiguity: *The athlete realised his goals ...*



- Evidence for object attachment: [Pickering, Traxler, Crocker, 2000]
  - The athlete realized his shoes were out of reach
    - Despite S-comp bias of verb, NP is initially attached as direct object
  - In ICMM, realized is initially tagged at S-comp, but the simpler DO analysis is then given higher probability, when NP is found:
    - In general, prefers simpler, flatter structures (e.g V-attachment of PPs)

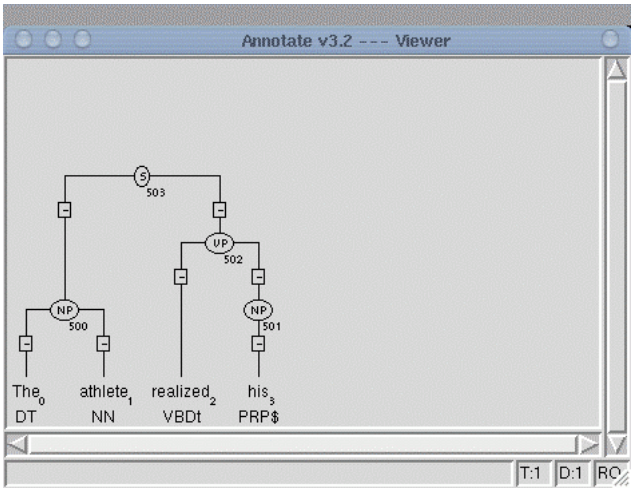
NP/S Ambiguity: 1

- *The athlete* *realized*



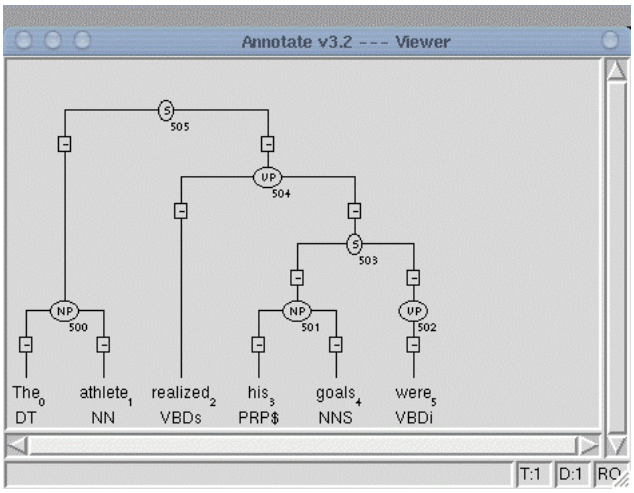
NP/S Ambiguity: 2

- *The athlete realized* *his*



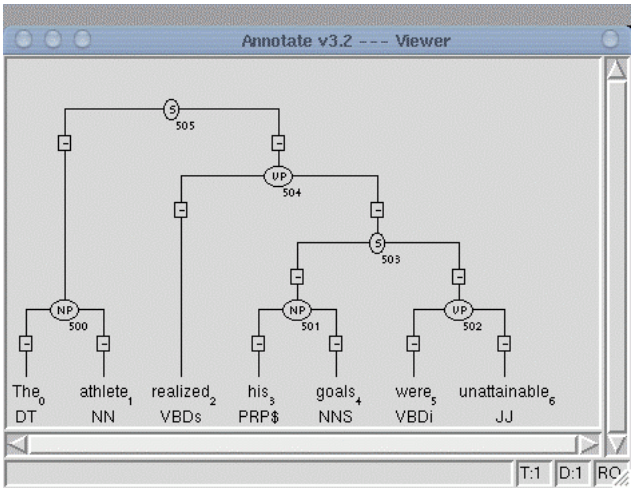
NP/S Ambiguity: 3

- *The athlete realized his goals* were



NP/S Ambiguity: 4

- *The athlete realized his goals were* unattainable



Probabilistic Context-free Grammars

- Context-free rules annotated with probabilities;
- Probabilities of all rules with the same left hand side sum to one;
- Probability of a parse is the product of the probabilities of all rules applied in the parse.

- 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

Jurafsky (1996)

- Jurafsky's (1996) approach:
  - probabilistic model of lexical and syntactic access and disambiguation;
  - accounts for psycholinguistic data using concepts from computational linguistics: probabilistic CFGs, Bayesian modeling frame probabilities;
  - focus here: syntactic disambiguation in human sentence processing.

- Overview of issues:
  - data to be modeled: frame preferences, garden paths;
  - architecture: serial, parallel, limited parallel;
  - probabilistic CFGs, frame probabilities;
  - examples for frame preferences, garden paths;
  - comparison with other models; problems and issues.

Frame Preferences

- (1) The women discussed the dogs on the beach.
  - a. The women discussed the dogs which were on the beach. (90%)
  - b. The women discussed them (the dogs) while on the beach. (10%)
- (2) The women kept the dogs on the beach.
  - a. The women kept the dogs which were on the beach. (5%)
  - b. The women discussed them (the dogs) while on the beach. (95%)

Parser Architecture

- Serial Parser
  - garden path means: wrong tree was selected at a choice point;
  - backtracking occurs, causes increased processing times.
- Parallel Parser
  - garden path means: correct tree was pruned;
  - backtracking occurs, causes increased processing times.
- Jurafsky (1996) assumes bounded parallelism in a parsing model based on probabilistic CFGs.
  - Pruning occurs if a parse tree is sufficiently improbable (beam search algorithm).