Computational Psycholinguistics

Lecture 8: **Probabilistic Parsing 2**

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

Jurafsky (1996)

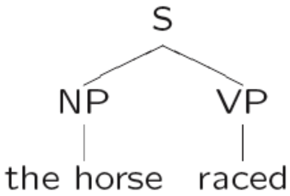
- Psycholinguistic model of lexical and syntactic access and disambiguation
- Exploits concepts from statistical parsing
    - Probabilistic CFGs
    - Bayesian modeling frame probabilities
- Architecture: Probabilistic, bounded, parallel parser
    - Parses are "pruned" (removed from memory) if they 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

Frames *and* Constructions

"The horse raced past the barn fell."

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

t_1:

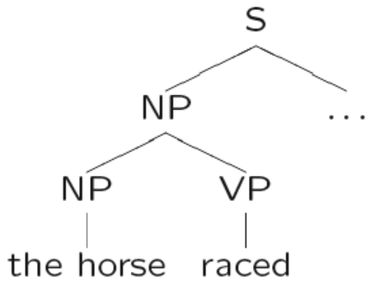


$p(t_1) = 0.92$ (preferred)

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

NP $\rightarrow$ NP XP   0.14

t_2:



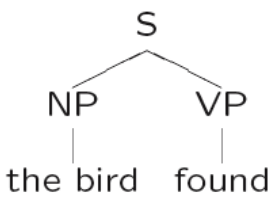
$p(t_1) = 0.0112$ (dispreferred)

Frame and Construction Probs

"The bird found died"

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

t_1:

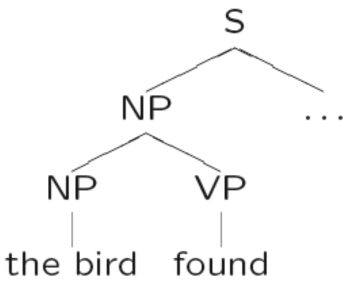


$p(t_1) = 0.38$ (preferred)

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

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

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 PCFG
  - 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
  - Adapted to operate incrementally

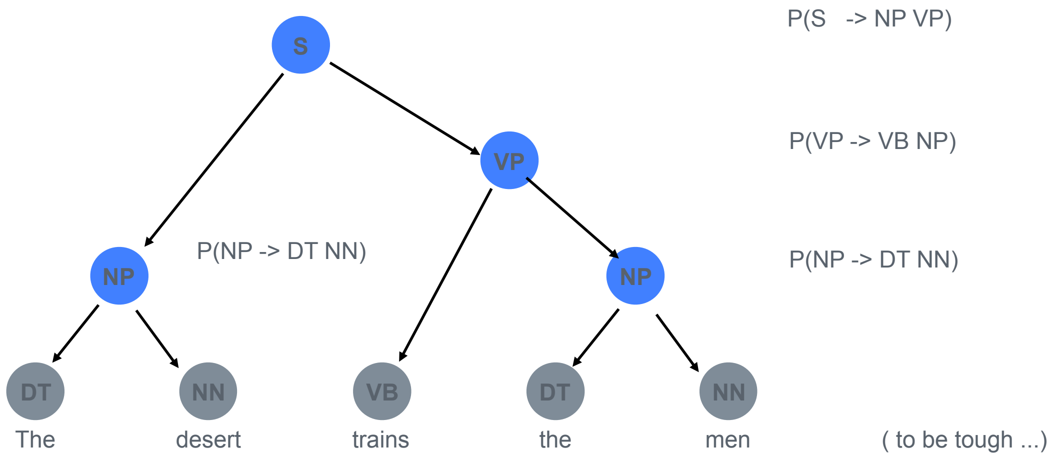
Crocker & Brants, *J.Psych.Res*, 2000

Stochastic Context-Free Parsing

- Probability of a parse is the product of the rules' probabilities

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

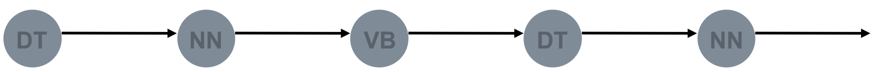
- Best parse:



Crocker & Brants, *J.Psych.Res*, 2000

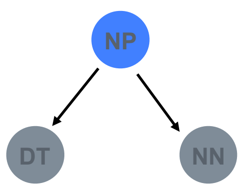
Probabilistic Tagging & Parsing

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



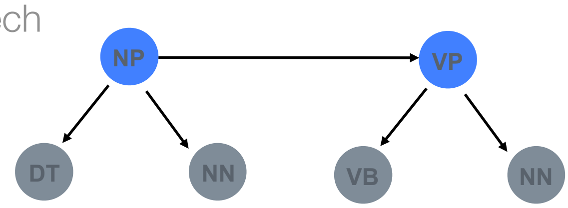
- Stochastic context-free grammars use

  "vertical" probabilities *P(DT,NN|NP)*



- Cascaded Markov Models apply "horizontal" probabilities

  to levels higher than parts-of-speech



Crocker & Brants, *J.Psych.Res*, 2000

Incremental Cascaded Markov Models

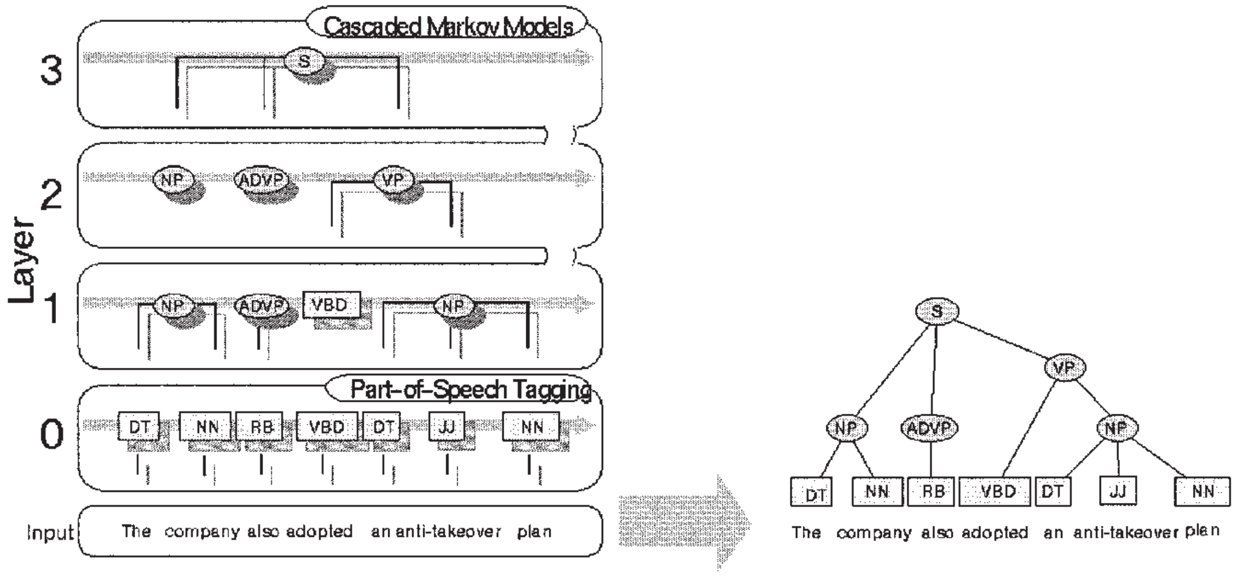
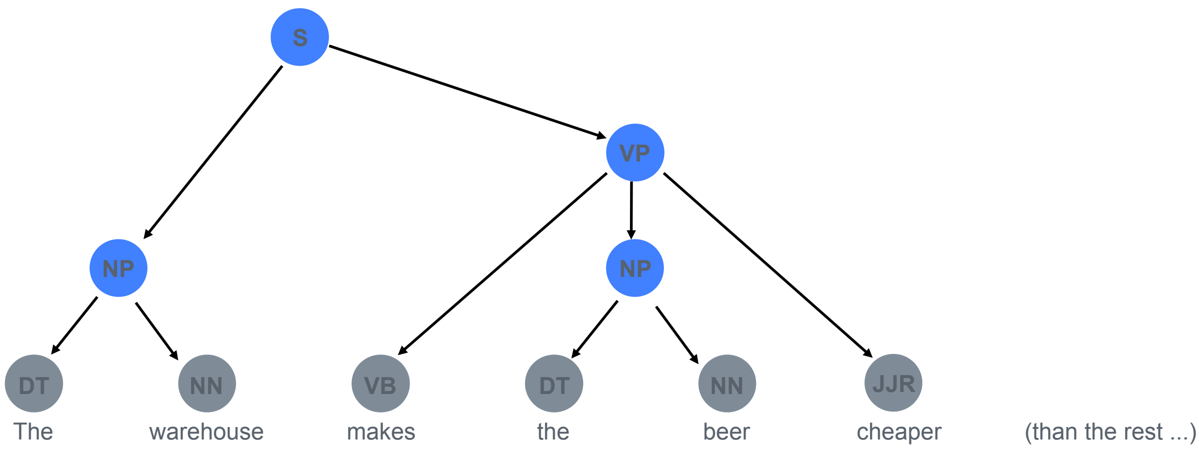


Fig. 1. The layered processing model. Starting with part-of-speech tagging (layer 0), possibly ambiguous output together with probabilities is passed to higher layers (only the best hypotheses are shown for clarity). At each layer, new phrases are added and filtered with a Markov model.

Crocker & Brants, *J.Psych.Res*, 2000

ICMM

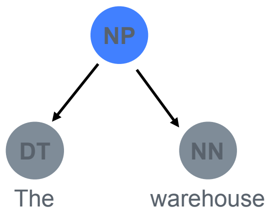
- Incrementally build hypotheses for all layers as soon as a word is read
- Filter hypotheses with Markov Models



ICMM

- Incrementally build hypotheses for all layers as soon as a word is read
- Filter hypotheses with Markov Models

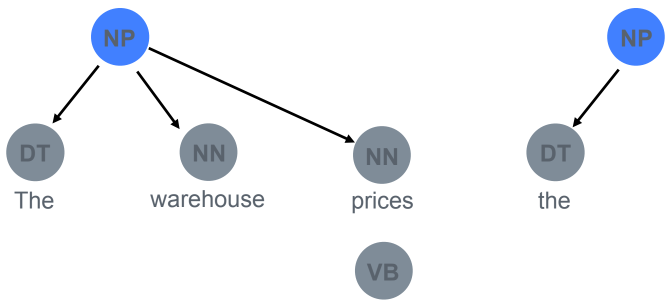
The warehouse prices the beer cheaper than the rest



ICMM

- Incrementally build hypotheses for all layers as soon as a word is read
- Filter hypotheses with Markov Models

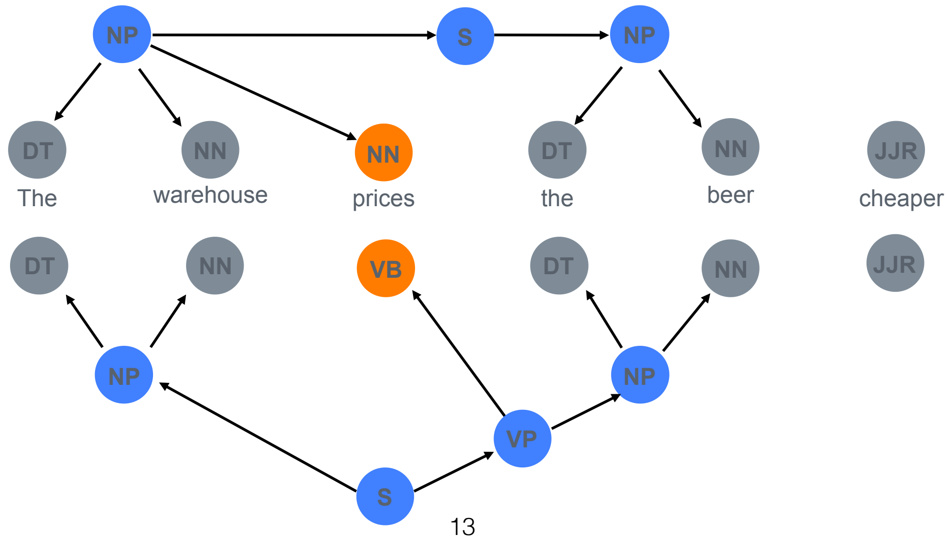
The warehouse prices the beer cheaper than the rest



ICMM

- Incremental build hypotheses for all layers as soon as a word is read
- Filter hypotheses with Markov Models

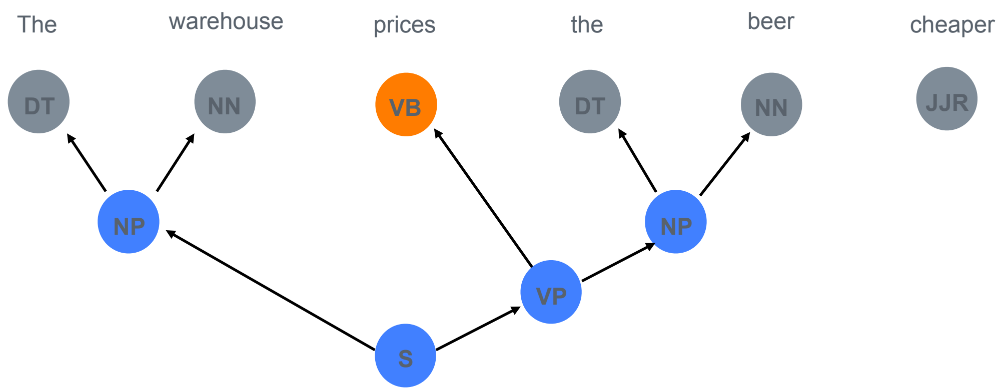
The warehouse prices the beer cheaper than the rest



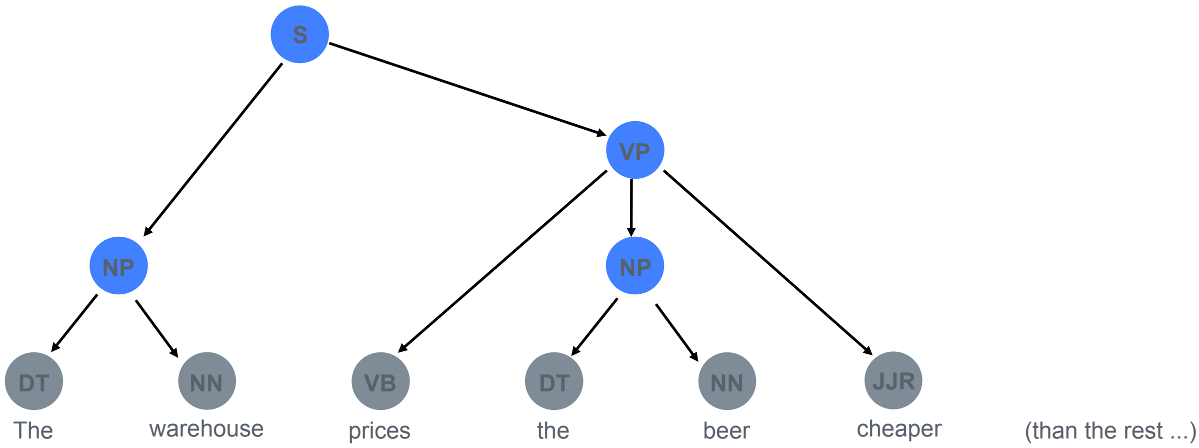
ICMM

- Incrementally build hypotheses for all layers as soon as a word is read
- Filter hypotheses with Markov Models

The warehouse prices the beer cheaper than the rest



Incremental Cascaded Markov Models

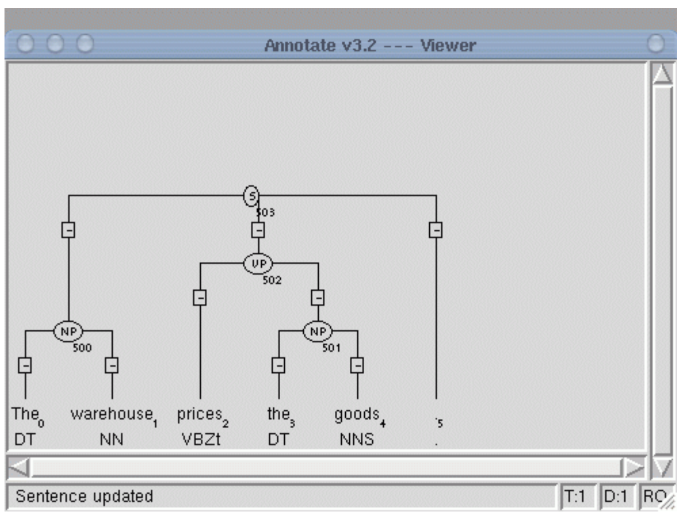


- Incremental (word-by-word) processing
- Build hypotheses for all layers as soon as a word is read
- Filter hypotheses with Markov Models

Crocker & Brants, *J.Psych.Res*, 2000

Noun-Verb Ambiguity

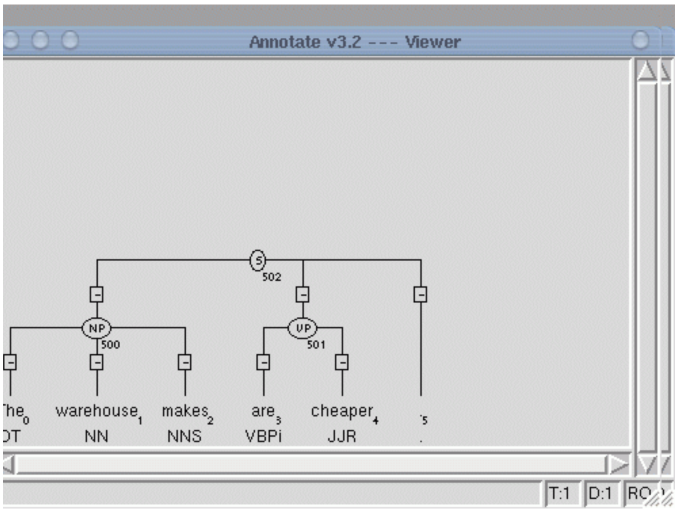
- Initial preference based mostly on SLCM component



"The warehouse prices the goods."

Noun-Verb Ambiguity

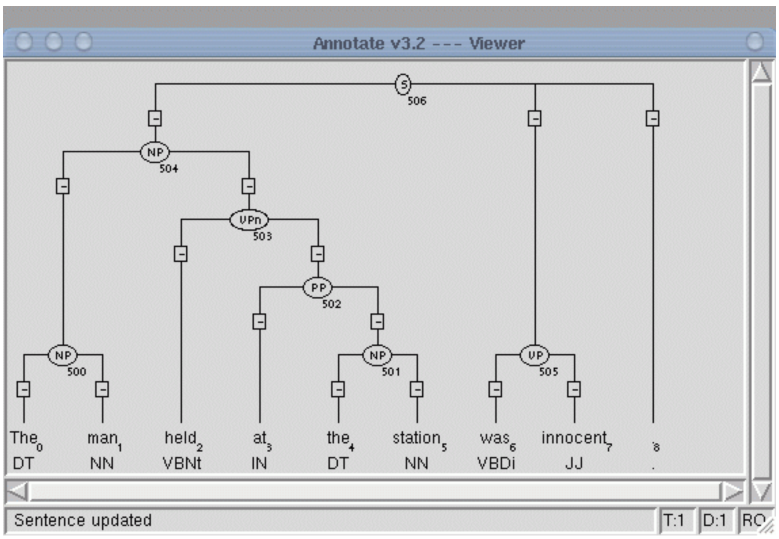
- Initial preference based mostly on SLCM component



"The warehouse makes are cheaper."

Reduced Relative Clause

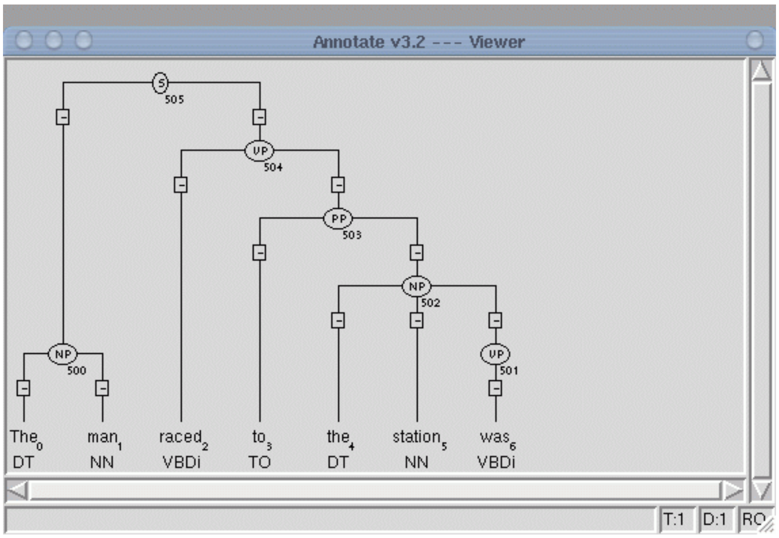
- Initial preference and reanalysis based purely on SLCM component



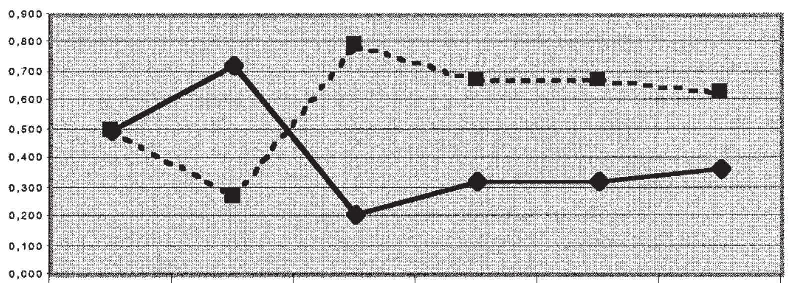
"The man held at the station was innocent"

Reduced Relative Clause

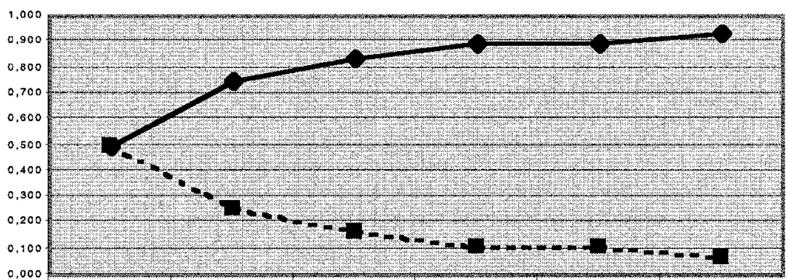
- Initial preference based purely on SLCM, leads to garden path



"The man raced to the station was innocent"



	The man	held	at	the	station	was
Main Clause	0,500	0,729	0,208	0,324	0,324	0,371
Red. Relative	0,500	0,271	0,792	0,676	0,676	0,629



	The man	raced	to	the	station	was
Main Clause	0,500	0,751	0,837	0,895	0,895	0,929
Red. Relative	0,500	0,249	0,163	0,105	0,105	0,071

Crocker & Brants, *J.Psych.Res*, 2000

Probabilistic Models, so far ...

- Three models, explain both good performance & many "pathologies"
  - **SLCM**: a hidden Markov model of lexical category disambiguation
  - **Jurafsky**: probabilistic models of parsing and lexical access
    - Combines structure & frame probabilities, not wide coverage.
  - **ICMM**: implementation of a wide-coverage probabilistic parser:
    - Combines "phrase structure", and "phrase sequence" probabilities
- Also: incremental, bounded probabilistic parsers don't lose accuracy, and are much more space/time efficient.

Challenges for Likelihood Models

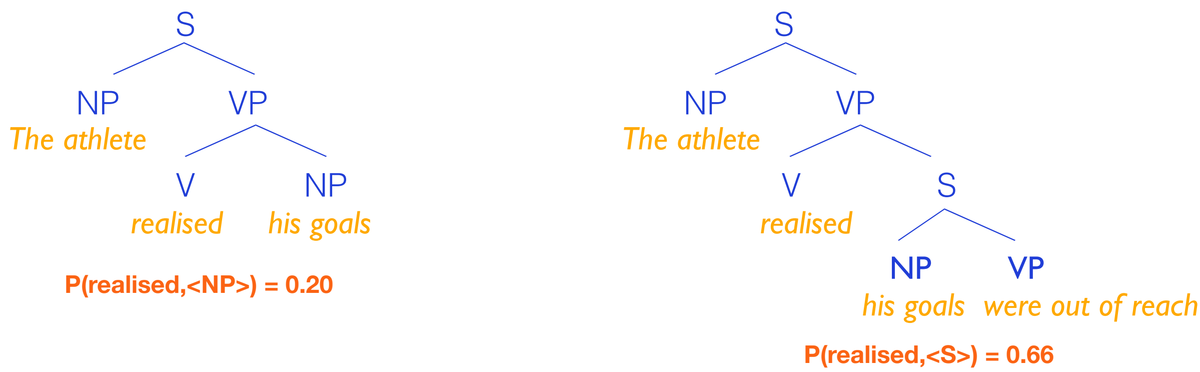
- So far, we've assume that the rational human syntactic processor seeks to optimise (incremental) parse likelihood:

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

- Is this always the case? Recall evidence from non-probabilistic models
  - Minimal attachment: prefer to attach PP to VP – though NP attachment is higher frequency in corpora.
  - Theta attachment: prefer maximal theta-grid – not the most likely
  - Evidence the people consider globally non-syntactic analyses
    - *The coach smiled at the player tossed the frisbee.*
  - Difficultly for grammatical & unambiguous – but memory intensive – structures.

Problem: Subcategorisation likelihood

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



- Evidence for object attachment: (Pickering, Traxler & Crocker 2000)
  - Despite S-comp bias of verb, NP is initially attached as direct object
  - Ideal likelihood models (e.g. Jurafsky) predict the opposite

Refining the Rational Analysis

- Cost: Local reanalysis is often easy, long-distance reanalysis is difficult
  - Decision should take this into account
- Interpretation: Only one (or few) analyses can be 'foregrounded'
  - I.e. only one interpretation is actively attended to, and evaluated
- Solution: Favour 'interpretable' dependencies
  - increase probability of locally backing out of a wrong analysis
  - avoid being led down the garden path by pure likelihood

Rational Analysis of Parsing

- The incremental parsing problem:
  - local ambiguities L_i must be resolved as they are encountered

$$P(\text{global success}) = \prod_{i=1}^{n} P(\text{success at } L_i)$$

  - success = settling on the globally correct analysis
    - Initially adopting an analysis, which is ultimately correct
    - Backing-out of a wrong analysis, and settling on the correct one
- Foreground the analysis which can be confidently rejected or confirmed.

Deriving the optimal function

- Informativity: f(P,T)
- P = prior probability
  - likelihood based on our experience
- T = testability $I(H_i) = P(H_i) \cdot S(H_i)$

$$S(H_i) = \frac{1}{P(\text{Confirm } H_i)}$$

  - measure of how useful new evidence

    E will be in estimating P(H|E).
- Maximise the chance of making the correct analysis, soon.

Consider two hypotheses H_1 & H_2:
$P(Correct_1) = P(H_1,Pass_1) + P(H_2,Fail_1)$
$= P(Pass_1|H_1)P(H_1) + P(Fail_1|H_2)P(H_2)$
$= P(H_1) + (1-1/S(H_1))P(H_2)$
$P(Correct_2) = P(H_2) + (1-1/S(H_2))P(H_1)$

Choose H where P(Correct) is greatest:
$P(Correct_1) > P(Correct_2)$
$P(H_1)+(1-1/S(H_1))P(H_2) > P(H_2)+(1-1/S(H_2))P(H_1)$
…
$S(H_1)P(H_1) > S(H_2)P(H_2)$

So, choose H_i where $P(H_i)S(H_i)$ is maximised

Deriving the optimal function

Consider two hypotheses H1 & H2:

$P(\text{Correct}_1) = P(H_1,\text{Pass}_1) + P(H_2,\text{Fail}_1)$

$= P(\text{Pass}_1|H_1)P(H_1) + P(\text{Fail}_1|H_2)P(H_2)$

$= 1 * P(H_1) + (1\text{-}P(\text{Pass}_1))P(H_2)$

$= P(H_1) + (1\text{-}1/S(H_1))P(H_2)$

$P(\text{Correct}_2) = P(H_2) + (1\text{-}1/S(H_2))P(H_1)$

Likelihood of success if we initially adopt H_1: is the sum of P for H_1 being true and confirmed, and P for H_2 being true and H_1 disconfirmed

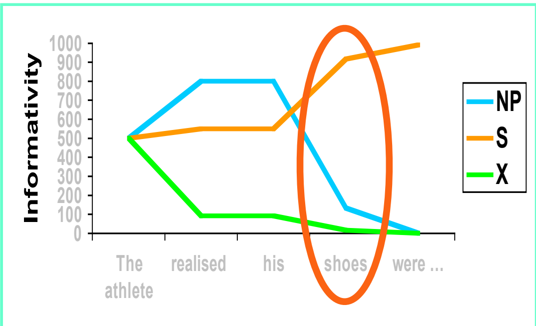
$$I(H_i) = P(H_i) \cdot S(H_i)$$

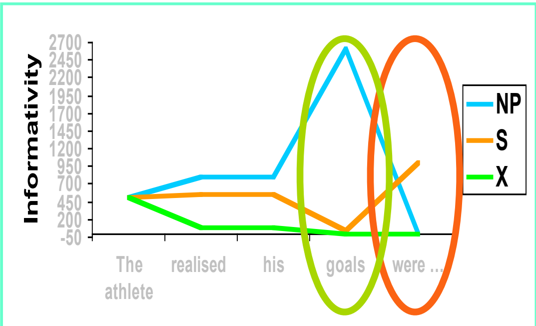
$$S(H_i) = \frac{1}{P(\text{Confirm } H_i)}$$

NP/S Revisited

A. *The athlete realised his goals* ***were out of reach***
B. *The athlete realised his* ***shoes*** *were out of reach*

- Using S-bias verbs (corpus & completion).
- Eye-tracking study revealed:
  - Increased RTs in coloured region
- Consistent with initial object attachment
- Confirms the prediction of the Informativity Model
- Falsifies the analysis based purely on Maximum Likelihood.





Summary of Informativity

- Optimal function incorporates aspects of earlier models:
  - Basic cognitive limitations: serial interpretation + reanalysis
  - Maximising success of reaching correct interpretation
- Explains why people don't always follow likelihood alone
  - Prefer to form interpretable dependencies
  - These can be evaluated as plausible, or trigger reanalysis *quickly*
- Informativity is an idealisation of what the HSPM should approximate

Rational Models

- **Motivation**: People process language: rapidly, robustly, and accurately
  - Experimental evidence for probabilistic mechanisms
- **Maximise Likelihood**: $\arg\max_i P(s_i) \text{ for all } s_i \in S$
  - **SLCM**: Simple, robust account of lexical category disambiguation
  - **Jurafsky**: Probabilistic parser that models a range of local ambiguities
  - **Crocker & Brants**: High accuracy and fast performance with beam search
  - **In common**: all models approximate a maximum likelihood function
  - **Differences**: the underlying symbolic model (n-gram, cfg), and what units of structure are associated with statistical parameters.
- **Informativity**: Motivates a optimal function that combines *Probability(S)* with *Specificity(S)*, where the latter is not unlike Pritchett's Theta-Attachment strategy, since role-receiving constituents are typically more constrained.

Open Problems

- Integrating plausible parsing mechanisms:
    - Either bounded parallel, or serial (momentary parallel) with reanalysis
- Better metrics for linking parser behavior with human processing complexity
- Implementing and evaluating more plausible “optimal functions”:
    - More linguistically informed probabilistic models (lexical, semantic ...)
    - Integration with non-probabilistic decision strategies (recency)
    - More sophisticated integration of memory load constraints