# Computational Psycholinguistics

# Lecture 5: **Probabilistic Acounts**

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

# Experience vs Rules

- The previous accounts adopt purely syntactic mechanisms for disambiguation
  - Assume a serial modular parser & the “primacy” of syntax
  - Initial parsing decisions are guided by syntax/theta-roles alone
- To what extent do non-syntactic **constraints** such as semantics, intonation, and context influence our resolution of ambiguity?
- Are syntactic and non-syntactic constraints probabilistic?
  - Does our prior **experience** with language, determine our preferences for interpreting the sentences we hear?

# Multiple constraints

"The doctor told the woman that ...

*story*

*diet was unhealthy*

*he was in love with her husband*

*he was in love with to leave*

*story was about to leave*

Prosody: intonation can assist disambiguation

Lexical preference: *that* = {Comp, Det, **RelPro**}

Subcat: *told* = { [ _ NP NP] **[ _ NP S]** [ _ NP S'] [ _ NP Inf] }

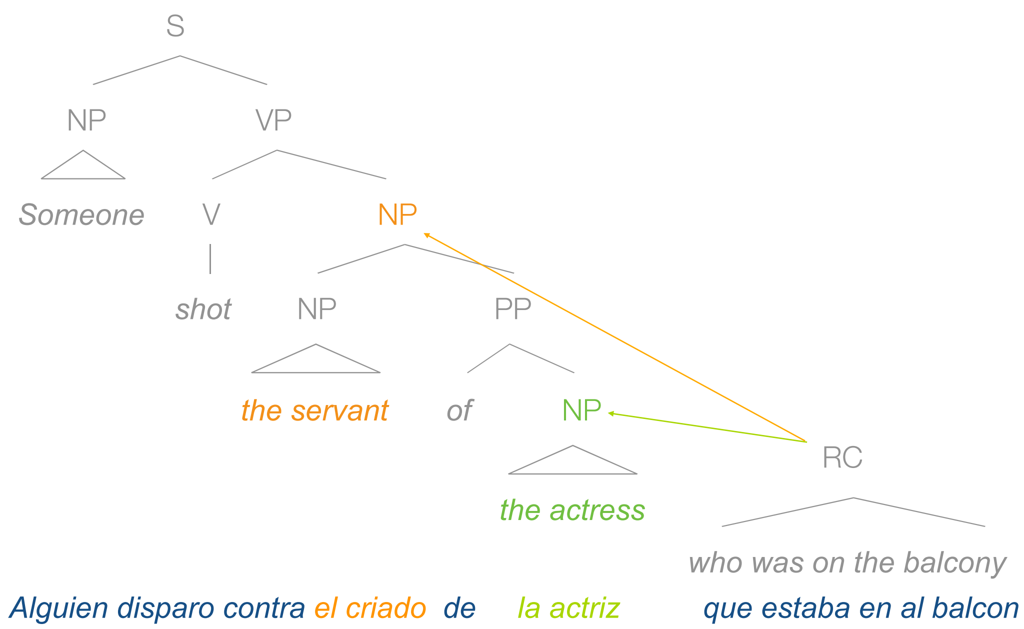
Semantics: Referential context, plausibility

- Reference may determine "argument attach" over "modifier attach"
- Plausibility of *story* versus *diet* as indirect object

# The Role of Experience

- Resolve ambiguities according to linguistic experience, early proposals:
  - **Lexical Guidance Hypothesis**: (Ford et al, 1982)
    - Resolve subcategorisation ambiguities using the most likely frame for the verb
  - **Linguistic Tuning Hypothesis**: (Cuetos et al, 1988;1996)
    - Resolve structural ambiguities according to the structure which has previously prevailed
- Relative clause attachment
  - *"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 | |

- Immediate low attachment, possibly revised quickly (even on-line) … seems the best account

# Probabilistic Models of Language

- Statistics in linguistics [Abney, 1996]
  - Acquisition, change, and variation
  - Ambiguity and graded acceptability
  - Brings ‘performance’ back into linguistics
- Statistics in computational linguistics
  - Effective: accurate and robust
  - Eschews ‘AI’ problem
  - Trainable & efficient

# Probabilistic Psycholinguistics

- Probabilistic models of sentence processing
  - Symbolic parsing models + probabilities (statistical)
  - Interactive, constraint-based accounts (connectionist)
- Probabilistic Models: Breadth and Depth
  - SLCM: Maximal likelihood for category disambiguation (Corley & Crocker)
  - Statistical models of human parsing (Jurafsky, Crocker & Brants)
  - Criticisms of likelihood & Information Theoretic Accounts (Hale, Levy, Demberg)

# Rational Analysis

- "An algorithm is likely understood more readily by understanding **the nature of the problem** being solved than by examining the mechanism (and the hardware) in which it is solved." (Marr, p27)

- **Principle of Rationality:** The cognitive system optimizes the adaptation of the behavior of the organism.

- If a cognitive processes is viewed as rational, then the *computational* theory should reflect optimal adaptation to the task & environment:

    1. Derive the Optimal Function

    2. Test against the empirical data

    3. Revise the Optimal Function

# Garden Path *vs.* Garden Variety

- Human Language Processing: **Garden Paths**

    ✗ Incremental disambiguation process can fail

    ✗ Memory limitations lead to breakdown

    ✗ Garden paths lead to misinterpretations, complexity or breakdown

- Human Language Processing: **Garden Variety**

    ✓ Accurate: typically recover the correct interpretation

    ✓ Robust: are able to interpret ungrammatical & noisy input

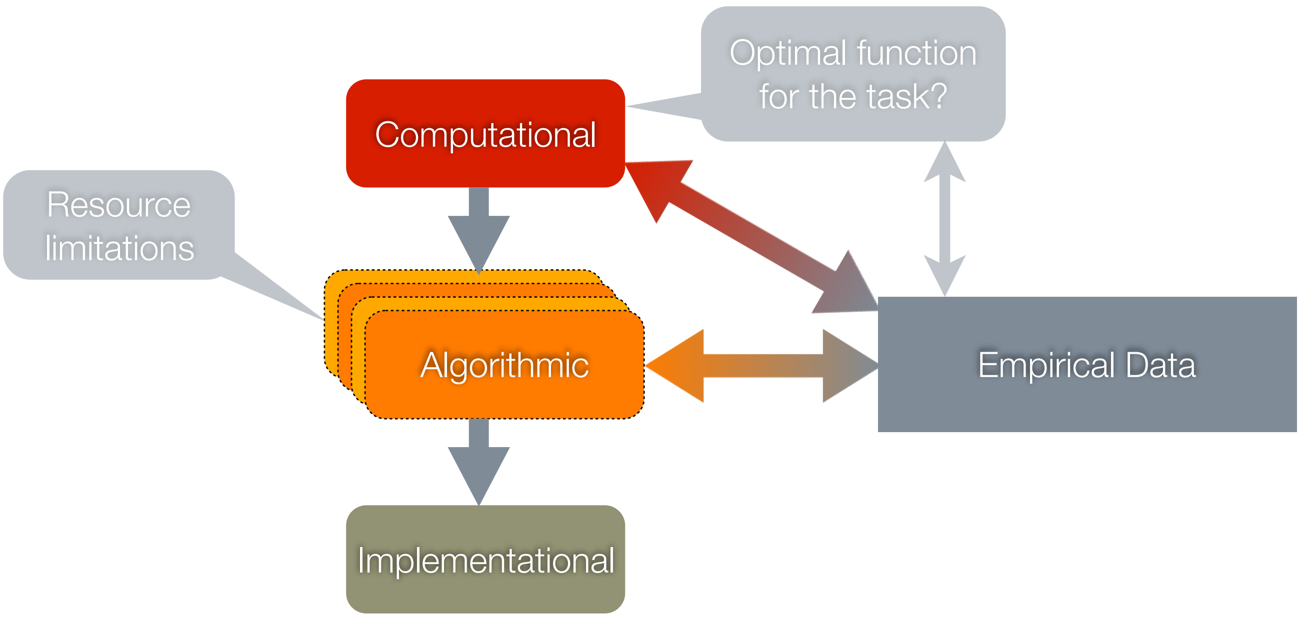
    ✓ Fast: people process utterances in real-time, incrementally

**Can we treat language as a rational cogntive system?**

# Marr's Levels of Modeling

- Theories/models can characterize processing at differing levels of abstraction
- Marr (1982) identifies three such levels:
  - *Computational* level: a statement of **what** is computed
  - *Algorithmic* level: specifies **how** computation takes place
  - *Implementational* level: is concerned with how algorithms are actually **neurally instantiated** in the brain
- The may be many algorithms for a given computational theory
- Many neural implementations could implement a given algorithm

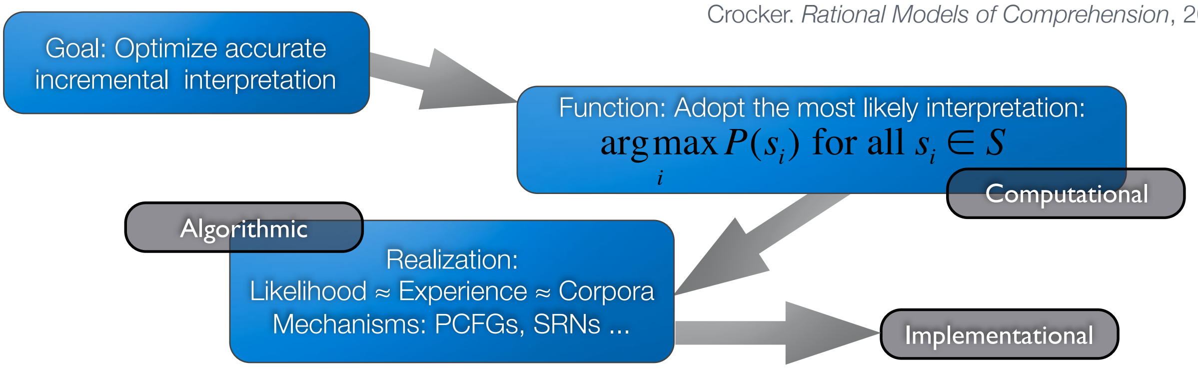
# Relating Models with Data



# Towards a Rational Analysis

- Hypothesis: In general people seem well-adapted for language.
  - Goal: Our models must account for, and explain:
    - Processing difficulty in specific circumstances
    - Effective performance in general
  - Method: Apply Rational Analysis
- Use probabilistic frameworks to reason about rational behaviour
  - Initial hypothesis: The optimal function is one which maximizes the likelihood of obtaining the correct interpretation of an utterance

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

# Motivating the Probabilistic HSPM

- Empirical: Evidence for the use of frequencies
  - Sense disambiguation [Duffy, Morris &Rayner]
  - Category disambiguation [Corley & Crocker]
  - Subcategorization frame selection [Trueswell et al,, Garnsey]
  - Structural preferences [Mitchell et al]
- Rational: Near optimal heuristic behaviour
  - Select the "most likely" analysis
  - Ideal for modular architectures, where full knowledge isn't available

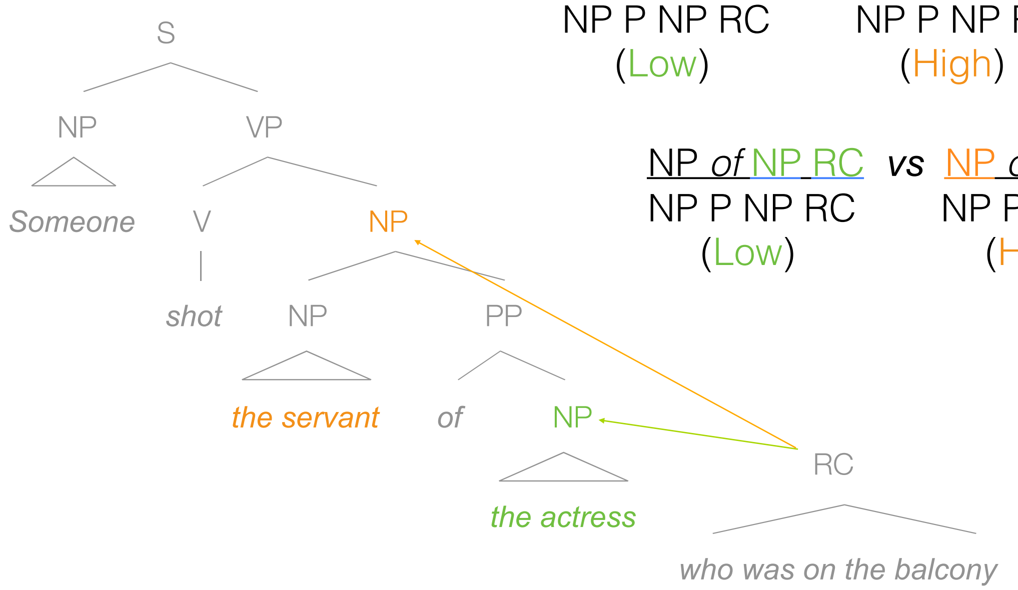
# The Grain Problem

- Experience-based models rely on frequency of prior linguistic exposure to determine preferences. What kinds of things do we count?
  - Actual sentence/structure occurrences? Data too sparse?
  - Lexical: Verb subcategorization frequencies. Do we distinguish tenses? Senses?
    - Word level: specific word forms or lemmas? Part-of-speech, how detailed?
  - Tuning is structural: NP P NP RC *vs* NP P NP RC
    NP P NP RC (Low) NP P NP RC (High)
- Does all experience have equal weight (old vs. new)?
- Are more frequent "words" or "collocations" (idioms) dealt with using finer grain statistics than rarer expressions?

# Relative Clause Attachment

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 P NP RC       NP P NP RC
(Low)       (High)




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

# Probabilistic Language Processing

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

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

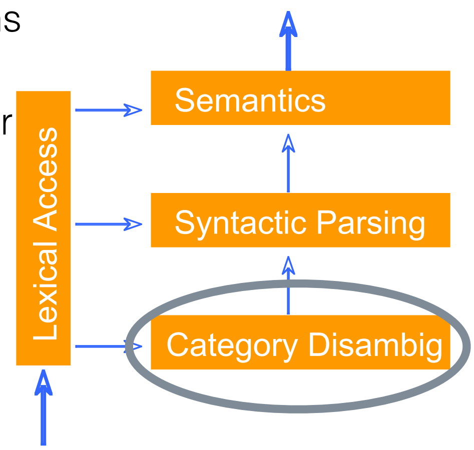
- ***P*** hides a multitude of sins:
    - ***P*** corresponds to the **degree of belief** in a particular interpretation
    - Influenced by recent utterances, experience, non-linguistic context
- ***P*** is usually determined by **frequencies in corpora** or **human completions**
- To compare probabilities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

# Implementation

- Interpretation of probabilities
  - Likelihood of structure occurring, ***P*** can be determined by frequencies in corpora or human completions
- Estimation of probabilities
  - Infinite structural possibilities = sparse data
  - Associate probabilities with **finite** description of language: e.g. PCFGs
- What mechanisms are required:
  - Incremental structure building and estimation of probabilities
  - Comparison of probabilities entails parallelism

# Lexical Category Disambiguation

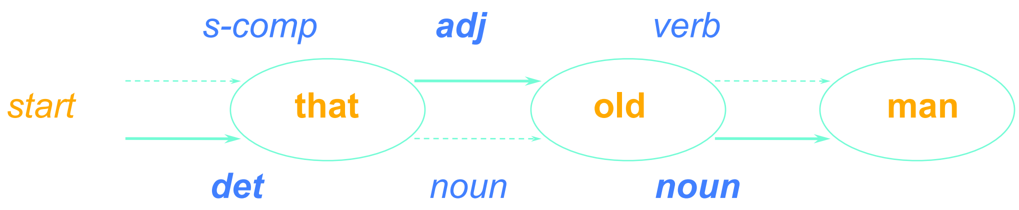
- Sentence processing involves the resolution of lexical, syntactic, and semantic ambiguity.
  - Solution 1: These are not distinct problems
  - Solution 2: Modularity, divide and conquer




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




# 2 Predictions

- The Statistical Hypothesis:
  - Lexical word-category frequencies, $P(w_i|t_i)$, are used for initial category resolution
- The Modularity Hypothesis:
  - Initial category disambiguation is modular, and not determined by (e.g. syntactic) context beyond $P(t_i|_{ti-1})$.
- Two experiments investigate
  - The use word-category statistics
  - Autonomy from syntactic context

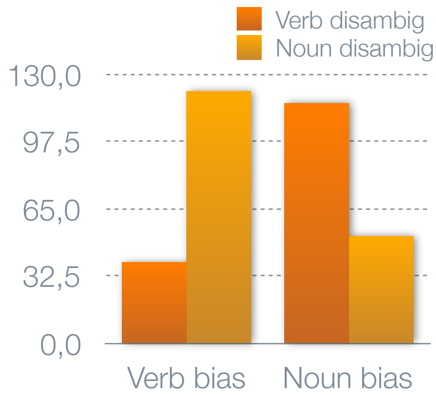
# Statistical Lexical Category Disambiguation

- 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: $P(w_i|t_i)$
- Category context: $P(t_i|t_{i-1})$ – constant!
- Trained on the Susanne corpus

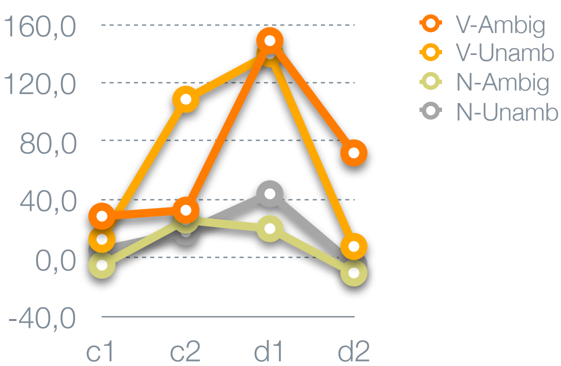
- Interaction between bias and disambiguation
- Category frequency determines initial decisions




# Modular Disambiguation?

- Do initial decisions reflect integrated use of 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.