Computational Psycholinguistics

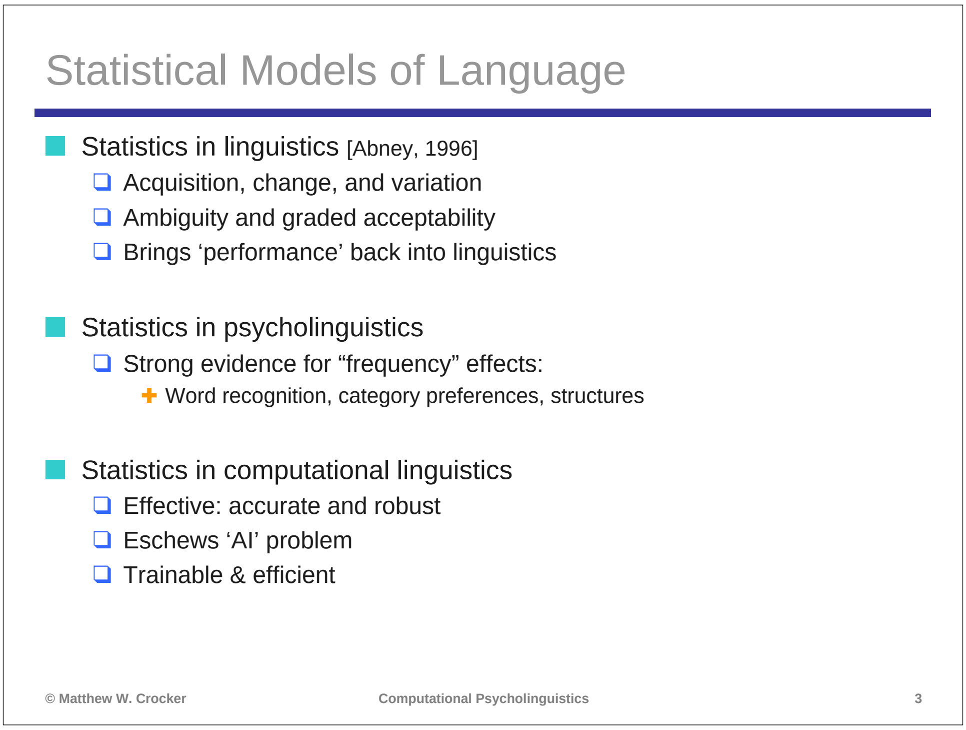
Lecture 6: Probabilistic Models
Lexical Category Resolution

Matthew W Crocker

Computerlinguistik
Universität des Saarlandes

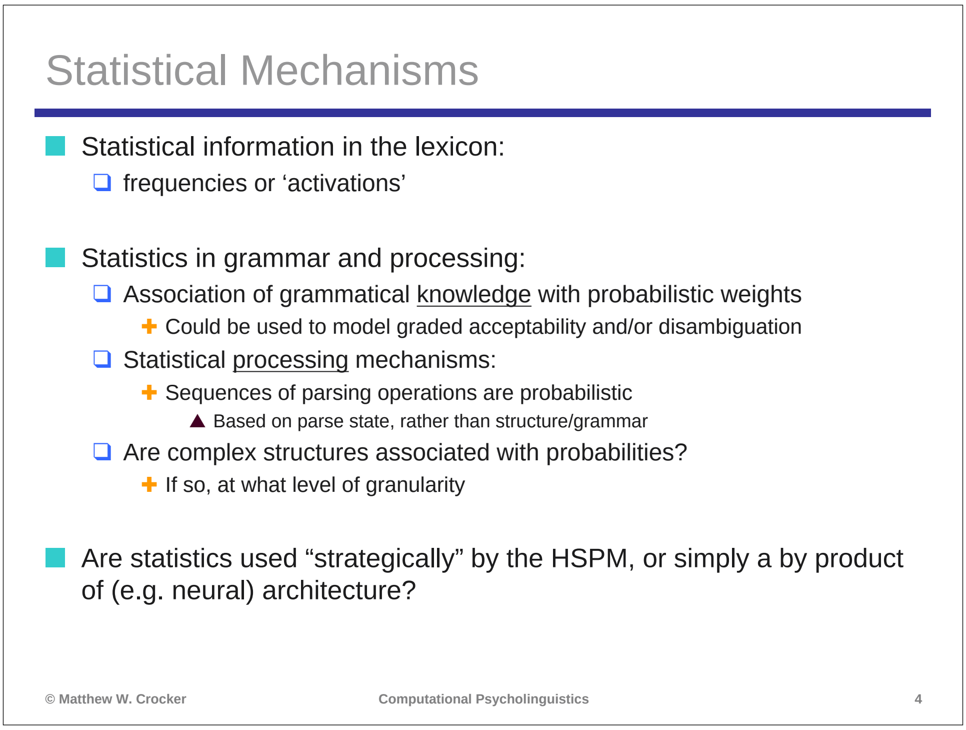
Overview

- General motivation for probabilistic models
- Empirical Support and Rational Analysis
- Probabilistic models of sentence processing
  - Interactive, constraint-based accounts (connectionist)
  - Symbolic parsing models (statistical)
- Probabilistic Models: Breadth and Depth
  - SLCM: Maximal likelihood for category disambiguation
  - Statistical models of human parsing (Jurafsky)
  - Wide coverage probabilistic sentence processing (Crocker & Brants)
  - Criticisms of likelihood, and possible alternative: Informativity



Statistical Models of Language

- Statistics in linguistics [Abney, 1996]
  - Acquisition, change, and variation
  - Ambiguity and graded acceptability
  - Brings ‘performance’ back into linguistics
- Statistics in psycholinguistics
  - Strong evidence for “frequency” effects:
    - Word recognition, category preferences, structures
- Statistics in computational linguistics
  - Effective: accurate and robust
  - Eschews ‘AI’ problem
  - Trainable & efficient



Statistical Mechanisms

- Statistical information in the lexicon:
  - frequencies or ‘activations’
- Statistics in grammar and processing:
  - Association of grammatical knowledge with probabilistic weights
    - Could be used to model graded acceptability and/or disambiguation
  - Statistical processing mechanisms:
    - Sequences of parsing operations are probabilistic
      - Based on parse state, rather than structure/grammar
  - Are complex structures associated with probabilities?
    - If so, at what level of granularity
- Are statistics used “strategically” by the HSPM, or simply a by product of (e.g. neural) architecture?

Motivating the Probabilistic HSPM

- Empirical: Evidence for the use of frequencies
  - Sense disambiguation [DM&R]
  - Category disambiguation [Corley&Crocker]
  - Subcategorization frame selection [TT&K, Garnsey]
  - Structural preferences [Mitchell et al]
- Rational: Near optimal heuristic behaviour
  - Adaptive: select the “most likely” analysis
    - Optimal category disambiguation [C&C]
    - Parsing [Jurafsky, Crocker & Brants]
  - Ideal for restricted (modular) architectures, where full knowledge-based decisions aren’t possible
- Methodological:
  - Transparently combine symbolic and stochastic mechanisms
  - Scaleable, predictive models
  - Blurring the boundary between rational and empirical ...

Garden Path versus 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
- 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

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

Maximal Likelihood Models

- Language Technology: Broad coverage, high-accuracy parsing
  - Parse with the highest probability is usually correct:
    E.g. Ratnarparki's Maximum Entropy parser: 86% parse accuracy
  - Also: speech recognition, POS tagging, semantic clustering, word sense
- Psycholinguistic evidence for the use of frequencies
  - Category disambiguation, word sense, subcategorization frame selection, structural preferences
- Psychological Models:
  - Constraint-based and connectionist (Tanenhaus, Macdonald, ...)
    - Probabilities contribute to determining activations
  - Jurafsky: probabilistic access and disambiguation
    - parallel parser with beam search, uses constituent and valence probabilities

➔ Determine the most likely analysis for a given input:

$$\boxed{\arg\max_{i} P(s_i) \text{ for all } s_i \quad S}$$

➔ Use estimates based on frequencies in prior experience

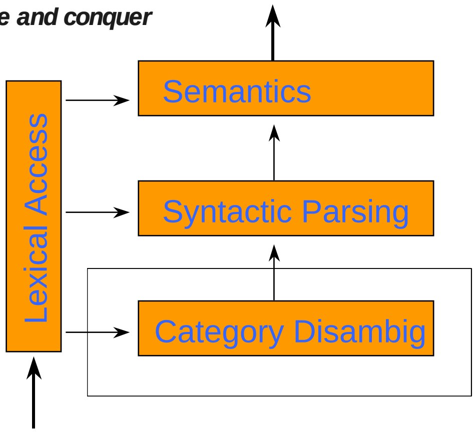
The Grain Problem

- Experience-based models rely on frequency of prior linguistic exposure to determine preferences.
- There are many ways to realise experience-based models
  - Possibilities: What kinds of things do we count?
    - Actual sentence/structure occurrences? Data too sparse?
    - Head driven: I.e. verb subcategorization frequencies
      - Do we distinguish tenses? Senses?
    - Word level, part-of-speech
    - Tuning is structural: NP P NP RC ***vs*** NP P NP RC
    - High Low
- Interesting issues:
  - Does all experience have equal weight (old vs. new)?
  - Are more frequent “words” or “strings” (idioms) dealt with using finer grain statistics that less frequent?

Statistical Lexical Category Module

- 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)



Lexical Access

Semantics

Syntactic Parsing

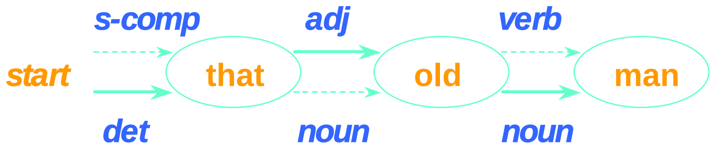
Category Disambig

The Model: A Simple POS Tagger

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

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

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



2 Predictions

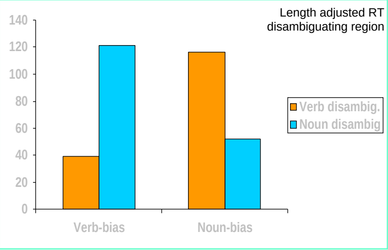
- The Statistical Hypothesis:
  - ❑ Lexical word-category frequencies are used for initial category resolution
- The Modularity Hypothesis:
  - ❑ Initial category disambiguation is modular, and not determined by (e.g. syntactic) context
- Two experiments investigate
  - ❑ The use word-category statistics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

Statistical Lexical Category Disambiguation

- Initially preferred category depends on:
  - Lexical bias: $P(w_i|t_i)$
  - Category context: $P(t_i|t_{i-1})$
  - Trained on the Susanne corpus

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

- 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! ...



➔ Interaction between bias and disambiguation

➔ Lexical category frequency determines initial category decisions

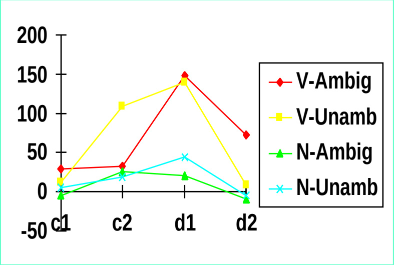
Modular Category Disambiguation ?

- Do initial decisions reflect integrated use both lexical and syntactic constraints/biases (e.g. Jurafsky) ?
- Do initial decisions prioritise lexical category biases (Corley&Crocker) ?
- N/V with immediate/late syntactic disambiguation

a) The foreman knows that the warehouse *prices* are cheaper than the rest. [N-bias, N-disamb]
b) The foreman knows that the warehouse *price* is cheaper than the rest. [N-bias, N-unamb]
c1 c2 d1 d2

➔ Main effect of bias in disambiguating region:

- Decisions are based on word bias, ignore syntactic constraints.
- Implies lexical category doesn't include number
- Problematic for lexicalist syntactic theories
- At c2, VA/VU difference is significant:
- Predicted by SLCM; contra integrated models
- Also accounted for by competition models



‘That’ Ambiguity (Juliano & Tanenhaus)

- ‘That’ ambiguity in syntactic context:
  - ***That* experienced** diplomat(s) would be very helpful ...
  - The lawyer insisted ***that* experienced** diplomat(s) would be very helpful

- 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

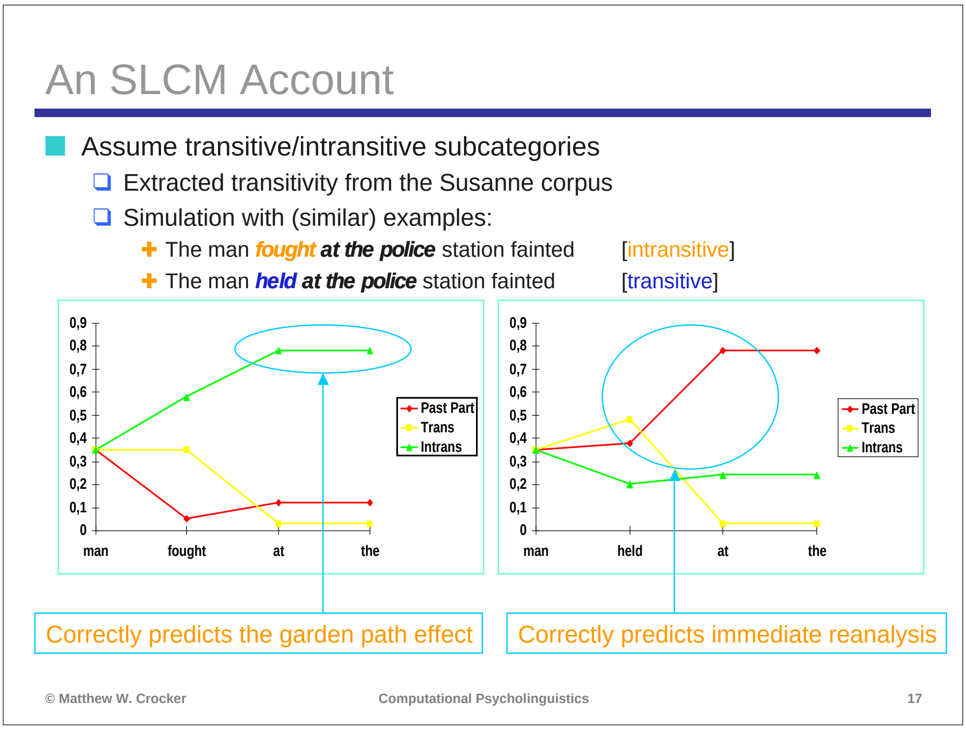
- The tagger model predicts internal reanalysis for some sequences.

- 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 bias is used

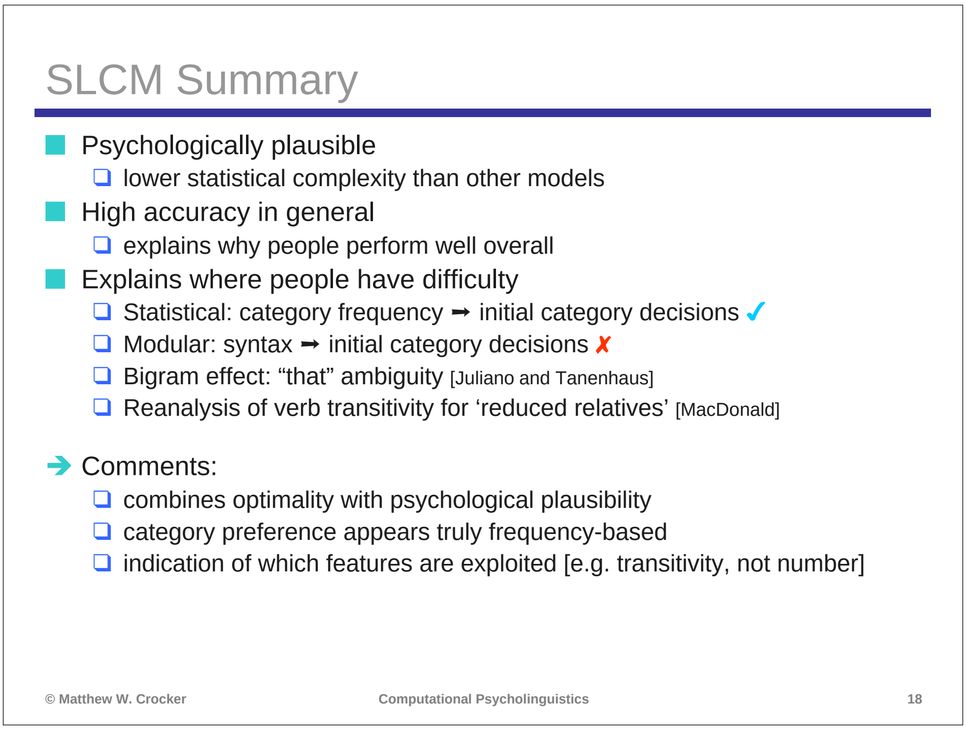


An SLCM Account

- 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]