Computational Psycholinguistics Lecture 7: Constraint-based Models

Matthew W. Crocker

Marty Mayberry

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

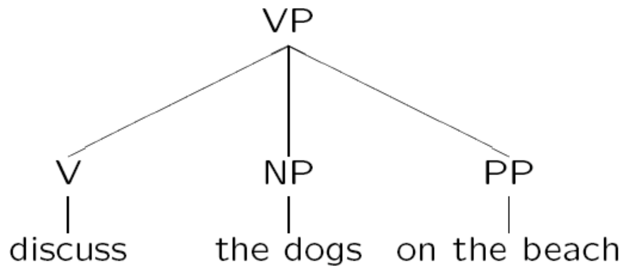
Frame Preferences

- 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:

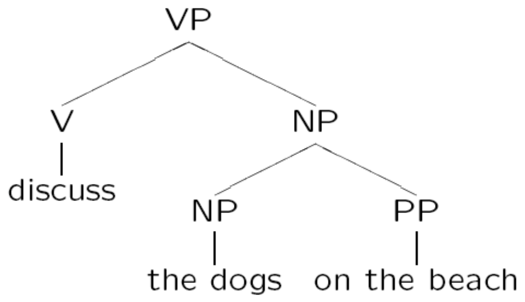


$p(t_1) = 0.15 \times 0.24 = 0.036$ (dispreferred)

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

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)

Probabilistic Models: Jurafsky

- What **architecture** is assumed?
  - Modular lexico-syntactic processor, no semantic knowledge
- What **mechanisms** is used to construct interpretations?
  - Incremental, bounded parallel parsing, with reranking
- What **information** is used to determine preferred structure?
  - Lexical and structural probabilities
- **Linking Hypothesis**:
  - Parse reranking causes increased RTs, if correct parse has been eliminated, predict a garden-path

Probabilistic Models, so far ...

- Three models, explain both good performance & “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
- Probabilistic parsers are typically massively parallel and also non-incremental.

Psychological Plausibility

- Are wide-coverage, probabilistic models cognitively plausible?
- Broad coverage probabilistic parsers:
    - High accuracy: over 90% 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

Summary of Experiment

- 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

Some Remaining Problems

- Integrating plausible parsing mechanisms:
  - Either bounded parallel, or serial (momentary parallel) with reanalysis
  - Monotonic parsing models (Sturt & Crocker)
- Better metrics for relating parser behaviour to 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

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

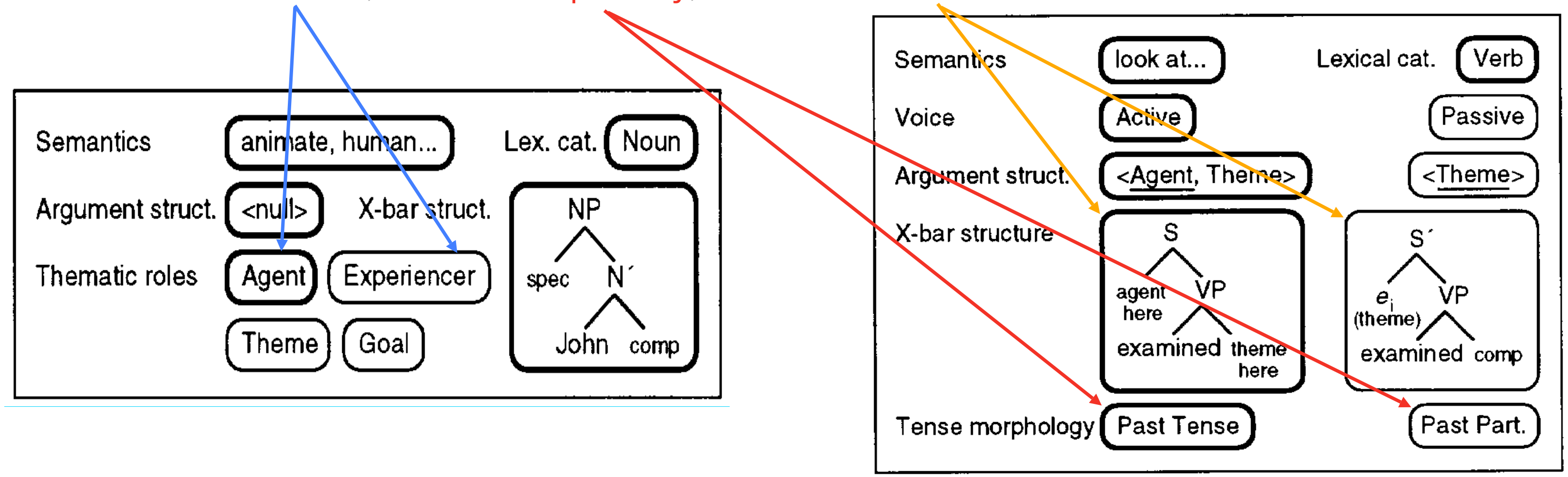
- Reference may determine “argument attach” over “modifier attach”
- Plausibility of *story* versus *diet* as indirect object

The Interactive Activation Model

(MacDonald et al, 1994)

- Rich lexical entries; frequency determines ‘activations’
- Consider: *“John examined the evidence”*
  - “examined” is either a simple past or past participle

➡ thematic fit, tense frequency, structural bias ...



Constraint-based Models

- What **architecture** is assumed?
  - Non-modular: all levels are constructed and interact simultaneously
- What **mechanisms** is used to construct interpretations?
  - Parallel: ranking based on constraint activations
- What **information** is used to determine preferred structure?
  - All relevant information and constraints use immediately
- **Linking Hypothesis**:
  - Comprehension is easy when constraints support a common interpretation, difficult when they compete

Interactive Activation

- The Interactive-Activation Model: In sum
  - Multiple access is possible at all levels of representation, simultaneously, constrained by frequency/context
  - Detailed lexical entries enriched with frequency information
  - Language processing is “constraint satisfaction”, between lexical entries, and across levels; No distinct parser
- Questions: Complex interaction behaviors are difficult to predict
  - Conflicting constraints should cause difficulty. Do they?
  - Difficult to actually implement, and estimate frequencies

The Competitive-Integration Model

(McRae et al, 1998)

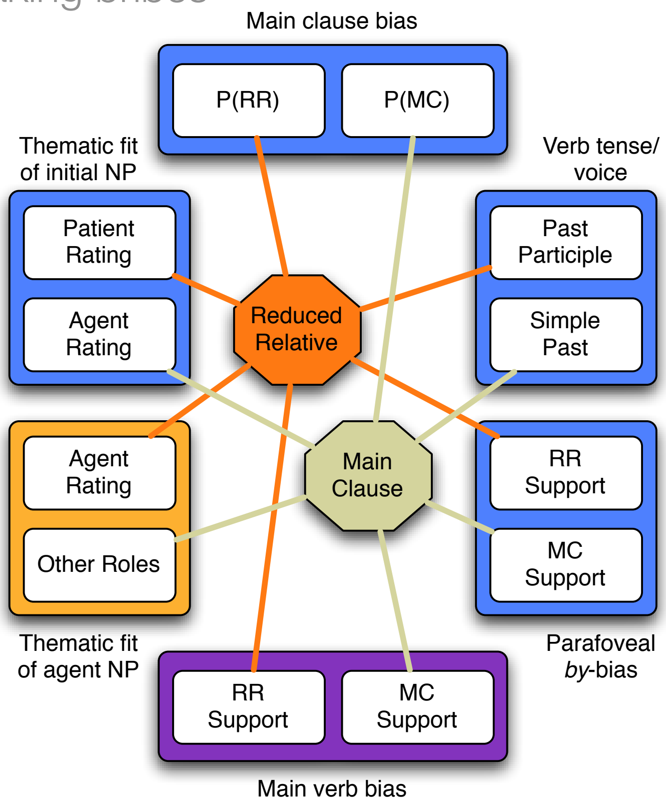
- Claim: Diverse constraints (linguistic and conceptual) are brought to bear simultaneously in ambiguity resolution.
- The Model: *Assumes the all analyses are constructed*
  - Constraints provide “probabilistic” support for analyses
    - Constraint are weighted and normalized
    - Lexical & structural bias, parafoveal cues, thematic fit ...
- Goal: Simulate reading times
  - RTs are claimed to correlate with the number of cycles required to settle on one of the alternatives

“No model-independent signature data pattern can provide definitive evidence concerning when information is used”

The Computational Model

The crook arrested by the detective was guilty of taking bribes

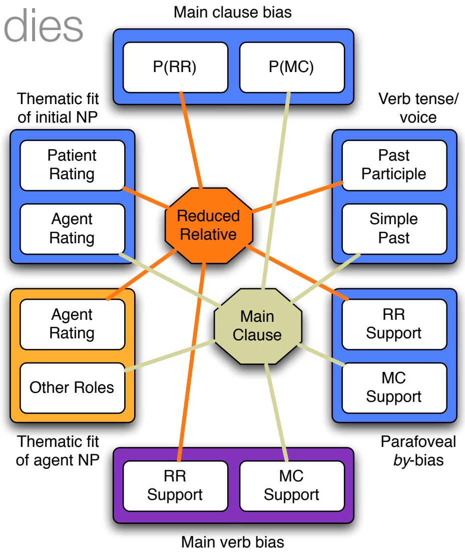
1. Combines constraints as they become available in the input
2. Input determines the probabilistic activation of each constraint
3. Constraints are weighted according to their strength
4. Alternative interpretations compete to a criterion
5. Cycles of competition mapped to reading times



Steps in the Experiment: (McRae et al 1998)

- Constraints contribute to the activation of competing analyses, over time

1. Identifying the relevant constraints
2. Computational model for the interaction of constraints
3. Estimate bias of each constraint from corpora & rating studies
4. Weight of each constraint: fit with off-line completions
5. Make predictions for reading times
6. Compare actual reading times with those predicted by:
   - Constraint-based model
   - Garden-path model



Constraint Parameters

“The crook/cop arrested by the detective was guilty of taking bribes”

Verb tense/voice constraint: is the verb preferentially a past tense (i.e. main clause) or past participle (reduced relative)

Relative log frequency is estimated from corpora: RR=.67 MC=.33

Main clause bias: general bias for structure for “NP verb+ed …”

Corpus: P(RR|NP + verb-ed) = .08, P(MC|NP + verb-ed) = .92

by-Constraint: extent to which ‘by’ supports the passive construction

Estimated for the 40 verbs from WSJ/Brown: RR= .8 MC= .2

Thematic fit: the plausibility of crook/cop as an agent or patient

Estimated using a rating study

by-Agent thematic fit: good Agent is further support for the RR vs. MC

Same method as (4).

Thematic Fit Parameters

"The crook/cop arrested by the detective was guilty of taking bribes"

- Estimating thematic fit with an off-line rating (1-7) study

```
How common is it for a
    crook   ______
    cop     ______
    guard   ______
    police  ______
    suspect______
To arrest someone?
To be arrested by someone?
```

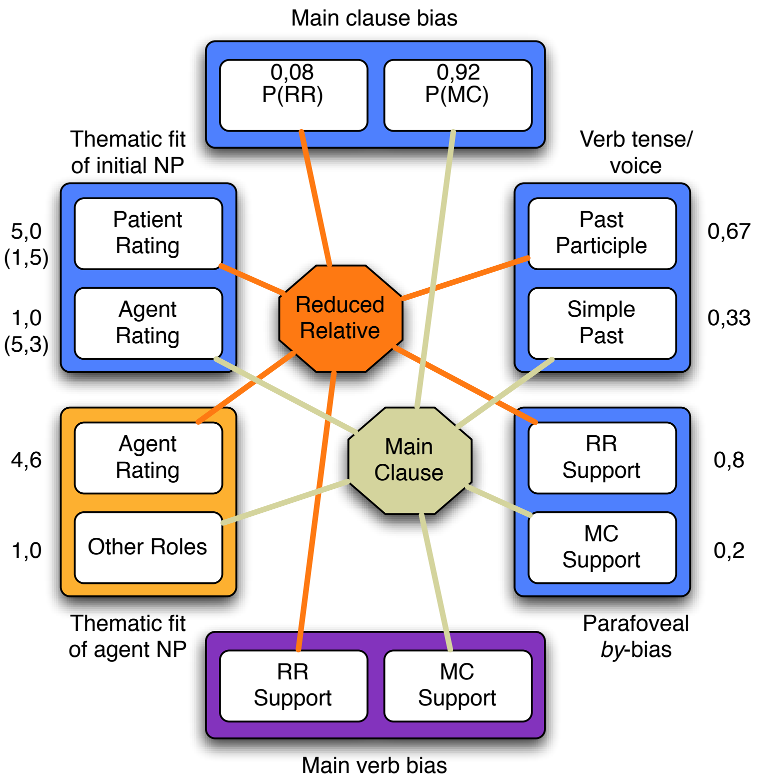
NP 1	Rel	Main
Agent	1,5	5,3
Patient	5,0	1,0

by NP	Rel	Main
Agent	4,6	1,0

The Computational Model

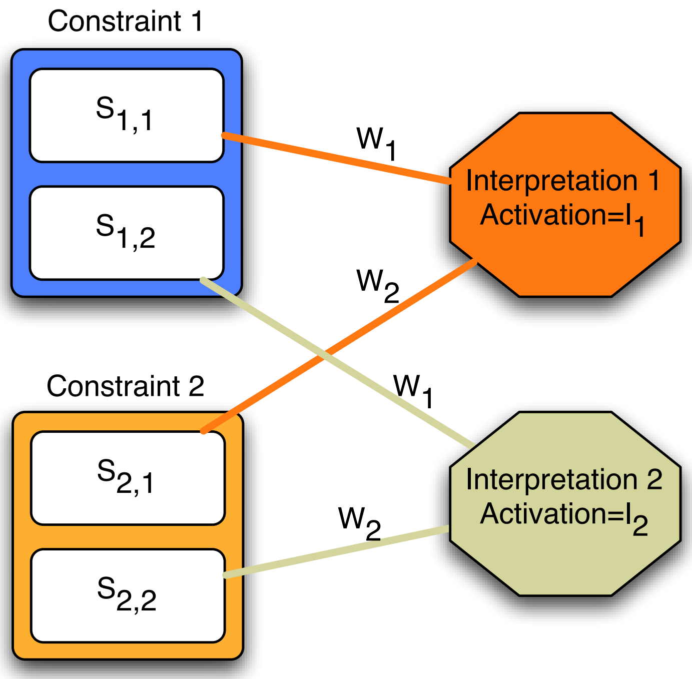
The crook arrested by the detective was guilty of taking bribes

1. Combines constraints as they become available in the input
2. **Input determines the probabilistic activation of each constraint**
3. Constraints are weighted according to their strength
4. Alternative interpretations compete to a criterion
5. Cycles of competition mapped to reading times



The recurrence mechanism

- $S_{c,a}$ is the raw activation of the node for the c^{th} constraint, supporting the a^{th} interpretation,
- w_c is the weight of the c^{th} constraint
- I_a is the activation of the a^{th} interpretation
- 3-step normalized recurrence mechanism:
  - Normalize: $S_{c,a}(norm) = \dfrac{S_{c,a}}{\sum_a S_{c,a}}$
  - Integrate: $I_a = \sum_c \left[w_c \cdot S_{c,a}(norm) \right]$
  - Feedback: $S_{c,a} = S_{c,a}(norm) + I_a \cdot w_c \cdot S_{c,a}(norm)$



$$\sum_i w_i = 1$$

A Gated Completion Study

- Establish that thematic fit does in fact influence “off-line” completion
- Use to adjust the model weights
- Manipulated the fit of NP1:
    - Good agents (and atypical patients)
    - Good patients (and atypical agents)
- Hypotheses: Effect of fit at verb
    - Additional effect at ‘by’
    - Ceiling effect after agent NP

Gated sentence completion study:

The cop/crook arrested ...

The crook arrested by ...

The crook arrested by the ...

The crook arrested by the detective...

Fitting Constraint Weights

- Adjust the weights to fit “off-line” data:
  - Brute force search of weights (~1M)
  - 20-40 cycles (step 2)
- Node activation predicts proportion of completions for each interpretation
  - Avg of activation from 20-40 cycles

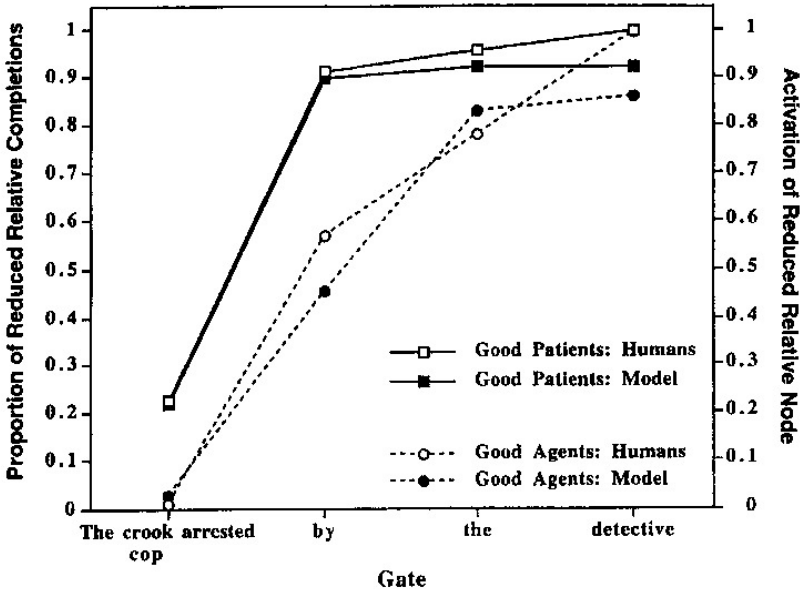


FIG. 2. Human and simulation results for fragment completions.

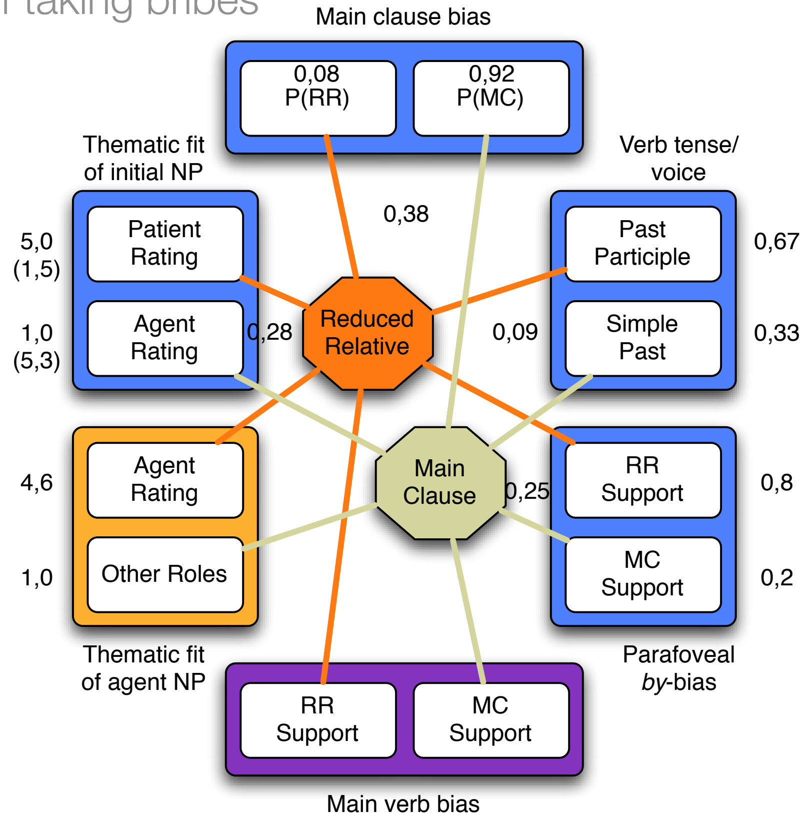
Counted “the crook arrested himself” as RR (!?)

The complete model

Constraint Based (CB) Model
MC bias: .5094 x .75
Thematic Fit: .3684 x .75
Verb tense: .1222 x .75
by-bias: .25

The crook arrested by the detective was guilty of taking bribes

1. Combines constraints as they become available in the input
2. Input determines the probabilistic activation of each constraint
3. **Constraints are weighted according to their strength**
4. Alternative interpretations compete to a criterion
5. Cycles of competition mapped to reading times



On-line study

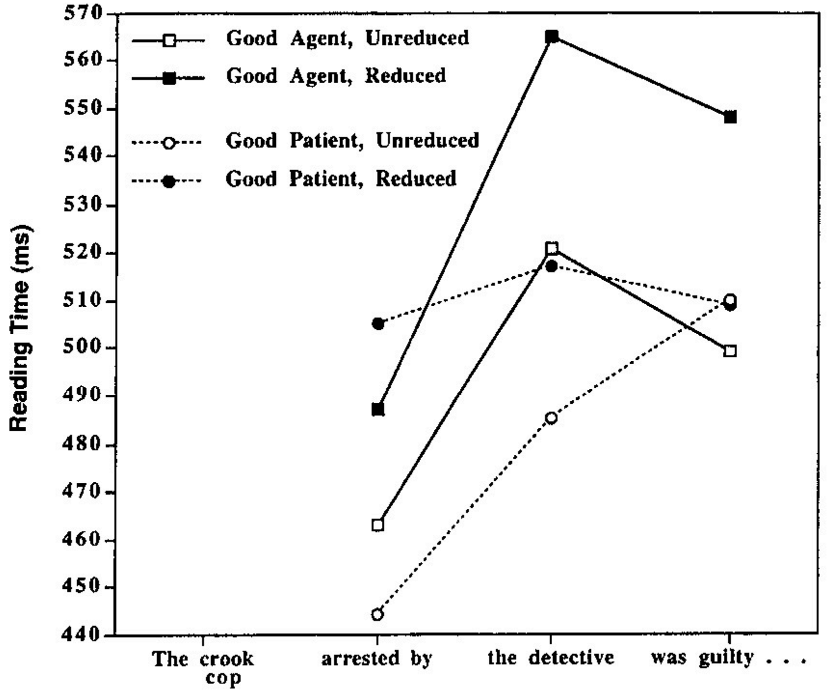


FIG. 5. Self-paced reading times for the Experiment.

- Two-word, self-paced presentation: (similar to completion studies)

The crook / arrested by / the detective / was guilty / of taking bribes

The cop / arrested by / the detective / was guilty / of taking bribes

The crook / that was / arrested by / the detective / was guilty / of taking bribes

The cop / that was / arrested by / the detective / was guilty / of taking bribes

Model Predictions

- Two “Versions” of the models:
  - Constraint-Based: constraints apply immediately for each region
  - GP: MC-bias & Main-Verb bias only, other constraints delayed
- Prediction Per-Region Reading times for each model:
  - Each region is processed until it reaches a (dynamic) criterion:

    *dynamic criterion = 1 - Δcrit*cycle*
  - As more cycles are computed, threshold is relaxed
  - *Δcrit=.01* means a maximum of *50* cycles

CB vs. GP predictions

Constraint Based (CB) Model
MC bias: .5094 x .75
Thematic Fit: .3684 x .75
Verb tense: .1222 x .75
by-bias: .25

Garden Path (GP) Model:
MC bias:1

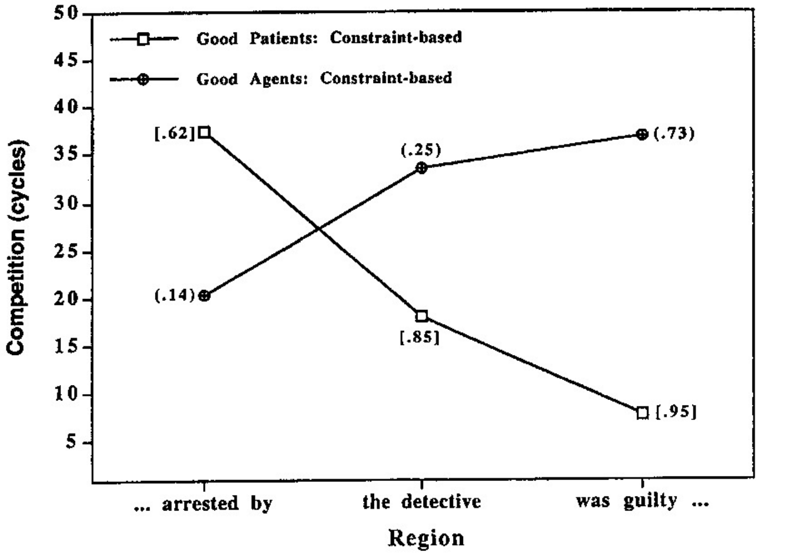


FIG. 3. Self-paced predictions derived from the constraint-based competition model. In this and all following model figures, the number beside each model datum is the mean activation of the reduced relative node after competition in that region for either (good agents) or [good patients].

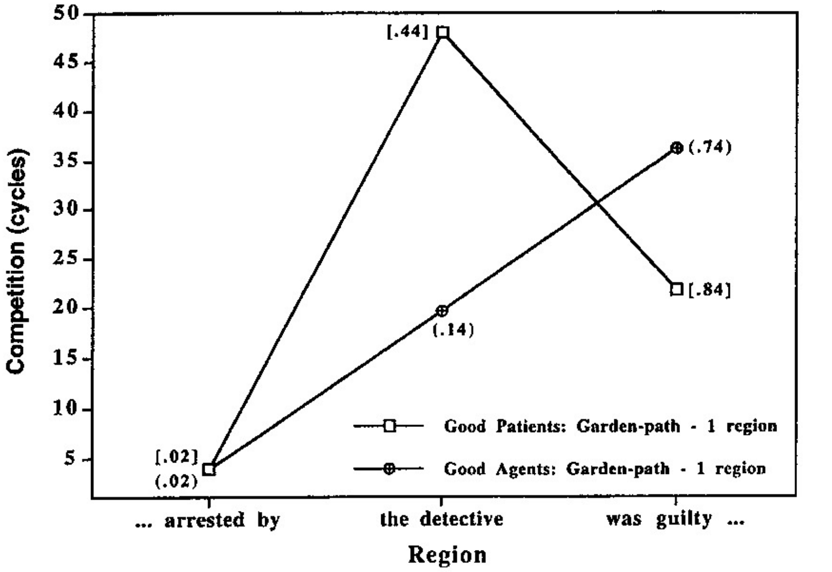


FIG. 4. Self-paced predictions as derived from the garden-path model when constraints other than the nain clause and main verb biases were delayed by a region.

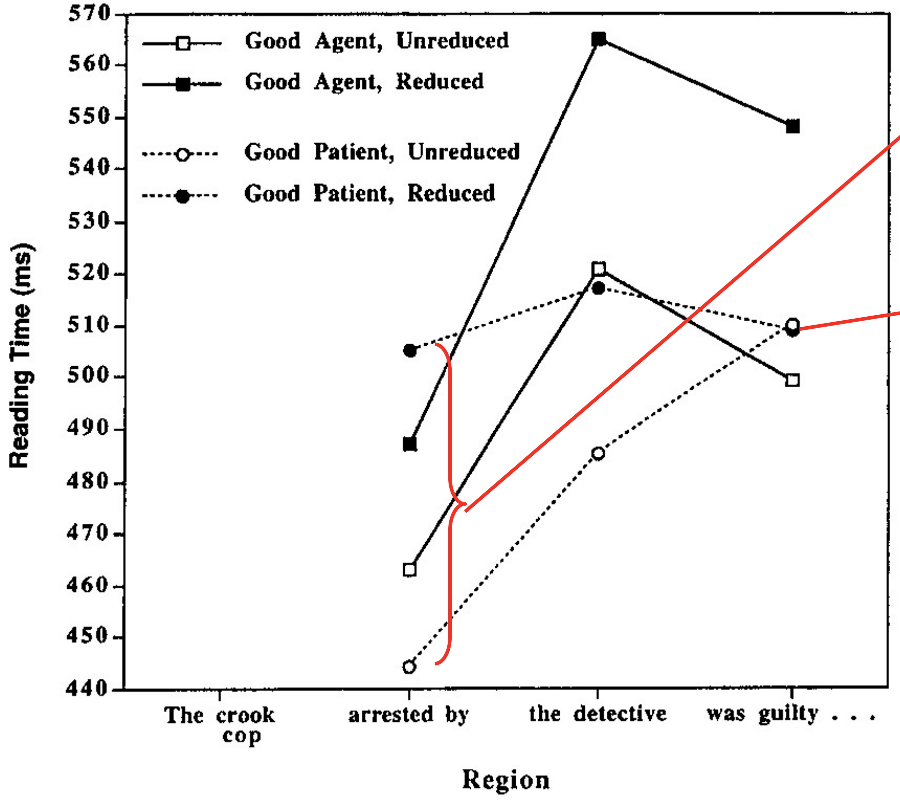


FIG. 5. Self-paced reading times for the Experiment.

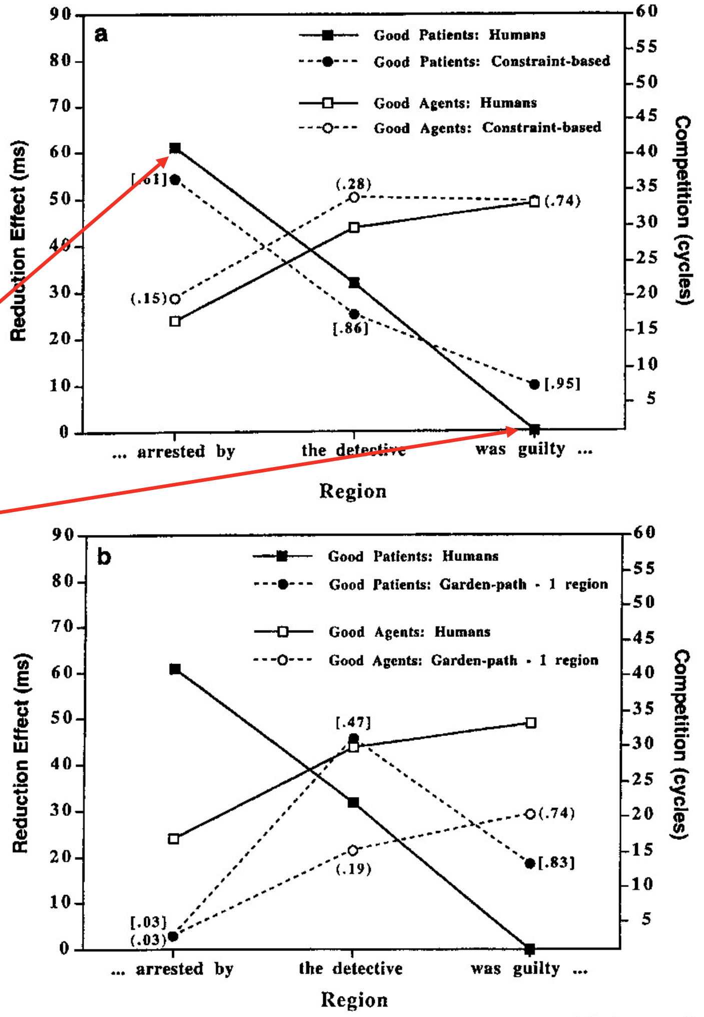


FIG. 6. Simulations of self-paced reading by (a) the constraint-based model, and (b) the one-region delay garden-path model.

3rd Model: Short Delay GP Theory

- The GP-model, has a 1-2 word delay in use of information, what if this delay is reduced? 4 cycles (10-25ms)
  - Better fit, but high reduction effect still predicted at main verb (good patient).
- Search for the (new) best weights:
  - MC bias: .2966 (.5094)
  - Thematic fit: .4611 (.3684)
  - V.tense: .0254
  - by-bias: .2199
- No-longer models completions

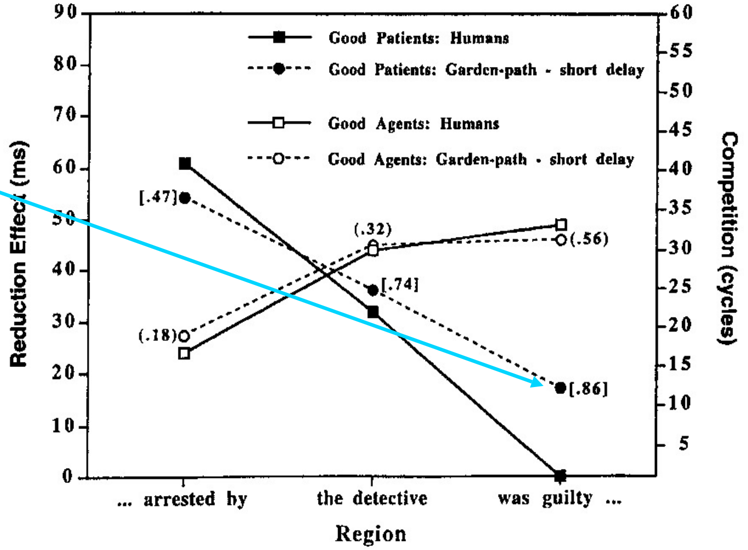


FIG. 7. Garden-path simulations of self-paced reading when constraints other than the main clause and main verb biases were delayed by 4 cycles of competition, or approximately 10–25 ms. Note the large predicted reduction effect at the main verb for the good patients.

Issues and Criticisms

- What “constraints” to include/exclude:
  - Ok if materials don’t vary w.r.t excluded constraint, or if excluded constraint correlates with included constraint
- Models constraint integration independent of parsing?
  - What is *really* being modelled? Can the approach scale?
- Is the implementation of the GP model a fair comparison
- Predicts long reading times when constraints compete
  - People are often *faster* at processing ambiguous regions!

Constraint-based vs. Probabilistic

- Similarities between constraint-based and probabilistic models:
  - Weighting of different constraints
  - Simultaneous integration of constraints
- Differences between constraint-based and probabilistic models:
  - Probabilistic models scale more easily, typically not “handcrafted”
  - Constraint-based models directly predict difficulty (competition among constraints), probabilistic models do not

Other Recent Approaches

- Narayanan & Jurafsky (1998, submitted): Use bayesian belief networks to combine SCFG like probabilities with other semantic and thematic probabilities
- Pado (2006): a wide-coverage model of role-assignment and thematic fit (plausibility), which can be integrated with a syntactic parser
- Expectation-based approaches (Hale, 2001; Levy 2006): Based on the probability distribution of all parses, processing difficulty is associated with its *surprisal*: a words conditional probability based on context.
- Stochastic models: Kempen & Vosse (2000), Tabor (2004) argue for mechanisms which emphasize local coherence, rather than “perfect” incremental parsing.

Readings so far ...

- Matthew Crocker. Mechanisms for Sentence Processing. In: Garrod & Pickering (eds), Language Processing, Psychology Press, London, UK, 1999.
- Dan Jurafsky. Probabilistic Modeling in Psycholinguistics. In Bod et al (eds.). Probabilistic Linguistics. The MIT Press, 2003.
- Ken McRae, Michael Spivey-Knowlton, Michael Tanenhaus. Modeling the Influence of Thematic Fit (and Other Constraints) in On-line Sentence Comprehension. Journal of Memory and Language, 38, 283–312 (1998).

Road Map

- Connectionist Approaches
  - Emphasize learning from experience
  - Develop own representations and “solutions” for language processing
  - Difficult to “inspect”: success of networks is usually established behaviorally
  - Cognitive properties: neurally inspired, robust,
- Basics of connectionist models and connectionist learning algorithms
- Lexical processing: past-tense formation, reading aloud
- Sentence Processing: simple recurrent networks