A Bayesian Model of Human Sentence Processing

Srini Narayanan
ICSI and UC Berkeley
Joint work with Dan Jurafsky, Stanford University

Talk Outline

- Introduction
  - Evidence for Probabilistic factors in sentence processing
  - Background and alternative models
  - Problems with current models
- A Bayesian Model of sentence processing
  - Basic Model/Result
  - Details
- Results on behavioral data
- Ongoing Work
- Conclusion

Suggestive facts about language comprehension

- Language is noisy, ambiguous, and unsegmented.
- How might humans interpret noisy input?
  - human visual processing: probabilistic models (Rao et al. 2001; Weiss & Fleet 2001)
  - categorization: probabilistic models (Tenenbaum 2000; Tenenbaum and Griffiths 2001b; Tenenbaum and Griffiths 2001a, Griffiths 2004)
  - human understanding of causation: probabilistic models (Rehder 1999; Glymour and Cheng 1998, Gopnik et al 2004)
- Why Probabilistic Models?

Why probabilistic models of language comprehension?

- The best normative solution to problems of decision-making under uncertainty
- Probability Theory
  - Principled methodology for weighing and combining evidence to choose between competing hypotheses/interpretations
  - Coherent semantics
  - Learnable from interaction with world
  - Bounded optimality

Controversial in early approaches

“ But it must be recognized that the notion “probability of a sentence” is an entirely useless one, under any known interpretation of this term.”

- Noam Chomsky 1969, p. 57

Can probabilistic models capture important aspects of language

- Probability: good normative model.
- Is it a good descriptive model of language?
  - Probabilistic models explain how *people* perform linguistic acts with incomplete information.
- Probability theory says:
  - If you have ever have to choose between things
    - Compute the probability of each choice given everything you know
    - And pick the most likely one (one with most utility)

When do we choose between multiple things

- **Comprehension:**
  - Segmenting speech input
  - Lexical ambiguity
  - Syntactic ambiguity
  - Semantic ambiguity
  - Pragmatic ambiguity
- **Production:** choice of words (or syntactic structure or phonological form or etc)
- **Learning**: choosing between:
  - Constraint rankings
  - Settings of parameters
  - Different grammars
  - Possible lexical entries for new words

Sentence Processing

- Modeling how humans build interpretations for sentences.
- Can probabilistic models play a role in this process?

Probabilistic Factors: Summary of evidence in comprehension

- Word Level
  - Lexeme frequencies (Tyler 1984; Salasoo and Pisoni 1985; inter alia)
  - Lemma frequencies (Hogaboam and Perfetti 1975; Ahrens 1998;)
  - Phonological probabilities (Pierrehumbert 1994, Hay et al (in press), Pitt et al(1998)).
- Word Relations
  - Dependency (word-word) probabilities (MacDonald (1993, 2001), Bod (2001)
  - Lexical category frequencies (Burgess; MacDonald 1993, Trueswell et al. 1996; Jurafsky 1996)
- Contructional/Semantic
  - Constructional probabilities (Mitchell *et al.* 1995; Croft 1995; Jurafsky 1996; (Corley and Crocker 1996, 2000; Narayanan and Jurafsky 1998, 2001; Hale 2001)
  - Sub- categorization probabilities (Ford, Bresnan, Kaplan (1982); Clifton, Frazier, Connine (1984), Trueswell *et al.* (1993)
  - Idiom frequencies (d'Arcais 1993)
  - Thematic role probabilities (Trueswell *et al.* 1994; Garnsey *et al.* 1997, McRae *et al.* (1998) McRae, Hare, Elman (2004))

Summary: Probabilistic factors and sentence comprehension

- What we know
  - Lots of kinds of knowledge interact probabilistically to build interpretations
- What we don't know
  - How are probabilistic aspects of linguistic knowledge represented?
  - How are these probabilities combined?
  - How are interpretations selected?
  - What's the relationship between probability and behavioral information like reading time?

Studying sentence comprehension: garden path sentences

- The horse raced past the barn stumbled.
- The horse ridden past the barn stumbled.
- The crook arrested by the police confessed.
- The cop arrested by the police confessed.
- The complex houses married and single students.
- The warehouse fires many employees in the spring.

Commonly studied ambiguities

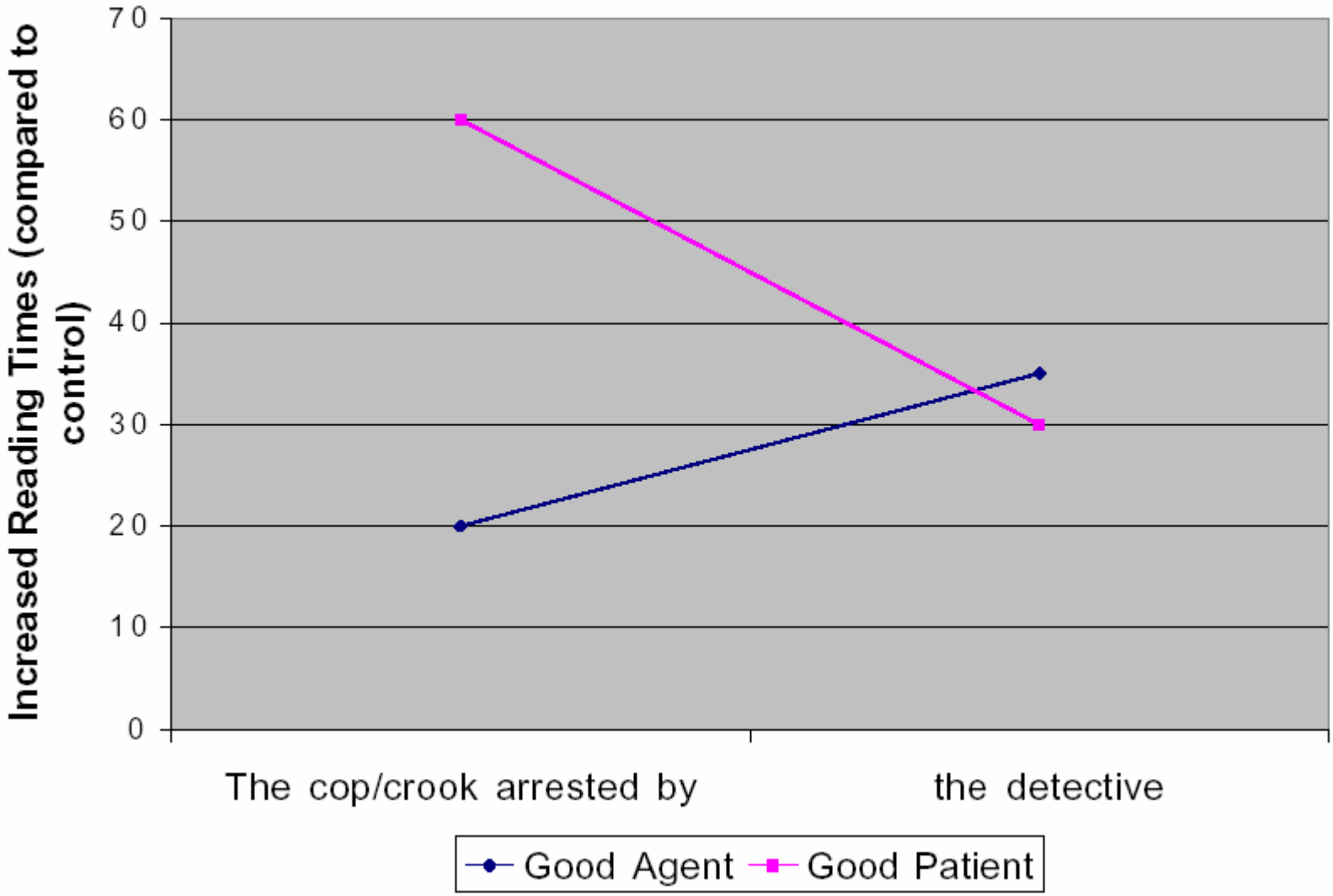
- An ambiguous prefix
  - 1. The witness examined...
  - 2. The witness examined by the lawyer turned out to be unreliable.
  - 3. The witness examined the evidence.
- Main-clause/reduced-relative ambiguity, first noticed by Bever(1970)
- Can cause processing difficulty (‘garden path effect’)
- Useful for testing sentence processing theories

A study of MV/RR ambiguity (McRae, Spivey, Tannenhaus)

- Off-line: sentence completion
  - The crook arrested
  - The crook arrested by
  - The crook arrested by the
- On-line: reading times in two-word moving window
  - The cop / arrested by / the detective / was guilty / of taking / bribes.
  - The cop / who was / arrested by / the detective / was guilty /of taking / bribes.
- Reading times at 3 regions: subject (*the cop*), V+P (*arrested by*), VG (*was guilty*).
- 40 sentences, 20 good agent (cop), 20 good patients (crook)

Factors in sentence processing

- prior probability that the verb (*arrested*) is preterite (simple past) versus participle
- general preference for main clause over reduced relative clause structures.
- syntactic subcategorization preference of verb (*arrested)*
- strong thematic constraints can ameliorate garden path effect.
  - (1) The witness examined by the lawyer turned out to be unreliable.
  - (2) The evidence examined by the lawyer turned out to be unreliable.
- thematic fit of subject head noun with verb:
  - *evidence*, is a better PATIENT than AGENT: GOOD PATIENT
  - *witness*, is a better AGENT: GOOD AGENT
  - Trueswell *et al.* (1994) showed more difficulty at 'by the lawyer' in (1) than (2)

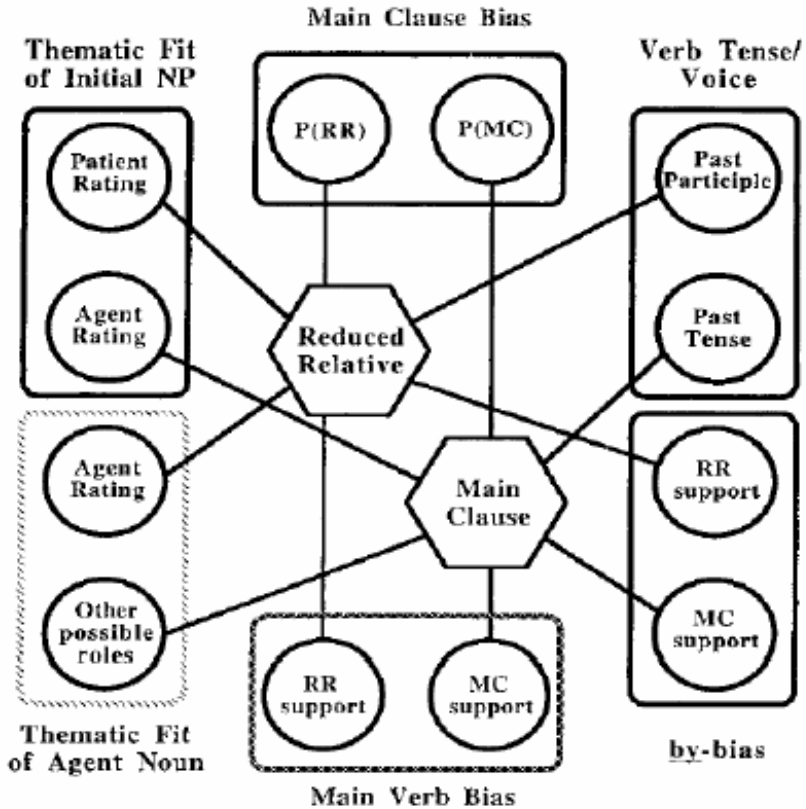


Good-agent sentences easier at initial NP; Good-patient sentences easier at post-by NP.

A Competition model (McRae et al. 1998, 2003)

- Connectionist model to combine constraints.
- Each parse represented by single pre-built localist node in network.
- Relevant weighted factors (from different sources) are activated at different stages of the input.
- The alternatives compete until one passes an activation threshold.

Spivey's competition model of the MV/RR data (McRae et al 1998, 03)



Features/Issues with the Competition Model

- Pro
  - captures probabilistic nature of constraints
  - integrates constraints to predict preference for ambiguous structures
  - makes predictions about reading time based on settling time
  - Correctly models MV/RR data.
- Con
  - Only models disambiguation, not building interpretations
  - Role of structural knowledge:
    - Model includes frequency-based main-verb/reduced relative structural constraint
    - Why this? And why no other?
  - Constraint values have no coherent semantics (Likert scales, counts, log ratios, probabilities, etc!)
  - No model of how weights could be learned

Summary so far

- Main problems with competition processing:
  - **structure**: Doesn't deal with structure
  - **probability**: Doesn't have a well-founded model of probability/weights
  - **integration**: Doesn't give us a good idea of how these integrate
- Alternatives: Human parsing as probabilistic parsing:
  - Jurafsky (1996)
  - Crocker and Brants (1996)
- Alternative: Structure and probability as separate systems
  - Townsend and Bever (2001)

Pickering Results (DO vs. SC)

- Reading time higher with implausible direct object sentences (DO)
  - After verb like *realize* which **expects a sentential complement (SC)**
  - With an implausible direct object
    - The young athlete *realized* **her potential** one day might make her a word-class sprinter.
    - The young athlete *realized* **her exercises** one day might make her a word-class sprinter.

- Q: Why should implausibility of **less-expected interpretation** affect reading time?

Previous Models cannot handle this

- Spivey competition model can't handle this
  - In competition model, reading time is proportional to 'settling time'
  - The closer two interpretations (in activation), the longer the settling time.
  - Thus competition model predicts: making worse interpretation worse should make competition easier, **speeding up** reading time!
  - Exact wrong prediction!
- Jurafsky (1996) model can't handle this
  - Reading time increases when correct parse is pruned, causing reanalysis
  - But no reason why making the worse parse worse should matter

Talk Outline

- Introduction
  - Evidence for Probabilistic factors in sentence processing
  - Background and alternative models
  - Problems
- A Bayesian Model of sentence processing
  - Basic Model/Result
  - Details
- Results on behavioral data
- Ongoing Work
- Conclusion

A Bayesian model of sentence comprehension

Narayanan and Jurafsky (2002, 2006(in press))

How do we do linguistic decision-making under uncertainty?

- Proposal: humans act as probabilistic reasoners.
- Bayesian approach tells us
  - How to combine structure and probability.
  - What probability to assign to a particular belief/interpretation/structure.
  - How these beliefs should be updated in the light of new evidence.
- Processing: In processing a sentence, humans:
  - consider possible interpretations (constructions) in parallel,
  - compute the probability of each interpretation,
  - continuously update probabilities as each piece of evidence arrives
  - prefer more probable interpretations

Reading Time

- Probability is a good predictor of human disambiguation preference.
- What about reading time?
- Many factors affect reading time:
  - plausibility
  - length
  - Prosody
  - Imageability
  - structural complexity
  - memory limitations
- What role does probability play?

Basic Predictions of the Model

- **Expectation:** Reading time is inversely proportional to the probability of what we read.
- **Attention:** Demoting our current best hypothesis causes increased reading time.

Expectation

- **Unexpected words/structure are reading slower.**

$$\text{reading time}(word) \propto \frac{1}{P(\text{word}|\text{context})}$$

  - High probability words are read faster than low probability words

- Background:
  - High frequency words are perceived more quickly (Howes 1951)
  - Improbable words (in context) take longer to read (Boland (1997), McDonald *et al.* 2003, inter alia)
  - Key insight of Hale (2001): reading time proportional to probabilistic information content of word, showed how to compute for SCFG

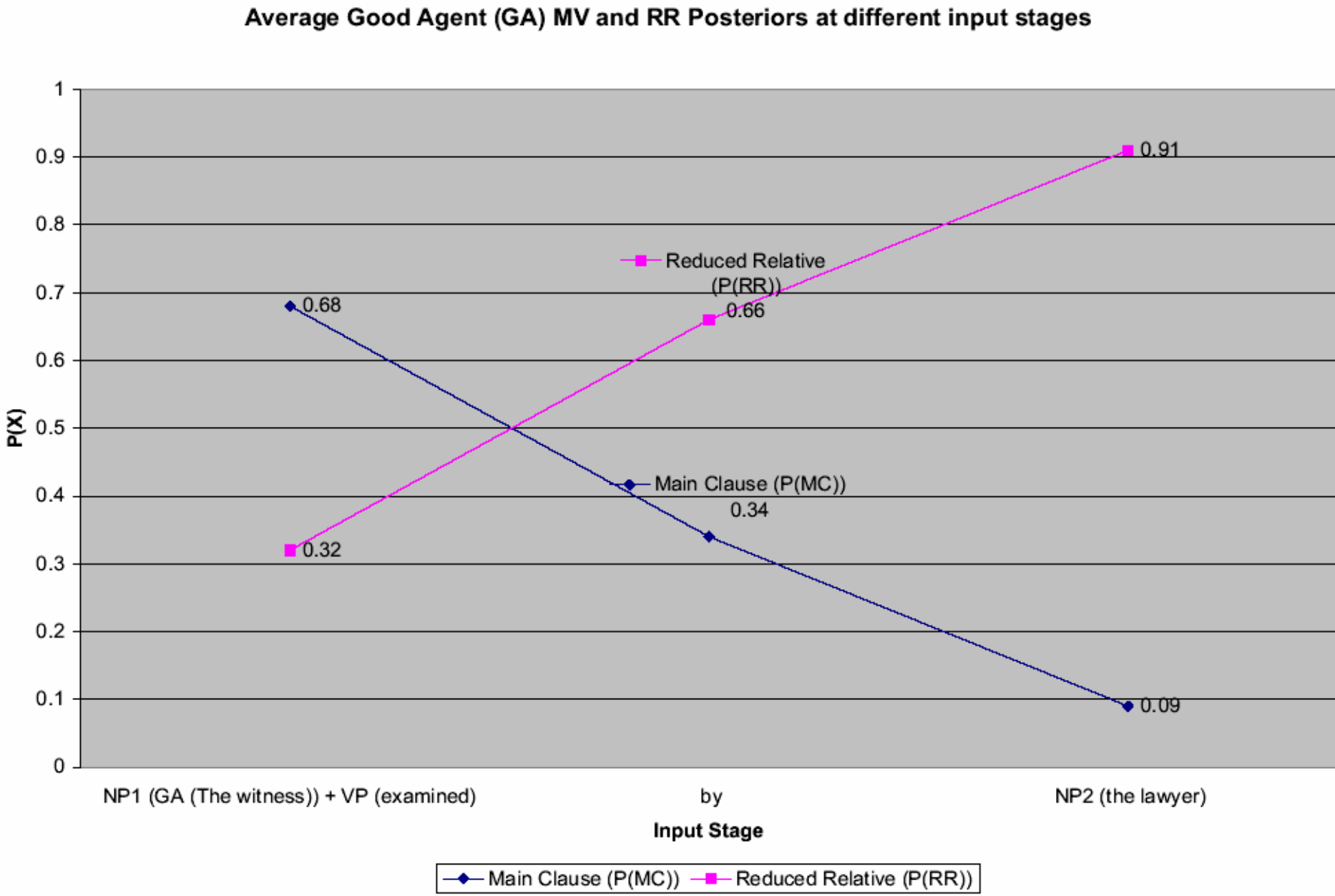
The Attention Principle

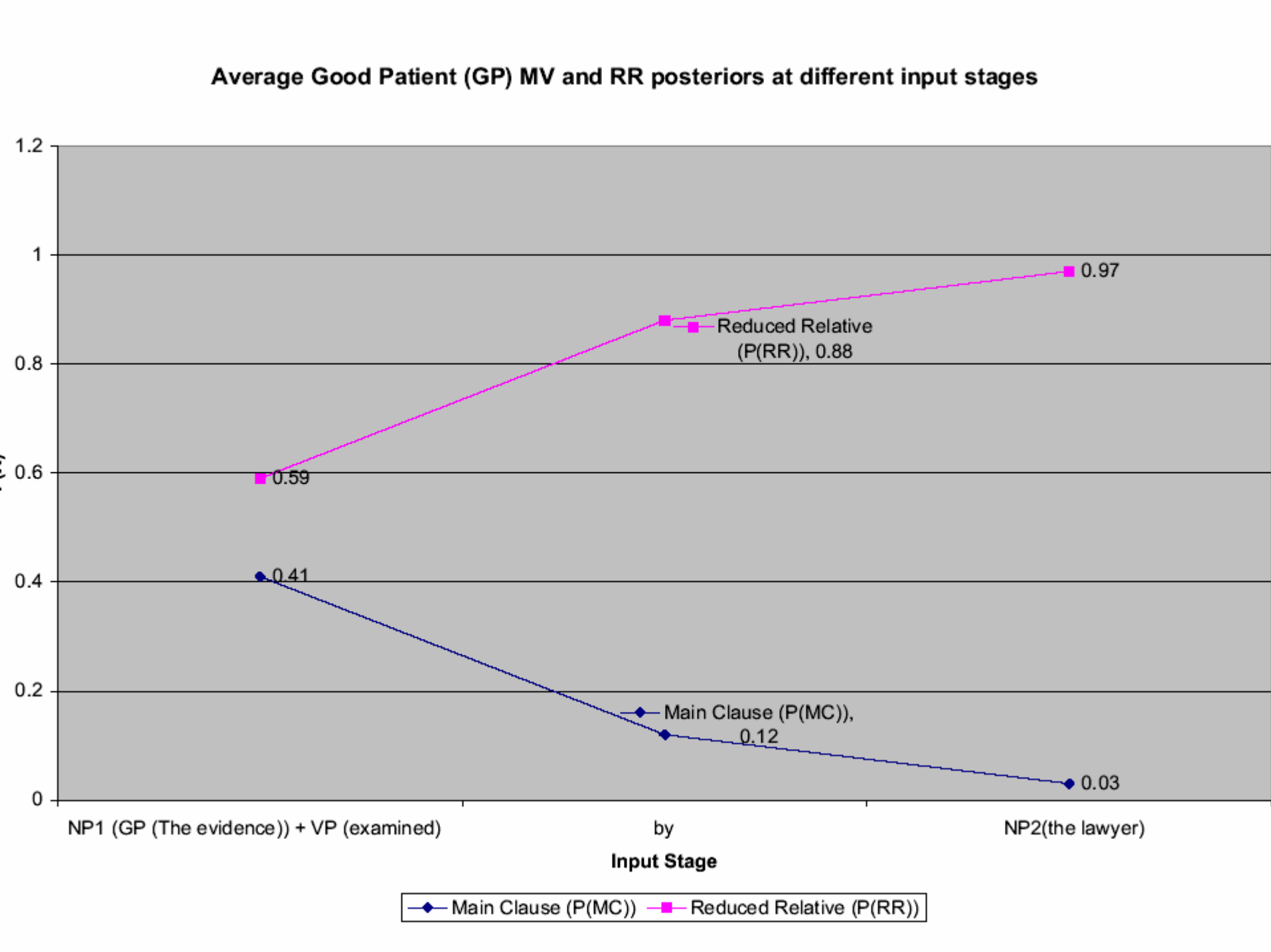
Narayanan and Jurafsky (2002)

- **Demotion of the interpretation in attentional focus causes increased reading time.**
- Architecture
  - limited parallelism, each interpretation ranked by probability
  - comprehender places attentional focus on the most-probable interpretation
  - new evidence may cause re-ranking of set of interpretations
  - reranking may cause an interpretation to drop out of attentional focus.

Basic result

- Expectation and attention principles match behavioral predictions about preference and reading time
- The Bayesian model offers a principled way of realizing key constraints on a sentence processing model:
  - Construction based (where construction is the unit of grammar that binds form (syntactic) and meaning (semantic) information.
  - probabilistic computation
  - incremental update
  - combination of structured and probabilistic knowledge


Average Good Agent (GA) MV and RR Posteriors at different input stages
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P(X)
Reduced Relative (P(RR))
0.68
0.66
0.91
Main Clause (P(MC))
0.34
0.32
0.09
NP1 (GA (The witness)) + VP (examined)
by
NP2 (the lawyer)
Input Stage
Main Clause (P(MC))
Reduced Relative (P(RR))


Average Good Patient (GP) MV and RR posteriors at different input stages
1.2
1
0.8
0.6
0.4
0.2
0
0.97
Reduced Relative (P(RR)), 0.88
0.59
0.41
Main Clause (P(MC)), 0.12
0.03
NP1 (GP (The evidence)) + VP (examined)
by
NP2(the lawyer)
Input Stage
Main Clause (P(MC))
Reduced Relative (P(RR))

Talk Outline

- Introduction
  - Evidence for Probabilistic factors in sentence processing
  - Background and alternative models
  - Problems
- A Bayesian Model of sentence processing
  - Basic Model/Result
  - Details
- Results on behavioral data
- Ongoing Work
- Conclusion

How to compute linguistic probabilities

- Humans choose the most probable interpretation.
- Problem: How to compute probability?
  - Can't just count how many times this interpretation occurred before! (Language is creative)
  - Humans must be breaking down the probability computation into smaller pieces!

Decomposing Probabilities

Two ways to break down probabilities

- **Independence:** Use linguistic intuitions to help come up with independence assumptions
  - Independence assumption: compute the probability of a more complex event by multiplying the probabilities of constituent events.
  - *Syntax* (Bod 2003): can't compute probabilities of whole complex parse tree just by counting (too rare). So assume that the pieces are independent, and multiply probability of tree fragments.
  - *Phonology* (Pierrehumbert 2003): can't compute probabilities of triphone events (too rare). So assume pieces are independent, and multiply probability of diphones.

Using Bayes rule

- **Bayes Rule** : sometimes it's easier to compute something else: (**generative model!!!!**):

$$P(\text{structure}|\text{input}) = \frac{\overbrace{P(\text{input}|\text{structure})}^{\text{likelihood}}\overbrace{P(\text{structure})}^{\text{prior}}}{P(\text{input})}$$

Our Model

- Three components of the probabilistic model:
  - Probabilistic models of word-word expectations
  - Probabilistic models of structure
  - Probabilistic models of valence expectations

Word-to-word expectations

- Lexical relations between neighboring words
  - *bigram probability*, *first-order Markov relation*, *transition probability*
  - *N*-gram probability is context-sensitive: P(havoc) is low, P(havoc| wreak) is high.

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})}$$

Syntactic Expectations

- Minimal assumption model: Stochastic Context-Free Grammar
  - Just augments phrase-structure grammars with probabilities

a. [.079] $VP \rightarrow VBD$

b. [.18] $VP \rightarrow VBD\ NP$

c. [.051] $VP \rightarrow VBD\ NP\ PP$

SCFG Trees

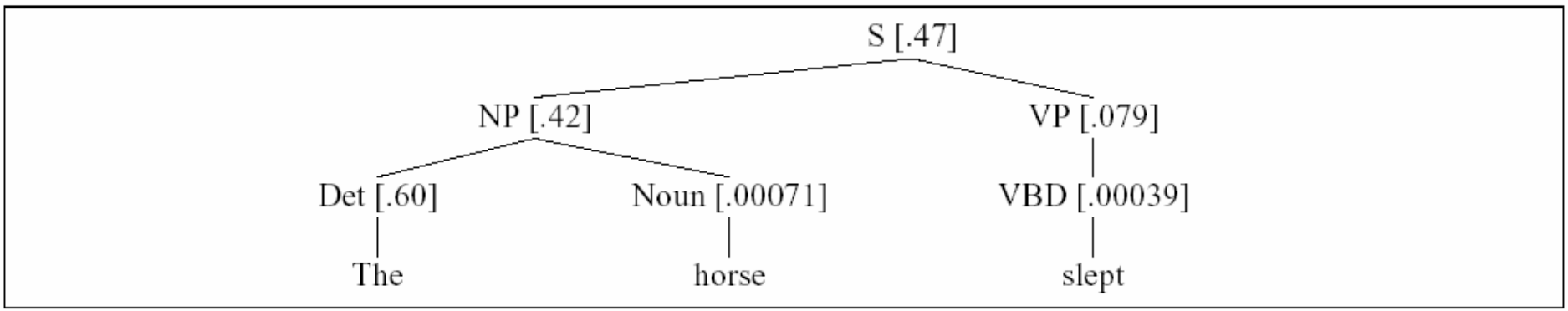


Figure 7: A parse tree for "The horse slept", with SCFG probabilities for the six rules.

$$P(\text{Tree}, S) = \prod_{n \in \text{Tree}} p(\text{rule\_expansion}(n) | n)$$

Valence: Syntactic Sub-cat Prob.

- Computing different syntactic subcategorizations
  - The doctor remembered **[*NP* the idea]**.
  - The doctor remembered [*S* that the idea had already been proposed].
  - The doctor suspected [*NP* the idea].
  - The doctor suspected **[*S* that the idea would turn out not to work].**

- Computed from corpora (Gahl, Roland, and Jurafsky (2005)):

$$P(\text{transitive}|\text{verb=melt}) = \frac{\text{Count(transitive instances of melt)}}{\text{Count(all instances of melt)}}$$

Thematic/Semantic fit

- Semantic fit of arguments with predicate:
  - ”cop” is a good AGENT of *arrest*
  - ”crook” is a good THEME of *arrest*
    - *P (Agent | verb = “arrest”, subject = “cop”)*
    - *P (Theme | verb = “arrest”, subject = “cop”)*
- How to compute probabilities:
  - Corpus counts for this are sparse
  - Eventual Method: count semantic features (frames, schemas) rather than words
  - Meanwhile, approximation: normalize from published human rating studies

Combining structured sources

- Requirements:
  - Combine information from multiple correlated features
  - Use structural relationships and independencies to minimize inter-feature correlations
  - Compact and clear representation
- Answer: Graphical Models (Bayes nets)

Bayes Nets

- Combines ideas from graph theory and probability theory to deal with *complexity* and *uncertainty*
- Basic expressions: statements about conditional probabilities *P (A | B)*
  - *If P(A|B) then A and B are independent*
  - *If P(A | B, C) = P(A | C), then A and B are conditionally independent, given B*
- A graphical model (nodes and edges)
  - nodes = variables (e.g., non-terminals in a parse)
  - edges = influences between variables (e.g., grammar rules)
  - strength of influences quantified by conditional probability

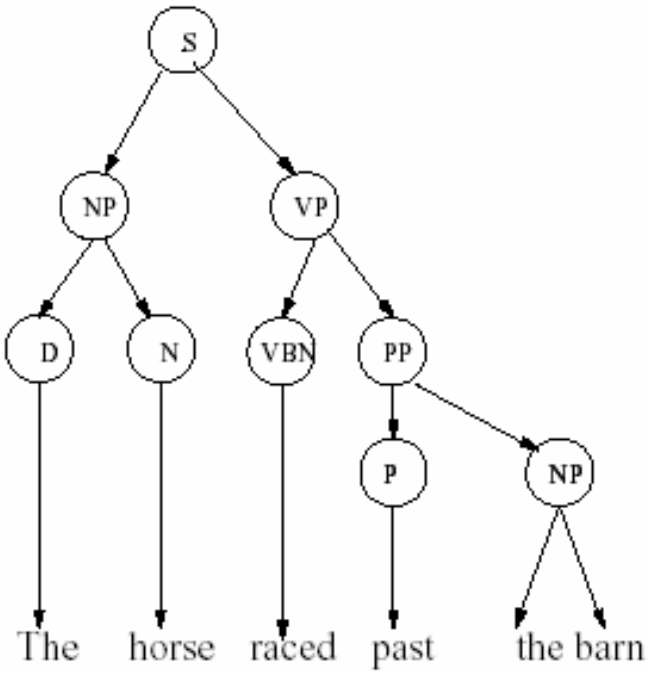
Why Bayes nets

- Makes explicit:
  - the sources of evidence
  - what their structure is
  - how they combine, what influences what, conditional independence
- flexible computation:
  - evidence can enter the network anywhere, bottom-up or top-down
  - any piece of the net can be instantiated as evidence,
  - can use the network to compute probabilities of different questions
- dynamic, on-line recomputation of probabilities as new words (evidence) appear.

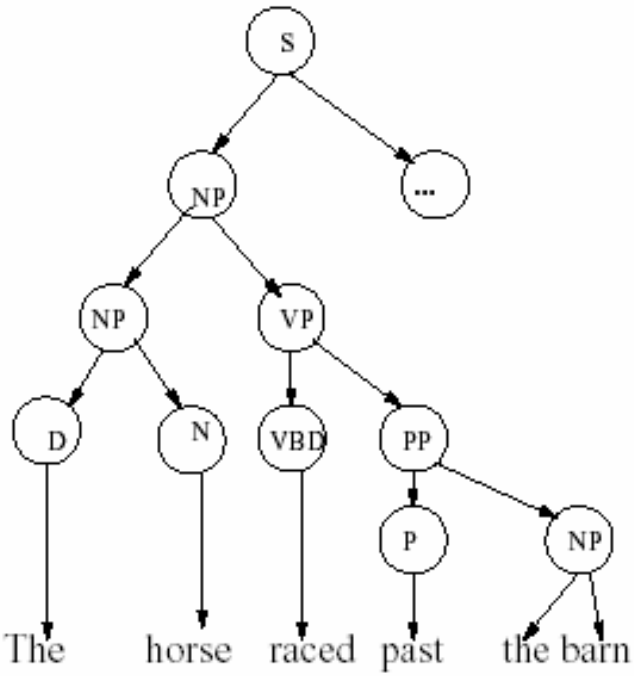
SCFG and Bayes nets

- Assumes SCFG skeleton.
  - Each bracketed string, such as *((The witness) (was examined))*}.
  - Corresponds to a fixed parse-tree structure which is a causal-tree graphical model.
  - The MV bracketing is a causal tree as is the RR bracketing.
- Propagating beliefs on the causal tree using the belief propagation algorithm (JLO) is identical to computing the parse probability using the inside-outside algorithm.

SCFG Parses as Bayes Nets

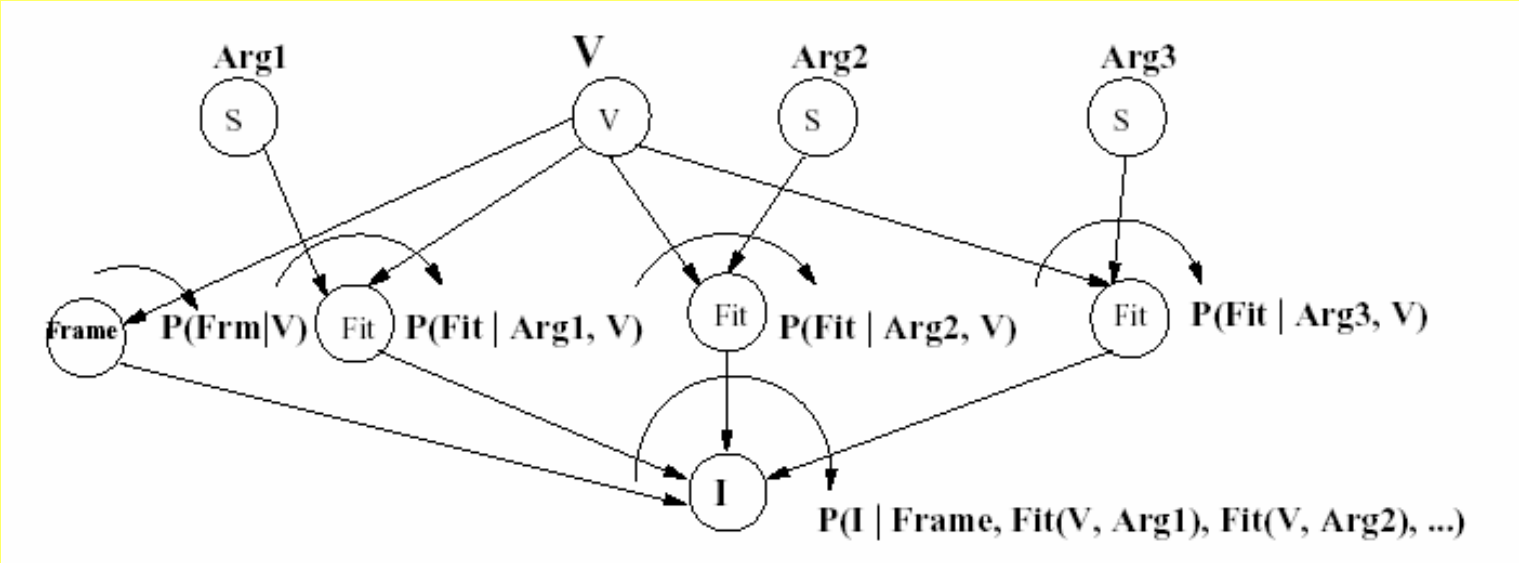


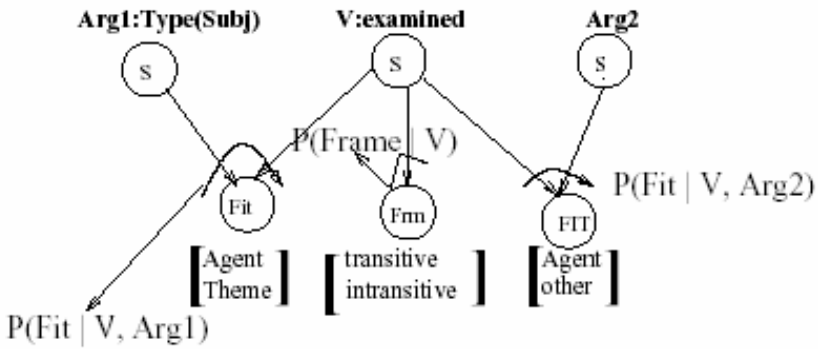
MV PARSE TREE



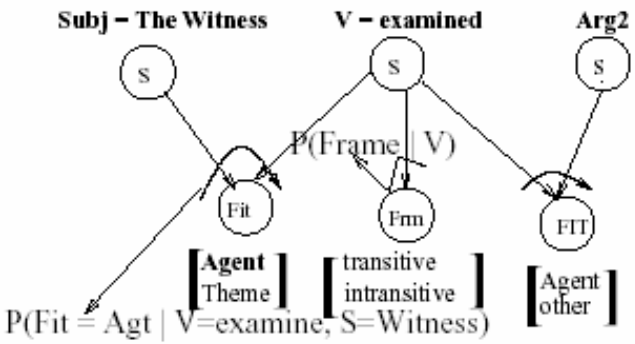
RR PARSE TREE

Thematic/Semantic fit

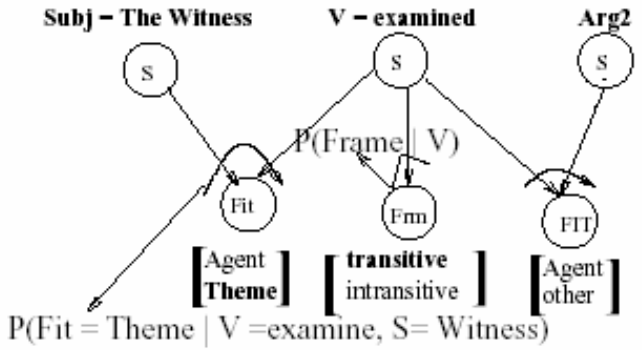




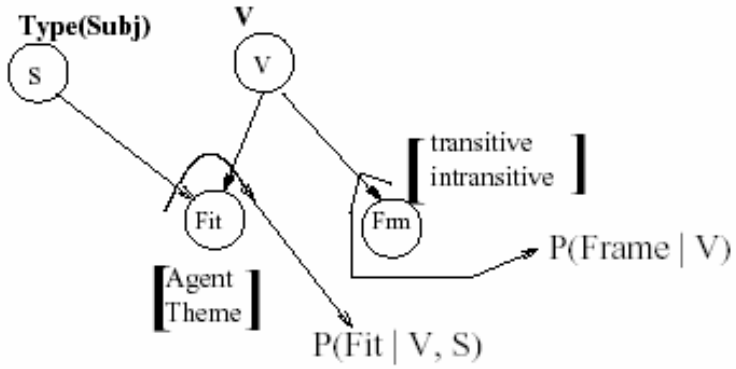
LEXICAL NETWORK AFTER "NP V"



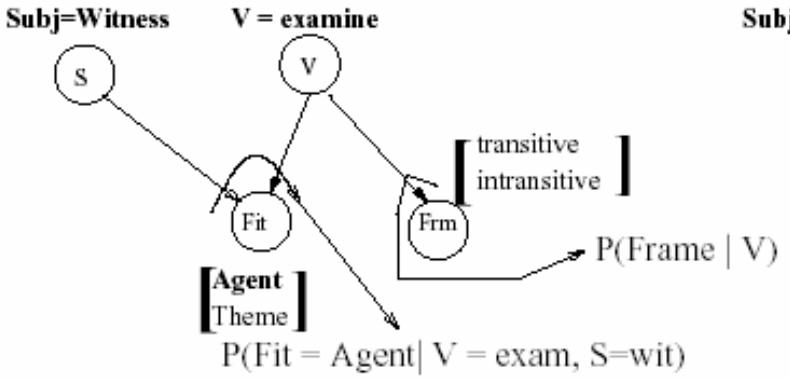
MAIN VERB INTERPRETATION



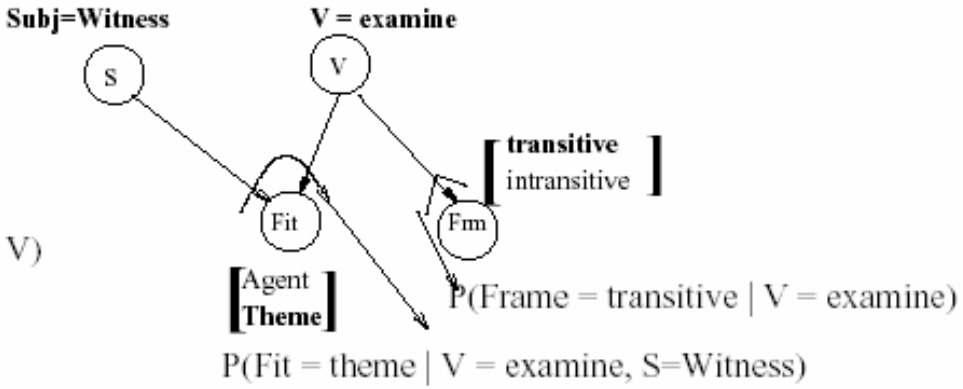
REDUCED RELATIVE INTERPRETATION



LEXICAL NETWORK AFTER "NP V"



MAIN VERB INTERPRETATION



REDUCED RELATIVE INTERPRETATION

Talk Outline

- Introduction
  - Evidence for Probabilistic factors in sentence processing
  - Background and alternative models
  - Problems
- A Bayesian Model of sentence processing
  - Basic Model/Result
  - Details
- Results on behavioral data
- Ongoing Work
- Conclusion

Testing the model

- Match behavioral data on the two most-studied ambiguities in sentence processing
  - Main Clause/Reduced Relative ambiguity
  - Direct Object/Sentential Complement ambiguity

Detailed Prediction: expectation

$$\text{reading time}(word) \propto \frac{1}{P(\text{word}|\text{context})}$$

$$P(\text{word}|\text{context}) = \frac{P(w_1 \ldots w_i)}{P(w_1 \ldots w_{i-1})} = \frac{P(I^t)}{P(I^{t-1})}$$

Summing over all interpretations at time t:

$$P(I_i^t) = \frac{\prod_{s=1}^{s=m} I_{i_s}^t}{\sum_{j=1}^{j=n} \prod_{s=1}^{s=m} I_{j_s}^t}$$

Expectation

$$\delta(I_{i^t}) = \frac{P(I_i^t)}{P(I_i^{t-1})}$$

$$\text{ProcessingTime}_{\text{Expectation}} \approx -\delta(I_{i^t})$$

Attention

- Most Likely interpretation

$$P^*(I^t) = \underset{i \in \text{interpretations}}{\text{argmax}} P(I_i^t)$$

- Reordering is a change in preference

$$P^*(I^t) \neq P^*(I^{t-1})$$

- We assume linear scaling cost for flip

Linear cost for flip

$$\text{ProcessingTime}_{\text{Reordering}} \approx$$

$$= \begin{cases} -w_{flip} \times \delta(I_{i^t}), & \text{if } P^*(I^t) \neq P^*(I^{t-1}) \\ -\delta(I_{i^t}), & \text{if } P^*(I^t) = P^*(I^{t-1}) \end{cases}$$

Summary of Predictions

- Expectation

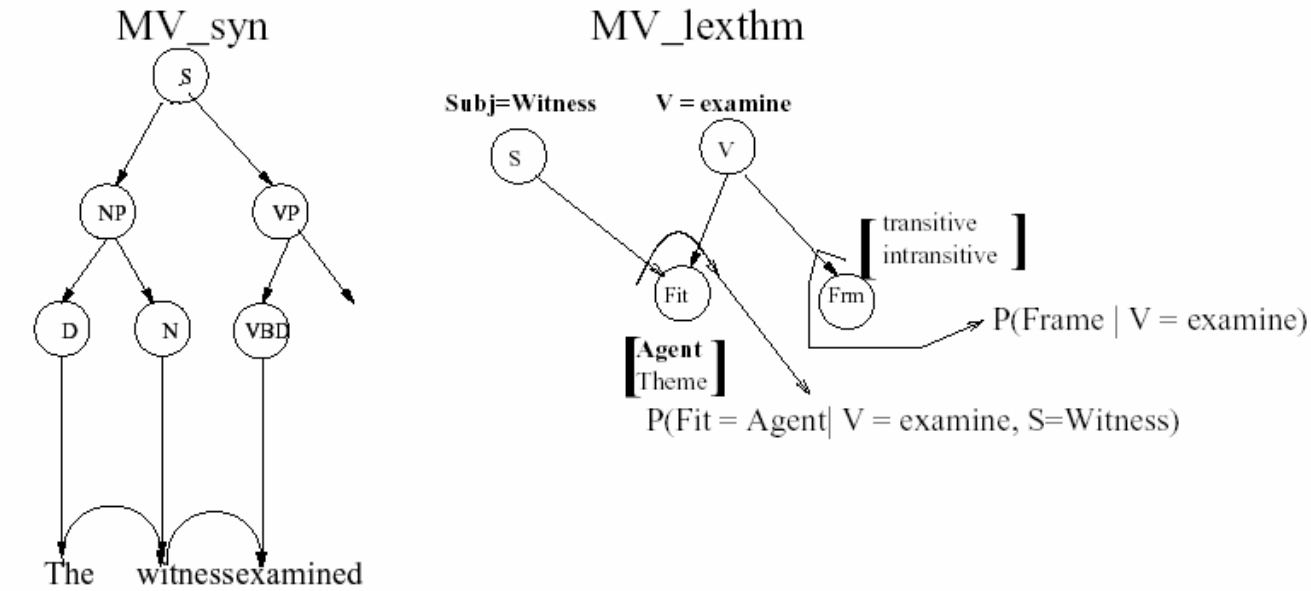
$$\text{ProcessingTime}_{\text{Expectation}} \propto \frac{1}{P(\text{word}|\text{context})}$$

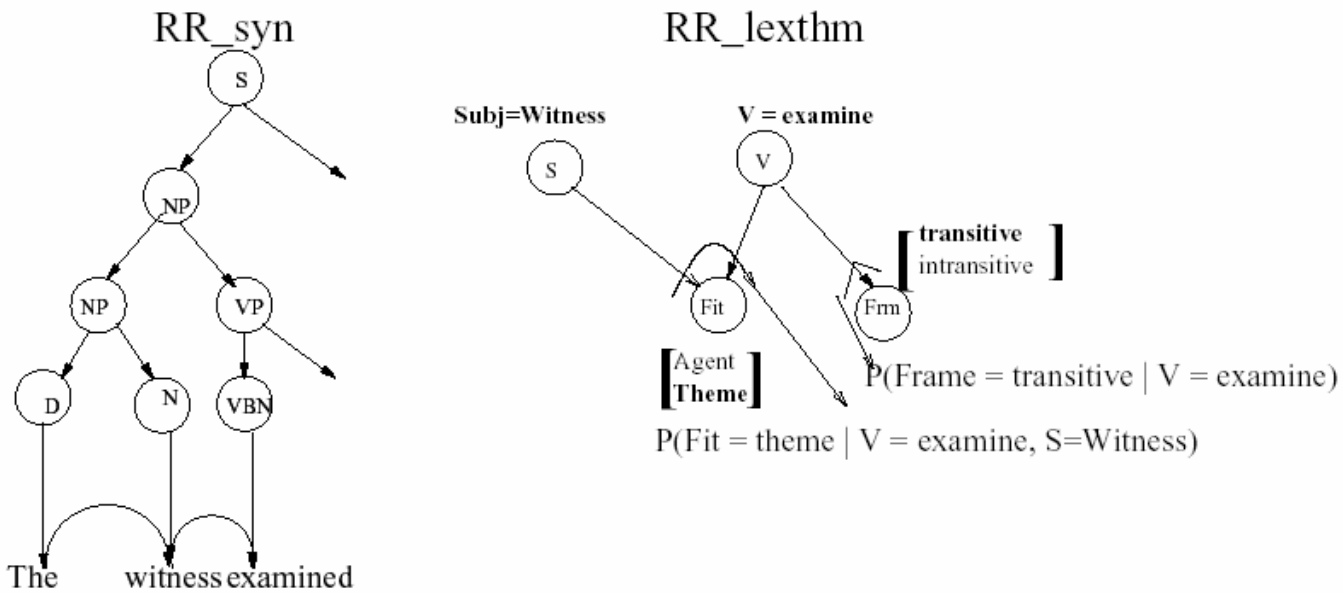
$$\text{ProcessingTime}_{\text{Expectation}} \propto -\delta(I_{i^t})$$

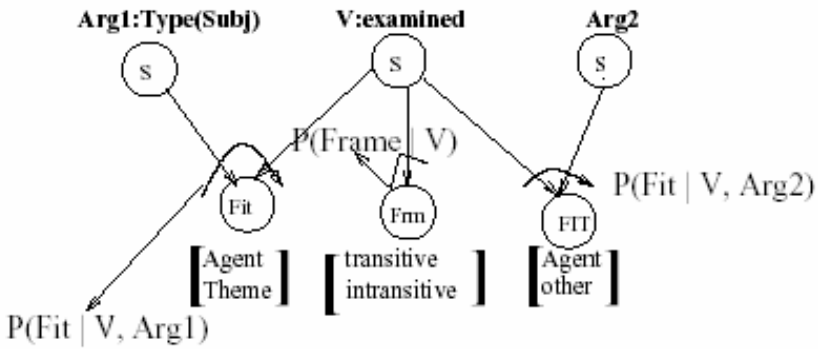
- Attention

$$\text{ProcessingTime}_{\text{Reordering}} \approx$$

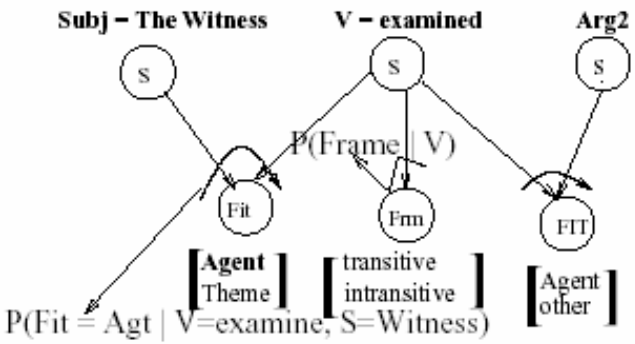
$$= \begin{cases} -w_{flip} \times \delta(I_{i^t}), & \text{if } P^*(I^t) \neq P^*(I^{t-1}) \\ -\delta(I_{i^t}), & \text{if } P^*(I^t) = P^*(I^{t-1}) \end{cases}$$


MV_syn
S
NP
VP
D
N
VBD
The
witness
examined
MV_lexthm
Subj=Witness
V = examine
S
V
transitive
intransitive
Fit
Frm
P(Frame | V = examine)
Agent
Theme
P(Fit = Agent| V = examine, S=Witness)

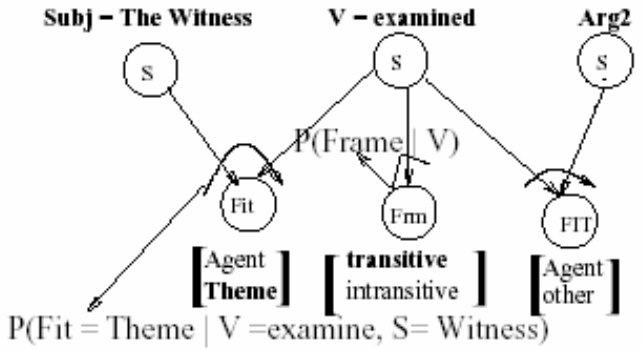

RR_syn
S
NP
NP
VP
D
N
VBN
The
witness
examined
RR_lexthm
Subj=Witness
V = examine
S
V
transitive
intransitive
Fit
Frm
Agent
Theme
P(Frame = transitive | V = examine)
P(Fit = theme | V = examine, S=Witness)



LEXICAL NETWORK AFTER "NP V"



MAIN VERB INTERPRETATION

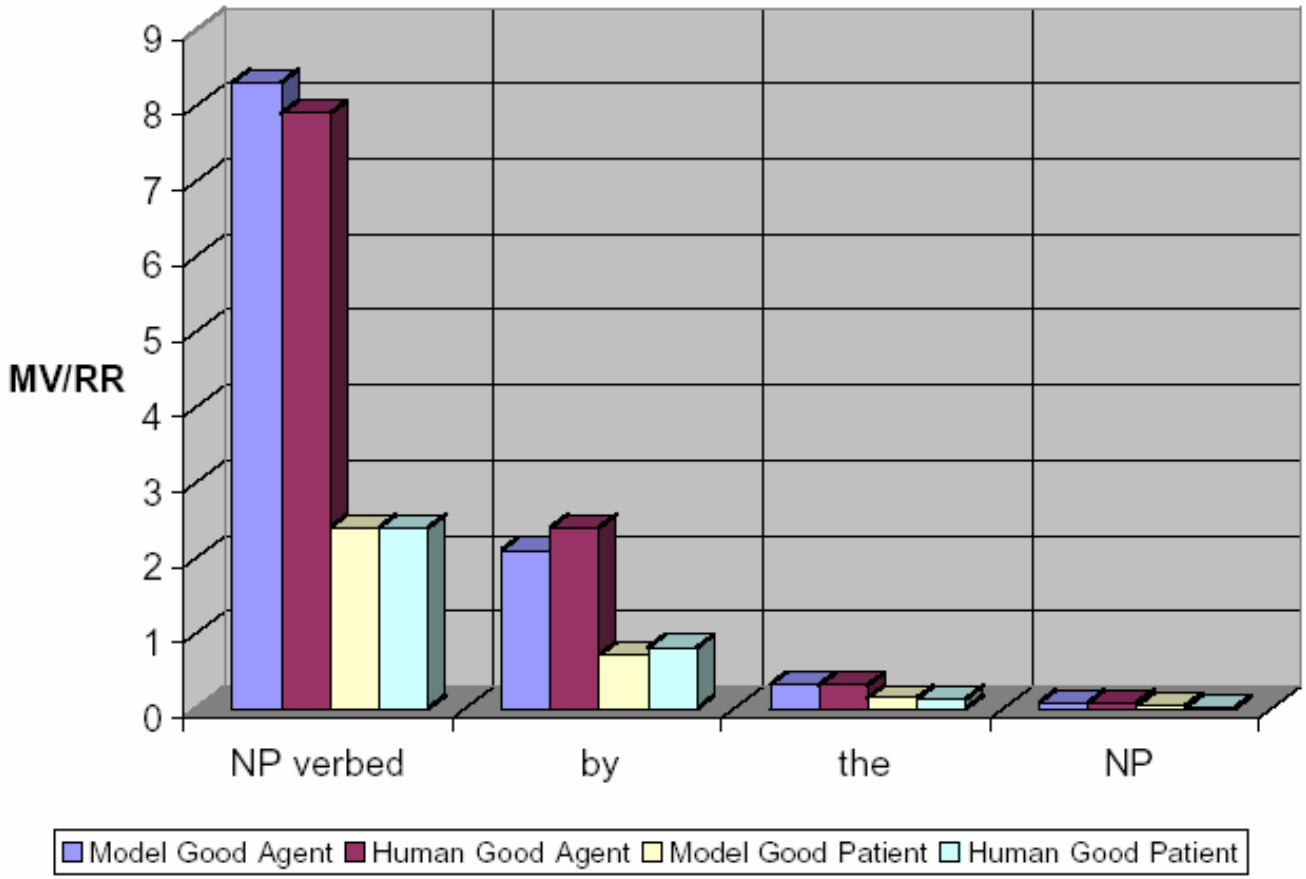


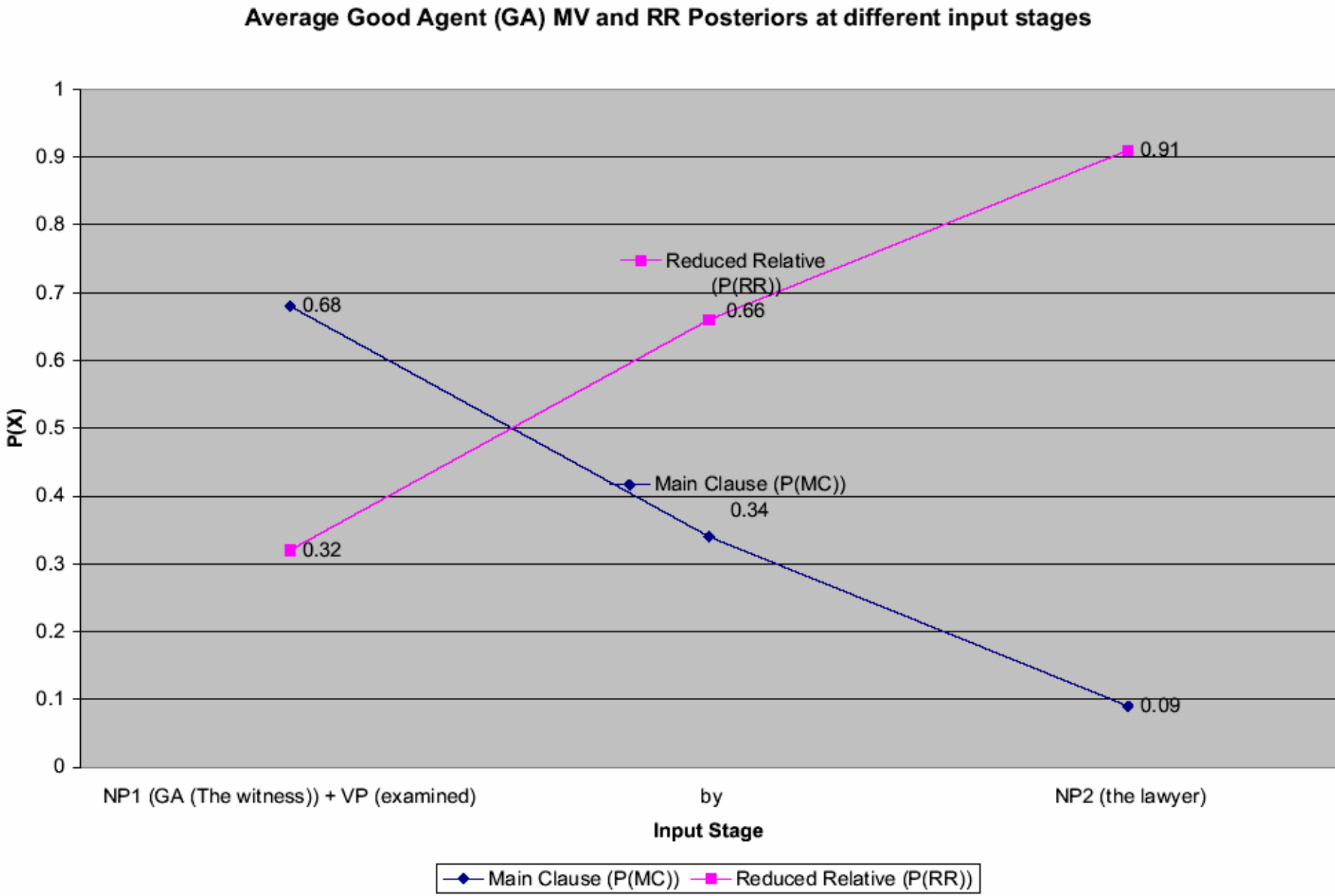
REDUCED RELATIVE INTERPRETATION

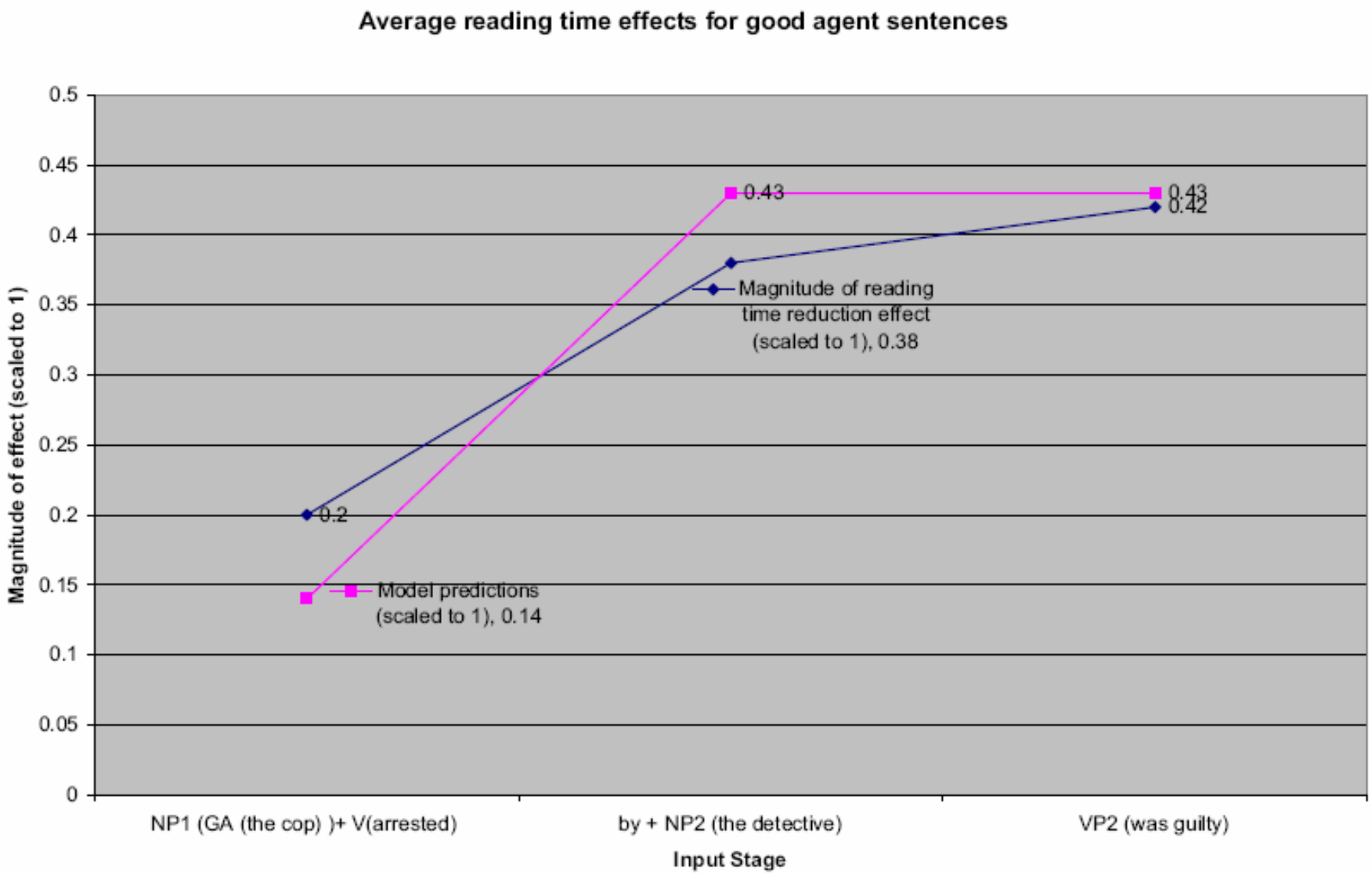
Data sources

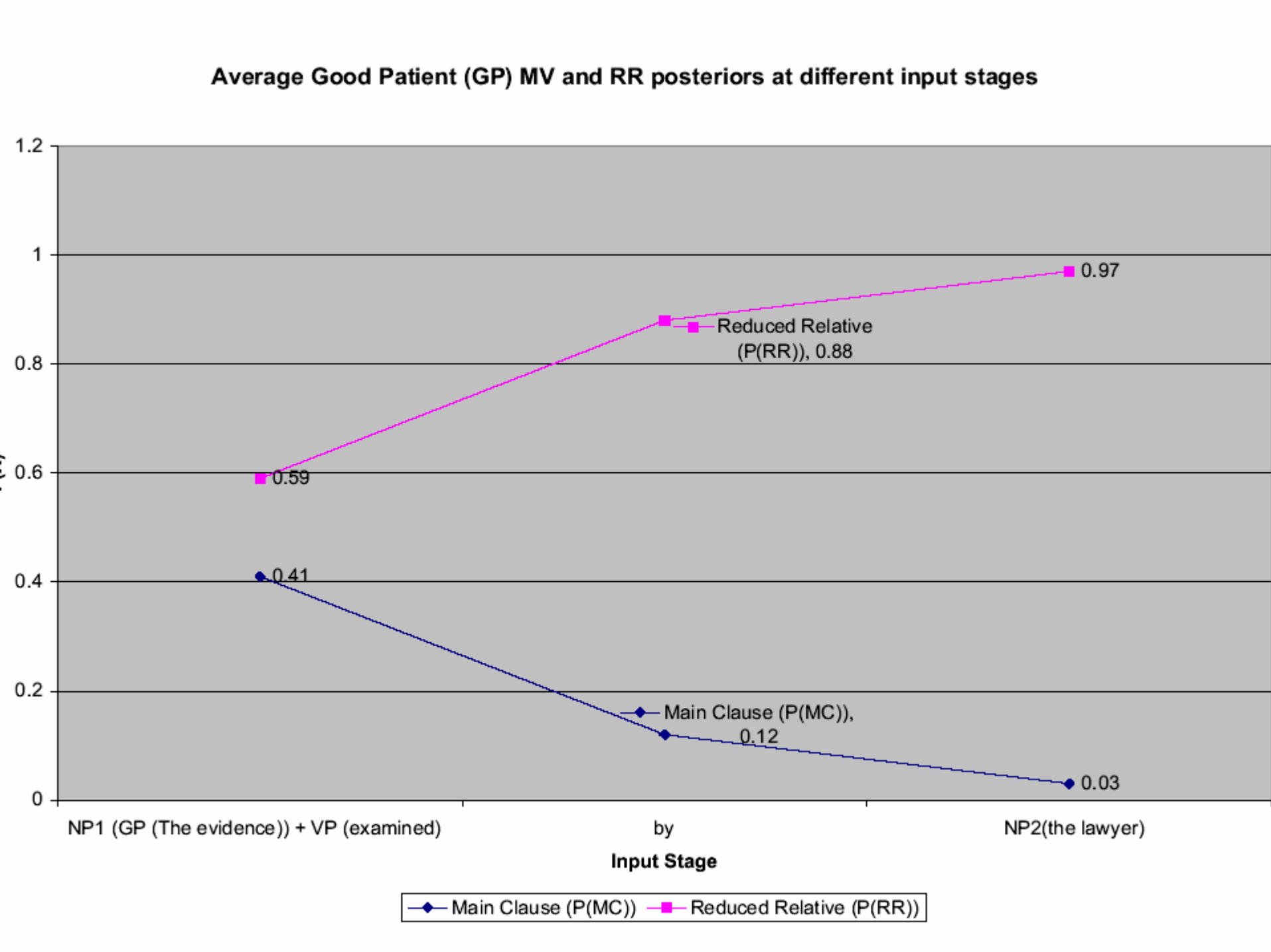
Data	Source
Valence Probabilities	
P(Agent|verb, initial NP)	McRae *et al.* (1998)
P(Theme|verb, initial NP)	McRae *et al.* (1998)
P(Theme|initial NP, verb-ed, by)	McRae *et al.* (1998) (.8, .2)
P(Agent|initial NP, verb-ed, by, the, NP)	McRae *et al.* (1998) (4.6 avg)
P(transitive|verb)	TASA corpus counts
P(intransitive|verb)	TASA corpus counts
SCFG Probabilities	
P(MC| SCFG prefix)	SCFG counts from corpora
P(RR| SCFG prefix)	SCFG counts from corpora
P(Participle|verb)	SCFG counts from corpora
P(SimplePast|verb)	SCFG counts from corpora

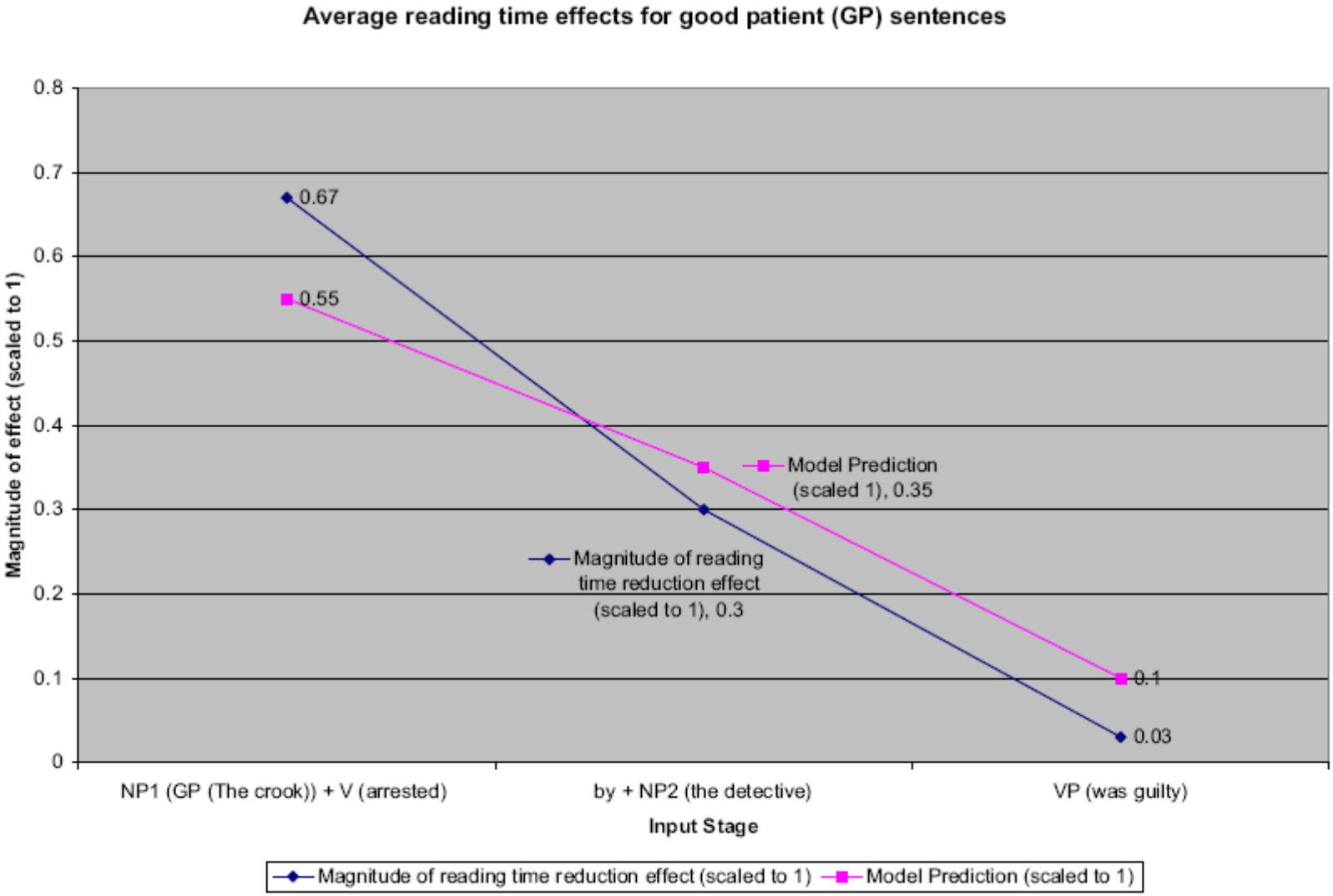
Sentence completion results




Average Good Agent (GA) MV and RR Posteriors at different input stages
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P(X)
Reduced Relative (P(RR))
0.68
0.66
0.91
Main Clause (P(MC))
0.34
0.32
0.09
NP1 (GA (The witness)) + VP (examined)
by
NP2 (the lawyer)
Input Stage
Main Clause (P(MC))
Reduced Relative (P(RR))


Average reading time effects for good agent sentences
Magnitude of effect (scaled to 1)
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.43
0.43
0.42
Magnitude of reading time reduction effect (scaled to 1), 0.38
0.2
Model predictions (scaled to 1), 0.14
NP1 (GA (the cop))+ V(arrested)
by + NP2 (the detective)
VP2 (was guilty)
Input Stage


Average Good Patient (GP) MV and RR posteriors at different input stages
1.2
1
0.8
0.6
0.4
0.2
0
0.97
Reduced Relative (P(RR)), 0.88
0.59
0.41
Main Clause (P(MC)), 0.12
0.03
NP1 (GP (The evidence)) + VP (examined)
by
NP2(the lawyer)
Input Stage
Main Clause (P(MC))
Reduced Relative (P(RR))


Average reading time effects for good patient (GP) sentences
Magnitude of effect (scaled to 1)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.67
0.55
0.1
0.03
Model Prediction (scaled 1), 0.35
Magnitude of reading time reduction effect (scaled to 1), 0.3
NP1 (GP (The crook)) + V (arrested)
by + NP2 (the detective)
VP (was guilty)
Input Stage
Magnitude of reading time reduction effect (scaled to 1)
Model Prediction (scaled to 1)

Results on MV/RR data

- Bayesian model successfully accounts for reading time data.
- Two factors:
  - Attention: (Good Agent sentences show demotion)
  - Expectation: (Combined effects of RR parse, verb bias, etc)

Modeling SC/DO data

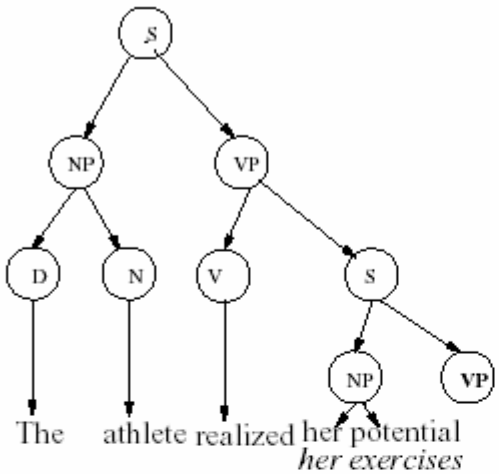
- Expectation Principle:

$$\text{reading time}(word) \propto \frac{1}{P(\text{word}|\text{context})}$$

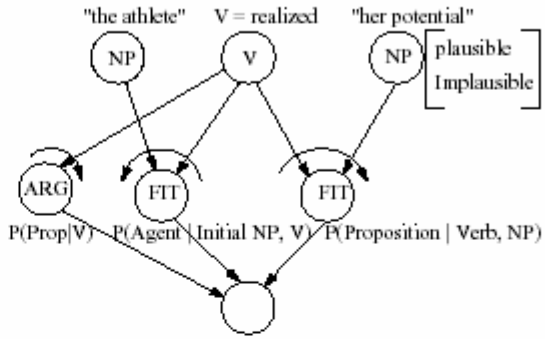
- *her exercises* is a low probability continuation to *realize*:
- – The young athlete realized **her potential** ...
- – The young athlete realized **her exercises** ...

Bayes net for SC and DO

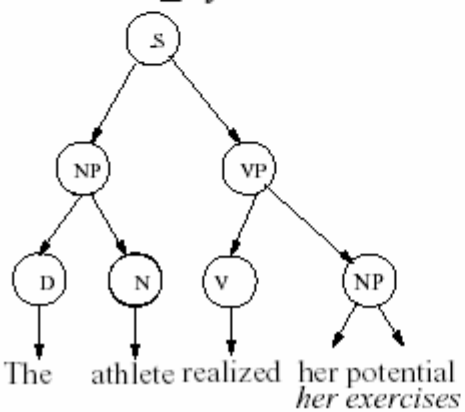
SC_Syn



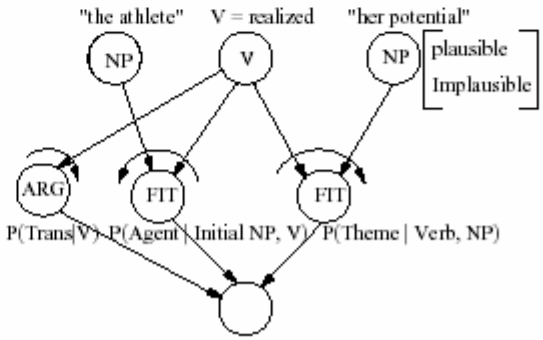
SC_lexthm



DO_Syn

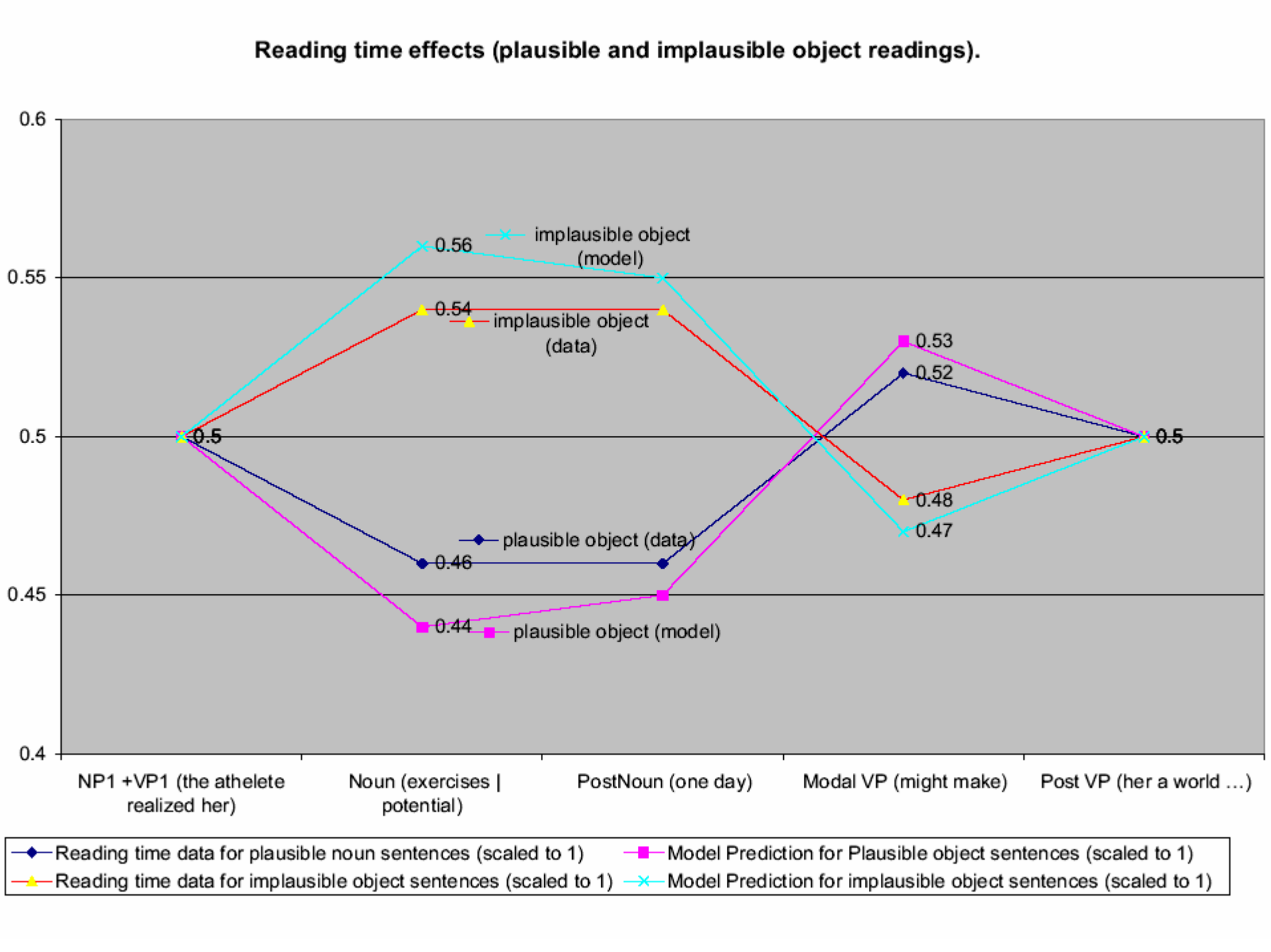


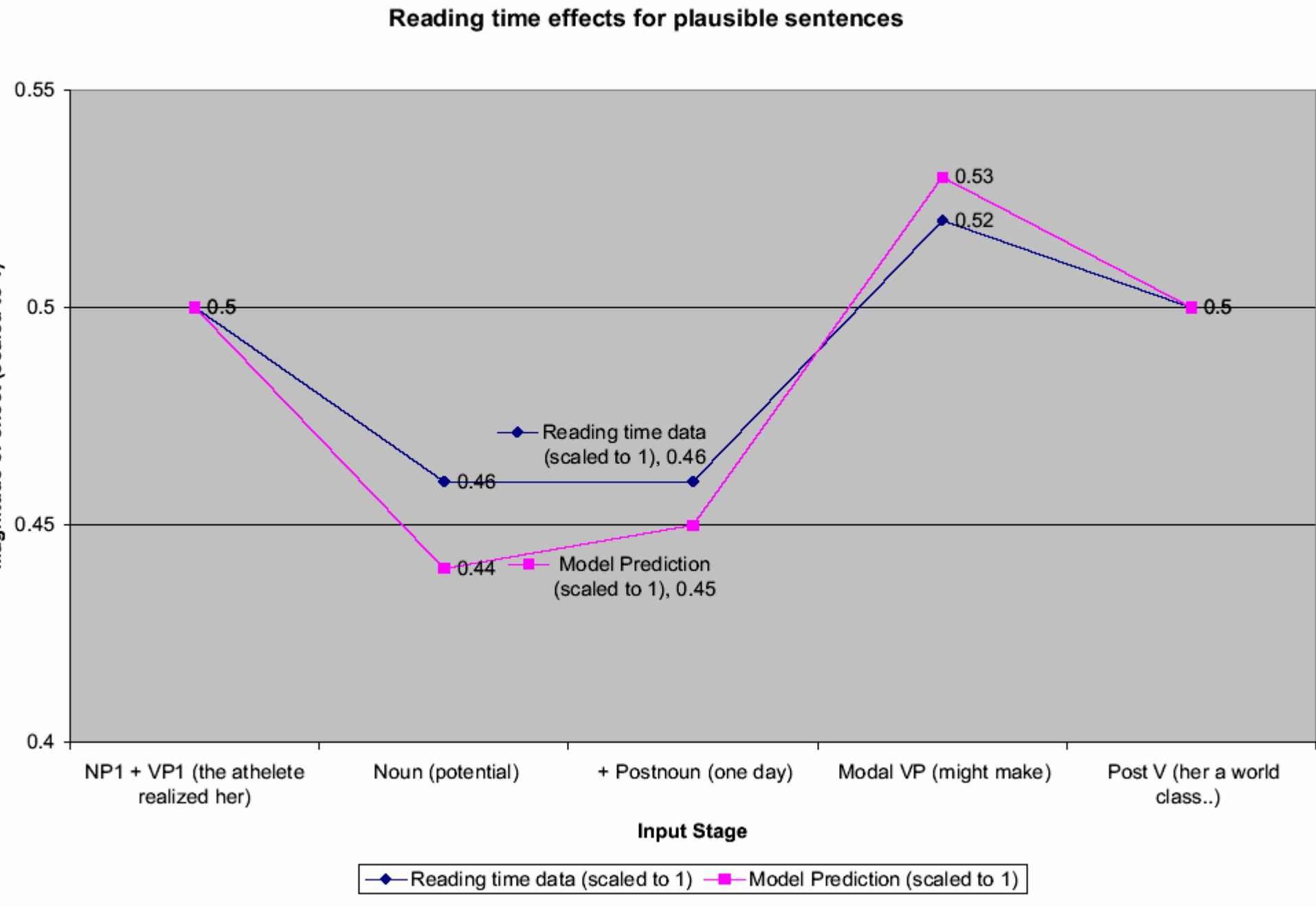
DO_lexthm

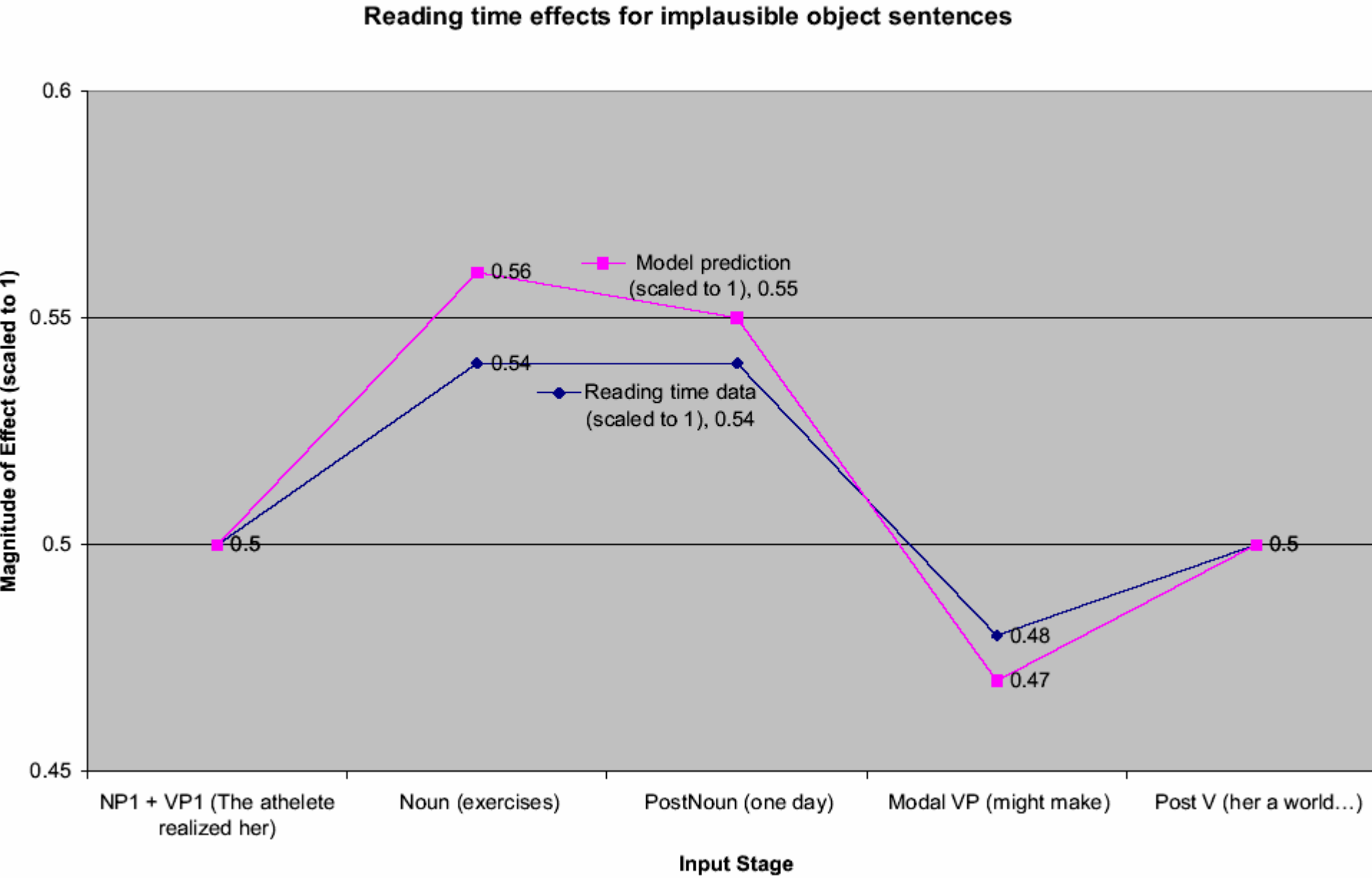


Data (Pickering et al. 2000)

Parameter	Value
$(P(SC)\|V = realized)$	.35
$(P(DO)\|V = realized)$	0.25
$(P(SC)\|VP = realized, [_{NP} her \ldots], InitialNP = The, young, athlete)$	.8
$(P(DO)\|VP = realized, [_{NP} her \ldots], InitialNP = The, young, athlete)$	0.2


Reading time effects (plausible and implausible object readings).
0.6
0.55
0.5
0.45
0.4
implausible object (model)
0.56
0.54
implausible object (data)
0.53
0.52
0.5
0.5
0.48
0.47
plausible object (data)
0.46
0.44
plausible object (model)
NP1 +VP1 (the athelete realized her)
Noun (exercises | potential)
PostNoun (one day)
Modal VP (might make)
Post VP (her a world …)
Reading time data for plausible noun sentences (scaled to 1)
Model Prediction for Plausible object sentences (scaled to 1)
Reading time data for implausible object sentences (scaled to 1)
Model Prediction for implausible object sentences (scaled to 1)


Reading time effects for plausible sentences
0.55
0.5
0.45
0.4
0.5
0.46
0.44
0.53
0.52
0.5
Reading time data (scaled to 1), 0.46
Model Prediction (scaled to 1), 0.45
NP1 + VP1 (the athelete realized her)
Noun (potential)
+ Postnoun (one day)
Modal VP (might make)
Post V (her a world class..)
Input Stage
Reading time data (scaled to 1)
Model Prediction (scaled to 1)


Reading time effects for implausible object sentences
Magnitude of Effect (scaled to 1)
0.6
0.55
0.5
0.45
Model prediction (scaled to 1), 0.55
Reading time data (scaled to 1), 0.54
0.56
0.54
0.5
0.48
0.47
0.5
NP1 + VP1 (The athelete realized her)
Noun (exercises)
PostNoun (one day)
Modal VP (might make)
Post V (her a world…)
Input Stage

Summary of SC/DO results

- Bayesian model successfully accounts for reading time data.
- Expectation principle: Unexpected words cause increased reading time.

Current work

- Build a scalable parser based on the model principles (John Bryant)
  - Combining evidence from multiple sources
  - Using construction grammar
  - Select best fitting construction in an incremental fashion
- Appears to match other behavioral data (Gibson, Hale).
- Do experiments (or better still data, anyone?)
- How does sentence processing integrate with semantics and inference (http://www.icsi.berkeley.edu/NTL)?

Future work

- We now have a framework to investigate
  - Big questions
    - Sentence processing as an on-line integration of
      - Discourse knowledge
      - Semantics
      - Prosody
      - Speaker intention/illocutionary foces
  - Smaller questions
    - Relationship between evidence combination and priming
    - Integrating probabilistic approaches with resource constraints such as working memory).

Conclusion

- Our model combines two basic ideas in Language Processing.
  - a) Linguistic Knowledge is highly structured and hierarchically organized (syntactic and argument structure knowledge).
  - b) Multiple sources of knowledge, conceptual and perceptual interact probabilistically in access and disambiguation (dynamic systems models and construction grammar).
- Using Graphical models allows us to compute the joint distribution of multiple, correlated features by using structural relationships to minimize inter-feature correlations. This has the dual advantage of compact representation and clarity of model.
- **Result:** A computational method that allows us model a wide range of psycholinguistic data and to systematically investigate the role of different knowledge sources on human language processing.