Computational Models
Lecture 9
Introduction to Psycholinguistics

Matthew W. Crocker
Pia Knoeferle

Department of Computational Linguistics
Saarland University

Models and Linking Hypotheses

Relate the theory/model to some observed measure

- Typically impossible to predict measures completely

Theories of parsing typically determine …

- what **mechanism** is used to construct interpretations?
- which **information** sources are used by the mechanism?
- which **representation** is preferred/constructed when ambiguity arises?

Linking Hypothesis:

- Preferred sentence structures should have faster reading times in the disambiguating region than dispreferred

Mechanisms for Language Processing

How structures are constructed

- Top-down
- Bottom-up
- Mixed strategy: Left-corner parsing

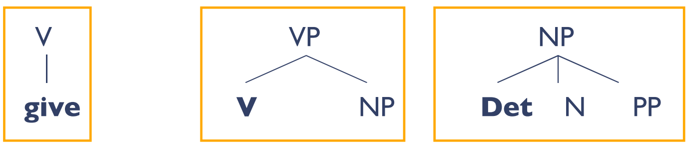
How local and global ambiguity are handled

- Serial (deterministic/non-deterministic)
- Parallel (bounded/unbounded)

A Psychologically Plausible Parser

Left-Corner Parsing

Rules are 'activated' by their 'left-corner'



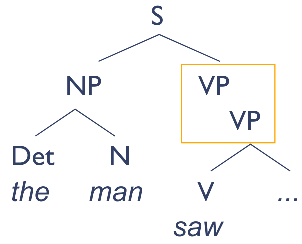
Combines input-driven with top-down

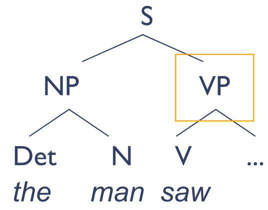
There is a 'class' of LC parsers

Evaluating the LC Parser

Not necessarily incremental:

- Variations: Arc-standard *versus* arc-eager





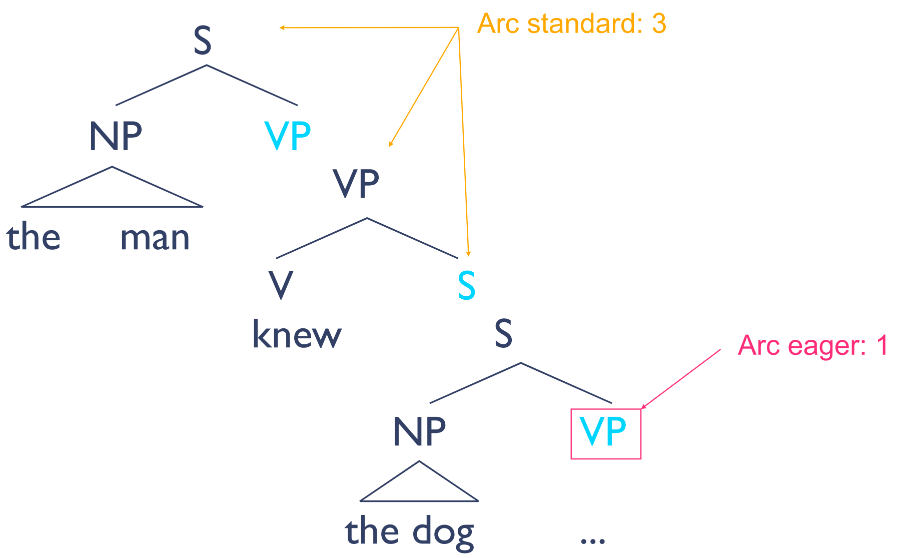
Affect on ambiguity resolution for arc-eager:

- Commitment to attachments is early, before daughters are completely built
- Top-down use of syntactic context and possible left-recursion problems

Evaluating the LC Parser

Variations:

- Arc-standard versus Arc-eager



Summary of Behaviour

Node	Arcs	Left	Centre	Right
Top-down	Either	O(n)	O(n)	O(1)
Shift-reduce	Either	O(1)	O(n)	O(n)
Left-corner	Standard	O(1)	O(n)	O(n)
Left-corner	Eager	O(1)	O(n)	O(1)
People		O(1)	O(n)	O(1)

A Theory of Sentence Processing: Frazier

What **architecture** is assumed?

- Modular syntactic processor, with restricted lexical (category) and semantic knowledge

What **mechanisms** is used to construct interpretations?

- Incremental, serial parsing, with reanalysis

What **information** is used to determine preferred structure?

- General syntactic principles based on the current phrase structure

Linking Hypothesis:

- Parse complexity and reanalysis cause increased RTs

Garden-Path Theory: Jurafsky (1996)

What **architecture** is assumed?

- Modular lexico-syntactic processor with lexical (category and subcategory), 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

But there are many models/theories ... Why?

Language: recursive = infinite search = many algorithms

Models *a la* Carte

Representations

- Levels, transformations, interfaces

Architectures

- Modules, information flow, bandwidth, dynamics/time

Mechanisms

- Serial; parallel structure building; competitive interaction

Reanalysis

- Serial repair; parallel re-ranking; activation-based

Multiple constraints in ambiguity resolution

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 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

Constraint-based Models

What **architecture** is assumed?

- Non-modular: all levels of representation are constructed and interact simulaneously

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

The Interactive Activation Model (MacDonald et al, 1994)

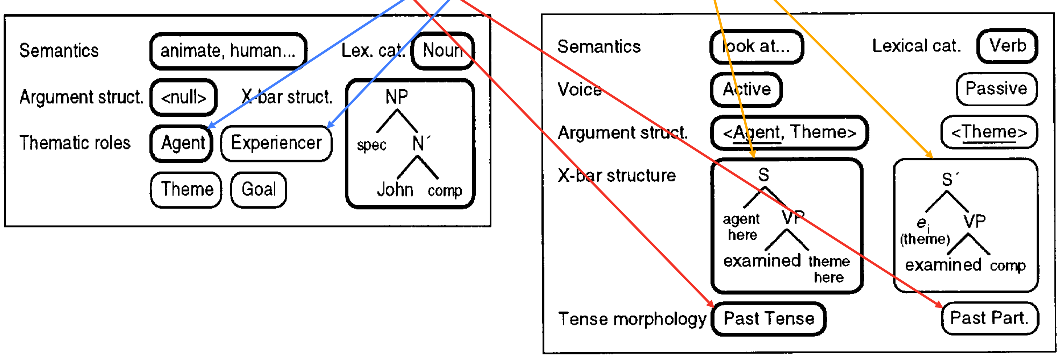
Rich lexical entries; frequency determines 'activations'

Consider: *"John examined the evidence"*

- "examined" is either a simple past or past participle

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

"John" *"Examined"*



MacDonald, Pearlmutter & Seidenberg

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 info
- Language processing is "constraint satisfaction", between lexical entries, and across levels; No distinct parser

Questions:

- Complex interaction behaviours 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.

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

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

Steps in the Experiment: (McRae et al 1998)

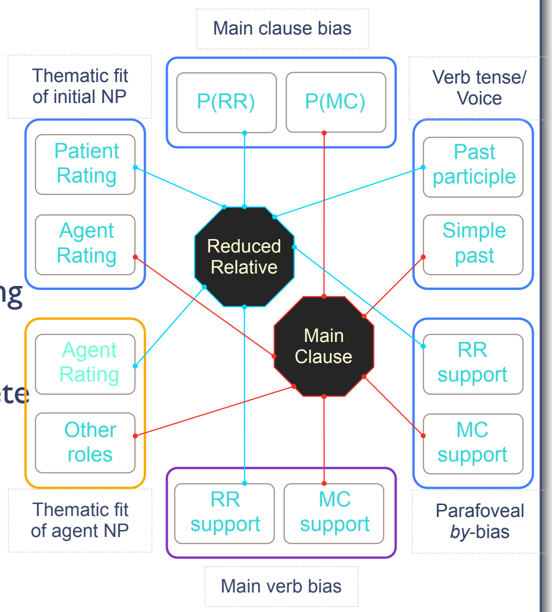
➡ How do constraints contribute to the activation of competing analyses, over time

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

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



Constraints/Parameters of the Model

"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

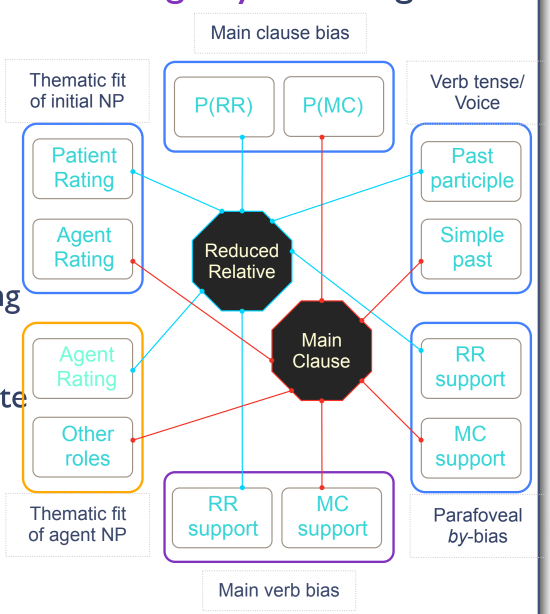
```
How common is it for a
    crook      _____
    cop        _____
    guard      _____
    police        _____
    suspect       _____
To arrest someone?
To be arrested by someone?
```

NP 1	Rel	Main	by NP	Rel	Main
Agent	1,5	5,3	Agent	4,6	1,0
Patient	5,0	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



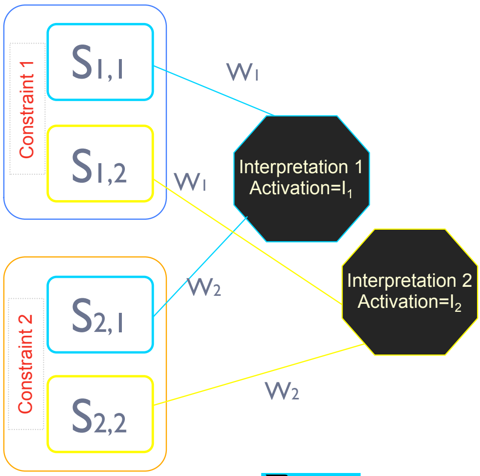
The recurrence mechanism

$S_{c,a}$ 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: $S_{c,a}(norm) = \frac{S_{c,a}}{\sum_a S_{c,a}}$

Integrate: $I_a = \sum_c [w_c \cdot S_{c,a}(norm)]$

Feedback: $S_{c,a} = S_{c,a}(norm) + I_a \cdot w_c \cdot S_{c,a}(norm)$

$\sum_i w_i = 1$

Fitting Constraint Weights using Completions

The 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

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

Gated sentence completion study:

The cop/crook arrested ...
The crook arrested by ...
The crook arrested by the ...
The crook arrested by the detective...

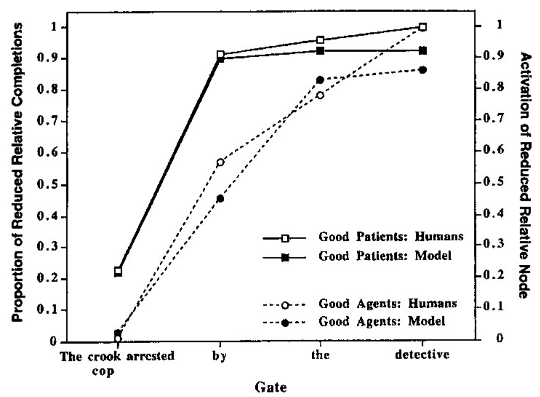


FIG. 2. Human and simulation results for fragment completions.

Counted "the crook arrested himself" as RR (!?)

Self-Paced Reading Study

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

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 (using the model)

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

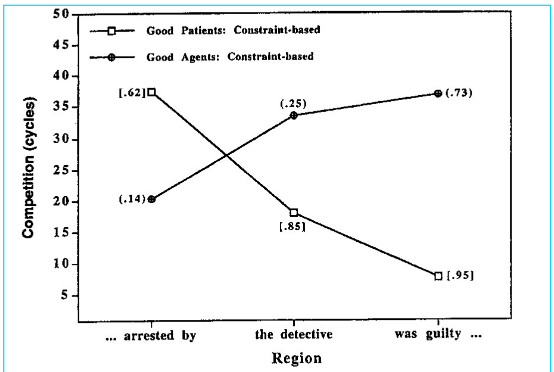


FIG. 3. Self-paced predictions derived from the constraint-based competition model. In this and all following model figures, the number beside each model datum is the mean activation of the reduced relative node after competition in that region for either (good agents) or [good patients].

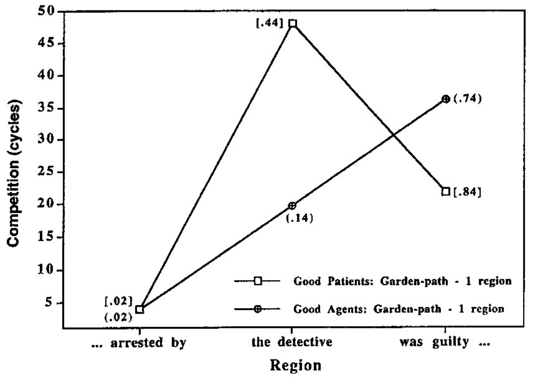


FIG. 4. Self-paced predictions as derived from the garden-path model when constraints other than the main clause and main verb biases were delayed by a region.

GP vs CB Modelling of the Reading

Reduction effect/cycles:

Human reading times:

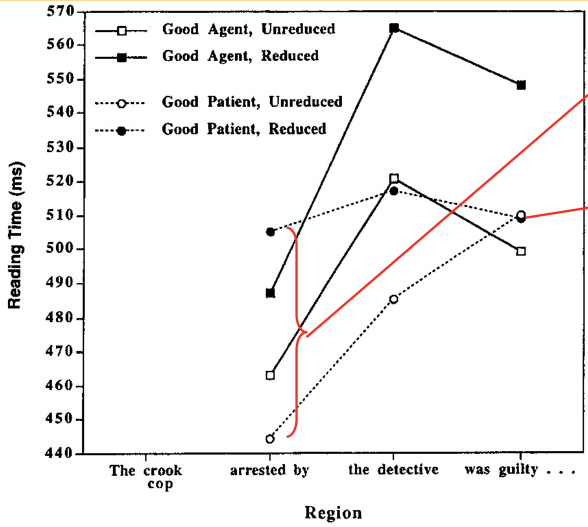


FIG. 5. Self-paced reading times for the Experiment.

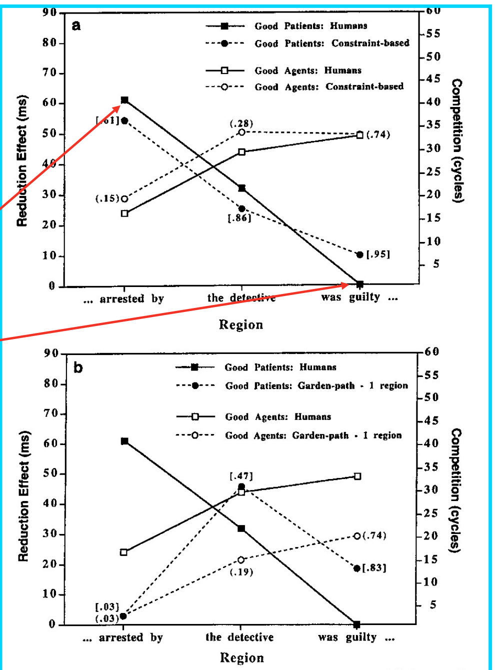


FIG. 6. Simulations of self-paced reading by (a) the constraint-based model, and (b) the one-region delay garden-path model.

3rd Version: Short Delay GP Model

The GP-model, has a 1-2 word delay in use of information, what if this delay is reduced?

- 4 cycles (10-25ms)
- Much better fit, except for the high reduction effect still predicted at main verb (good patient).
- RMS error 5.5

Search for the best assignment of weights:

MC bias: .2966 (.5094)

Th. fit: .4611 (.3684)

V.tense: .0254

by-bias: .2199

- RMS error 2.77
- (but no-longer models completions)

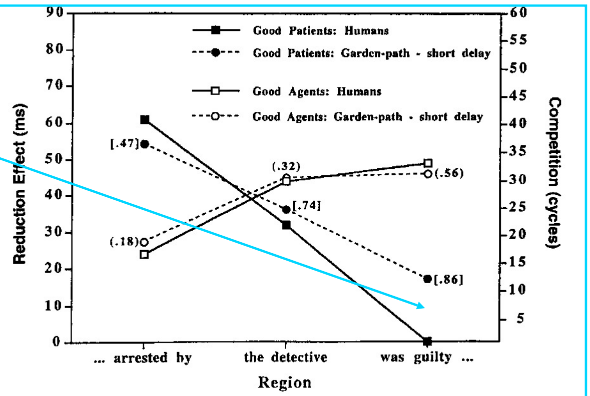


FIG. 7. Garden-path simulations of self-paced reading when constraints other than the main clause and main verb biases were delayed by 4 cycles of competition, or approximately 10–25 ms. Note the large predicted reduction effect at the main verb for the good patients.

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:
  - E.g. tense bias (incl) correlates with transitivity (excl)

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

- What other syntactic constraints might be included?

Predicts long reading times when constraints compete

- People are often *faster* at processing ambiguous regions!

The Competitive Attachment Model

CAPERS: A Hybrid Model

- A symbolic-connectionist model of sentence parsing
- A competition-based model of parsing and reanalysis

Models parsing, disambiguation, and reanalysis via competitive activation among structural alternatives

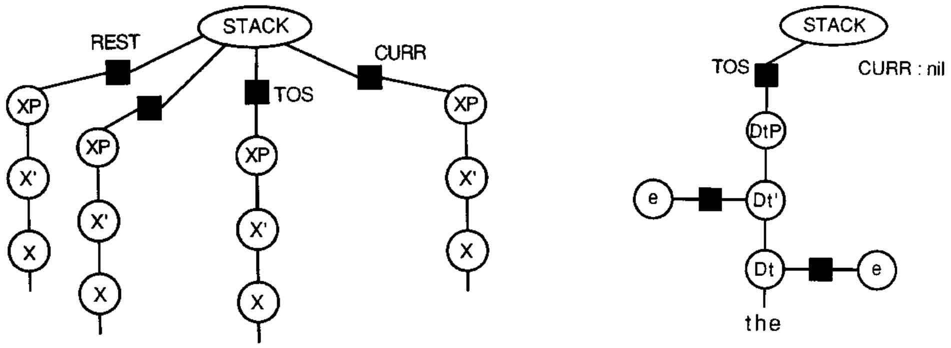
- Direct, symbolic encoding of linguistic representations
- Distributed decision-making
- Competition-based spreading activation
  - Competition is indirect, nodes vie for output activation from their neighbors

Building blocks of CAM

Words instantiate X' projection templates

Lexical item determine valency of projections:

- specifiers, complements, and modifiers



Implementation of the Model

Nodes in the tree correspond to p-nodes, and are only projected on the basis of lexical input.

Attachments between sisters are formed by a-nodes:

- Mediate feature agreement between p-nodes
- Each p-node uses constraint-based spreading activation (CBSA) to allocate activation to it's a-nodes:
  - Proportional to the current activation of the a-nodes
- The degree of satisfaction of grammatical constraints determines the a-nodes state-value, which in turn contributes to the activation
- A-nodes "AND" their inputs, to ensure that they "agree"
- Null (phi-nodes) are inserted for attachments which are yet to be made.

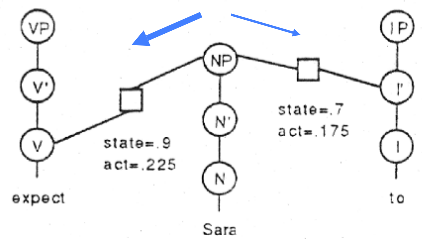
CBSA

The CBSA Function

- o_{ji}: output from n_i to n_j
- a_j: activation of n_j
- k: ranges over nodes connected to n_i

$$o_{ji} = \frac{a_j}{\sum_k a_k} \bullet a_i$$

Consider: *Mary expected Sarah to leave*



A-nodes "state" reflects degree to which grammar constraints are satisfied

The output activation of p-nodes:

- Shared to it's a-nodes, proportional to their current activation

When Kiva eats food gets thrown

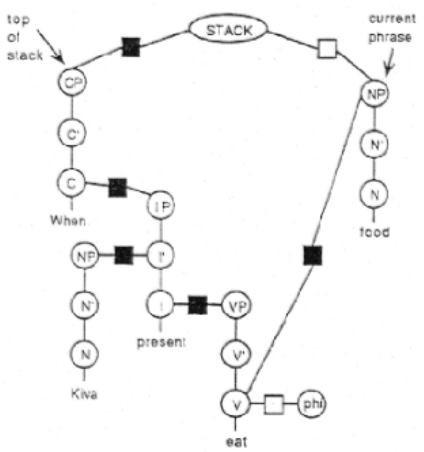
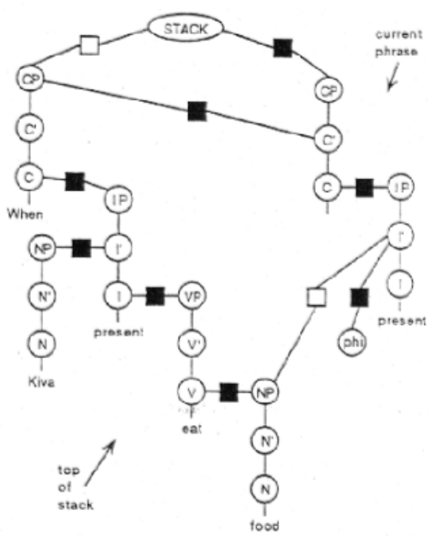


Figure 10: The NP *food* has a single valid attachment to the parse tree.



Examples in detail

Mary expected Sara to leave

Sara:

- Default stack attachment is not competitive, since a1 is highly activated

to:

- IP is projected
- a2, a3 simultaneously present a revised structure possibility
- a1 competes with a2, a3 for activation from the V & NP nodes
- IP and I' can put all activation into a2, a3 respectively

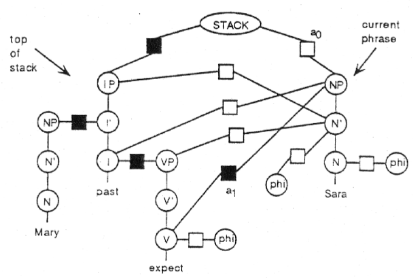
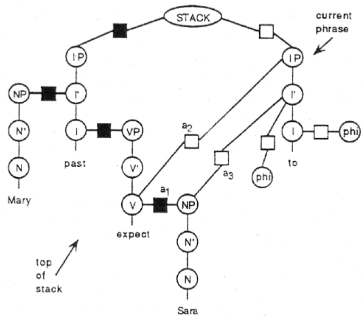
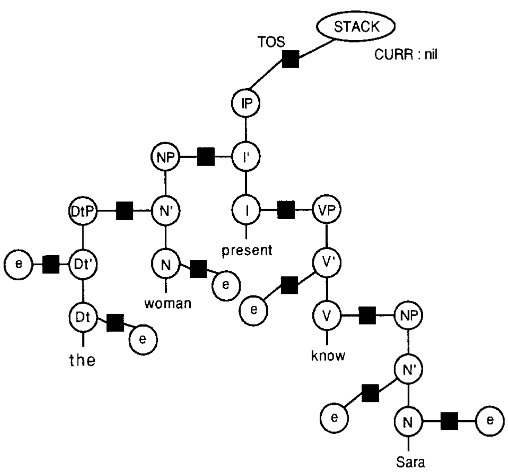
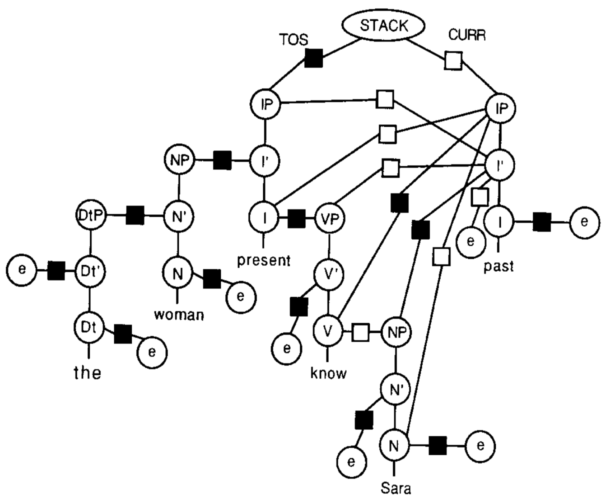


Figure 7: The network after attaching the NP *Sara*.

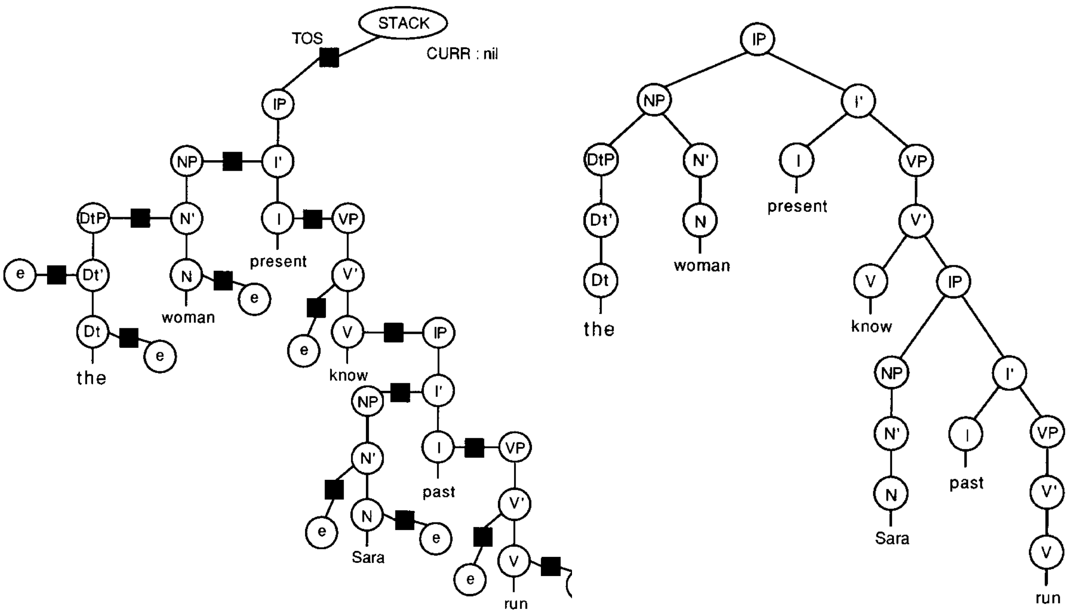


Stevenson's Model: Example 1

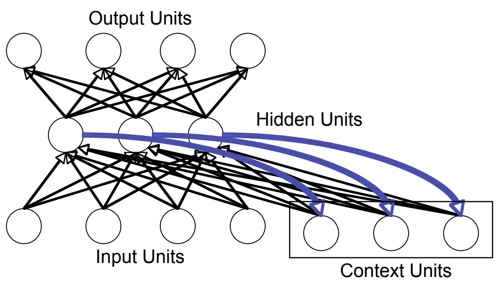




Example 1: continued



Connectionist models



Purely connectionist models: Simple Recurrent Networks

- Learn based on experience
  - Supervised: trained to output a meaning representation
  - Unsupervised: trained to predict the next word
- Good a learning a range of linguistic constraints, and integrating multiple information sources

Summary of SRNs

Successes:

- Automatic learning of syntactic constraints
- Analysis suggests lexical categories (syn, sem) are also implicitly represented
- Limited generalization to unseen words and structures

Weaknesses

- No explicit representation of linguistic structure
- Emphasis on learning, rather than processing
  - Difficult to measure reanalysis difficulty
- Issues of scaleability

Summary

Theories of sentence processing are distinguished by their

- <u>the information sources which resolve ambiguity</u>
- <u>mechanisms for building representations</u>

Processing complexity is typically predicted when

- <u>the preferred interpretation must be revised</u>
- <u>the processing load associated the structure is high</u>

Theories offer different underlying motivations:

- Minimize cognitive resources, and/or
- Maximize comprehension
- Cognitive (neural) plausibility, and acquistion