Introduction to Psycholinguistics

Lecture 5: Experience- and Constraint-based Theories of Disambiguation

Matthew W Crocker

Computerlinguistik
Universität des Saarlandes

So far …

- Ambiguity in sentence processing:
  - Using reading-times to establish the preferred interpretation
  - Used evidence about the preferred interpretation to determine underlying parsing mechanisms
- Theories of syntactic parsing and disambiguation:
  - Garden Path Theory (Frazier): minimal attachment + late closure
  - Theta-Attachment (Pritchett): maximise role reception + assignment
- Parsing mechanisms: Arc-eager Left-corner parsing
  - Incremental: attaches each word into a connected (partial) parse tree
  - Mixes top-down and bottom-up strategies
  - Provides a reasonable account of memory load: why centre-embeddings are harder than left- or right- embeddings
  - Implementation of disambiguation strategies is still required

But what's missing …

- The previous accounts focus on
  - Syntactic (and lexico-syntactic) ambiguity
  - Purely syntactic mechanisms for disambiguation
  - Thus assume a modular parser, or at least the "primacy" of syntax
- Other factors: Experience and non-sytactic constraints
- Experience: is it possible that our prior experience with language, determines our preferences for interpreting the sentences we hear?
  - Tuning hypothesis: disambiguate structure according to how it has been most frequently disambiguated in the past.
- Non-syntactic constraints: to what extent do semantics, intonation, and context influence our resolution of ambiguity?

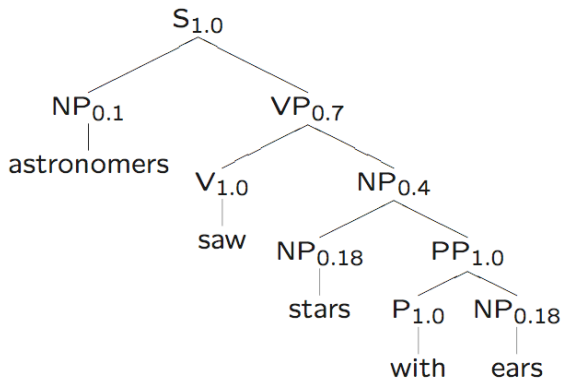
Probabilistic Context-free Grammars

- Context-free rules annotated with probabilities;
- Probabilities of all rules with the same left hand side sum to one;
- Probability of a parse is the product of the probabilities of all rules applied in the parse.
- Example (Manning and Schütze 1999)

S → NP VP	1.0	NP → NP PP	0.4
PP → P NP	1.0	NP → astronomers	0.1
VP → VP NP	0.7	NP → ears	0.18
VP → VP NP	0.3	NP → saw	0.04
P → with	1.0	NP → stars	0.18
V → saw	1.0	NP → telescopes	0.1

Example 1a

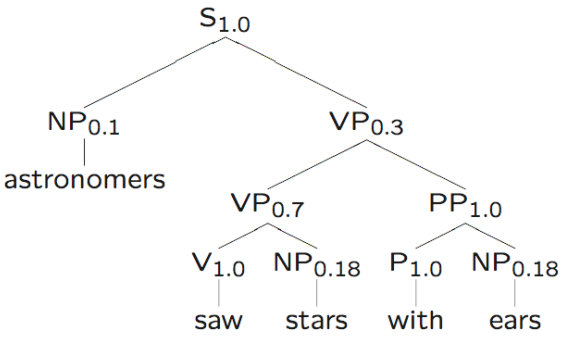
t_1:



$$P(t_1) = 1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0009072$$

Example 1b

t_2:



$$P(t_1) = 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0006804$$

Jurafsky (1996)

- Jurafsky's (1996) approach:
  - probabilistic model of lexical and syntactic access and disambiguation;
  - accounts for psycholinguistic data using concepts from computational linguistics: probabilistic CFGs, Bayesian modeling frame probabilities;
  - focus here: syntactic disambiguation in human sentence processing.
- Overview of issues:
  - data to be modeled: frame preferences, garden paths;
  - architecture: serial, parallel, limited parallel;
  - probabilistic CFGs, frame probabilities;
  - examples for frame preferences, garden paths;
  - comparison with other models; problems and issues.

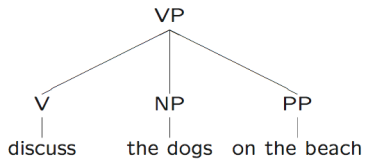
Frame Preferences

- (1) The women discussed the dogs on the beach.
  - a. **The women discussed the dogs which were on the beach.**
  - b. The women discussed them (the dogs) while on the beach.

$p(\text{discuss}, \langle \text{NP PP} \rangle) = 0.24$

VP → V NP XP 0.15

t_1:

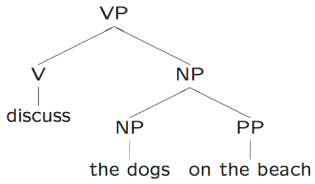


$p(t_1) = 0.15 \times 0.24 = 0.036$ (dispreferred)

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

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)

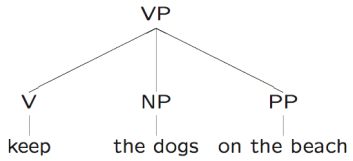
Frame Preferences

- (2) The women kept the dogs on the beach.
  - a. The women kept the dogs which were on the beach.
  - b. **The women discussed them (the dogs) while on the beach.**

$p(\text{keep}, \langle \text{NP XP[pred } +] \rangle) = 0.81$

VP → V NP XP 0.15

t_1:

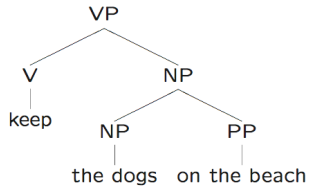


$p(t_1) = 0.15 \times 0.81 = 0.12$ (preferred)

$p(\text{keep}, \langle \text{NP} \rangle) = 0.19$

VP → V NP 0.39
NP → NP XP 0.14

t_2:



$p(t_2) = 0.19 \times 0.39 \times 0.14 = 0.01$ (dispreferred)

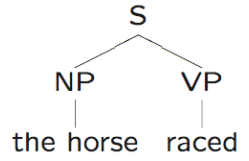
Modeling Garden Paths

- The reduced relative clause often cause irrecoverable difficulty, but nor always:
  - The horse raced past the barn fell (irrecoverable)
  - The bird found died (recoverable)
- We can use probabilities to distinguish the two cases, in a way a purely structural account (Frazier, or Pritchett) cannot.
- Assume a parallel parser …
  - The parse with the highest probability is preferred
  - Only those parsers which are within some "beam" of the preferred parse are kept, others are discarded

The horse raced past the barn fell

$p(\text{race}, \langle \text{NP} \rangle) = 0.92$

t_1:

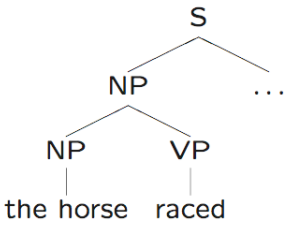


$p(t_1) = 0.92$ (preferred)

$p(\text{race}, \langle \text{NP NP} \rangle) = 0.08$

NP → NP XP 0.14

t_2:

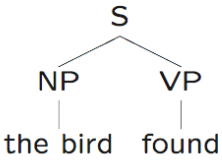


$p(t_1) = 0.0112$ (dispreferred)

The bird found died

$p(\text{find}, \langle \text{NP} \rangle) = 0.38$

t_1:

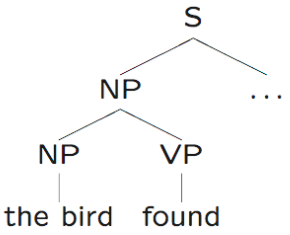


$p(t_1) = 0.38$ (preferred)

$p(\text{find}, \langle \text{NP NP} \rangle) = 0.62$

NP → NP XP 0.14

t_2:



$p(t_1) = 0.0868$ (dispreferred)

The Jurafsky Model

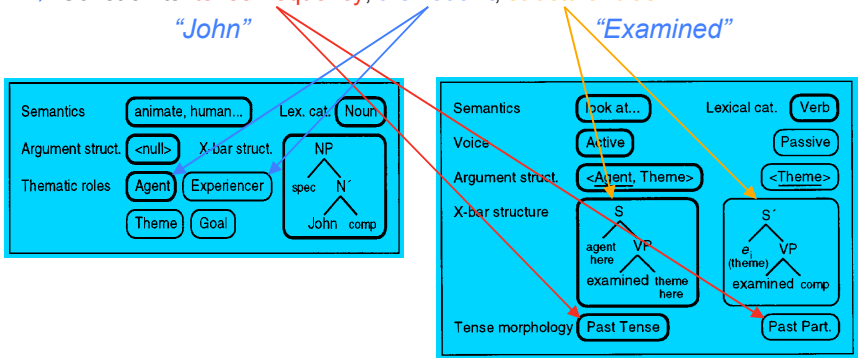
- Setting the beam width:
  - "The horse raced past the barn fell" 82:1
  - "The bird found died" 4:1
- Jurafsky assumes a garden path occurs (I.e. a parse is pruned) if its probability ratio with the best parse is greater than 5:1
- Open issues:
  - Where do we get the probabilities?
  - Does the model work for other languages?
  - How do we account for memory load phenomena?
  - Still purely syntactic (?): what about other constraints?

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 category ambiguity:
  - *that* = {Comp, Det, RelPro}
- Subcategorization ambiguity:
  - *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

- Rich syntactic/thematic features
- Frequency determines 'activations'
- Consider: *"John examined the evidence"*
  - "examined" is ambiguous, as either a simple past or past participle
  - Constraints: tense frequency, thematic fit, structural bias ...



MacDonald, Pearlmutter & Seidenberg

- The Interactive-Activation Model: In sum
  - Multiple access is possible at all levels of representation, constrained by frequency/context
  - All levels are of representation are available to the language processor, simultaneously
  - Highly lexicalist, entries enriched with frequency and syntactic info, "built" not accessed
  - Language processing is "constraint satisfaction", between lexical entries, and across levels
  - No distinct parser
- Questions:
  - Acquisition of the model: where does the linguistic knowledge come from, and the probabilities/activations?
  - Implementation: does such a model work in practice?
    - ✚ Complex interaction behaviours are difficult to predict

The Competitive-Integration Model

- Claim: Diverse constraints (linguistic and conceptual) are brought to bear simultaneously in ambiguity resolution.
  - Contra: modular models with distinct syntactic processing and delayed influence of conceptual constraints
- Problem: *"No model-independent signature data pattern can provide definitive evidence concerning when information is used"*
- The Model:
  - Not a parser: *assumes the competing analyses have been constructed*
  - Constraints provide "probabilistic" evidence to support alternatives
    - Each constraint has a weight, these are normalised to sum to 1
    - Lexical frequency bias, structure bias, parafoveal cues, thematic fit ...
  - Constraints activations, *C*, are integrated to activate each interpretation, *I*
  - I-activation is fed-back to the C-activation; then next cycle begins
- 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)

➡ Goal: investigate the time-course with which constraints contribute to the activation of competing analyses

1. Identifying the relevant constraints
2. Computational model for the interaction of constraints
3. Determining the bias of each constraint
   - From corpora: frequency used to determine probability
   - From off-line study: norms used to determine probability
4. Determining the weight of each constraint
   - Fit with off-line completions
5. Make predictions for reading times
6. Compare predicted reading times of:
   - Constraint-based model
   - Garden-path model
   - Short-delay garden path model
   - ... With the actual reading times from on-line studies

Constraints/Parameters of the Model

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

① 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

② Main clause bias: general bias for structure for "NP verb+ed …"
- Corpus estimate: P(RR|NP + verb-ed) = .08, P(MC|NP + verb-ed) = .92

③ *by*-Constraint: extent to which 'by' supports the passive construction
- Estimated for the 40 verbs from WSJ/Brown: RR= .8 MC= .2

④ Thematic fit: the plausibility of crook/cop as an agent or patient
- Estimated using a norming study

⑤ *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

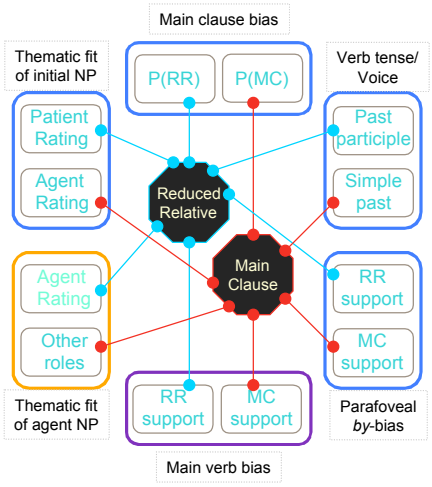
```
How common is it for a
    crook          _____
    cop            _____
    guard          _____
    police         _____
    suspect        _____
To arrest someone?
To be arrested by someone?
```

The results: Initial NP	Relative	Main
Good Agents (e.g. *the cop*):	1.5	5.3
Good Patients (e.g. *the crook*):	5.0	1.0
The results: Agent NP	Relative	Main
Good Agents (e.g. *the detective*):	4.6	1.0 (constant)

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



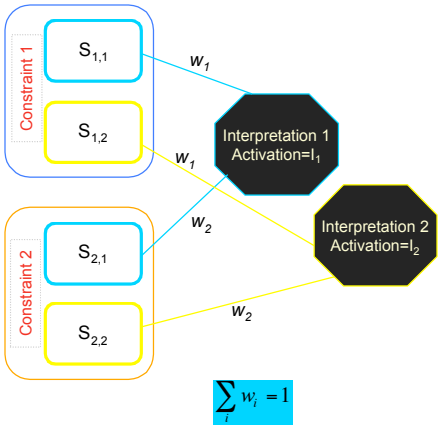
The recurrence mechanism

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

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



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...

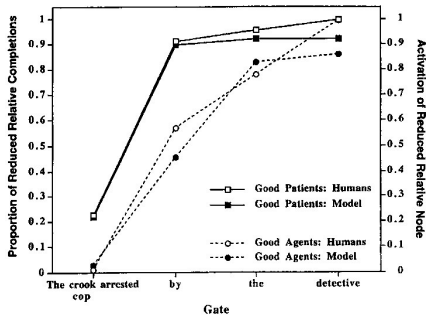


FIG. 2. Human and simulation results for fragment completions.

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

Self-Paced Reading Study

- Two-word, self-paced presentation:

  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
- Same beginning as the completion studies
- Three Models
  - Constraint-Based: constraints apply immediately for each region
  - Garden-Path: MC-bias & Main-Verb bias only, other constraints (lexical specific, and conceptual) are delayed one region
  - Short-Delay Garden Path
- 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

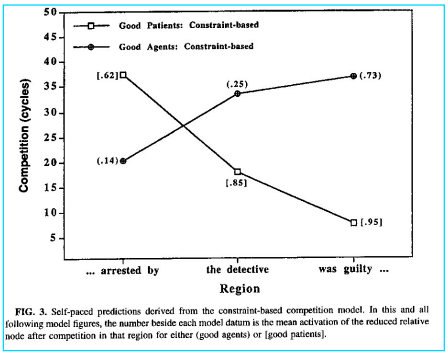


FIG. 3. Self-paced predictions derived from the constraint-based competition model. In this and all following model figures, the number beside each model datum is the mean activation of the reduced relative node after competition in that region for either (good agents) or [good patients].

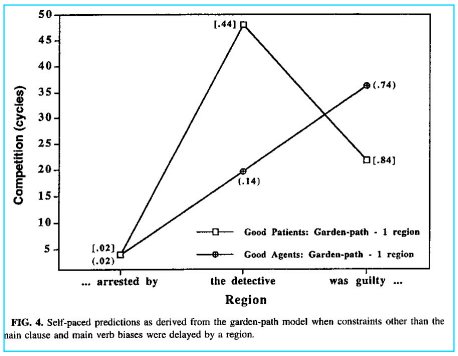


FIG. 4. Self-paced predictions as derived from the garden-path model when constraints other than the main clause and main verb biases were delayed by a region.

GP vs CB Modelling of the Reading

Reduction effect/cycles:

Human reading times:

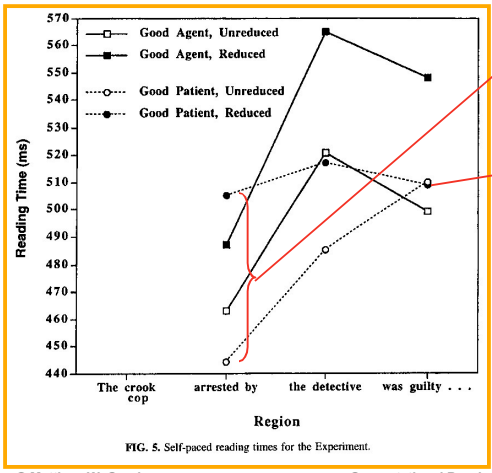


FIG. 5. Self-paced reading times for the Experiment.

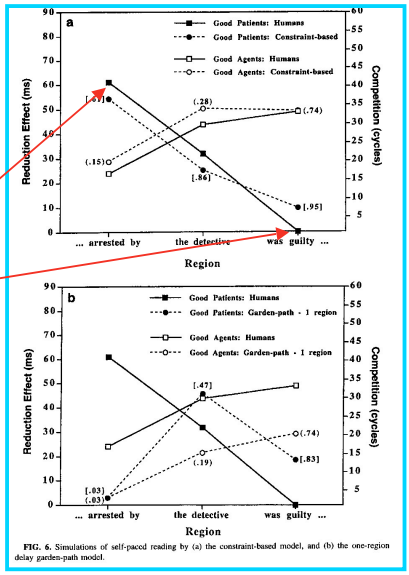


FIG. 6. Simulations of self-paced reading by (a) the constraint-based model, and (b) the one-region delay garden-path model.

Simulating a 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

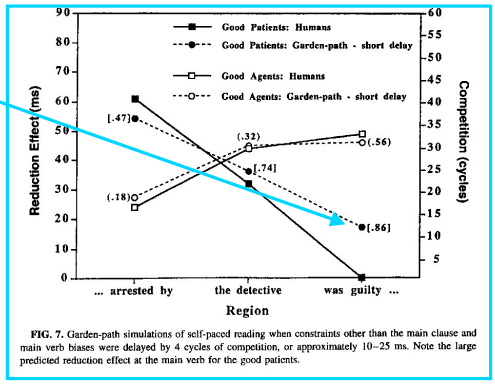


FIG. 7. Garden-path simulations of self-paced reading when constraints other than the main clause and main verb biases were delayed by 4 cycles of competition, or approximately 10–25 ms. Note the large predicted reduction effect at the main verb for the good patients.

- 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)

Ambiguities revisited: [preferred/dis-preferred]

- NP/VP Attachment Ambiguity:
  - "The cop [saw [the burglar] [with the binoculars]]"
  - "The cop saw [the burglar [with the gun]]"
- NP/S Complement Attachment Ambiguity:
  - "The athlete [realised [his goal]] last week"
  - "The athlete realised [[his shoes] were across the room]"
- Clause-boundary Ambiguity:
  - "Since Jay always [jogs [a mile]] the race doesn't seem very long"
  - "Since Jay always jogs [[a mile] doesn't seem very long]"
- Red. Relative-Main Clause Ambiguity:
  - "[The woman [delivered the junkmail on Thursdays]]"
  - "[[The woman [delivered the junkmail]] threw it away]"
- Relative/Complement Clause Ambiguity:
  - "The doctor [told [the woman [that he was in love with]] [to leave]]"
  - "The doctor [told [the woman] [that he was in love with her]]"

Issues and Criticisms

- Decision about what constraints to include/exclude, McRae et al:
  - Less important if materials don't vary w.r.t excluded constraint, or,
  - If bias of excluded constraint correlates well with included constraints:
    - E.g. tense bias (included) correlates well with transitivity (excluded)
- Not a model of language *processing*:
  - Is it legitimate to characterise information flow separate from the structure building mechanism.
  - What is *really* being modelled? Can the approach be scaled up?
- Garden-path: A straw man
  - Is the implementation of the GP model fair, for purposes of comparison
  - What other constraints might be considered purely syntactic.
- Predicts long reading times when constraints are in close competition
  - In fact, people are often *faster* at processing ambiguous regions!
- Not truly probabilistic: activations only *begin* as probabilities
  - Also, many probabilities are derived from ratings (not frequencies)