Language understanding and Bayesian inference

**Topic 2**: noisy channels, audience design, modularity & Bayesian inference

Daniel Lassiter
Stanford Linguistics
NASSLLI ‘14 UMD
course plan

1. probability and Bayesian inference
2. noisy channels, audience design, modularity & Bayesian inference
3. anchoring interpretation in world models
4. Bayesian pragmatics I
5. Bayesian pragmatics II
today’s plan

1. architectural issues
2. evidence for the noisy-channel model
3. audience design
4. (modularity & Bayesian inference)
levels of analysis

<table>
<thead>
<tr>
<th>Computational theory</th>
<th>Representation and algorithm</th>
<th>Hardware implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?</td>
<td>How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?</td>
<td>How can the representation and algorithm be realized physically?</td>
</tr>
</tbody>
</table>

*Figure 1–4. The three levels at which any machine carrying out an information-processing task must be understood.*

(Marr, 1982)
importance of computational level

Although algorithms and mechanisms are empirically more accessible, it is the top level, the level of computational theory, which is critically important from an information-processing point of view. The reason for this is that the nature of the computations that underlie perception depends more upon the computational problems that have to be solved than upon the particular hardware in which their solutions are implemented. To phrase the matter another way, an algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embodied.

In a similar vein, trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: It just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds’ wings make sense. More to the point, as we shall

(Marr, 1982)
rational analysis

The rational level of analysis … is not “psychologically real,” in the sense that it does not assert that any specific computation is occurring in the human head. Rather, it is an attempt to do an analysis of the criteria that these computations must achieve in order to assure the rationality of the system. … [I]t is about constraints on the behavior of that system in order for that behavior to be optimal. If we assume that cognition is optimized, these behavioral constraints are constraints on the mechanisms.

(Anderson 1991)
Anderson’s six steps

1. Specify the goals of the cognitive system.
2. Develop a formal model of the environment to which the system is adapted.
3. Make minimal assumptions about computational limitations.
4. Derive optimal behavior given (1-3).
5. Examine empirical literature: predictions confirmed?
6. If predictions are off, iterate.

(Anderson, 1991)
Application to language

1. Goal:
   - proximal: communication
   - ultimate: promoting non-linguistic ends
     (ultimately, survival & reproduction)

2. Model: linguistic architecture + Bayes

3. limitations: memory, epistemic access, …
Shannon (1948)

Fig. 1 — Schematic diagram of a general communication system.

• Model-based Bayesian before it was cool
• Bayesian because it’s intuitive and it works

“What message was encoded at the source, given the evidence received at the destination?”
Shannon (1948)

Fig. 1 — Schematic diagram of a general communication system.

“What message was encoded at the source, given the evidence received at the destination?”

\[ P(\text{message}|\text{signal}) = \frac{P(\text{signal}|\text{message}) \times P(\text{message})}{\sum_{m' \in \text{messages}} P(\text{signal}|m') \times P(m')} \]

inference by model inversion
Shannon (1948)

Fig. 1 — Schematic diagram of a general communication system.

“What message was encoded at the source, given the evidence received at the destination?”

\[ P(\text{message}|\text{signal}) \propto P(\text{signal}|\text{message}) \times P(\text{message}) \]

inference by model inversion
very noisy channel model
very noisy channel, simplified

Bayes net interpretation: causal connections are generally noisy.
adaptation for language

Noise possible, inference needed everywhere

Background model for much work in AI, psych
speaker-oriented model

(Zeevat, 2014)
listener-oriented model

(Frank & Goodman, 2012; Lassiter & Goodman, 2013; Goodman & Lassiter, 2014)
having it both ways?
recursive pragmatics
today’s plan

1. architectural issues
2. evidence for the noisy-channel model
3. audience design
4. modularity & Bayesian inference
psycholinguistic evidence

\[ P(\text{message}|\text{signal}) \propto P(\text{signal}|\text{message}) \times P(\text{message}) \]

- \( P(\text{message}) \): semantic prior
- \( P(\text{signal}|\text{message}) \): uncertainty about encoding (production) and transmission noise

- Do people trade off these components as model predicts?
Gibson et al. (2013)

\[ P(s_i | s_p) \propto P(s_p | s_i) \times P(s_i) \]

simple noise model:

\[ P(\text{insertion}) = P(\text{deletion}) = q \]

*but*: all deletions are the same (null)

probability of a **specific** insertion is very small
Inference

For any candidate reconstructions $s_1$ and $s_2$:

\[
\frac{P(s_1 | s_p)}{P(s_2 | s_p)} = \frac{P(s_p | s_1)}{P(s_p | s_2)} \times \frac{P(s_1)}{P(s_2)}
\]

- relative strength
- noise model
- relative plausibility
Gibson et al. (2013)

Experiment:

– implausible vs. plausible sentences, e.g.
  The woman gave the candle (to) the girl
– Ps answer comprehension questions
– manipulate noise rate, expectations about plausibility of sentences
<table>
<thead>
<tr>
<th>English constructions</th>
<th>Plausible version</th>
<th>Change</th>
<th>Implausible version</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Active/passive</td>
<td>a. The girl kicked the ball. (active)</td>
<td>Two insertions</td>
<td>c. The girl was kicked by the ball. (passive)</td>
</tr>
<tr>
<td></td>
<td>b. The ball was kicked by the girl. (passive)</td>
<td>Two deletions</td>
<td>d. The ball kicked the girl. (active)</td>
</tr>
<tr>
<td>2. Subject-locative/</td>
<td>a. Onto the table jumped a cat. (subject-locative)</td>
<td>One deletion,</td>
<td>c. The table jumped onto a cat. (object-locative)</td>
</tr>
<tr>
<td>object-locative</td>
<td>b. The cat jumped onto a table. (object-locative)</td>
<td>one insertion</td>
<td>d. Onto the cat jumped a table. (subject-locative)</td>
</tr>
<tr>
<td>3. Transitive/intransitive</td>
<td>a. The tax law benefited the businessman. (transitive)</td>
<td>One insertion</td>
<td>c. The tax law benefited from the businessman. (intransitive)</td>
</tr>
<tr>
<td></td>
<td>b. The businessman benefited from the tax law. (intransitive)</td>
<td>One deletion</td>
<td>d. The businessman benefited the tax law. (transitive)</td>
</tr>
<tr>
<td>4. DO/PO goal</td>
<td>a. The mother gave the daughter the candle. (DO-goal)</td>
<td>One insertion</td>
<td>c. The mother gave the daughter to the candle. (PO-goal)</td>
</tr>
<tr>
<td></td>
<td>b. The mother gave the candle to the daughter. (PO-goal)</td>
<td>One deletion</td>
<td>d. The mother gave the candle the daughter. (DO-goal)</td>
</tr>
<tr>
<td>5. DO/PO benefactive</td>
<td>a. The cook baked Lucy a cake. (DO-benef)</td>
<td>One insertion</td>
<td>c. The cook baked Lucy for a cake. (PO-benef)</td>
</tr>
<tr>
<td></td>
<td>b. The cook baked a cake for Lucy. (PO-benef)</td>
<td>One deletion</td>
<td>d. The cook baked a cake Lucy. (DO-benef)</td>
</tr>
</tbody>
</table>

The five alternations that are investigated in this paper are as follows: 1, active/passive; 2, subject-locative/object-locative; 3, transitive/intransitive; 4, double-object/prepositional phrase object goals; and 5, double-object/prepositional phrase object benefactives. The number of insertions and deletions that are needed to form an implausible alternation from the plausible version is provided for each plausible/implausible pair, as a proposed hypothesis for how the implausible versions might be generated. benef, benefactive; DO, double object; PO, prepositional phrase object.
noise model: 1 deletion

\( s_p: \) “The businessman benefited the tax law”

- \( s_1 = s_p \)
- \( s_2 = \) “The businessman benefited \textbf{from} the tax law”

- Insertion/deletion rate = \( q = .1 \)

\[ \text{L’hood : } \frac{P(s_p|s_1)}{P(s_p|s_2)} = \frac{1 - q}{q} = \frac{.9}{.1} = 9 \]

- prefer \( s_2 \) if > 9x as semantically plausible
noise model: 1 insertion

$s_p$: “The tax law benefited from the businessman”

- $s_1 = s_p$
- $s_2 = “The tax law benefited the businessman”$

$q=.1$, but the **specific** insertion must be chosen

$$\frac{P(s_p \mid s_1)}{P(s_p \mid s_2)} = \frac{(1 - .1)/|\text{ALT}|}{.1} = 9 \times |\text{ALT}|$$

$s_2$ has to be way more plausible (size principle)
qualitative predictions

1. preference for plausible meanings depends on # of edits required
2. deletions preferred over insertions
3. increasing noise rate increases preference for plausible meanings
4. greater rate of anomalous sentences increases preference for literal readings
## results

<table>
<thead>
<tr>
<th>Experiment 1.1-1.5</th>
<th>Passive / Active (1c) / (1d)</th>
<th>Obj-Loc / Subj-Loc (2c) / (2d)</th>
<th>Intrans / Trans (3c) / (3d)</th>
<th>PO-goal / DO-goal (4c) / (4d)</th>
<th>PO-ben / DO-ben (5c) / (5d)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No syntactic errors or implausible materials in the fillers</strong></td>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
<td><img src="image5.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment 2.1-2.5</th>
<th>Passive / Active (1c) / (1d)</th>
<th>Obj-Loc / Subj-Loc (2c) / (2d)</th>
<th>Intrans / Trans (3c) / (3d)</th>
<th>PO-goal / DO-goal (4c) / (4d)</th>
<th>PO-ben / DO-ben (5c) / (5d)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High proportion of syntactic errors in the fillers</strong></td>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
<td><img src="image5.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment 3.1-3.5</th>
<th>Passive / Active (1c) / (1d)</th>
<th>Obj-Loc / Subj-Loc (2c) / (2d)</th>
<th>Intrans / Trans (3c) / (3d)</th>
<th>PO-goal / DO-goal (4c) / (4d)</th>
<th>PO-ben / DO-ben (5c) / (5d)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High proportion of implausible materials</strong></td>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
<td><img src="image5.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

1. # edits matters
2. deletion pref.
3. noise rate matters
4. plausibility manip. matters

Fig. 2. Percentage of trials in which participants relied on the literal syntax for the interpretation of the implausible syntactic constructions. Error bars are 95% confidence intervals. Examples of each implausible construction and its closest edit to a plausible construction are given in Table 1, e.g., an example of
discussion

comprehenders

– trade off semantic priors, prob. of corruption as predicted by noisy channel model
– actively model noise, adjust to noise rate
– actively model and adjust to speaker reasonability
unpacking ‘semantic prior’

encoding preferences: \[ P(s_i) = \sum_m P(s_i|m) \times P(m) \]

message preferences: \[ P(m) = \sum_{b,d} P(m|b,d) \times P(b) \times P(d) \]
encoding preferences

\[ P(m|s_i) \propto P(s_i|m) \times P(m) \]

- inferred message
- encoding preferences
- message preferences

\( P(s_i|m) \) describes production system

many possible constraints, e.g.:

- grammatical constraints
- memory-induced preferences (\( \Rightarrow \) frequency effects?)
- ceteris paribus pref. for shorter encoding
message preferences

\[ P(m) = \sum_{b,d} P(m|b,d) \times P(b) \times P(d) \]

• source of ‘decision-theoretic’ character
• do L’s maintain rich models of speaker beliefs and desires, or relatively generic ones?
• ‘successful communication’ is a likely desire.
  – Do speakers tailor messages to an audience?
  – Do listeners believe that they do?
today’s plan

1. architectural issues
2. evidence for the noisy-channel model
3. audience design
4. modularity & Bayesian inference
listeners model speakers

Grice: Ls actively model S beliefs, desires and use them to infer intended messages

– inference: S uttered \( u \) instead of something else

– prediction built into Bayesian noisy channel model
Speakers help listeners along? Clark & Murphy ‘82:

- ‘When people talk, they tailor what they say to their audience.’
- ‘They expect their audience to use that fact in figuring out what they meant.’
- ‘Listeners … can reason through to the speaker’s meaning accordingly.’
- ‘People have to reason from the design assumption even to get syntax and word meanings right.’
architecture for AD, S-side
architecture for AD, L-side

BELIEF RECOGNITION

other inferences

SPEAKER

ACOUSTIC SIGNAL PRODUCED

ACOUSTIC SIGNAL RECEIVED

GRAMMATICAL DECODING

INTENTION RECOGNITION

DESIRE RECOGNITION

other stuff
definite descriptions & demonstratives:
  – choice of the, this, that, a depends on S beliefs about mutual identifiability – S~L mutual knowledge
  – Ls interpret descriptions on the assumption that Ss are making reasoning choices

Clark et al. 1982, w/ photo of Reagan & Stockman:
  – “You know who this man is, don’t you?”
  – “Do you have any idea at all who this man is?”
AD example: accommodation

Mary took some picnic supplies out of the basket. The beer was warm.

“The listener … assumes that there ought to be some beer there if the speaker designed the utterance so that he could find such a referent uniquely.”

Clark & Murphy, 1982
AD example: bridging

The front of the car crashed into the wall.
The window was destroyed.

• S would only use the window if she thought L could work out which window was intended
• L infers it’s the window that she and S could most easily coordinate on: the one attached to the car mentioned
• coordination of beliefs is crucial (cf. GT pragmatics)
AD example: anaphora

Bill has a new car.
Mary gave it to him.

Rough AD account:

– S would only use *it, him* if she thought L could work out the referents
– L knows this, uses as basis to infer …

[Lots more to be said here! Day 5]
“Hand me the cake mix” with 2 boxes on table
  – one near speaker
  – one near listener
• Ls infer that the cake mix nearer to them is intended: otherwise S would pick it up himself
• Ls do not make this inference if S’s hands are full: instead they display confusion about referent
Error and confusion arise and fidelity decreases, when, no matter how good the coding, one tries to crowd too much over a channel … a general theory at all levels will surely have to take into account not only the capacity of the channel but also … the capacity of the audience.

(Weaver, 1948)
quantifying information

informativity (surprisal) of an observation:

\[ I(q) = \log_2 \frac{1}{P(q)} \]

\[ = -\log_2 P(q) \]

low prior probability => high informativity

(Shannon, 1948)
AD and information density

• Goal: maximize rate of information exchange
• avoid overloading receiver with too-high rate
  1. Ss should strive for constant info rate
  2. close to the max channel/receiver capacity
• How to measure channel/receiver capacity??
• We can at least get at (1) …
Uniform Information Density

Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density). Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density (ceteris paribus).

Jaeger 2010; cf. Levy & Jaeger 2006
In-class experiment

Which sounds more natural, (1) or (2)?

1. My boss confirmed we are absolutely crazy

2. My boss confirmed that we are absolutely crazy
In-class experiment

Which sounds more natural, (1) or (2)?

1. My boss thinks we are absolutely crazy

2. My boss thinks that we are absolutely crazy
UID predictions

optional *that*

• I know *(that)* he is happy.
• I saw the man *(that)* you mentioned.

UID predicts:

– unpredictable (highly informative) contexts prefer *that* to keep transmission rate within capacity
– omission more likely w/ predictable continuations

Related: *uh/um*, optional *to*, phonetic reduction, …
Estimating informativity

CC = ‘complement clause’

\[ I(\text{CC}|\text{verb}) = -\log_2 P(\text{CC}|\text{verb}) \]

\[ \approx -\log_2 \left( \frac{\text{Freq(verb + CC)}}{\text{Freq(verb)}} \right) \]
My boss confirmed that we were absolutely crazy.

My boss thinks that I am absolutely crazy.
Jaeger’s corpus experiment

- Parsed spoken corpus (from Switchboard)
- 642 conversations, 7369 CCs
- *that*-biases ranging from
  - 1% (*guess*), to
  - 75% (*worry*)
- Linear mixed-effects model including numerous previously proposed factors
  (dependency length, production difficulty, …)
results

- Information density is
  - a highly significant predictor (p < .0001)
  - strongest predictor by far
- ‘speakers are less likely to produce … that, the lower the information density of the CC onset. This effect holds even while other variables that previous studies found to be correlated … are controlled for.’
results
Lexical structure

• Piantidosi et al. 2011: information content highly predictive of word lengths
  – ‘lexicon optimized for efficient communication’
  – but this is likely a diachronic effect
Do Ss reduce contextually uninformative elements when grammar & lexicon allow?

1. **supportive**: Jim was bad at algebra so he hated (math / mathematics).

2. **neutral**: Jim introduced himself as someone who hated (math / mathematics).

UID: predictability => pref. for shorter form
Lexical choice

neutral contexts:
56% short form

supportive contexts:
67% short form

difference: $p < .001$
interpretation

• speakers’ choices track expected **predictability** by **listeners** in **context**
• supports UID as audience design principle
  – predictable elements tend to be reduced
  – increase what would be low transmission rate
• extends previous work with similar conclusions for phonetic reduction
  (e.g., Aylett & Turk ’04)
problems with AD?

Ss take L into account at least sometimes when formulating utterances. But still ...

How do people use language to convey and understand intentions? We started investigating this question with the accepted assumption that, unlike young children, adult language users are not egocentric. Instead, we assumed that adults rely on a “model of the other person’s mind” when they use language. To our surprise, our experiments provided evidence against this fundamental assumption. In this article, we explain why we have come to the conclusion that adults routinely process language egocentrically, adjusting to the other’s perspective only when they make an error.

Keysar, Barr, & Horton, 1998
Fig. 1. The array of objects from the director’s and the matcher’s perspectives. The critical instructions were, “Put the bottom block below the apple.” The director was referring to the second block from the top, but the question of interest was whether the matcher would consider the lower block as the referent even though that block was occluded from the director’s perspective (Keysar, Barr, Balin, & Brauner, 1996).
possible explanation

optimality vs. resource-bounded optimality
(Simon ‘82, Anderson ‘91)

Keysar et al.:
• use own epistemic state as proxy for other’s, monitor for major divergence
• cheap approximation of audience design

‘optimal’ given goal of conserving resources?
another account

Krauss & Fussell ’91:
– much evidence Ss take L beliefs into account
– but usu. assume L beliefs fairly similar to own

different ‘bounded rational’ account:
– Ss search for a reasonable guess at L beliefs
– but tend to use own beliefs as a starting point
– extended review: Nickerson ’99

cf. ‘rationality of anchoring’ (Lieder et al. ‘13)
today’s plan

1. architectural issues
2. evidence for the noisy-channel model
3. audience design
4. modularity & Bayesian inference
modularity

• design principle for good programming
• ‘a program is modular with respect to a change in a given part if the change can be done without changing the rest of the program’ (e.g., no state) (Van Roy & Haridi, 2004)
• in cogsci: ‘any large computation should be split up into a collection of small, nearly independent, specialized subprocesses’ (Marr)
Fodorian modularity

- ‘informationally encapsulated … access to background information is constrained by general features of cognitive architecture’

- non-modularity entails ‘anything that the organism knows … is ipso facto available as a premise in perceptual inference’
Fodorian modularity

‘local, stupid, and extremely nervous’

Inherent variability has been effectively filtered. What perceptual systems typically “know about” is how to infer current distal layouts from current proximal stimulations: the visual system, for example, knows how to derive distal form from proximal displacement, and the language system knows how to infer the speaker’s communicative intentions from his phonetic productions. Neither mechanism, on the present account, knows a great deal else, and that is entirely typical of perceptual organization. Perceptual systems have to be simple.

(Fodor, 1985)
reflexes

Suppose that you and I have known each other for many a long year . . . and you have come fully to appreciate the excellence of my character. In particular, you have come to know perfectly well that under no conceivable circumstances would I stick my finger in your eye. Suppose that this belief of yours is both explicit and deeply felt. You would, in fact, go to the wall for it. Still, if I jab my finger near enough to your eyes, and fast enough, you'll blink. . . . [The blink reflex] has no access to what you know about my character or, for that matter, to any other of your beliefs, utilities [or] expectations. For this reason the blink reflex is often produced when sober reflection would show it to be uncalled for. . . . (p. 71)
Müller-Lyer
visual illusions

The apparent difference in length of the Mueller–Lyer figures, for example, doesn’t disappear when one learns that the arrows are in fact the same size. It seems to follow that at least some perceptual processes are insensitive to at least some of one’s beliefs. Very much wanting the Mueller–Lyer illusion to go away doesn’t make it disappear either; it seems to follow that at least some perceptual processes are insensitive to at least some of one’s utilities. The ecological good sense of this arrangement is surely self-evident. Prejudiced and wishful seeing makes for dead animals.
is language modular?

A parser for [a language] L contains a grammar of L. … [i]t infers from certain acoustic properties of a token to a characterization of certain of the distal causes of the token (e.g., to the speaker's intention that the utterance should be a token of a certain linguistic type). Premises of this inference can include … information about the acoustics of the token … information about the linguistic types in L … and nothing else.

(Fodor, 1984)
is language modular?

Lg. understanding is much smarter than vision

- **parsing**: plausibility, visual context affects on-line and off-line measures (Crain & Steedman ‘85; Altmann & Steedman ’88, Tanenbaum et al. ’95; many others)

- **lexical access**: recognition, polysemy affected by plausibility, even identity of speaker …

- **semantics/pragmatics**: ambiguity resolution, anaphora, scalar implicature, …
is language architecture modular?

• evidence against **any** kind of modularity?
  – cf. PDP, constraint-based, embodied, …

• no threat to modularity as design principle
  – modules as functions, no ref. to global variables

• orthogonal to question of innateness

• motivated by need for piecemeal evolution?
  – cf. modular programming for debugging, tinkering

(cf. Carruthers, 2006)
modular but interactive

- information flows along isolated paths
- modules communicate because interpretation is a complex joint inference
- ‘language is modular; inference is not’ (N.G.)
what’s next

• build some simple modules & a fragment
  – a simple world model
  – a small lexicon interpreted in the world model
  – principles of syntactic, semantic composition
  – pragmatics of literal interpretation
• explore predictions about ambiguity resolution, type-shifting, etc.