Computational Psycholinguistics
Lecture 5: uncertain input and uniform information density

Roger Levy
University of California – San Diego

ESSLLI
24 July 2009
Incrementality and Rationality

- Online sentence comprehension is hard
- But lots of information sources can be usefully brought to bear to help with the task
- Therefore, it would be *rational* for people to use *all the information available*, whenever possible
- This is what *incrementality* is
- We have lots of evidence that people do this often

“Put the apple on the towel in the box.”  (Tanenhaus et al., 1995)
Anatomy of _ye olde garden path sentence_

_The horse raced past the barn fell._

(The horse _that was_ raced past the barn fell.)

(The criminal _that was_ arrested by the detective confessed.)

(The criminal arrested by the detective confessed.)

- It’s weird
- People fail to understand it most of the time
- People are more likely to _misunderstand_ it than to understand it properly
  - “What’s a _barn fell_?”
  - The horse _that_ raced past the barn fell
  - The horse raced past the barn _and_ fell
- Today I’m going to talk about three outstanding puzzles involving garden-path sentences
Garden paths: What we do understand

• We have decent models of how this sentence is not understood
  • Incremental probabilistic parsing with beam search (Jurafsky, 1996)
  • Surprisal (Hale, 2001; Levy, 2008):
    • disambiguating word *fell* is extremely unexpected
    • ➔ alarm signal: “this shouldn’t be happening”
    • parser quits
  
• These models are based on *rational use of evidential information* (data-driven probabilistic inference)
  • Also compatible with gradated garden-path difficulty (Garnsey et al., 1997; McRae et al., 1998)

*The horse raced past the barn fell > The criminal arrested by the detective confessed*
Surprisal and garden paths: theory

- Revisiting *the horse raced past the barn fell*
- After *the horse raced past the barn*, assume 2 parses:

  ![Tree diagrams showing two parses of the sentence.]

- Jurafsky 1996 estimated the probability ratio of these parses as 82:1
- The surprisal differential of *fell* in reduced versus unreduced conditions should thus be \(\log_2 83 = 6.4\) bits

*(assuming independence between RC reduction and main verb)*
Surprisal and garden paths: practice

- An unlexicalized PCFG (from parsed Brown corpus of English) gets right monotonicity of surprisals at disambiguating word “fell”

Aside: These are way too high, but that’s because the grammar’s crude

this is the key comparison; the difference is small, but in the right direction
Garden Paths: What we don’t understand so well

- How do people arrive at the misinterpretations they come up with?
- What factors induce them to be more or less likely to come up with such a misinterpretation
Puzzle 1: global inference

• Try to read & comprehend this sentence:
• Question: if that sentence is true, must it follow…
  • that the man hunted the deer???
  • that the deer ran into the woods??
• This was the sentence (Christianson et al., 2001):
  While the man hunted the deer ran into the woods
  This comma wasn’t there!
• Readers answer yes to both questions!!!
• They’re garden-pathed at first…
• …and then recover wrongly into some hybrid meaning
• Major problem for rational sentence processing theories: *inferences incompatible with the complete sentence*
Puzzle 1: global inference

• Try to read & comprehend this sentence:
  *While the man hunted the deer ran into the woods.*

• And now let’s do a little math:
  8095 – 5107 + 4043 = 7031

• Question: if the sentence was true, must it follow…
  • …that *the man hunted the deer***?
  • …that *the deer ran into the woods***?

• Readers answer yes to both questions!!!

• They’re garden-pathed at first…

• …and then recover wrongly into some hybrid meaning

• *Major problem for rational sentence processing theories: inferences incompatible with the complete sentence* (Christianson et al., 2001)
Puzzle 2: incremental inference

• Try to understand this sentence:

(a) The coach smiled at the player tossed the frisbee.

...and contrast this with:

(b) The coach smiled at the player thrown the frisbee.
(c) The coach smiled at the player who was thrown the frisbee.
(d) The coach smiled at the player who was tossed the frisbee.

• Readers boggle at “tossed” in (a), but not in (b-d)

(Tabor et al., 2004)
Why is *tossed/thrown* interesting?

- As with classic garden-paths, part-of-speech ambiguity leads to misinterpretation
  - *The horse* raced past the barn…fell
  - *verb?*  
  - *participle?*

- But now context “should” rule out the garden path:
  - *The coach smiled at the player* tossed…
  - *verb?*  
  - *participle?*

- Another challenge for rational models: **failure to condition on relevant context**
Uncertain input in language comprehension

• State of the art models for ambiguity resolution ≈ probabilistic incremental parsing
• Simplifying assumption:
  • Input is *clean* and *perfectly-formed*
  • No uncertainty about input is admitted
• Intuitively seems patently wrong…
  • We sometimes *misread* things
  • We can also *proofread*
• Leads to two questions:
  1. What might a model of sentence comprehension under uncertain input look like?
  2. What interesting consequences might such a model have?
Today: a first-cut answer

1. **What might a model of sentence comprehension under uncertain input look like?**
2. **What interesting consequences might such a model have?**

- **First**: a simple noisy-channel model of rational sentence comprehension under uncertain input
- **Then**: we’ll solve the two psycholinguistic puzzles
  1. *global inference*
  2. *incremental inference*
- We use probabilistic context-free grammars (PCFGs) and weighted finite-state automata (WFSAs) to instantiate the model
- In each case, input uncertainty solves the puzzle
The noisy-channel model

- Say we use a weighted generative grammar $G$ to parse a sentence $w$. We get a posterior over structures $T$:

$$P_G(T|w) = \frac{P(T, w)}{P(w)} \propto P(T, w)$$

- If we don’t observe a sentence but only a noisy input $I$:

$$P_G(T|I) \propto \sum_w P(I|T, w)P(w|T)P(T)$$

- Posterior over possible sentences:

$$P_G(w|I) \propto \sum_T P(I|T, w)P(w|T)P(T)$$

Levy, 2008 (EMNLP)
The noisy-channel model (II)

- This much is familiar from the parsing of speech (Hall & Johnson, 2003, 2004; Johnson & Charniak, 2004)

- Alternative scenario: we know the true sentence \( w^* \) but not observed input \( I \) (e.g., the study of reading)

- *Expected inferences of the comprehender* marginalize over the input \( I \):

\[
P(w|w^*) = \int_I P_C(w|I, w^*) P_T(I|w^*) dI
\]

\[\propto Q(w, w^*)\]
The noisy-channel model (II)

- This much is familiar from the parsing of speech (Hall & Johnson, 2003, 2004; Johnson & Charniak, 2004)

- Alternative scenario: we know the true sentence $w^*$ but not observed input $I$ (e.g., the study of reading)

- *Expected inferences of the comprehender* marginalize over the input $I$:

$$P(w|w^*) = \int_I P_C(w|I, w^*) P_T(I|w^*) dI$$

$$= P_C(w) \int_I \frac{P_C(I|w) P_T(I|w^*)}{P_C(I)} dI$$

For a rational comprehender, these are the same! So this is a kernel function!
Representing noisy input

- How can we represent the type of noisy input generated by a word sequence?
- *Probabilistic finite-state automata* (pFSAs; Mohri, 1997) are a good model

\[
\text{vocab} = a, b, c, d, e, f
\]

**Input symbol**

**Log-probability (surprisal)**

- "Word 1 is a or b, and I have no info about Word 2"
A generative probabilistic grammar determines beliefs about which strings are likely to be seen.

- Probabilistic Context-Free Grammars (PCFGs; Booth, 1969)
- Probabilistic Minimalist Grammars (Hale, 2006)
- Probabilistic Finite-State Grammars (Mohri, 1997; Crocker & Brants 2000)

In position 1, \{a,b,c,d\} equally likely; but in position 2:
- \{a,b\} are usually followed by e, occasionally by f
- \{c,d\} are usually followed by f, occasionally by e
Combining grammar & uncertain input

- Bayes’ Rule says that the evidence and the prior should be combined (multiplied)
- For probabilistic grammars, this combination is the formal operation of intersection (see also Hale, 2006)
Revising beliefs about the past

- When we’re uncertain about the future, grammar + partial input can affect beliefs about what will happen.
- With uncertainty of the past, grammar + future input can affect beliefs about what has already happened.
word 1
{b,c} {?}

words 1 + 2
{b,c} {f,e}
Flexibility of pFSAs

- Probabilistic FSAs can also allow us to represent inputs of *variable length*
- **ε-transitions** allow for the possibility of generating fewer than two input symbols
- **Loops** allow for the possibility of generating more than two input symbols*

- This pFSA gives probability to infinitely many strings, but the most likely are \{a,b\}{a,b}
The noisy-channel model (FINAL)

\[ P(w|w^*) \propto P_C(w)Q(w, w^*) \]

- For \( Q(w, w^*) \): a WFSA based on Levenshtein distance between words (\( K_{LD} \)):

Cost(a cat sat) = 0

Cost(sat a sat cat) = 8

Result of \( K_{LD} \) applied to \( w^* = \text{a cat sat} \)
Puzzle 1 recap: global inference

While the man hunted the deer ran into the woods.

- Readers tend to answer “yes” to both:
  1. Did the man hunt the deer?
  2. Did the deer run into the woods?
What does our uncertain-input theory say?

• In near-neighbor sentences the man does hunt the deer:

  (a) While the man hunted the deer it ran into the woods.

  (b) While the man hunted it the deer ran into the woods.

• (a-b) are “near” $w^*$ in Levenshtein-distance space

• Our theory may then explain this result…

• …if the comprehender’s grammar can push them into inferring structures more like (a-b)
Testing the intuition of our theory

\[ P(w | w^*) \propto P_C(w) Q(w, w^*) \]

- The Levenshtein-distance kernel \( K_{LD} \) gives us \( Q(w, w^*) \).
- A small PCFG can give us \( P_C(w) \).
- Recall that \( K_{LD} \) is a WFSA
  - So \( P(w | w^*) \) is a \textit{weighted intersection} of \( K_{LD} \) with \( P_C \).
- \textbf{Metric of interest}: % of 100-best parses (Huang & Chiang, 2005) in which “the man really does hunt the deer”:
  - While the man hunted the deer [pronoun] ran into the woods.
  - While the man hunted [pronoun] the deer ran into the woods.
Testing the intuition: results

- **GardenPath**: *While the man hunted the deer ran into the woods*
- **Comma**: *While the man hunted, the deer ran into the woods*
- **Transitive**: *While the man hunted the pheasant the deer ran into the woods*

% Misleading parses

![Graph showing misleading parses vs. noise level]

Model & human misinterpretations match
Relevance to contemporary theories

- As noted by Tabor et al., these results are problematic for theories requiring *global contextual consistency*
  - Garden-path theory (Frazier 1987)
  - “Unrestricted race” theories (Traxler et al., 1998)
  - Fully incremental probabilistic theories (Jurafsky 1996; Hale 2001, 2006; Levy 2008)

- [what do we need instead???]
Contextual constraint & rationality

• Let’s recast the problem in even more general terms

• *Rational* models of comprehension: the comprehender uses *all the information currently available*

• In local-coherence sentences, the comprehender seems to be systematically *ignoring* available information

• Local-coherence effects’ challenge: *to what extent is human sentence comprehension rational?*
Existing proposed theories

- Proposed models posit a context-ignoring, *bottom-up* component of comprehension:
  - Gibson, 2006
    \[ P(cat_i|w_i, \text{context}) \propto P(cat_i|\text{context})P(cat_i|w_i) \]
  - Tabor & Hutchins, 2004; Tabor, 2006
  - Hale, 2007

- To the extent that these models are rational, it can only be in terms of “bounded rationality” (Simon 1957)

- *To what extent do we want to bound the rationality of human sentence comprehension?"*
Today’s proposal

• I simply want to argue that it is premature to conclude from local-coherence effects that the parser’s rationality must be bounded in this respect
• There is another possibility that has been overlooked thus far
• Instead of relaxing the assumption about rational use of context, we may instead have misspecified the input representation
Puzzle 2 recap: Incremental inference

- Near-neighbors make the “incorrect” analysis “correct”:
  - Hypothesis: the boggle at “tossed” involves what the comprehender wonders whether she might have seen

Any of these changes makes **tossed** a main **verb**!!!
The core of the intuition

• Grammar & input come together to determine two possible “paths” through the partial sentence:

  • \textit{tossed} is more likely to happen along the bottom path
  • This creates a large shift in belief in the \textit{tossed} condition
  • \textit{thrown} is very unlikely to happen along the bottom path
  • As a result, there is no corresponding shift in belief
Incremental inference under uncertain input

- Traditionally, the input to a sentence-processing model has been a *sequence of words*
- But really, input to sentence processor should be more like the output of a word-recognition system

The coach smiled at the player tossed the frisbee

- That means that the possibility of *misreading/mishearing* words must be accounted for
- On this hypothesis, local-coherence effects are about *what the comprehender wonders whether she might have seen*

these changes would make *main-verb tossed globally coherent!!!*
Inference through a noisy channel

- So how can we model sentence comprehension when the input is still noisy?
- A generative probabilistic grammatical model makes inference over uncertain input possible
  - This is the noisy channel from NLP/speech recognition
  - Inference involves Bayes’ Rule

\[
P(\text{words}|\text{input, grammar}) \propto P(\text{input}|\text{words, grammar})P(\text{words}|\text{grammar})
\propto P(\text{input}|\text{words})P(\text{words}|\text{grammar}) \quad [\text{by assumption}]
\]

**Prior:** Comprehender’s knowledge of language

**Evidence:** Noisy input probability, dependent only on the “words” generating the input
Back to local-coherence effects

- How does this relate to local-coherence effects?
- Here’s an oversimplified noisy-input representation of the offending sentence

*The coach smiled at the player tossed the frisbee.*
Here's a hand-written finite-state grammar of reduced relative clauses
Ingredients for the model

\[ P(w | w^*) \propto P_C(w)Q(w, w^*) \]

- \( Q(w, w^*) \) comes from \( K_{LD} \) (with minor changes)
- \( P_C(w) \) comes from a probabilistic grammar (this time finite-state)
- We need one more ingredient:
  - a quantified signal of the alarm induced by word \( w_i \) about changes in beliefs about the past
1. Probabilistic Grammatical Knowledge

- We can transform a (strongly regular) PCFG into a weighted FSA
- We use the following grammar with surprisal values estimated from the parsed Brown corpus

<table>
<thead>
<tr>
<th>Surprisal</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.30</td>
<td>$S \rightarrow S$-base Conj $S$-base</td>
</tr>
<tr>
<td>0.01</td>
<td>$S \rightarrow S$-base</td>
</tr>
<tr>
<td>0.00</td>
<td>$S$-base $\rightarrow$ NP-base VP</td>
</tr>
<tr>
<td>3.71</td>
<td>$NP \rightarrow NP$-base RC</td>
</tr>
<tr>
<td>0.11</td>
<td>$NP \rightarrow NP$-base</td>
</tr>
<tr>
<td>0.00</td>
<td>NP-base $\rightarrow$ Det N</td>
</tr>
<tr>
<td>2.02</td>
<td>VP $\rightarrow$ V PP</td>
</tr>
<tr>
<td>0.69</td>
<td>VP $\rightarrow$ V NP</td>
</tr>
<tr>
<td>2.90</td>
<td>VP $\rightarrow$ V</td>
</tr>
<tr>
<td>1.32</td>
<td>S/NP $\rightarrow$ NP-base VP/NP</td>
</tr>
<tr>
<td>2.04</td>
<td>RC $\rightarrow$ VP-pass/NP</td>
</tr>
<tr>
<td>4.90</td>
<td>RC $\rightarrow$ WP FinCop VP-pass/NP</td>
</tr>
<tr>
<td>0.47</td>
<td>RC $\rightarrow$ WP S/NP</td>
</tr>
<tr>
<td>0.74</td>
<td>S/NP $\rightarrow$ VP</td>
</tr>
<tr>
<td>3.95</td>
<td>VP/NP $\rightarrow$ V NP</td>
</tr>
<tr>
<td>0.10</td>
<td>VP/NP $\rightarrow$ V</td>
</tr>
<tr>
<td>2.18</td>
<td>VP-pass/NP $\rightarrow$ VBN NP</td>
</tr>
<tr>
<td>0.36</td>
<td>VP-pass/NP $\rightarrow$ VBN</td>
</tr>
</tbody>
</table>
2. Sentence ➔ noisy input mapping

- There are lots of possibilities here
- Our implementation: start with the sequence of actually observed words

  \[
  \text{Dist}(\text{dog, cat}) = 3 \quad \text{Dist}(\text{the, toe}) = 1 \\
  \text{Dist}(\langle \text{eps} \rangle, \text{toes}) = 4 \quad \text{Dist}(\text{got, hot}) = 2
  \]
Quantifying alarm about the past

- **Relative Entropy** (KL-divergence) is a natural metric of change in a probability distrib. (Storck et al., 1995; Itti & Baldi, 2005; Levy, 2008)
- Our distribution of interest is *probabilities over the previous words in the sentence*
- Call this distribution $P_i(w_{[0,j]})$
- Conditions on words 0 through $i$
- Event space: words before position $j$
- The change induced by $w_i$ is the **error identification signal** $EIS_i$, defined as

$$D \left( P_i(w_{[0,i]}) \right) \left\| P_{i-1}(w_{[0,i]}) \right)$$

**new distribution**  **old distribution**
Error identification signal: example

- Measuring change in beliefs about the past:
  
  No change: \( EIS_2 = 0 \)

\[ \sum_{w_1} P_2(w_1) \log \frac{P_2(w_1)}{P_1(w_1)} = \begin{cases} 0.28 (1 - 1.82) + 0.72 (1 - 0.48) = 0.14 \end{cases} \]
Results on local-coherence sentences

• Locally coherent:  *The coach smiled at the player tossed the frisbee*
• Locally incoherent: *The coach smiled at the player thrown the frisbee*

_EIS greater for the variant humans boggle more on_

(All sentences of Tabor et al. 2004 with lexical coverage in model)
Experimental confirmation

- Novel prediction: changing the neighborhood of the context can change the EIS

The coach smiled at the player tossed the frisbee

The coach smiled toward the player tossed the frisbee

- Substituting toward for at should reduce the EIS
- In free reading, we should see less tendency to regress from tossed when the EIS is small

(Levy, Bicknell, Slattery & Rayner, CUNY 2009)
Model predictions

(The coach smiled at/toward the player tossed/thrown the frisbee)
Experimental design

• In a free-reading eye-tracking study, we crossed at/toward with tossed/thrown:

The coach smiled at the player tossed the frisbee
The coach smiled at the player thrown the frisbee
The coach smiled toward the player tossed the frisbee
The coach smiled toward the player thrown the frisbee

• Prediction: interaction between preposition & ambiguity in some subset of:
  • Early-measure RTs at critical region tossed/thrown
  • First-pass regressions out of critical region
  • Go-past time for critical region
  • Regressions into at/toward
Procedure

- 24 items+36 fillers, 40 participants
- Isolated-sentence reading
- Each sentence followed by a comprehension question
- Eye movements monitored with Eyelink 2000 tower setup
Experimental results

First-pass RT

Proportion of trials

at ambig
at unambig
toward ambig
toward unambig

Proportion correct answers

at ambig
at unambig
toward ambig
toward unambig

The coach smiled at the player tossed

Go-past RT

Proportion of trials

at ambig
at unambig
toward ambig
toward unambig

Proportion correct answers

at ambig
at unambig
toward ambig
toward unambig

Comprehension accuracy

The coach smiled at the player tossed...?
What this result tells us

• Readers must have residual uncertainty about word identity
  • Word misidentification alone won’t get this result in a fully incremental model:

  The coach smiled toward the player. \textcolor{red}{thrown}
  The coach smiled at the player. \textcolor{red}{thrown}
  The coach smiled as the player. \textcolor{red}{thrown}
  The coach smiled toward the player. \textcolor{red}{tossed}
  The coach smiled at the player. \textcolor{red}{tossed}
  The coach smiled as the player. \textcolor{red}{tossed}

• Also, readers respond to changes in uncertainty in a sensible way
What this result does not tell us

- Doesn’t show that Levy (2008)’s error-identification model is the only model that can get this results
  - But any model without uncertainty about prior-word identities of prior words has its work cut out for it
- Doesn’t show that local-coherence effects are exclusively a consequence of rational inference
  - But provides more supporting evidence that rational inference may be at least part of the story
Error identification signal: local coherences

- Full experiment:
  - Probabilistic grammar with rule probabilities estimated from parsed Brown corpus
  - Lexicon with all \(<\text{tag}, \text{word}\>\) combinations of frequency $>500$ in parsed Brown corpus (plus sentence wds)
  - Error identification signal as defined above

$EIS = 0.07$
$EIS = 0.0001$

*The coach smiled at the player* tossed
*The coach smiled at the player* thrown

- The important part of the change is that *at* can be re-interpreted as *and* or other near-neighbors
But, you may protest...

- Most items in Tabor et al., 2004 did not involve the preposition *at* before the modified noun.
- For example:

  The manager watched a waiter *served/given* pea soup by a trainee.

- But these sentences can also involve revisions of past beliefs—specifically, *that a word has been missed*.
Missed words

- Modeling beliefs about missed words requires only a minor modification to the noisy-input representation.

hallucinated word insertions
Missed words (II)

- Experiment 2: stimulus without the preposition *at*

  - The difference in error-identification signal is much smaller, but we still get it

  The manager watched a waiter *served*...
  The manager watched a waiter *given*...

  - *EIS = 0.0168*
  - *EIS = 0.0117*

- The difference in error-identification signal is much smaller, but we still get it
Other potential applications of theory

• “Good-enough” processing representations (Ferreira et al., 2002)

  While Anna dressed the baby played in the crib.

• “Morphological mismatch” processing effects in cases of superficial semantic mismatch (Kim & Osterhout, 2005)

  The meal devoured…

• Modeling longer-distance regressions in reading of naturalistic text
What the model is still missing

• Lots of things! But a couple of things most sorely missed:
  • Trans-finite-state probabilistic rules (technical, not theoretical shortcoming)
  • Richer probabilistic information sources, such as *plausibility* of noun-verb match (statistical, not theoretical shortcoming)

The bandits worried about the *prisoner* transported…
The bandits worried about the *gold* transported…

*more difficult* (Tabor et al., 2004)
Probability in language production

• Why do people talk the way they do?
• Linguistic communication involves transactions in uncertainty
• But it takes place under adverse conditions:
  • Auditory environment is noisy
  • People’s working memory is limited
  • Environment competes for attention
  • Interlocutors have incomplete knowledge of each other
• Yet communication seems to work most of the time
• How is redundancy achieved?
Hypothesis about language use

• Surprisal: predictable (=less informative) words easier
• *If* processing difficulty is at all superlinear in surprisal
  ⇒ provable (through Jensen’s Inequality) that spreading out information evenly in a sentence minimizes total comprehension difficulty (Levy & Jaeger 2007)
• We call this idea *Uniform Information Density* (UID)
• Same idea can also be motivated through noisy-channel view of linguistic communication
The empirical phenomenon

- Certain types of *relative clauses* (RC) in English are optionally introduced by the “meaningless” word *that*

  \[ \text{How big is the family \( (\text{that}) \) you cook for \( __ \)}? \]

  • *RC* modifies the noun *family*

  "you cook for the family"

- Relative clauses are an important part of the infinite expressive capacity of human language (recursion)
- What governs use of the optional function word *that*?

*Levy & Jaeger 2007 (see also Jaeger 2006)*
UID/constant entropy

- The idea of spreading out information equally has also appeared as a noisy “channel capacity” argument previously.

- **Probabilistic reduction hypothesis** for phonetic realization (Jurafsky et al., 2001; Bell et al. 2003)
  - See also Aylett, 1999; Aylett & Turk, 2004: *Smooth signal redundancy hypothesis*

- **Entropy rate constancy** throughout a discourse (Genzel & Charniak 2002, 2003)

- Ours is the first study to examine a specific linguistic speaker-choice variable *above the phonetic level*
Spreading out information in RCs

• In an RC without *that*, the first word does two things:

  How big is the family *you*…

  1) It signals that a relative clause has begun
  2) It signals some information about the contents of the relative clause

• Inserting *that* separates these two things:

  How big is the family *that you*…

(1) (2)

• Hypothesis: speakers should use *that* more when the RC’s onset is informationally dense
Spreading out information in RCs (2)

(1) (2)

How big is the family *that you*…

- We want to measure the quantity of information (1) and (2) *literally* using information theory

- (1) is $P(\text{that} \mid \text{context}) \ P(\text{RC} \mid \text{context})$

- (2) is $P(w_i \mid \text{context,RC}) \ [\text{you}]$
Dataset

- Corpus of spontaneous telephone conversation by speakers of American English (Switchboard corpus)
- Roughly 1 million words of conversation have been annotated for linguistic structure
- Contains 3,452 datapoints (relative clauses for which that can potentially be omitted)
Probabilistic model of structural production

- We use tree structures to represent natural language structure and ambiguity as a sentence unfolds…

```
NP-SBJ  BES  S
   NP-PRD(1)
      NP(2)
         NP(3)
          RC
          you’d ever want to do …
```

```
PRP  ’s
  it
```

```
CD  PP
  one
```

```
IN
  of
```

```
DT  JJ  JJ  NNS  PP-LOC
  the  last  few  things
```

```
IN
  in
```

```
DT  NN
  world
```

```
DT  the
```
Calculating phrasal predictability

- The use of tree structure also gives us a recurrence relation expressing the predictability of an upcoming phrase in the tree:

\[
P(\text{RC}_{n+1} \ldots | w_{1 \ldots n}, T_{1 \ldots n}) = \sum_{i=0}^{k} \left[ P(\text{RC} | N_i) \prod_{j=0}^{i-1} P(*END* | N_j) \right]
\]

we need to estimate these model parameters
The statistical problem

- There are two statistical questions to be addressed:
  1. How do we choose the phrasal predictability model $P(X|N_i)$?
  2. How do we assess whether phrasal predictability is associated with speakers' behavior in *that*-use?

- These correspond to two somewhat different types of statistical question:
  1. prediction: designing an accurate model of an outcome (machine learning)
  2. hypothesis testing: assessing a particular factor’s association with an outcome (classical statistics)
The statistical problem (2)

- In both cases, there are huge numbers of features that may potentially affect the outcome
  - e.g., each English noun may have distinctive tendencies for RC modification (way, apple)
- Problem of *model selection*: which features to put into the model?
- The answer differs for each statistical question:
  1. Prediction: a very large, overparameterized model is OK, as long as it accurately predicts outcomes
  2. Hypothesis testing: test the factor of interest in a small model with carefully developed control factors
Two-step model

\[ P( \text{RC} \mid \text{context} ) \rightarrow P( \text{that} \mid \text{RC}) \]

Control factors

- three outcomes (RC, *END*, other)
- regularized\(^*\) multinomial logistic regression (MaxEnt model)
- large number of surface & structural features of context (~3.3\(\times\)10\(^6\); \(n\approx10^6\))
- binary outcome
- unregularized logistic regression (bootstrapped by speaker cluster)
- phrasal predictability is a single covariate
- a select set of controls \(^*\) constitutes another 27 parameters (\(n=3,452\))
Linguistic theory suggests many types of features that may be important:

- Semantically empty words tend to be elucidated relative clauses.
- Definite articles and superlative adjectives, especially together, like RCs.
- Postmodifiers of the noun tend to fill this need for elucidation.
Regularized multinomial logit

- We need to compute $P(X_{ij}|N_i)$ for $X_{ij}=${RC,NULL,other}
- Using all these (overlapping, sparse) features to do so
- Regularized logit models handle this nicely
- *Featurize* each context (a node $N_i$ with its tree) as a vector $f(N_i)$; the probability is set to be

$$P(X_{ij}|N_i) = \frac{1}{Z} e^{\lambda_j \cdot f(N_i)}$$

- Learning problem is now finding parameter vector $\lambda$
- *Regularize* (=keep small) parameters by maximizing penalized likelihood:

$$\text{Lik}(X;\lambda) = \left[ \prod_i P(X_i|N_i) \right] - \sum_{jk} \frac{1}{\sigma^2} (\lambda_{jk})^2$$
Investigating control factors

- Separate studies (Jaeger 2006) investigated the role of many other factors in *that*-use:
  - Length of the relative clause and distance of long-distance extraction site
  - Disfluency (production difficulty)
  - Adjacent identical segments (i.e., tendency to avoid saying *that that*)...
  - Speaker gender

- These factors & others were selected from a larger set using backward AIC optimization
Putting the two models together

• Hypothesis test: enter \(-\log P(\text{RC}|\text{context})\) as covariate with the control factors in a logistic regression
• Result: phrasal predictability is associated with \textit{that}-omission at \(p<0.0001\) (Wald statistic)
• We can also run backward model selection using AIC again on the new model
• Result: several control factors drop out of the model
  • adjacent identical segments seem not to matter
  • speaker gender effect goes away
• \textit{Phrasal predictability helps us make sense of that-use}
Production study: conclusion

- Speakers seem sensitive to information density as a principle of communicative optimality
- An optional function word like *that* acts as a “pressure valve” for speakers to regulate information flow
- Leads to a rather unconventional view of grammar
  - conventional: a set of categorical rules reflecting universal, innate principles
  - new view: a set of statistically-oriented tools to achieve communicative ends
- Are these views irreconcilable?
  - *I think this is one of the major issues facing the field*
- Methodology: combine different statistical modeling principles to gain insights about human language