Computational Psycholinguistics
Lecture 3, part 1: language production and grammatical choice

Klinton Bicknell & Roger Levy

2015 LSA Summer Institute
University of Chicago
14 July 2015
• Our model of comprehension, in a nutshell: infer and use

\[ P(\text{Meaning} | \text{Input}, \text{Context}) \]

• Today’s proposal for modeling language production: for the interpretation I want to convey, I want to minimize

\[ \text{Cost(Utterance} | \text{Meaning}, \text{Context}) \]
Factors determining utterance “cost” given context and intended meaning:

- Utterance should be likely to successfully convey intended meaning!
- Utterance should be *efficiently structured*
  - Succinct wherever possible (wastes minimal time)
  - Organized so as to minimize comprehender’s effort
  - Organized to be as easy as possible to produce
  - ...
- If the utterance achieves subsidiary goals of the speaker and/or addressee, that’s a bonus
  - Signaling speaker group affiliation
  - Teaching the listener something indirectly
  - ...
In principle, you could use *any* utterance (or extra-linguistic signal) to try to convey *any* meaning!

- e.g., intended meaning: *I’d like a beer*
- Possible utterances:

  - *I’d like a beer*
  - *Where can I get a beer?*
  - *It’s pilsville time*
  - *I’m in Germany*
  - *Grass is green*
  - *Garr!*
  - *[Mime beer-drinking]*

Not all these alternatives are equally likely to successfully communicate the intended meaning!
Grammatical choice

• It has proven productive to focus on how speakers choose among tightly delimited sets of alternative utterances
  • Example: the *dative alternation*
    
    *Terry gave the exhausted traveller from France a silver dollar.*
    
    *Terry gave a silver dollar to the exhausted traveller from France.*

• Basic question: what factors of general theoretical interest successfully predict which variant the speaker chooses?

• Implicit assumptions:
  • The variants are *(near-)*meaning equivalent
  • The variants are generally *available together* to the speaker
Case studies in grammatical choice

• The dative alternation

  Terry gave the exhausted traveler from France a silver dollar.
  Terry gave a silver dollar to the exhausted traveler from France.

• Optional that-deletion in relative clauses

  I know the family you feed.
  I know the family that you feed.

• Optional to-deletion in the DoBe construction

  The least we should do is make it as much fun as possible.
  The least we should do is to make it as much fun as possible.

• Methods:
  • Multivariate statistical corpus analysis
  • Probabilistic computational modeling
  • Controlled behavioral experiments
Case study #1: the dative alternation

- Terms used with the dative alternation:

<table>
<thead>
<tr>
<th>Structure Description</th>
<th>Example</th>
<th>Syntactic Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepositional dative structure:</td>
<td>...gave [toys] [to the children]</td>
<td>V NP PP</td>
</tr>
<tr>
<td>Double object structure:</td>
<td>...gave [the children] [toys]</td>
<td>V NP NP</td>
</tr>
<tr>
<td>Dative PP:</td>
<td>...gave [toys] [to the children]</td>
<td>V NP PP</td>
</tr>
<tr>
<td>Dative NP:</td>
<td>...gave [the children] [toys]</td>
<td>V NP NP</td>
</tr>
<tr>
<td>Theme:</td>
<td>...gave [toys] [to the children]</td>
<td>V NP PP</td>
</tr>
<tr>
<td></td>
<td>...gave [the children] [toys]</td>
<td>V NP NP</td>
</tr>
<tr>
<td>Recipient:</td>
<td>...gave [toys] [to the children]</td>
<td>V NP PP</td>
</tr>
<tr>
<td></td>
<td>...gave [the children] [toys]</td>
<td>V NP NP</td>
</tr>
</tbody>
</table>

(Bresnan et al., 2007; Goldberg, 2006; Kako, 2006; Myslin & Levy, in prep)
Case study #1: the dative alternation

- Two schools of thought

1. The two variants *subtly differ in meaning*
   - Prepositional dative signals *transfer of location*
   - Double object signals *transfer of possession*
   - Evidence:
     - Introspective (e.g., Goldberg, 2006):
       - *I sent storage a book.* → *storage* refers metonymically to something animate
       - *I sent a book to storage.* → no such inference
       *That movie gave me the creeps.*
       ✗ *That movie gave the creeps to me.*
     - Experimental (Kako, 2006):
       - *The rom gorped the blick to the dax.* (more likely!)
       - *The rom gorped the dax the blick.* (less likely!)
     - How likely is *gorping* to involve moving something?

PD: ...gave [toys] [to the children]
DO: ...gave [the children] [toys]
Case study #1: the dative alternation

2. General *processing preferences* govern the alternation
   - Alignment of the following preferences with linear order
     - discourse-given < discourse-new
     - short < long
     - definite < indefinite
     - animate < inanimate
     - pronoun < full NP

*Evidence: univariate corpus analysis* (Collins, 1995)
Case study #1: the dative alternation

- Two schools of thought
  - PD: …gave \text{[toys]} \text{[to the children]}
  - DO: …gave \text{[the children]} \text{[toys]}

2. General processing preferences govern the alternation

**Evidence: multivariate corpus analysis**

\[
b_{\text{verb}} \sim N(0, \sigma_{b}^{2})
\]

\[
\eta \sim \alpha + \sum_{i} \beta_{i} x_{i} + b_{\text{verb}}
\]

<table>
<thead>
<tr>
<th>Predictor (x_{i})</th>
<th>Coefficient (\beta_{i})</th>
</tr>
</thead>
<tbody>
<tr>
<td>log Recipient Length</td>
<td>1.31</td>
</tr>
<tr>
<td>log Theme Length</td>
<td>-1.17</td>
</tr>
<tr>
<td>Recipient Animacy</td>
<td>2.14</td>
</tr>
<tr>
<td>Theme Animacy</td>
<td>-0.92</td>
</tr>
<tr>
<td>Recipient Discourse Status</td>
<td>1.33</td>
</tr>
<tr>
<td>Theme Discourse Status</td>
<td>-1.76</td>
</tr>
<tr>
<td>Recipient Pronominality</td>
<td>-1.54</td>
</tr>
<tr>
<td>Theme Pronominality</td>
<td>2.2</td>
</tr>
<tr>
<td>Recipient Definiteness</td>
<td>0.8</td>
</tr>
<tr>
<td>Theme Definiteness</td>
<td>-1.09</td>
</tr>
</tbody>
</table>

(My implementation of the analysis of Bresnan et al., 2007)
Face-off between two theories

- The “construction grammar” theory: subtle meaning differences govern the dative alternation
- The “processing-optimality” theory: general syntactic processing ease governs the dative alternation
- How do we distinguish among the two theories?
- How would you do an experiment to tell?
Our approach

• Perhaps *both* are right!
• What would such a theory look like?

• Let’s go one step farther and treat this as a *directed graphical model*, or Bayes Net!

(Myslin & Levy, in prep)
When processing explains form away

- Joint influence of meaning and processing on syntactic form:

- We could test this theory by directly modeling $P(F|M,P)$
- But we’ll go one step further
  - Observing the chosen syntactic form renders meaning and processing conditionally dependent
  - Strong processing preferences can explain form away and reduce the association with meaning
Experiment 1

The zarg prolted the cherid to a really gromious flig.

Which is more likely?

- The cherid is in a new place.  
  *LOCATIVE* inference
- The cherid has a new owner.  
  *POSSESSIVE* inference

(Myslin & Levy, in prep)
### Experiment 1

| Sentence                                                                 | $S$ | $P(S|G)$ |
|--------------------------------------------------------------------------|-----|---------|
| The zarg prolted [the cherid] to [a really gromious flig].                | PO  | high    |
| The zarg prolted [the flig] [a really gromious cherid].                  | DO  | high    |
| The zarg prolted [a really gromious cherid] to [the flig].               | PO  | low     |
| The zarg prolted [a really gromious flig] [the cherid].                  | DO  | low     |

Which is more likely?

- The cherid is in a new place.  
  LOCATIVE inference
- The cherid has a new owner.  
  POSSESSIVE inference

(Myslin & Levy, in prep)
Experiment 1: results

The bar chart shows the proportion of LOCATIVE inferences as a function of grammatical probability. The chart includes data for two structures: DO and PO. The error bars indicate the variability in the data.

| Sentence                                                                 | Structure | S   | P(S|G) |
|--------------------------------------------------------------------------|-----------|-----|-------|
| The zarg prolted [the cherid] to [a really gromious flig].                |           | PO  | high  |
| The zarg prolted [the flig] [a really gromious cherid].                  |           | DO  | high  |
| The zarg prolted [a really gromious cherid] to [the flig].               |           | PO  | low   |
| The zarg prolted [a really gromious flig] [the cherid].                  |           | DO  | low   |

(Myslin & Levy, in prep)
Experiment 2: length only

![Bar chart showing proportion of LOCATIVE inferences for high and low grammatical probability, with error bars.]

| Sentence                                           | Structure | $S$  | $P(S|G)$ |
|----------------------------------------------------|-----------|------|----------|
| The zarg prolted [the cherid] to [the really gromious flig]. | DO        | high |          |
| The zarg prolted [the flig] [the really gromious cherid].   | DO        | high |          |
| The zarg prolted [the really gromious cherid] to [the flig]. | PO        | low  |          |
| The zarg prolted [the really gromious flig] [the cherid].  | DO        | low  |          |

(Myslin & Levy, in prep)
Experiment 3: definiteness only

Definiteness only

n.s.

high

low

Grammatical probability

(Myslin & Levy, in prep)
Overall results

(Myslin & Levy, in prep)
Summary

• Everyone wins!
  • There are subtle meaning differences between PD and DO structures
  • But preference for fidelity of form-meaning mapping is defeasible
  • Processing considerations can explain form away and affect strength of inferences regarding meaning
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
  \[ \Rightarrow \text{provable (through Jensen’s Inequality) that spreading out information evenly in a sentence minimizes total comprehension difficulty} \] (Levy & Jaeger 2007)
- We call this idea \textit{Uniform Information Density} (UID)
- Same idea can also be motivated through noisy-channel view of linguistic communication
Ramifications for production

- Consider a Bayesian picture of recovering any aspect of language structure from a signal

\[ P(\text{Structure}|\text{Signal}) = \frac{P(\text{Signal}|\text{Structure})}{P(\text{Signal})} \frac{P(\text{Structure})}{P(\text{Signal})} \]

- In general, a trade-off between (top-down) prior and (bottom-up) evidence

- The stronger the prior expectations for the structure, the less signal needs to be given

- Level of sound → word: vowel duration in function words is modulated by word predictability (Jurafsky et al., 2001)

<table>
<thead>
<tr>
<th>High-predictability</th>
<th>Low-predictability</th>
</tr>
</thead>
<tbody>
<tr>
<td>been a</td>
<td>compost a</td>
</tr>
<tr>
<td>with a</td>
<td>field a</td>
</tr>
<tr>
<td>where a</td>
<td>costs a</td>
</tr>
<tr>
<td>select a</td>
<td>children a</td>
</tr>
</tbody>
</table>
Case study #2: *that*-deletion in RCs

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

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

  - 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)*
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* *(1)* *y*(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 \( P(\text{that} | \text{context}) \cdot P(\text{RC} | \text{context}) \)

- (2) is \( P(w_i | \text{context,RC}) \) [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)
A first test
Probabilistic model of structural production

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

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

$$P(RC_{n+1}...|w_1...n, T_1...n) = \sum_{i=0}^{k} \left[ P(RC|N_i) \prod_{j=0}^{i-1} P(*END*|N_j) \right]$$

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 (~3.3×10^6; \( n \approx 10^6 \))

- 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)
Feature space for prediction model

- 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,NUL,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:

$$\left(\prod_i P(X_i|N_i)\right) - \sum_{jk} \frac{1}{\sigma^2} (\lambda_{jk})^2$$
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(RC|context) as covariate with the control factors in a logistic regression
• Result: phrasal predictability is associated with 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
• *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
Introducing the construction by examples

**Post-Copular Verb** (PCV)

*what* the CBO does is **takes** Congress’s promises at face value

*what we have done* is **taken** military action in Bosnia through NATO

*all he’s been doing* is **going** over legal papers

*all the government does* is **send** out checks

*the thing that I tried to do* was **to keep** the score close

*the thing I’m doing* is **trying** to learn from my mistakes

*the least we should do* is **make it** as much fun as possible

(Wasow, Levy, et al., forthcoming)

There is almost no literature discussing this construction
Some Corpus Examples

1. what we're here on earth to do is (to) celebrate humanity
2. what I would do is (to) call upon the press to police yourselves
3. the other thing that it’ll do is (to) facilitate getting Chinese troops into Tibet as well
4. the most important thing that Bretton Woods did was (to) create two institutions for international cooperation on monetary international problems
5. all they can do is (to) circumvent themselves
6. all I want to do is (to) go to work

• Can you tell which ones had to in the original?
Some Initial Expectations

Based on previous work on optional *that*, we expected that things that might make production and/or comprehension harder would increase rate of *to* use. Specifically:

- less frequent head words in subject
- added length/complexity both before and after the critical position (between copula and PCV)
- material intervening between *do* and *be* or between *be* and PCV
- less frequent forms of *do*
- less frequent forms of the copula
- less frequent PCV

*Uniform Information Density*
In-construction PCV frequency effect robust

\begin{align*}
\log_{10} \text{in-construction frequency} & \quad \beta
\end{align*}
Other UID-related work

- UID effects in spontaneous production at multiple levels
  - auxiliary contraction (Frank & Jaeger, 2008)
  - *that*-reduction in complement clauses (Jaeger, 2010)
  - clausal planning (Gomez Gallo & Jaeger, 2009)

- Organization of *grammar*: clausal word order typology
  (Maurits, Perfors, & Navarro, 2010)
Zipf’s law 1

- A word’s log-frequency and log-rank-frequency are linearly related

(graph showing the relationship between rank word frequency and word log-frequency)

(Zipf, 1935)
Zipf’s law 2

- There is a (more or less) linear relationship (albeit noisy) between word log-frequency and word length

(Zipf, 1935; figure from Piantadosi et al., 2011)
Imagine that a monkey hits the keys of a typewriter at random, subject only to these constraints: (1) he must hit the space bar with a probability of \( p^* \) and all the other keys with a probability of \( p(L) = 1 - p^* \), and (2) he must never hit the space bar twice in a row. I wish to examine the monkey's output, not because it is interesting, but because it will have some of the statistical properties considered interesting when humans, rather than monkeys, hit the keys.

Miller (1957)

- **With** \( M \) letters on the typewriter, and \( q = (1 - p^*)/M \), then
  \[
P(w = \text{any word of length } K) = q^K
  \]
  \[
  \log P(w) = K \log q \quad \text{Zipf's law 2}
  \]

- **For the “average-frequency” word of length** \( K \) **there will be approximately** \( M^K/2 \) **higher-frequency words**
  \[
  \log \text{rank-freq}(w) \approx K \log \frac{M}{2}
  \]
  \[
  \log \text{rank-freq}(w) \propto \log P(w) \quad \text{Zipf's law 1}
  \]
Miller’s monkeys

• There are two conclusions that one could draw from this state of affairs:

  1. If so simple and un-language-like a process as typewriter monkeys could give rise to Zipf’s law(s), then the fact that language happens to follow these laws could not be of possible scientific interest

  2. Given how thoroughly unlike monkey-typing human language is, the fact that it exhibits deeply similar statistical properties is remarkable and merits careful study
Miller’s monkeys

• Conclusion (1) carried the day, and the issue became marginal:

Research workers in statistical linguistics have sometimes expressed amazement that people can follow Zipf's Law so accurately without any deliberate effort to do so. We see, however, that it is not really very amazing, since monkeys typing at random manage to do it about as well as we do....It seems...that Zipf's rule can be derived from simple assumptions that do not strain one's credulity (unless the random placement of spaces seems incredible), without appeal to least effort, least cost, maximal information, or any branch of the calculus of variations.

Miller (1957)
According to principles of UID, it should be expected word *surprisal*, not word *frequency*, that word length is optimized for.

Mathematically, the expected surprisal of a word \( w \) is:

\[
\sum_{Ctxt} P(Ctxt|w) \log \frac{1}{P(w|Ctxt)}
\]

Seyfarth (2014) calls this quantity *word informativity*.
Informativity versus frequency

(Seyfarth, 2014)
UID in the lexicon and Miller’s monkeys

- Piantadosi, Tily, & Gibson (2011) show this is true in 11 out of 11 languages investigated!
UID in the lexicon and Miller’s monkeys

- In detail, for English: 

  (Piantadosi, Tily, & Gibson, 2011)

- Crucially, *monkeys on a typewriter will not give this result*
Informativity effects on acoustic duration

- Earlier: word predictability affects acoustic duration
- Seyfarth (2014) found that word *informativity* affects duration in the same way!
Noisy-channel models in sentence comprehension

Klinton Bicknell & Roger Levy

LSA Institute 2015
U Chicago
Overview of the day

• A puzzle for surprisal theory

• Theory to handle the puzzle by revising some foundational assumptions about sentence comprehension

• Empirical study on syntactic comprehension arising from the new theory
Puzzle: 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)
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…*

- A 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?
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
- We use probabilistic context-free grammars (PCFGs) and weighted finite-state automata (WFSAs) to instantiate the model
- **Then**: we show how the introduction of 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 \( I \):

\[
P(w|w^*) = \int_I PC(w|I, w^*) PT(I|w^*) dI
\]

\( \propto Q(w, w^*) \)
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

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”
Probabilistic Linguistic Knowledge

- 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 *weighted intersection*

Grammar affects beliefs about the future
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\}

grammar
The noisy-channel model (FINAL)

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

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

Cost(\textit{a cat sat}) = 0

Cost(\textit{sat a sat cat}) = 8  \quad \text{Result of } K_{LD} \text{ applied to } w^* = \textit{a cat sat}
Incremental inference under uncertain input

• Near-neighbors make the “incorrect” analysis “correct”:

   Any of these changes makes \textit{tossed} a main verb!!

   \textit{The coach smiled at the player tossed the frisbee}

• Hypothesis: the boggle at “tossed” involves \textit{what the comprehender wonders whether she might have seen}
The core of the intuition

- Grammar & input come together to determine two possible “paths” through the partial sentence: (line thickness ≈ probability)

- *tossed* is more likely to happen along the bottom path
  - This creates a large shift in belief in the *tossed* condition

- *thrown* is very unlikely to happen along the bottom path
  - As a result, there is no corresponding shift in belief
Ingredients for the model

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

- \(Q(w, w^*)\) comes from Levenshtein distance
- \(P_C(w)\) 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
Quantifying alarm about the past

- Relative Entropy (KL-divergence) is a natural metric of change in a probability distrib. (Levy, 2008; Itti & Baldi, 2005)
- Our distribution of interest is probabilities over the previous words in the sentence
- Call this distribution $P_i(w_{[0,j]}))$
- The change induced by $w_i$ is the error identification signal $EIS_i$, defined as

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

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

- Measuring change in beliefs about the past:

  \{a,b\} \{?\} \quad \text{No change: } EIS_2 = 0 \quad \{a,b\} \{f,e\}

\[ \sum_{w_1} P_2(w_1) \log \frac{P_2(w_1)}{P_1(w_1)} = 0.28 (1 - 1.82) + 0.72 (1 - 0.48) = 0.14 \]
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)
Novel predictions of the model

• Theoretical recap:
  • Comprehension inferences involve trade-offs between *uncertain perception* and *prior grammatical expectations*
  • We saw how model may account for Tabor et al.'s (2004) local coherence sentences

• Novel prediction:
  • Uncertain-input effects should be *dependent on the perceptual neighborhood* of the sentence
Prediction: neighborhood manipulation

- Uncertain-input effects should be dependent on the perceptual neighborhood of the sentence
- Resulting novel prediction: changing neighborhood of the context can affect EIS & thus comprehension behavior

The coach smiled at the player tossed the frisbee

(that?) (who?)

The coach smiled toward the player tossed the frisbee

(and?) (that?) (who?)

- Substituting toward for at should reduce the EIS

(Levy, Bicknell, Slattery, & Rayner, 2009, PNAS)
(The coach smiled at/toward the player tossed/thrown the frisbee)
There are advantages and disadvantages of both electronic and hardcopy journals. Hardcopy journals are more easily browsed, more portable and, of course people are very much used to their format. Electronic journals save on paper and their format has improved considerably over the past few years, but there are still problems over managing copyright restrictions and persuading people to use electronic instead of hardcopy journals. There is also the problem of portability. More and more journals are now being published in electronic format, although some publishers will only let you subscribe to an electronic journal provided you also subscribe to the hardcopy (more money for the same thing). Some electronic journals cost over 100% more than their equivalent hardcopy. With all these factors in mind I have been discussing individual and shared-subscriptions with the Biochemistry Department, the RSL and Blackwell’s. Whilst I feel that a move from hardcopy to electronic journals will be a very slow process in the ULP Library, electronic publishing is being carefully monitored and I would hope to introduce a few electronic texts into the Library alongside the journals which are already available for free over the Internet.
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*
Experimental results

The coach smiled at the player tossed...

First-pass RT

Regressions out

Go-past RT

Go-past regressions

Comprehension accuracy

Noise level (low=noisy)

EIS

Proportion correct answers

Proportion of trials

Proportion of trials

Proportion of trials

Proportion of trials

Comprehension accuracy
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...thrown
  The coach smiled at the player...thrown
  The coach smiled as the player...thrown

  Should be about equally hard

  The coach smiled toward the player...tossed
  The coach smiled at the player...tossed
  The coach smiled as the player...tossed

  Should be easier, if anything

• Also, readers respond to changes in uncertainty in a sensible way
Summary

great evidence for noisy channel comprehension

• readers maintain uncertainty about the past

• readers update that uncertainty on the basis of new input

• something like EIS seems to explain (at least part of) Tabor et al.'s (2004) local coherences effects
Surprisal and EIS

processing difficulty and eye movements

- two types of processing difficulty: surprisal and EIS

- unclear how these types of difficulty affect eye movements?
  - more fixations? longer fixations? regressions?

- unclear why these types of difficulty affect eye movements
  - why keep looking at a word in proportion to surprisal?
  - why make regressions when EIS is high?

next up: a principled view of reading that answers these questions: reading as efficient visual information gathering