Computational Psycholinguistics
Lecture 4: language production and grammatical choice

Klinton Bicknell & Roger Levy

ESSLLI 26
University of Tübingen
21 August 2014
• 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

\[ \text{Cost(utterance|Meaning, Context)} \]
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
  - ...
Language production and grammatical choice

• 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*  *It’s pilsville time*  *Grass is green*
  *Where can I get a beer?*  *I’m in Germany*  *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 Type</th>
<th>Example</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepositional dative</td>
<td>...gave [toys] [to the children]</td>
<td>V NP PP</td>
</tr>
<tr>
<td>structure:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double object</td>
<td>...gave [the children] [toys]</td>
<td>V NP NP</td>
</tr>
<tr>
<td>structure:</td>
<td></td>
<td></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?
Case study #1: the dative alternation

• Two schools of thought

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 [toys] [to the children]
  - DO: ...gave [the children] [toys]

2. General processing preferences govern the alternation

**Evidence: multivariate corpus analysis**

\[
\begin{align*}
\eta & \sim \alpha + \sum_i \beta_i x_i + b_{\text{verb}} \\
\eta & \sim \alpha + \sum_i \beta_i x_i + b_{\text{verb}} \\
\end{align*}
\]

<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.
- The cherid has a new owner.

(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
Experiment 1: results

![Bar chart showing the proportion of LOCATIVE inferences for different grammatical probability and structure conditions.]

| 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      |

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

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

Definiteness only

n.s.

grammatical probability

high low

(Myslin & Levy, in prep)
Overall results

**Length & definiteness**

- ***

**Length only**

- ***

**Definiteness only**

- n.s.

(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
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
Ramifications for production

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

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

- 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

  \[
  \text{“you cook for the family”}
  \]

- 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:
  
  \[
  \text{How big is the family you...} \quad \text{phrasal predictability}
  \]

  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:

  \[
  \text{How big is the family that you...}
  \]

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

• (2) is $P(w_i \mid \text{context, RC})$  [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:

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

\[
N_3
\quad
N_2
\quad
N_1
\quad
N_0
\]

we need to estimate these model parameters
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)
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)
- \textit{regularized} multinomial logistic regression (MaxEnt model)
- large number of surface & structural features of context (\(3.3 \times 10^6\); \(n \approx 10^6\))
- binary outcome
- unregularized logistic regression (bootstrapped by speaker cluster)
- phrasal predictability is a single covariate
- \textit{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:

\[
\left[ \prod_i P(X_i|N_i) \right] - \sum_{jk} \frac{1}{\sigma^2} (\lambda_{jk})^2
\]
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*
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
In-construction PCV frequency effect robust

The graph shows the relationship between the log base 10 of the in-construction frequency and the parameter β. The x-axis represents the log base 10 of the in-construction frequency, ranging from 0 to 3, and the y-axis represents the parameter β, ranging from 0 to -5.0. The graph includes multiple lines indicating different levels of frequency effect.
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)
• A word’s log-frequency and log-rank-frequency are linearly related

(Zipf, 1935)
Zipf’s law 2

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

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)
UID in the lexicon and Miller’s monkeys

• 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:

$$\sum_{\text{Ctx}} P(\text{Ctx}|w) \log \frac{1}{P(w|\text{Ctx})}$$

• 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!*

---

**Correlation with length**

- **ENGLISH**
- **GERMAN**
- **CZECH**
- **DUTCH**
- **FRENCH**
- **ITALIAN**
- **POLISH**
- **PORTUGUESE**
- **ROMANIAN**
- **SPANISH**
- **SWEDISH**

- **Average Information Content**
- **Frequency**
- **Partial correlation**
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!