The nature of linguistic knowledge

Mark Liberman
University of Pennsylvania
A framing story

The place: LSA 2008 in Chicago, a talk by Greg Guy
The topic: t/d deletion (e.g. best buy ⇔ bes’ buy)
The conflict: a crisis of faith, over “rules” vs. “exemplars”
The trigger:

-- deletion by association --
words that often occur in deletion-favorable environments
(e.g. before consonants)
are more likely to show deletion even in unfavorable environments
(e.g. before vowels)
[…according to a corpus study in progress with Jennifer Hay]

Should Greg abandon his faith?

Chicago 1/22/2008
The mind’s dictionary

What do we know about…

how to pronounce a word?
how to form a plural or a past tense?
what a word means?
how to use a word in a sentence?
The obvious answer…

It’s a dictionary entry. Dictionaries are designed for this, and do a good job. But this is science, not pragmatic lexicography. So we elaborate the answer…

In theory:
  a formalized dictionary entry

In practice:
  a psychologized dictionary entry
Representational building blocks...

• symbolic categories
  – phonemes or phonological features
  – parts of speech or morpho-syntactic features
  – semantic features (or, mostly, not)

• in structured combinations
  – segments, syllables, feet, …
  – morphemes, words, …
The classical solution

• “Underlying forms” of morphemes are symbolic strings like dictionary pronunciations (or string-like structures of features, perhaps in layers)
• These are combined into words by (quasi-)syntax
• Surface “pronunciation” is also a symbol string, computed by rules and/or constraints
• This defines the lexicon’s claims on sound.

Speech production and perception obviously involve many other factors.
Some issues

• Redundancy
  – “figure it out” vs. “look it up”
  – complete paradigms vs. “principal parts”
  – what redundancy to remove? and how?

• Variation
  – geographical, social, individual
  – stylistic choices
  – lists of options? (statistical) rules? dueling grammars?
The insurgent faith: “exemplar theory”

… every token of experience is classified and placed in a vast organizational network… new tokens of experience are not decoded and then discarded …

At first it might seem rather implausible to suppose that every token of language use encountered by a speaker/hearer is registered in memory. …

[But] human memory capacity is quite large.

Joan Bybee,
“From usage to grammar: The mind’s response to repetition”, 2006
Linguists:
“exemplar theory” = impact of usage

A usage-based view takes grammar to be the cognitive organization of one’s experience with language. Aspects of that experience, for instance, the frequency of use of certain constructions or particular instances of constructions, have an impact on representation that are evidenced in speaker knowledge of conventionalized phrases, and in language variation and change.

Bybee, op. cit.
Caching, frequency, prototypes…

… almost everything about the existence of variation and the effects of experience …

Exemplar representations allow specific information about instances of use to be retained in representation.

Exemplar representation provides a natural way to allow frequency of use to determine the strength of exemplars.

Exemplar clusters are categories that exhibit prototype effects. They have members that are more or less central to the category, rather than categorical features.

Bybee, op. cit.
The psychological tradition is more sophisticated


(1) In exemplar-memory models, the learner stores mental representations of exemplars, grouped by category, then classifies new instances on the basis of their similarity to the remembered ensembles.

(2) In feature-frequency (or cue-validity) models, the learner records the relative frequencies of occurrence of individual features of exemplars, then classifies new instances on the basis of estimates of the likelihood that the vector of features in a test pattern arose from each alternative category.

(3) In prototype models, the learner forms an abstract representation of each category represented in a series of learning experiences, then classifies new instances on the basis of their distances from the category prototypes in a psychological space.
Estes’ “array framework”

(common to all three types of models)

… when a learner observes exemplars of categories, information about each exemplar is stored in the form of a vector of feature or attribute values. … a learner’s memory representation of a sequence of trials might take the form

A 11011101
A 10101111
B 10000001

where the rows correspond to individual exemplars, and A and B denote category tags.
Two cheers for exemplars: stories from a misspent youth


1975: Landauer’s landfill theory of memory -- exemplars in space and time.

1985: Stress by table lookup -- learning by counting exemplars.

1990: Letter-to-sound by analogy to exemplars -- 500% better than rules.

2000: Pavlov’s revenge -- I finally learn about learning theory.
Landauer’s landfill


A very simple spatial model of memory storage and retrieval is described, analyzed, and discussed. The postulated memory is without organization in the sense that neither the place of storage nor the order of search during retrieval is influenced by the nature of the information being stored or retrieved. The memory consists of a three-dimensional space containing a large number of homogeneously distributed loci at which data may be stored. Data received near each other in time are stored at nearby locations. Access is by an undirected expanding-sphere search. The model exhibits a wide variety of quantitatively and qualitatively humanlike behavior with respect to both standard learning and forgetting paradigms and with respect to frequency effects and other phenomena in word processing.
Leaving:

Fig. 1. Schematic diagram of memory structure and movement of the recording pointer. The storage space is assumed to be three-dimensional with a large number of small storage loci homogeneously distributed within it. The point at which data can be entered or a search initiated at any moment describes a random walk. When an edge is encountered the pointer reenters the space from a symmetrically opposite point. Dots in the figure represent successive pointer locations.
Retrieving:

Fig. 2. The possible search space for retrieval at a given moment is defined by a region with radius $r$. Search proceeds outward from the current locus of the pointer as the surface of an expanding spherical shell. The pointer's location is not influenced by either the occurrence or outcome of a search.
What’s missing

• Short-term/long-term distinction
• Semantic/episodic distinction
• Linguistic/nonlinguistic distinctions
• and all other distinctions…

“…the coffee grounds are next to the morning paper, on top of the breakfast dishes …”
Some results:

FIG. 5. Simulated forgetting curves for a fact encountered on one to eight successive occasions, along with data from a similar experiment with humans. [Total memory equals 7 × 7 × 7; search radius equals three steps (α = .18); two random walk steps between trials; N = 1000.]
Some more results:

FIG. 10. The time required to say the name of an object represented by a line drawing. The smooth curve is a least-squares fit of the equation $RT = a + kn^{-1/3}$, an approximation to the model for a large storage space. Frequencies are taken from the Lorge magazine count (Thorndike & Lorge, 1944).
Fig. 12. Data for RT to press a key indicating the correct of three possible responses in a PA task as a function of number of previous equally spaced trials. The solid line is a least-squares fit derived from the same equation used in Figs. 10 and 11. The filled circles are for the human data, the open circles for simulation "Ss".
Spaced practice

Fig. 8. Illustration of the effects of spacing for short retention intervals. As in Fig. 6, shaded regions represent the areas from which one or the other of two storage loci can be reached by a search. The area within the dotted lines defines the possible loci of the pointer some arbitrary short time after the second of two massed (a) or spaced (b) trials. Unlike the situation after long retention intervals, the probability of retrieval may be greater for massed trials.
Learning stress

• Sequence of rules, e.g.
  \([+\text{syll}] \rightarrow [\text{l stress}] / \rightarrow [-\text{syll}]_0 [+\text{syll}] [-\text{syll}]_0 \#\).

• Interaction of constraints, e.g.
  *ss \sim *www \sim h=s

• Table look-up, e.g.
  weight-string \leftrightarrow stress-string

  \[3^4 = 81 \quad 3^5 = 243\]

• Analogy to known cases
English letter-to-sound mapping


Phone directories list millions of different surnames, more every year.

Other sorts of names can also be a problem.

Letter-to-sound rules: 50% “good” pronunciation, 85% “acceptable”

For proper names, a 50,000-word name list has only 60-70% coverage of name tokens (15-25% coverage of types)

What to do?
Analogy to the rescue

Rhyming: Califano ⇒ Talifano, Alifano
Stress-neutral endings: Stephen ⇒ Stephenville
Compounding: Abdulhussein, Baumgaertner

Results on new proper names:
89% “good”, 95% acceptable pronunciations
(compared to 50% “good”, 85% “acceptable” for our rules)
A practical test

Test of 50,000 word dictionary + analogy + rules

on 250,000 marketing-list names:

<table>
<thead>
<tr>
<th>Method</th>
<th>Raw Counts</th>
<th>Percentage</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type</td>
<td>Token</td>
<td>Type</td>
</tr>
<tr>
<td>Direct Hit</td>
<td>40,208</td>
<td>40,354,563</td>
<td>16.08%</td>
</tr>
<tr>
<td>Name + Stress-Neutral Ending</td>
<td>42,454</td>
<td>16,996,558</td>
<td>16.98%</td>
</tr>
<tr>
<td>Name + Primary-Stress Ending</td>
<td>627</td>
<td>69,016</td>
<td>.25%</td>
</tr>
<tr>
<td>Name + ity-class Ending</td>
<td>6,760</td>
<td>984,662</td>
<td>2.70%</td>
</tr>
<tr>
<td>Name + al-class Ending</td>
<td>2,393</td>
<td>341,301</td>
<td>.96%</td>
</tr>
<tr>
<td>Name + Name (Compound)</td>
<td>16,619</td>
<td>1,663,444</td>
<td>6.65%</td>
</tr>
<tr>
<td>Name + Suffix-Exchange</td>
<td>14,523</td>
<td>1,010,813</td>
<td>5.81%</td>
</tr>
<tr>
<td>Rhyme with Name</td>
<td>29,924</td>
<td>1,579,499</td>
<td>11.97%</td>
</tr>
<tr>
<td>Partial Letter-to-Sound</td>
<td>13,796</td>
<td>500,923</td>
<td>5.52%</td>
</tr>
<tr>
<td>Prefix</td>
<td>1,230</td>
<td>117,857</td>
<td>.49%</td>
</tr>
<tr>
<td>Combinations of Above</td>
<td>43,874</td>
<td>3,023,149</td>
<td>17.55%</td>
</tr>
<tr>
<td>All Dictionary-based Methods</td>
<td>212,408</td>
<td>66,641,785</td>
<td>84.96%</td>
</tr>
<tr>
<td>Remainder (to be handled by namsa)</td>
<td>37,592</td>
<td>1,364,118</td>
<td>15.04%</td>
</tr>
<tr>
<td>Totals</td>
<td>250,000</td>
<td>68,005,903</td>
<td>100.00%</td>
</tr>
<tr>
<td>Method</td>
<td>Ordinary Words (non-names)</td>
<td>Capitalized Words (names)</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>----------------------------</td>
<td>---------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Raw Counts</td>
<td>Raw Counts</td>
<td>Percentage</td>
</tr>
<tr>
<td></td>
<td>Type</td>
<td>Token</td>
<td>Type</td>
</tr>
<tr>
<td>Direct Hit</td>
<td>13,356</td>
<td>20,646,656</td>
<td>18.27%</td>
</tr>
<tr>
<td>Stress-Neutral</td>
<td>24,050</td>
<td>4556078</td>
<td>32.90%</td>
</tr>
<tr>
<td>Primary-Stress</td>
<td>679</td>
<td>76,222</td>
<td>.93%</td>
</tr>
<tr>
<td>itary-Class</td>
<td>1,943</td>
<td>332,587</td>
<td>2.66%</td>
</tr>
<tr>
<td>al-Class</td>
<td>1,174</td>
<td>237,979</td>
<td>1.61%</td>
</tr>
<tr>
<td>Suffix-Exchange</td>
<td>497</td>
<td>36,763</td>
<td>.68%</td>
</tr>
<tr>
<td>Rhyme</td>
<td>2,907</td>
<td>436,839</td>
<td>3.98%</td>
</tr>
<tr>
<td>Prefix</td>
<td>5,133</td>
<td>23,833</td>
<td>4.35%</td>
</tr>
<tr>
<td>Partial L-to-S</td>
<td>5,591</td>
<td>79,753</td>
<td>4.74%</td>
</tr>
<tr>
<td>Compound</td>
<td>3,718</td>
<td>170,877</td>
<td>5.09%</td>
</tr>
<tr>
<td>Combinations</td>
<td>15,151</td>
<td>966,033</td>
<td>23.46%</td>
</tr>
<tr>
<td>All Methods</td>
<td>63,475</td>
<td>27,460,034</td>
<td>89.58%</td>
</tr>
<tr>
<td>Remainder</td>
<td>9,618</td>
<td>21,076</td>
<td>10.42%</td>
</tr>
<tr>
<td>Totals</td>
<td>73,093</td>
<td>27,481,110</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Chicago 1/22/2008
Another alternative

(Similar to Estes’ feature-frequency models)

- Basic representations are not symbol strings, but distributions (of activation or probability) over such strings
- Morphology and phonology combine and modify these
- The phonological “surface” is likewise a random variable whose instances are classical pronunciations.
A motivating example: Emergence of shared pronunciations

[agent-based modeling experiments from Liberman (2000)]

- Definition of success:
  - Social convergence
    (“people are mostly the same”)
  - Lexical differentiation
    (“words are mostly different”)
- These two properties are required for successful communication
A simplest model

• Individual belief about word pronunciation:
  vector of binary random variables
  e.g. feature #1 is 1 with p=.9, 0 with p=.1
  feature #2 is 1 with p=.3, 0 with p=.7
  ...

• (Instance of) word pronunciation: (random) binary vector
  e.g. 1 0 ...

• Initial conditions: random assignment of values to beliefs of N agents
• Additive noise (models output, channel, input noise)
• Perception: assign input feature-wise to nearest binary vector
  i.e. categorical perception
• Social geometry: circle of pairwise naming among N agents
• Update method: linear combination of belief and perception
  belief is “leaky integration” of perceptions
Coding words as bit vectors

Morpheme template
\[ C_1 V_1 (C_2 V_2 ) (\ldots) \]
Each bit codes for one feature in one position in the template, e.g. “labiality of \( C_2 \)”

Some 5-bit morphemes:
- 11111 \( g^w u \)
- 00000 \( tæ \)
- 01101 \( g a \)
- 10110 \( b i \)

<table>
<thead>
<tr>
<th>Feature</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 ) labial?</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( C_1 ) dorsal?</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( C_1 ) voiced?</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( more C_1 features \ldots )</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>( V_1 ) high?</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( V_1 ) back?</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( more V_1 features \ldots )</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>( g^w u \ldots )</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>( tæ \ldots )</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
</tbody>
</table>
Belief about pronunciation as a random variable

Each pronunciation instance is an N-bit vector
(= feature vector = symbol sequence)
but belief about a morpheme’s pronunciation is a probability distribution over symbol sequences,
encoded as N independent bit-wise probabilities.
Thus [01101] encodes /ga/
but < .1 .9 .9 .1 .9 > is
  [ 0 1 1 0 1 ] = ga with p≈.59
  [ 0 1 1 0 0 ] = gæ with p≈.07
  [ 0 1 0 0 1 ] = ka with p≈.07
  etc. ...
“lexicon”, “speaking”, “hearing”

Each agent’s “lexicon” is a matrix
- whose columns are template-linked features
  - e.g. “is the first syllable’s initial consonant labial?”
- whose rows are words
- whose entries are probabilities
  - “the 3\textsuperscript{rd} word’s 2\textsuperscript{nd} syllable’s vowel is back with p=.973”

**MODEL 1:**
To “speak” a word, an agent “throws the dice”
  to chose a pronunciation (vector of 1’s and 0’s)
  based on that row’s p values
Noise is added (random values like .14006 or .50183)
To “hear” a word, an agent picks the nearest vector of 1’s and 0’s
  (which will eliminate the noise if it was < .5 for a given element)
Updating beliefs

When a word $W_i$ is heard, 
hearer “accomodates” belief about $W_i$ 
in the direction of the perception. 
New belief is a linear combination 
of old belief and new perception: 

$$B_t = \alpha B_{t-1} + (1- \alpha)P_t$$

Old belief = < .1 .9 .9 .1 .9 >
Perception = [ 1 1 1 0 1 ]
New belief = [ .95*.1+.05*1 .95*.9+.05*1 . . . ] 
= [ .145 .905 ... ]
Other issues

• Who talks to whom when?
• How accurate is communication of reference?
• When are beliefs updated?
• Answers don’t seem to be crucial
• In the experiments discussed today:
  – N (imaginary) people are arranged in a circle
  – On each iteration, each person “points and names” for her clockwise neighbor
  – Everyone changes positions randomly after each iteration
• Other geometries (grid, random connections, etc.) produce similar results
• Simultaneous learning of reference from collection of available objects (i.e. no pointing) is also possible
It works!

- Channel noise = gaussian with $\sigma = .2$
- Update constant $\alpha = .8$
- 10 people
- one bit in one word for people #1 and #4 shown:
Gradient output = faster convergence

Instead of saying 1 or 0 for each feature, speakers emit real numbers (plus noise) proportional to their belief about the feature. Perception is still categorical. Result is faster convergence, because better information is provided about the speaker’s internal state.
Gradient input = no convergence

If we make perception gradient (i.e. veridical), then (whether or not production is categorical) social convergence does not occur.
Divergence with population size

With gradient perception, it is not just that pronunciation beliefs continue a random walk over time. They also diverge increasingly at a given time, as group size increases.

20 people:

40 people:
What’s going on?

- Input categorization creates “attractors” that trap beliefs despite channel noise and initially random assignments
- Positive feedback creates social consensus
- Random effects generate lexical differentiation
- Assertions: to achieve social consensus with lexical differentiation, any model of this general type needs
  - stochastic (random-variable) beliefs
    - to allow learning
  - categorical perception
    - to create attractor to “trap” beliefs
Back to the future

- An instance of the “array framework” from Estes 1986
- Categorization by a “feature-frequency model”
- “Linear operator” learning model (e.g. Bush & Mosteller 1951)
- New twist: subjects generate each other’s training material
Linear operator model

• The animal maintains an estimate of resource density for each patch (or response frequency in p-learning)
• At certain points, the estimate is updated
• The new estimate is a linear combination of the old estimate and the “current capture quantity”

Updating equation:

\[ E_n = wE_{n-1} + (1 - w)C \]

w “memory constant”
C “current capture quantity”

Bush & Mosteller (1951), Lea & Dow (1984)
Do the pseudo-probabilities really matter?

How about a model in which belief about the pronunciation of a word is a binary vector rather than a discrete random variable -- or in more anthropomorphic terms, a string of symbols rather than a probability distribution over strings of symbols.

If we have a very regular and reliable arrangement of who speaks to whom when, then success is trivial. Adam tells Eve, Eve tells Cain, Cain tells Abel, and so on. There is a perfect chain of transmission and everyone winds up with Adam's pronunciation.

The trouble is that less regular less reliable conversational patterns, or regular ones that are slightly more complicated, result in populations whose lexicons are blinking on and off like Christmas tree lights.
Consider a circular world, permuted randomly after each conversational cycle, with values updated at the end of each cycle so that each speaker copies exactly the pattern of the "previous" speaker on that cycle. Here's the first 5 iterations of a single feature value for a world of 10 speakers. Rows are conversational cycles, columns are speakers (in "canonical" order).

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Here's another five iterations after 10,000 cycles -- no signs of convergence:

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Even with a combination of update algorithm and conversational geometry that converges, such a system will be fragile in the face of occasional incursions of rogue pronunciations.
Conclusions

For “naming without Adam”, it’s sufficient that
- perception of pronunciation be categorical
- belief about pronunciation be stochastic

Are these conditions also necessary?
No -- there are other ways to get convergence.
But this picture is simple, general, and plausible.
An obvious point

- The same discrete probability distribution can be represented in many ways, e.g.
  - a mathematical recipe
  - a procedural recipe (some PDP models)
  - an empirical histogram of binned counts
  - an urn full of tokens
- Any successful feature-frequency model can be re-implemented as an exemplar model, by calculating feature frequencies as needed

(Leaving dynamics aside)
An equally obvious point

• Most (all?) linguistic arguments for “exemplar theory” are based on:
  – frequency effects (at many levels)
  – context effects (linguistic and non-)
  – analogical effects (type-based)

• There are many alternative ways to model these

• Nearly all formal questions remain on the table
Greg’s crisis of faith…

… was not necessary.

t/d deletion is more likely to happen before vowels
   (an unfavorable environment)
in words that occur often before consonants
   (a favorable environment)

If we pronounce a word by reaching into the urn of all the tokens we’ve ever heard, and picking one at random, this would be true.

But there are many other paths to the same result.

In particular, a simple feature-frequency model would naturally adjust its distribution in the direction of the pronunciations that it experiences.

And such a model trivially allows the kind of stochastic deletion rule that Greg used to believe in.
Conclusion

The classical model of lexical representations is simple, clear, useful, … and untenable.

The turn within linguistics to “exemplar theory” is healthy, fruitful, interesting, … and misleading to many people,

since the research engages many phenomena and many different (implicit) models but the name implies a memory-based model that is rarely necessary and sometimes empirically wrong

Chicago 1/22/2008