Remembering prosody in discourse: Verbatim memory and regeneration

Jennifer Bellik & Tom Roberts
University of California, Santa Cruz
CUNY 2019, CU Boulder | March 30, 2019
What was the title of this talk?
Verbatim vs. gist memory

In memory tasks:

- Memory for verbatim form of expressions (‘surface features’) is good at first, but decays rapidly
- Memory for gist (broad semantic content) is retained much longer

Why is gist memory more durable?

Verbatim memory is illusory. People extract gist from verbatim expressions...

...then discard them, so only gist memory remains. Apparent memory for surface features in recall tasks is reconstructed from gist.

Regeneration Hypothesis (Potter & Lombardi)

Problems for Regeneration

- In P&L’s recall tasks, participants reproduced surface syntax even when ‘meaning-equivalent’ alternatives were available.
- P&L attribute this to syntactic priming, not verbatim memory.

Problems for Regeneration

- Other research suggests a need for both verbatim and gist memory (e.g., Fuzzy Trace Theory)
- Previously encountered fine-grained phonetic detail impacts later processing (cf. Exemplar Theory)
- Many manipulations tested (e.g. ditransitive/double object alternation, active/passive alternation) plausibly differ in information structure

Is there memory for truly surface features?
If so, what role does regeneration play in memory?

**Fuzzy Trace Theory**: Reyna & Brainerd 1995
Research questions

How does memory for prosody work in **spoken** English discourse?

Is there verbatim memory for surface (non-contentful) prosody?

How does prosody impact gist memory?

**Method:** Recognition tasks with auditory presentation

**Experiment 1:** Rhythm Rule

**Experiments 2 & 3:** Regenerating questions from Focal Pitch Accent
Experiment 1: Rhythm Rule

- Is there verbatim memory for features with no interpretive impact?
- Such a surface prosodic feature: Rhythm Rule (RR)
- RR: Move main stress leftward to maintain ‘strong’ beats at regular intervals. (*Optional.*)
- In an XP consisting of two words, the RR can be triggered if the second word has initial primary stress and the first has final or penultimate primary stress

\[
\text{ThirTEEN MEN} \rightarrow \text{THIRteen MEN}
\]

\[
\text{hypoDERmic NEEdle} \rightarrow \text{HYpodermic NEEdle}
\]

Liberman & Prince 1977, Hayes 1984, et seq
Procedure: Recall Task

**Stimulus**
- Listen to the dialogue.

**Distractor task**
- 15 - 9 = ?

**Probe**
- Did you hear this?

**Confidence rating**
- How confident are you of your response? (1-4 scale)
Exp. 1 Materials

Items (24) = 3-6 line dialogues between speakers A & B.
- Last line (probe) was spoken by B and ended with the RR target region
- RR target region = Adj+N pair, utterance final
- New vs. Old probes differed *only* in the last two words (spliced in)

Fillers (48):
- Counterbalanced position of probe in dialogue (2-4 sentences back)
- whether probe was spoken by A or B
- Manipulated both local prosody and sentence-level tunes, but new probes were still string-identical
B: How is your landscaping project going?
A: I'm having trouble finding good plants for this climate.
B: There are actually a lot of shrubs that can survive cold weather.
A: What's an example?

B: Well, there's the MINnesotan YUCca [+RR]
    MinneSOtan YUCca [-RR]
Exp. 1 Method

2x2 design: [±RR] x [Probe New/Old]

Participants (n=36):
- UCSC undergraduates
- 3 excluded for not learning English before age 5
- 33 total analyzed
Exp. 1 Predictions

If surface stress is stored verbatim in memory:

Participants will be sensitive to changes in application of the rhythm rule.

If surface stress is not stored verbatim in memory:

Participants will be insensitive to these changes.

If +/-RR application is in free variation:

No effect of ±RR.
Analytical Method: Detection Theory

AUC = measure of sensitivity
- Area under the Receiver Operating Characteristic curve
- ROC calculated using the Ratings Method
- AUC > 0.5 = sensitivity

$c$ = criterion = measure of response bias
- Negative $c$ → ‘yes’ bias

Macmillan & Creelman 2005
Exp. 1 Results: New vs. Old Probes

- **AUC > 0.5**: Participants can discriminate between new and old probes.
- **c < 0**: Bias to respond “yes, I heard this before”

<table>
<thead>
<tr>
<th></th>
<th>Experimental Items</th>
<th>Fillers</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC (95% CI)</td>
<td>0.590 (0.557-0.623)</td>
<td>0.733 (0.708-0.758)</td>
</tr>
<tr>
<td>c</td>
<td>-1.05</td>
<td>0.1</td>
</tr>
</tbody>
</table>

MacMillan & Creelman 2005, Robin et al. 2011
Exp. 1 Discussion

- Although the deck was stacked against participants, they were significantly (albeit weakly) sensitive to changes in RR application.
- No effect of whether initial presentation was +RR or -RR: Sensitivity really reflects access to a verbatim representation, not regeneration.

**Upshot:** Non-contentful surface prosody can be remembered, with some difficulty.

**Next question:** Can prosody also have an impact on gist memory?
Experiments 2 & 3: Focal Pitch Accent

Prosody prototypically conveys “higher” discourse information.

A prosodic feature with interpretive impact: Focal Pitch Accent (FPA)

- FPA indicates that a constituent is interpreted under exhaustive focus.
- Felicitous answers must put FPA only on the constituent corresponding to the question’s wh-word. (ex. Gertrude in SA, kale in S2)

<table>
<thead>
<tr>
<th>SQ: Who massaged the kale?</th>
<th>OQ: What did Gertrude massage?</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA: [F\text{Gertrude}] massaged the kale.</td>
<td>OA: Gertrude massaged the [F\text{Kale}].</td>
</tr>
</tbody>
</table>
Experiments 2 & 3: Focal Pitch Accent

- In case of a mismatch between answer FPA and question wh-word, will questions be reconstructed to match answer FPA?
- Experiment 2: Memory for interrogatives
- Experiment 3: Prosodic mismatch + regeneration

<table>
<thead>
<tr>
<th>Hear:</th>
<th>Q1: Who massaged the kale?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A2: Gertrude massaged the ([<em>F^KALE</em>]).</td>
</tr>
<tr>
<td>Regenerate:</td>
<td>Q2: What did Gertrude massage?</td>
</tr>
</tbody>
</table>

Selkirk 1995
Experiment 2: Interrogatives vs. Declaratives

- Previous work focuses on “gist” memory for declarative sentences.
- But what is the “gist” of a question?
- Prevailing theoretical accounts of question meaning treat it as a set of propositions: i.e., a set of declarative denotations.
  - Translates to penalty for remembering interrogatives as opposed to declaratives?

Exp. 2 Method

Auditory recognition task, same procedure as Exp. 1

2x2 design: [Interrogative / Declarative] x [Probe New/Old]

Participants (n=55):
  - UCSC undergraduates who learned English by age 5
Exp. 2 Materials

Items (24) consisted of 5-8 line conversations between A & B
● Target = penultimate line; Interrogative or Declarative
● Last line = felicitous response; same in both conditions

Fillers (48) counterbalanced
● Position of the probe appeared in the dialogue (1, 3, or 5 sentences back)
● Whether probe was spoken by A or B

<table>
<thead>
<tr>
<th>Target + felicitous response</th>
<th>Old probe</th>
<th>New probe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interrogative</td>
<td>A: Who massaged the kale? B: [Gertrude]$_F$ massaged the kale.</td>
<td>Who massaged the kale?</td>
</tr>
</tbody>
</table>
Exp. 2 Predictions

If gist is related to **propositional content**...

Changed interrogatives will be harder to remember than changed declaratives, because interrogative meaning is strictly more complex.

Both declarative and interrogative probe changes involve changes in lexical items, so there are strong cues to recognition of changes in both.
Exp. 2 Results

Recall of declaratives and interrogatives both extremely good and not significantly different.

→ Established baseline for Exp. 3
Experiment 3 (pilot): Regeneration of questions

**Experiment 1**: Stress is remembered independently of interpretative impact.

**Experiment 2**: Recognition memory for interrogatives is very good.

**Experiment 3**: Will questions be regenerated based on answer prosody?

Test case: Q-A focus mismatch.

<table>
<thead>
<tr>
<th>Q1: Who massaged the kale?</th>
<th>Q2: What did Gertrude massage?</th>
</tr>
</thead>
<tbody>
<tr>
<td>![fGERTRUDE] massaged the kale.</td>
<td>A2: Gertrude massaged the ![fKALE].</td>
</tr>
</tbody>
</table>
Exp. 3 Method

2x2 design: [Felicitous/Infelicitous] x [Probe New/Old]

Participants (n=16): UCSC graduate and undergraduate students

Items (n=40):
- Question-Answer pairs produced by 2 speakers
  - Questions involve wh-extraction of the subject or object.
  - Answers put FPA on the subject or object.
- Presented via Latin square without fillers.

Procedure differed from Expt. 1 & 2 in using a distractor melody (no math problem).
Exp. 3 Results (pilot) - Question Probes

Sensitivity to probe newness was better for questions followed by felicitous answers (though not by much).

<table>
<thead>
<tr>
<th></th>
<th>Felicitous Response</th>
<th>Infelicitous Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC (95% CI)</td>
<td>0.928 (0.880-0.969)</td>
<td>0.890 (0.836-0.938)</td>
</tr>
<tr>
<td>c</td>
<td>-0.47</td>
<td>-0.36</td>
</tr>
</tbody>
</table>
Exp. 2 & 3 Discussion

- Exp 2 suggests that interrogatives are easily recognized when their follow-up response is felicitous.
  - Possible that lexical/syntactic changes are simply an extremely strong cue.
  - A recall task might show less of a ceiling effect.
- Exp 3 suggests that infelicitous prosody on a response may interfere with recall of a question.
- This suggests that prosodic information (at least FPA) may also be used to facilitate memory of prior discourse moves.
Conclusion

Prosody appears to play a role in both verbatim and gist memory:

★ Surface (prosodic) features of language in spoken dialogues are stored in memory, not just gist + regeneration
★ However, incongruous answer prosody appears to be able to trigger erroneous recall of questions, suggesting that remembered prosody may impact gist memory.

Future work:
● Comparison of reconstruction of interrogatives vs. declaratives
● Effects on memory of highly interpretative semantic content, e.g. intonational tunes
Thank you!

Special thanks to Steven Foley, Jed Pizarro-Guevara, Amanda Rysling, Matt Wagers, audiences at UCSC S/Lab and CAMP 2, and RAs Madeleine King and Koy Rugama.

Contact: {jbellik,rotom}@ucsc.edu
References


References

Exp. 1 Item Creation

- For target sentences, two versions were recorded: one with stress shift and one without
- The authors selected whichever version sounded more natural (15/24 originally carried the stress shift) and spliced out the final two words for items with the other stress patterns
- Double-checked that splices did not sound obvious
- Anticipation of upcoming stress might affect prosody in item production
  - **BUT**: No significant difference between % correct for items whose carrier phrase was from +RR or -RR recordings ($\chi^2(1) = 0.026$)
# Exp. 1 Response results

<table>
<thead>
<tr>
<th>Stress shift in stimulus [+RR]</th>
<th>Did you hear this before?</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Target = “MINnesotan YUCca”</td>
<td>“Yes”</td>
<td>“No”</td>
</tr>
<tr>
<td>Old Probe “MINnesotan YUCca”</td>
<td>197 (Hits)</td>
<td>17 (Misses)</td>
</tr>
<tr>
<td>New Probe “MinneSOtan YUCca”</td>
<td>157 (False Alarms)</td>
<td>55 (Correct Rejections)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No stress shift in stimulus [-RR]</th>
<th>Did you hear this before?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target = “MinneSOtan YUCca”</td>
<td>“Yes”</td>
</tr>
<tr>
<td>Old Probe “MinneSOtan YUCca”</td>
<td>195 (Hits)</td>
</tr>
<tr>
<td>New Probe “MINnesotan YUCca”</td>
<td>167 (False Alarms)</td>
</tr>
<tr>
<td></td>
<td>45 (Correct Rejections)</td>
</tr>
</tbody>
</table>
Exp. 1 Confidence Results

- “Yes” bias:
  - Accuracy is much higher for Old items (92%) than New (24%).
  - “No” responses were more likely to be correct, but received lower confidence ratings

<table>
<thead>
<tr>
<th></th>
<th>1 (Not sure)</th>
<th>2</th>
<th>3</th>
<th>4 (Very sure)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Yes”</td>
<td>2</td>
<td>28</td>
<td>156</td>
<td>530</td>
</tr>
<tr>
<td>“No”</td>
<td>1</td>
<td>37</td>
<td>57</td>
<td>41</td>
</tr>
</tbody>
</table>

- Overall proportion correct (57%) is not greater than chance (By items: $X^2(23)=17.9, p=0.76$; by subjects: $X^2(32) = 32.15, p=0.46$), but proportion correct doesn’t capture sensitivity well with such a large bias.
Exp. 1 d’ by item

Overall d’=0.66
Exp. 1 $d'$ by subject

Overall $d'=0.66$
Why not just report chi-squared?

Not as useful a metric as AUC when there is bias

Chi-squared tests of proportion correct were not significant by subject or item for Exp. 1 (p>>0.05), due to the huge yes-bias.

But AUC + c shows that there is significant sensitivity here.