An Analysis of Frequency- and Memory-Based Processing Costs

Marten van Schijndel  William Schuler
Department of Linguistics
The Ohio State University

June 10, 2012
Observation isn’t explanation

Many current metrics predict complexity with no cognitive explanation.

- Surprisal and entropy reduction reflect corpus statistics.
Motivation

Observation isn’t explanation

Many current metrics predict complexity with no cognitive explanation.

- Surprisal and entropy reduction reflect corpus statistics.

Goal: An Explanation

- How do current theories of working memory fit with current theories of language processing?
- Do memory effects predict difficulty over frequency effects?
- Provide a rationale for why humans have certain difficulties
Hypothesis

Memory effects cause processing difficulty beyond frequency effects
Overview

Hypothesis
Memory effects cause processing difficulty beyond frequency effects

1. Working memory primer
2. Memory and language processing theories
3. Introduce connected component parser
4. Eye-tracking evaluation
5. Results
**Working Memory**

**Temporal and Sequential Cueing**

Temporal Context Model [Howard and Kahana, 2002]
Hierarchic Sequential Prediction [Botvinick, 2007]

- Learned *sequential* associations
- Contextual *temporal* associations
**Working Memory**

**Temporal and Sequential Cueing**

Temporal Context Model [Howard and Kahana, 2002]
Hierarchic Sequential Prediction [Botvinick, 2007]

- Learned *sequential* associations
- Contextual *temporal* associations

Diagram:

```
  Making Tea
    /   \\   /
   Heat Water    Brush Teeth    Steep Tea
```

Temporal Cueing in the Morning
## Working Memory

### Temporal and Sequential Cueing

- Temporal Context Model [Howard and Kahana, 2002]
- Hierarchic Sequential Prediction [Botvinick, 2007]
  - Learned *sequential* associations
  - Contextual *temporal* associations

### Focus

Attended vs Passive States [McElree, 2006]
**Working Memory**

**Temporal and Sequential Cueing**

Temporal Context Model [Howard and Kahana, 2002]
Hierarchic Sequential Prediction [Botvinick, 2007]

- Learned *sequential* associations
- Contextual *temporal* associations

**Focus**

Attended vs Passive States [McElree, 2006]
**Working Memory**

**Temporal and Sequential Cueing**

Temporal Context Model [Howard and Kahana, 2002]
Hierarchic Sequential Prediction [Botvinick, 2007]

- Learned *sequential* associations
- Contextual *temporal* associations

**Focus**

Attended vs Passive States [McElree, 2006]
**Working Memory**

**Temporal and Sequential Cueing**

Temporal Context Model [Howard and Kahana, 2002]
Hierarchic Sequential Prediction [Botvinick, 2007]

- Learned *sequential* associations
- Contextual *temporal* associations

**Focus**

Attended vs Passive States [McElree, 2006]
Temporal and Sequential Cueing

Temporal Context Model [Howard and Kahana, 2002]
Hierarchic Sequential Prediction [Botvinick, 2007]
- Learned *sequential* associations
- Contextual *temporal* associations

Focus

Attended vs Passive States [McElree, 2006]

Difficulty with
- Temporal cueing (Accessing non-focused information)
- Resolving embedded dependencies

Key: Inhibition Facilitation
Dependency Locality Theory [Gibson, 2000]

**Difficulty with**
- Unresolved dependencies

**Storage cost**
- Beginning dependencies
- Maintaining dependencies

**Integration cost**
- Resolving dependencies
ACT-R [Lewis et al., 2006]

**Difficulty with**
- Activation decay
- Similarity interference

**Encoding cost**
- Beginning a new dependency

**Retrieval cost**
- Resolving a dependency

Retrieval can be *facilitated* by re-activations.
Dynamic Recruitment [Just and Varma, 2007]
Difficult constructions $\rightarrow$ extra processing resources

**Difficulty with**
- Center embeddings

**Recruitment**
- Beginning embeddings

**Release**
- Completing embeddings
Embedding Difference [Wu et al., 2010]

*Increased embedding depth* \{ Beginning embeddings \\
*Reduced embedding depth* \{ Completing embeddings
Connected Components

'S/NP' and 'NP/N' represent unresolved dependencies
Predictions

<table>
<thead>
<tr>
<th>Theory</th>
<th>Encoding</th>
<th>Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hier. Sequential Prediction</td>
<td>positive</td>
<td>positive</td>
</tr>
<tr>
<td>Dependency Locality Theory</td>
<td>positive</td>
<td>positive</td>
</tr>
<tr>
<td>ACT-R</td>
<td>positive</td>
<td>positive</td>
</tr>
<tr>
<td>Dynamic Recruitment</td>
<td>positive</td>
<td>negative</td>
</tr>
<tr>
<td>Embedding Difference</td>
<td>positive</td>
<td>negative</td>
</tr>
</tbody>
</table>

Predicted correlation of parse operations to reading times under each theory
**Connected Component Parsing**

```
S
     /\  
    NP  VP
   /   /
  D   N
    |   |
   the studio
```

**Working Memory:**

```
S/VP
```
S

NP     VP

D       N     V     NP

the    studio  bought

Working Memory:

S/NP
Connected Component Parsing

S

NP

D

the

N

studio

VP

I

bought

NP

D

np

the

Working Memory:

S/NP

NP/N
**Connected Component Parsing**

van Schijndel, Schuler

**Frequency and Memory Costs**

June 10, 2012 11 / 31
The diagram illustrates a parse tree for the sentence "the studio bought the author's rights." The tree is structured with a main clause (S) containing a noun phrase (NP) and a verb phrase (VP). The noun phrase is further divided into a determiner (D), a noun (N), and a possessed noun phrase (NP) indicating the author's rights. The verb phrase consists of a verb (V) "bought." The diagram represents how sentences are broken down into constituent parts for analysis in the context of working memory.
F and L binary decisions (+,−) made at each timestep

- **F(irst)**: Current word is the first element of a new embedding
- **L(ast)**: Current word is the last element of an embedding

Only one F, only one L [van Schijndel et al, 2013]
**Parser Operations**

F and L binary decisions (+,−) made at each timestep

- **F(irst):** Current word is the first element of a new embedding
- **L(ast):** Current word is the last element of an embedding

Only one F, only one L [van Schijndel et al, 2013]

- **F+L− (Encode):** Create a new connected component

```
       S
      / \
 NP   VP
 / \  /  \\
 D   N V   NP
  the studio bought

       S
      / \
 NP   VP
 / \  /  \\
 D   N V   NP
  the studio bought

NP
  the
```
**Parser Operations**

F and L binary decisions (+,–) made at each timestep

- **F(irst):** Current word is the *first* element of a new embedding
- **L(ast):** Current word is the *last* element of an embedding

Only one F, only one L [van Schijndel et al, 2013]

- F+L– (Encode): Create a new connected component
- F–L+ (Integrate): Combine two connected components

---

![Diagram]

Integrate
Eye Tracking

- Assumption: Slower reading = difficulty
- How much can be processed up to a given point?
- Many different metrics (fixation duration, regression, etc)
Eye Tracking

- Assumption: Slower reading = difficulty
- How much can be processed up to a given point?
- Many different metrics (fixation duration, regression, etc)

Measure of choice: Go-Past Duration [Clifton et al., 2007]
Go-past durations:

John went to the shop today

Cumulative factors are summed over the go-past region
Non-cumulative factors are based on the initial word in a region (shop)
Go-past durations:

John went to the shop today

Cumulative factors are summed over the go-past region
Non-cumulative factors are based on the initial word in a region (shop)
Training

Parser accuracy is comparable to Berkeley [van Schijndel et al., 2012]

- Parser and Lexicon: WSJ02-21 [Marcus et al., 1993]
  - 39,832 sentences
  - 950,028 words
- Ngrams: Brown [Francis and Kucera, 1979], WSJ02-21, BNC, Dundee [Kennedy et al., 2003]
  - 5,052,904 sentences
  - 87,302,312 words

Ngrams calculated using SRILM [Stolcke, 2002] with modified Kneser-Ney smoothing [Chen and Goodman, 1998]
Evaluation

- Dundee corpus [Kennedy et al., 2003]
  - 10 subjects
  - 2,388 sentences
  - 58,439 words
  - 260,124 go-past durations
- Filtered Dundee corpus
  - 154,168 go-past durations

Exclusions: UNK-threshold 5, first and last of a line, fixations skipping more than 4 words (track/attention loss)

Metric Calculations: Probability-weighted, parallel model
**Baseline Metrics**

Fitting a linear mixed effects model (*lmer* in R)

### Fixed Effects

- Word length
- Sentence position
- Prev, Next word fixated?
- Unigram and bigram probs
- Surprisal
- Region length
- Cumulative surprisal
- Cumulative entropy reduction
- Joint interactions
- Spillover predictors

### By-subject random slopes (Note: Not in paper)

- Effect of interest (e.g. Encode)
- Prev word fixated?
- Cumulative surprisal
- Region length

With Subject and Item random intercepts

Fit to log-transformed durations
## Predictions - Revisited

<table>
<thead>
<tr>
<th>Theory</th>
<th>Encoding</th>
<th>Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hier. Sequential Prediction</td>
<td>positive</td>
<td>positive</td>
</tr>
<tr>
<td>Dependency Locality Theory</td>
<td>positive</td>
<td>positive</td>
</tr>
<tr>
<td>ACT-R</td>
<td>positive</td>
<td>positive</td>
</tr>
<tr>
<td>Dynamic Recruitment</td>
<td>positive</td>
<td>negative</td>
</tr>
<tr>
<td>Embedding Difference</td>
<td>positive</td>
<td>negative</td>
</tr>
</tbody>
</table>

Predicted correlation of parse operations to reading times under each theory.
## Results

<table>
<thead>
<tr>
<th>Operation</th>
<th>Factor</th>
<th>Coeff</th>
<th>Std. Error</th>
<th>t-score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoding</td>
<td>F+L–</td>
<td>0.023</td>
<td>0.005</td>
<td>4.238</td>
<td>0.001</td>
</tr>
<tr>
<td>Integration</td>
<td>F–L+</td>
<td>-0.015</td>
<td>0.005</td>
<td>-3.215</td>
<td>0.007</td>
</tr>
<tr>
<td>Cue Active</td>
<td>F–L–</td>
<td>0.002</td>
<td>0.003</td>
<td>0.800</td>
<td>0.437</td>
</tr>
<tr>
<td>Cue Awaited</td>
<td>F+L+</td>
<td>-0.004</td>
<td>0.003</td>
<td>-1.298</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Significance of Improvement over Baseline

Each FL factor is cumulative
• No positive integration cost with frequency
Conclusion

• No positive integration cost with frequency
• Significant negative integration cost
Conclusion

- No positive integration cost with frequency
- Significant negative integration cost
- Supports: Dynamic Recruitment, Embedding Difference
**Conclusion**

- No positive integration cost with frequency
- Significant negative integration cost
- Supports: Dynamic Recruitment, Embedding Difference

- No evidence of DLT’s maintenance cost
CONCLUSION

- No positive integration cost with frequency
- Significant negative integration cost
- Supports: Dynamic Recruitment, Embedding Difference

- No evidence of DLT’s maintenance cost
- Confounds assumption of Slow = Difficult
Conclusion

- No positive integration cost with frequency
- Significant negative integration cost
- Supports: Dynamic Recruitment, Embedding Difference
- No evidence of DLT’s maintenance cost
- Confounds assumption of Slow = Difficult
- Remaining inhibition suggests difficulty beyond frequency effects (perhaps a cause of frequency effects)
Thanks!
Thanks to Kodi Weatherholtz and Rory Turnbull for their assistance with R-wrangling and working with linear mixed effect models!

Thanks to Peter Culicover, Micha Elsner, and the OSU CompLing group for feedback on the project.

Questions?
**Frequency Effects**

**Surprisal** [Hale, 2001]

Predictability of a word given the context:

\[
surprisal(x_t) = -\log_2 \left( \frac{\sum_{s \in S(x_1...x_t)} P(s)}{\sum_{s \in S(x_1...x_{t-1})} P(s)} \right)
\] (1)

**Entropy Reduction** [Hale, 2003]

Entropy is a measure of uncertainty:

\[
H(x_1...t) = \sum_{s \in S(x_1...x_t)} -P(s) \cdot \log_2 P(s)
\] (2)

The reduction in uncertainty caused by observing \(x_t\):

\[
\Delta H(x_1...t) = \max(0, H(x_1...t-1) - H(x_1...t))
\] (3)

\(S(x_1...x_t) = \) trees whose leaves have \(x_1...x_t\) as a prefix
Eye Tracking

Go-past durations:

John went to the shop today

Cumulative factors are summed over the go-past region
Non-cumulative factors are based on the initial word in a region (shop)
Go-past durations:

John went to the shop today

\( X \) = Go-past region

Red = Fixation in go-past duration

Cumulative factors are summed over the go-past region
Non-cumulative factors are based on the initial word in a region (shop)
Eye Tracking

Go-past durations:

John went to the shop today

\( X \) = Go-past region

Red = Fixation in go-past duration

Cumulative factors are summed over the go-past region
Non-cumulative factors are based on the initial word in a region (shop)
Eye Tracking

Go-past durations:

John went to the shop today

X = Go-past region

Red = Fixation in go-past duration

Cumulative factors are summed over the go-past region
Non-cumulative factors are based on the initial word in a region (shop)
Go-past durations:

John went to the shop today

$X$ = Go-past region

Red = Fixation in go-past duration

Cumulative factors are summed over the go-past region
Non-cumulative factors are based on the initial word in a region (shop)
Transforming the response variable

Histogram of `data.dev$fdur`

Frequency

0 10000 20000 30000 40000 50000

`data.dev$fdur`

0 1000 2000 3000 4000
TRANSFORMING THE RESPONSE VARIABLE

Histogram of $\log(\text{data.dev$fdur$})$


Bibliography II

The dependency locality theory: A distance-based theory of linguistic complexity.
In Image, language, brain: Papers from the first mind articulation project symposium, pages 95–126, Cambridge, MA. MIT Press.

A probabilistic earley parser as a psycholinguistic model.
In Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics, pages 159–166, Pittsburgh, PA.

Grammar, Uncertainty and Sentence Processing.
Bibliography III


