HIERARCHIC SYNTAX IMPROVES READING TIME PREDICTION

Marten van Schijndel and William Schuler
Department of Linguistics
The Ohio State University
June 3, 2015
Previous studies have debated whether humans use hierarchic syntax.
Previous studies have debated whether humans use hierarchic syntax.
Previous studies have debated whether humans use hierarchic syntax.

But standard baseline predictors may be deficient.
This work shows that:

1. Baselines can be greatly improved (accumulation).
2. Hierarchic syntax is still predictive over stronger baseline.
3. Hierarchic syntax not improved by accumulation.
4. Long distance dependencies independently improve model.
This work shows that:
Baselines can be greatly improved (accumulation)
This work shows that:
Baselines can be greatly improved (accumulation)
Hierarchic syntax is still predictive over stronger baseline
This work shows that:
Baselines can be greatly improved (accumulation)
Hierarchic syntax is still predictive over stronger baseline
Hierarchic syntax not improved by accumulation
This work shows that:
Baselines can be greatly improved (accumulation)
Hierarchic syntax is still predictive over stronger baseline
Hierarchic syntax not improved by accumulation
Long distance dependencies independently improve model
The red apple that the girl ate ...
The red apple that the girl ate ...
The red apple that the girl ate ...

**Frank & Bod (2011)**

Baseline:
- Sentence Position
- Word length
- \(N\)-grams (Unigram, bigram)
The red apple that the girl ate ...

**FRANK & BOD (2011)**

Baseline:
- Sentence Position
- Word length
- $N$-grams (Unigram, bigram)
The red apple that the girl ate ...

FRANK & BOD (2011)

Baseline:
- Sentence Position
- Word length
- $N$-grams (Unigram, bigram)
The red apple that the girl ate ...

**Frank & Bod (2011)**

**Baseline:**
- Sentence Position
- Word length
- N-grams (Unigram, bigram)

**Test POS Predictors:**
- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)
Hierarchic Syntactic in Reading?

Frank & Bod (2011)

Baseline:
- Sentence Position
- Word length
- N-grams (Unigram, bigram)

Test POS Predictors:
- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)
Frank & Bod (2011)

Baseline:
- Sentence Position
- Word length
- N-grams (Unigram, bigram)

Test POS Predictors:
- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)
## Frank & Bod (2011)

### Baseline:
- Sentence Position
- Word length
- $N$-grams (Unigram, bigram)

### Test POS Predictors:
- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

Outcome:
- $PSG < ESN + PSG$
- $ESN = ESN + PSG$
- Hierarchic doesn’t help over sequential

---

van Schijndel and Schuler

Hierarchic Syntactic Reading Times

June 3, 2015
**Frank & Bod (2011)**

Baseline:
- Sentence Position
- Word length
- $N$-grams (Unigram, bigram)

Outcome:
- $\text{PSG} < \text{ESN} + \text{PSG}$
- $\text{ESN} = \text{ESN} + \text{PSG}$

Test POS Predictors:
- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)
<table>
<thead>
<tr>
<th><strong>FRANK &amp; BOD (2011)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline:</strong></td>
</tr>
<tr>
<td>- Sentence Position</td>
</tr>
<tr>
<td>- Word length</td>
</tr>
<tr>
<td>- N-grams (Unigram, bigram)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Test POS Predictors:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>- Echo State Network (ESN)</td>
</tr>
<tr>
<td>- Phrase Structure Grammar (PSG)</td>
</tr>
</tbody>
</table>

**Outcome:**
- PSG $<$ ESN + PSG  
  Sequential helps over hierarchic
- ESN $=$ ESN + PSG
## Frank & Bod (2011)

**Baseline:**
- Sentence Position
- Word length
- $N$-grams (Unigram, bigram)

**Test POS Predictors:**
- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

**Outcome:**
- $\text{PSG} < \text{ESN + PSG}$
- $\text{ESN} = \text{ESN + PSG}$  
  Hierarchic doesn’t help over sequential
Fossum & Levy (2012)

Replicated Frank & Bod (2011):
PSG < ESN + PSG
ESN = ESN + PSG
**FOSSUM & LEVY (2012)**

Replicated Frank & Bod (2011):

\[ \text{PSG} \prec \text{ESN} + \text{PSG} \]
\[ \text{ESN} = \text{ESN} + \text{PSG} \]

Better \( n \)-gram baseline (more data) changes result:

\[ \text{PSG} \equiv \text{ESN} + \text{PSG} \]
\[ \text{ESN} = \text{ESN} + \text{PSG} \]
**Fossum & Levy (2012)**

Replicated Frank & Bod (2011):
PSG < ESN + PSG  
ESN = ESN + PSG

Better \(n\)-gram baseline (more data) changes result:
PSG \(\equiv\) ESN + PSG  
Sequential doesn’t help over hierarchic  
ESN = ESN + PSG
Replicated Frank & Bod (2011):
PSG < ESN + PSG
ESN = ESN + PSG

Better $n$-gram baseline (more data) changes result:
PSG $\equiv$ ESN + PSG  Sequential doesn’t help over hierarchic
ESN = ESN + PSG

Also: lexicalized syntax improves PSG fit
Previous reading time studies:

- Unigrams/Bigrams/Trigrams
  Trained on WSJ, Dundee, BNC
Previous reading time studies:

- Unigrams/Bigrams/Trigrams
  Trained on WSJ, Dundee, BNC
- Only from region boundaries
**B I G R A M  E X A M P L E**

Reading time of *girl* after *red*

The red apple that the **girl** ate ...

\[ \text{region} \]

\[ X: \text{bigram target} \quad X: \text{bigram condition} \]
**BIGRAM EXAMPLE**

Reading time of *girl* after *red*

1. The red apple that the *girl* ate ...

   - Region

   - $X$: bigram target
   - $X$: bigram condition

• Fails to capture entire sequence;
BIGRAM EXAMPLE

Reading time of *girl* after *red*

1. The red apple that the *girl* ate ...

   region

   X: bigram target  X: bigram condition

- Fails to capture entire sequence;
- Conditions never generated;
**BIGRAM EXAMPLE**

Reading time of *girl* after *red*

The red apple that the **girl** ate ...

\[ \text{region} \]

**X**: bigram target  \quad **X**: bigram condition

- Fails to capture entire sequence;
- Conditions never generated;
- Probability of sequence is deficient
**Cumulative Bigram Example**

Reading time of *girl* after *red*:

The red **apple** that **the** **girl** ate ...  

X: bigram targets         X: bigram conditions

---

*van Schijndel and Schuler    Hierarchic Syntactic Reading Times*  
*June 3, 2015*  8 / 1
CUMULATIVE BIGRAM EXAMPLE

Reading time of girl after red:

The \textbf{red apple that the girl ate ...}

\textbf{X}: bigram targets \hspace{1cm} \textbf{X}: bigram conditions

- Captures entire sequence;
- Well-formed sequence probability;
- Reflects processing that must be done by humans
Previous reading time studies:

- Unigrams/Bigrams/Trigrams
- Trained on WSJ, Dundee, BNC
- Only from region boundaries
Previous reading time studies:

- Unigrams/Bigrams/Trigrams
- Trained on WSJ, Dundee, BNC
- Only from region boundaries

This study:

- 5-grams (w/ backoff)
- Trained on Gigaword 4.0
- Cumulative and Non-cumulative
Dundee Corpus (Kennedy et al., 2003)

- 10 subjects
- 2,388 sentences
- 58,439 words
- 194,882 first pass durations
- 193,709 go-past durations

Exclusions:

- Unknown words (5 tokens)
- First and last of a line
- Regions larger than 4 words (track loss)
Baseline:
Fixed Effects
- Sentence Position
- Word length
- Region Length
- Preceding word fixated?

Random Effects
- Item/Subject Intercepts
- By Subject Slopes:
  - All Fixed Effects
  - $N$-grams (5-grams)
  - $N$-grams (Cumu-5-grams)
Baseline:
Fixed Effects
- Sentence Position
- Word length
- Region Length
- Preceding word fixated?

Random Effects
- Item/Subject Intercepts
- By Subject Slopes:
  - All Fixed Effects
  - $N$-grams (5-grams)←
  - $N$-grams (Cumu-5-grams)←
CUMU-\textit{n}-GRAMS PREDICT READING TIMES

First Pass

Log-likelihood

Baseline

+N-gram

+Cumu-N-gram

+Both
CUMU-\textit{n}-GRAMS PREDICT READING TIMES

First Pass

Go-Past
FOLLOW-UP QUESTIONS

• Is hierarchic surprisal useful over the better baseline?
• If so, can it be similarly improved through accumulation?

van Schijndel & Schuler (2013) found it could over weaker baselines
• Is hierarchic surprisal useful over the better baseline?
FOLLOW-UP QUESTIONS

- Is hierarchic surprisal useful over the better baseline?
- If so, can it be similarly improved through accumulation?
FOLLOW-UP QUESTIONS

- Is hierarchic surprisal useful over the better baseline?
- If so, can it be similarly improved through accumulation?
  van Schijndel & Schuler (2013) found it could over weaker baselines

Grammar:
Berkeley parser, WSJ, 5 split-merge cycles (Petrov & Klein 2007)
Baseline:
Fixed Effects
• Same as before
• $N$-grams (5-grams)
• $N$-grams (Cumu-5-grams)
Baseline:

Fixed Effects
- Same as before
- $N$-grams (5-grams)
- $N$-grams (Cumu-5-grams)

Random Effects
- Same as before
- By Subject Slopes:
  - Hierarchic surprisal
  - Cumu-Hierarchic surprisal
Baseline:
Fixed Effects
• Same as before
• $N$-grams (5-grams)
• $N$-grams (Cumu-5-grams)

Random Effects
• Same as before
• By Subject Slopes:
  • Hierarchic surprisal ←
  • Cumu-Hierarchic surprisal ←
Hierarchic surprisal predicts reading times

First Pass and Go-Past

Diagram:

```
Log-likelihood

Base → +CumuSurp
+Surp → +Both
```

First Pass and Go-Past
• Suggests previous findings were due to weaker $n$-gram baseline
CUMULATIVE SURPRISAL DOESN’T HELP?!

- Suggests previous findings were due to weaker $n$-gram baseline
- Suggests only local PCFG surprisal affects reading times
CUMULATIVE SURPRISAL DOESN’T HELP?!

- Suggests previous findings were due to weaker n-gram baseline
- Suggests only local PCFG surprisal affects reading times

But... long-distance dependencies should affect reading times!
CUMULATIVE SURPRISAL DOESN’T HELP?!

- Suggests previous findings were due to weaker \( n \)-gram baseline
- Suggests only local PCFG surprisal affects reading times

But... long-distance dependencies should affect reading times!

Let’s try a PCFG that tracks long-distance deps
Nguyen et al. (2012)
Nguyen et al. (2012)
Baseline:
Fixed Effects
• Same as before

Random Effects
• Same as before
• By Subject Slopes:
  • Hierarchic PTB surprisal
  • Hierarchic GCG surprisal
Baseline:

Fixed Effects

• Same as before

Random Effects

• Same as before

• By Subject Slopes:
  • Hierarchic PTB surprisal ←
  • Hierarchic GCG surprisal ←
LONG-DISTANCE SURPRISAL PREDICTS READING TIMES

First Pass and Go-Past

Log-likelihood

+Both

+PTB +GCG

Baseline
First Pass and Go-Past

Both help independently
Hierarchic syntax predicts reading times over strong linear baseline
Hierarchic syntax predicts reading times over strong linear baseline

Long-distance dependencies *do* affect reading times
Hierarchic syntax predicts reading times over strong linear baseline

Long-distance dependencies *do* affect reading times

Studies should use cumu-*n*-grams in their baselines
FUTURE WORK

Compare to Echo State Networks
FUTURE WORK

Compare to Echo State Networks

Test anticipatory accumulation
Thanks to:

- Stefan Frank
- Attendees of CUNY 2015
- National Science Foundation (DGE-1343012)
First Pass Evaluation (Log-Likelihood):

<table>
<thead>
<tr>
<th></th>
<th>Base</th>
<th>Base+(N)-gram</th>
<th>Base+(Cumu-n)-gram</th>
<th>Base+Both</th>
<th>Base+Both</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(-1212399)</td>
<td>(-1212396) ((p &lt; 0.05))</td>
<td>(-1212392) ((p &lt; 0.01))</td>
<td>(-1212387) ((p &lt; 0.01))</td>
<td>(-1212387) ((p &lt; 0.01))</td>
</tr>
</tbody>
</table>
CUMU-\(n\)-GRAMS PREDICT READING TIMES

Comparable with go-past durations

<table>
<thead>
<tr>
<th>Go-Past Evaluation (Log-Likelihood):</th>
<th>Base</th>
<th>Base+Cumu-(n)-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>-1261582</td>
<td>-1261576 ((p &lt; 0.01))</td>
</tr>
<tr>
<td>Base+N-gram</td>
<td>-1261577 ((p &lt; 0.01))</td>
<td></td>
</tr>
<tr>
<td>Base+Both</td>
<td>-1261570 ((p &lt; 0.01))</td>
<td>-1261570 ((p &lt; 0.01))</td>
</tr>
</tbody>
</table>
First Pass Evaluation (Log-Likelihood):

<table>
<thead>
<tr>
<th></th>
<th>Base</th>
<th>Base+CumuSurp</th>
<th>Base+Both</th>
<th>Base+Both</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1212260</td>
<td>-1212259</td>
<td>-1212253</td>
<td>-1212253</td>
</tr>
<tr>
<td>Base+Surp</td>
<td>-1212253</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(p &lt; 0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(p < 0.01\)
Comparable with go-past durations

<table>
<thead>
<tr>
<th>Go-Past Evaluation (Log-Likelihood):</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base</strong></td>
<td></td>
</tr>
<tr>
<td>-1261488</td>
<td></td>
</tr>
<tr>
<td><strong>Base+Surp</strong></td>
<td><strong>Base+CumuSurp</strong></td>
</tr>
<tr>
<td>-1261481 ( (p &lt; 0.01) )</td>
<td>-1261487</td>
</tr>
<tr>
<td><strong>Base+Both</strong></td>
<td><strong>Base+Both</strong></td>
</tr>
<tr>
<td>-1261481</td>
<td>-1261481 ( (p &lt; 0.01) )</td>
</tr>
</tbody>
</table>
First Pass Evaluation (Log-Likelihood):

<table>
<thead>
<tr>
<th></th>
<th>Base</th>
<th>Base+PTB</th>
<th>Base+GCG</th>
<th>Base+Both</th>
<th>Base+Both</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>−1212242</td>
<td>−1212239 (p &lt; 0.01)</td>
<td>−1212239 (p &lt; 0.05)</td>
<td>−1212235 (p &lt; 0.05)</td>
<td>−1212235 (p &lt; 0.01)</td>
</tr>
</tbody>
</table>

Both help independently
PCFG surprisal helps more with go-past durations

<table>
<thead>
<tr>
<th></th>
<th>Base</th>
<th>Base+PTB</th>
<th>Base+GCG</th>
<th>Base+Both</th>
<th>Base+Both</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1261474</td>
<td>1261468 (p &lt; 0.01)</td>
<td>1261470 (p &lt; 0.01)</td>
<td>1261465 (p &lt; 0.01)</td>
<td>1261465 (p &lt; 0.01)</td>
</tr>
</tbody>
</table>

Again, both help independently.
## Fixed Effect Coefficients for Base+PTB+GCG

<table>
<thead>
<tr>
<th>Predictor</th>
<th>First Pass</th>
<th>Go-Past</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coef</td>
<td>t value</td>
</tr>
<tr>
<td>sentpos</td>
<td>-2.47</td>
<td>-3.59</td>
</tr>
<tr>
<td>wlen</td>
<td>25.90</td>
<td>8.67</td>
</tr>
<tr>
<td>prevfix</td>
<td>-30.16</td>
<td>-7.81</td>
</tr>
<tr>
<td>n-gram</td>
<td>-2.39</td>
<td>-1.81</td>
</tr>
<tr>
<td>cumu-n-gram</td>
<td>-14.69</td>
<td>-7.36</td>
</tr>
<tr>
<td>rlen</td>
<td>-5.67</td>
<td>-1.31</td>
</tr>
<tr>
<td>surp-GCG</td>
<td>4.97</td>
<td>2.87</td>
</tr>
<tr>
<td>surp-PTB</td>
<td>4.20</td>
<td>3.23</td>
</tr>
</tbody>
</table>