FREQUENCIES THAT MATTER IN SENTENCE PROCESSING

Marten van Schijndel ¹
September 7, 2015

¹Department of Linguistics, The Ohio State University
Occurrence frequencies describe languages well

- Zipf
- Statistical NLP (esp. vector spaces)

Occurrence frequencies have major influence on sentence processing

- Behavioral measures (e.g., reading times)
- Processing measures (e.g., ERPs)

Uniform Information Density

Saarland SFB-1102
Occurrence frequencies describe languages well

- Zipf
- Statistical NLP (esp. vector spaces)
Occurrence frequencies describe languages well

- Zipf
- Statistical NLP (esp. vector spaces)

Occurrence frequencies have major influence on sentence processing

- Behavioral measures (e.g., reading times)
- Processing measures (e.g., ERPs)
- Uniform Information Density
- Saarland SFB-1102
NULL HYPOTHESIS DEMANDS CONTROL

Linguists must control for frequencies to do research.

How do people try to account for frequencies?
Linguists must control for frequencies to do research
Linguists must control for frequencies to do research

How do people try to account for frequencies?
Case Study 1: Cloze Probabilities
van Schijndel, Culicover, & Schuler (2014)

Pertains to: Pickering & Traxler (2003), inter alia
Ask subjects to generate distribution
Ask subjects to generate distribution

Sentence generation norming:
Write sentences with these words
landed, sneezed, laughed, ...
Ask subjects to generate distribution

Sentence generation norming:
Write sentences with these words

landed, sneezed, laughed, ...

Cloze norming:
Complete this sentence

The pilot landed ________
Ask subjects to generate distribution

Sentence generation norming:
Write sentences with these words
landed, sneezed, laughed, ...

Cloze norming:
Complete this sentence
The pilot landed the plane.
Ask subjects to generate distribution

Sentence generation norming:
Write sentences with these words

landed, sneezed, laughed, ...

Cloze norming:
Complete this sentence

The pilot landed the plane. The pilot landed in the field.
Ask subjects to generate distribution

Sentence generation norming:
Write sentences with these words

landed, sneezed, laughed, ...

Cloze norming:
Complete this sentence

NP: The pilot landed ____ plane. PP: The pilot landed in the field.

Pickering & Traxler (2003) used 6 cloze tasks to determine frequencies
Ask subjects to generate distribution

Sentence generation norming:
Write sentences with these words
landed, sneezed, laughed, ...

Cloze norming:
Complete this sentence

NP: The pilot landed the plane.  PP: The pilot landed in the field.

25%  40%

Pickering & Traxler (2003) used 6 cloze tasks to determine frequencies
Stimuli

(1) That’s the plane that the pilot landed behind in the fog.
(2) That’s the truck that the pilot landed behind in the fog.

Readers slow down at *landed* in (2)
Stimuli

(1) That’s the plane that the pilot landed behind in the fog.
(2) That’s the truck that the pilot landed behind in the fog.

Readers slow down at *landed* in (2)

Suggests they try to link *truck* as the object of *landed* despite:

- *landed* biased for PP complement
  - 40% PP complement
  - 25% NP complement
Readers initially adopt a transitive interpretation despite subcat bias
Readers initially adopt a transitive interpretation despite subcat bias

... Early-attachment processing heuristic
READERS initially adopt a transitive interpretation despite subcat bias.

:: Early-attachment processing heuristic

But what about syntactic frequencies?
Nguyen et al. (2012)

\[
\begin{array}{c}
\text{N} \\
\text{D} \quad \text{N-aD} \quad \text{N-rN} \\
\text{the} \quad \text{apple} \quad \text{that} \\
\text{D} \quad \text{N-aD} \quad \text{V-aN-bN} \\
\text{the} \quad \text{girl} \quad \text{ate}
\end{array}
\]
Nguyen et al. (2012)

The diagram illustrates the structure of a sentence in Generalized Categorial Grammar (GCG). The sentence is "The apple that the girl ate." The diagram shows the syntactic structure of the sentence, with nodes representing categories such as nouns (N), adjectives (A-aN), and verbs (V-gN). The diagram helps to visualize the relationships between the words in the sentence and how they are categorized in GCG.
Pickering & Traxler (2003)

(1) That’s the plane that the pilot landed behind in the fog.
(2) That’s the truck that the pilot landed behind in the fog.

Transitive

Intransitive
(1) That’s the plane that the pilot landed behind in the fog.
(2) That’s the truck that the pilot landed behind in the fog.
What about syntactic frequencies?

Pickering & Traxler (2003)

(1) That’s the plane that the pilot landed behind in the fog.
(2) That’s the truck that the pilot landed behind in the fog.

(a) \[\text{VP-gNP} \quad \text{VP-gNP} \quad \text{PP} \quad \text{TV} \quad t_i \quad \text{P} \quad \text{NP} \quad \text{landed} \quad \text{behind}\] Transitive

(b) \[\text{VP-gNP} \quad \text{VP} \quad \text{PP-gNP} \quad \text{IV} \quad \text{P} \quad t_i \quad \text{landed} \quad \text{behind}\] Intransitive
**WHAT ABOUT SYNTACTIC FREQUENCIES?**

---

**Pickering & Traxler (2003)**

1. That’s the plane that the pilot landed behind in the fog.
2. That’s the truck that the pilot landed behind in the fog.

---

van Schijndel et al. (2014)

Using syntactic probabilities with cloze data:
Pickering & Traxler (2003)

(1) That’s the plane that the pilot landed behind in the fog.
(2) That’s the truck that the pilot landed behind in the fog.

van Schijndel et al. (2014)
Using syntactic probabilities with cloze data:

\[
P(\text{Transitive} \mid \text{landed}) \propto 0.016
\]
\[
P(\text{Intransitive} \mid \text{landed}) \propto 0.004
\]
WHAT ABOUT SYNTACTIC FREQUENCIES?

Pickering & Traxler (2003)

(1) That’s the plane that the pilot landed behind in the fog.
(2) That’s the truck that the pilot landed behind in the fog.

van Schijndel et al. (2014)

Using syntactic probabilities with cloze data:

\[
P(\text{Transitive} \mid \text{landed}) \propto 0.016
\]
\[
P(\text{Intransitive} \mid \text{landed}) \propto 0.004
\]

Transitive interpretation is 300% more likely!
Subcat processing accounted for by hierarchic syntactic frequencies
Early attachment heuristic unnecessary

Also applies to heavy-NP shift heuristics (Staub, 2006), unaccusative processing (Staub et al., 2007), etc.

Suggests cloze probabilities are insufficient as a frequency control
But do people use hierarchic syntactic probabilities?
Subcat processing accounted for by hierarchic syntactic frequencies
Early attachment heuristic unnecessary

Also applies to heavy-NP shift heuristics (Staub, 2006), unaccusative processing (Staub et al., 2007), etc.
Subcat processing accounted for by hierarchic syntactic frequencies
Early attachment heuristic unnecessary
Also applies to heavy-NP shift heuristics (Staub, 2006), unaccusative processing (Staub et al., 2007), etc.
Suggests cloze probabilities are insufficient as a frequency control
Subcat processing accounted for by hierarchic syntactic frequencies
Early attachment heuristic unnecessary

Also applies to heavy-NP shift heuristics (Staub, 2006), unaccusative processing (Staub et al., 2007), etc.

Suggests cloze probabilities are insufficient as a frequency control

But do people use hierarchic syntactic probabilities?
Case Study 2: N-grams and Syntactic Probabilities
van Schijndel & Schuler (2015)

Pertains to: Frank & Bod (2011), inter alia
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 how robust were their models?
This work shows that:

N-gram models can be greatly improved (accumulation)
Hierarchic syntax is still predictive over stronger baseline
(Long distance dependencies independently improve model)
This work shows that:

$N$-gram models can be greatly improved (accumulation)
This work shows that:

* N-gram models can be greatly improved (accumulation)

Hierarchic syntax is still predictive over stronger baseline

(Long distance dependencies independently improve model)
The red apple that the girl ate ...
HIERARCHIC SYNTAX IN READING?

1 The red apple that the girl ate ...

FRANK & BOD (2011)

Baseline:
• Sentence Position
• Word length
• N-grams (Unigram, bigram)
HIERARCHIC SYNTAX IN READING?

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

\begin{itemize}
\item $w_1$
\item $w_2$
\item $w_3$
\item $w_4$
\item $w_5$
\item $w_6$
\end{itemize}

FRANK \& BOD (2011)

Baseline:
\begin{itemize}
\item Sentence Position
\item Word length
\item $N$-grams (Unigram, bigram)
\end{itemize}
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 ...
HIERARCHIC SYNTAX IN READING?

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 SYNTAX 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)
HIERARCHIC SYNTAX 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)
HIERARCHIC SYNTAX 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)

Outcome:
PSG < ESN + PSG
ESN = ESN + PSG
Hierarchic doesn't help over sequential
FRANK & BOD (2011)

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

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

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

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:
- PSG < ESN + PSG
- ESN = 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):
PSG < ESN + PSG
ESN = ESN + PSG

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

Also: lexicalized syntax improves PSG fit
Most previous reading time studies:

- $N$-grams trained on WSJ, Dundee, BNC (or less)
Most previous reading time studies:

- N-grams trained on WSJ, Dundee, BNC (or less)

This study:

- N-grams trained on Gigaword 4.0
Most previous reading time studies:

- N-grams trained on WSJ, Dundee, BNC (or less)
- Unigrams/Bigrams

This study:

- N-grams trained on Gigaword 4.0
IMPROVED $n$-GRAM BASELINE

\begin{figure}
\centering
\includegraphics[width=\textwidth]{n-gram_perplexity.png}
\caption{N-gram Perplexity}
\end{figure}
Improved $n$-gram baseline

Most previous reading time studies:

- $N$-grams trained on WSJ, Dundee, BNC
- Unigrams/Bigrams

This study:

- $N$-grams trained on Gigaword 4.0
- 5-grams
Most previous reading time studies:

- $N$-grams trained on WSJ, Dundee, BNC
- Unigrams/Bigrams
- Only from region boundaries

This study:

- $N$-grams trained on Gigaword 4.0
- 5-grams
**BIGRAM EXAMPLE**

Reading time of *girl* after *red*

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

region

**X**: bigram target  **X**: bigram condition
**BIGRAM EXAMPLE**

Reading time of *girl* after *red*

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

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

- Fails to capture entire sequence;
**BIGRAM EXAMPLE**

Reading time of *girl* after *red*

The red apple that the **girl** ate …

1

2

---

*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 ...

1

2

region

X: bigram target  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 ...

\[
\text{X: bigram targets} \quad \text{X: bigram conditions}
\]


**CUMULATIVE BIGRAM EXAMPLE**

Reading time of *girl* after *red*:

\[
\text{The red } \underline{\text{apple}} \text{ that } \underline{\text{the girl}} \text{ ate } \ldots
\]

- **X**: bigram targets
- **X**: bigram conditions

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

- $N$-grams trained on WSJ, Dundee, BNC
- Unigrams/Bigrams
- Only from region boundaries

This study:

- $N$-grams trained on Gigaword 4.0
- 5-grams
Most previous reading time studies:

- *N*-grams trained on WSJ, Dundee, BNC
- Unigrams/Bigrams
- Only from region boundaries

This study:

- *N*-grams trained on Gigaword 4.0
- 5-grams
- Cumulative and Non-cumulative
Dundee Corpus (Kennedy et al., 2003)

- 10 subjects
- 2,388 sentences
- First pass durations (≈ 200,000)
- Go-past durations (≈ 200,000)

Exclusions:

- Unknown words (<5 tokens)
- First and last of each 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 and Go-Past
• 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?
• 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
CUMULATIVE SURPRISAL DOESN’T HELP?!

- Suggests previous findings were due to weaker $n$-gram baseline

Follow-up work shows long distance dependencies independently influence reading times.
Suggests previous findings were due to weaker \( n \)-gram baseline
Suggests only local PCFG surprisal affects reading times
• Suggests previous findings were due to weaker $n$-gram baseline
• Suggests only local PCFG surprisal affects reading times

Follow-up work shows long distance dependencies independently influence reading times
Hierarchic syntax predicts reading times over strong linear baseline
Hierarchic syntax predicts reading times over strong linear baseline

Long-distance dependencies help over hierarchic syntax
Hierarchic syntax predicts reading times over strong linear baseline

Long-distance dependencies help over hierarchic syntax

Studies should use cumu-$n$-grams in their baselines
We need to carefully control for:

- Cloze probabilities (Smith 2011)
WHAT DOES THIS MEAN FOR OUR MODELS?

We need to carefully control for:

- Cloze probabilities (Smith 2011)
- $N$-gram frequencies (local and cumulative)
- Hierarchic syntactic frequencies
- Long distance dependency frequencies
- …(discourse, etc.)

Then we can try to interpret experimental results.

What do we do about convergence? Is there a way to avoid this explosion of control predictors?
We need to carefully control for:

- Cloze probabilities (Smith 2011)
- $N$-gram frequencies (local and cumulative)
- Hierarchic syntactic frequencies
We need to carefully control for:

- Cloze probabilities (Smith 2011)
- $N$-gram frequencies (local and cumulative)
- Hierarchic syntactic frequencies
- Long distance dependency frequencies
WHAT DOES THIS MEAN FOR OUR MODELS?

We need to carefully control for:

• Cloze probabilities (Smith 2011)
• $N$-gram frequencies (local and cumulative)
• Hierarchic syntactic frequencies
• Long distance dependency frequencies
• ...(discourse, etc.)

Then we can try to interpret experimental results.

What do we do about convergence?
Is there a way to avoid this explosion of control predictors?
Case Study 3: Evading Frequency Confounds
van Schijndel, Murphy, & Schuler (2015)
Can we measure memory load with fewer controls?
Can we measure memory load with fewer controls?

Why do so many factors influence results?
Can we measure memory load with fewer controls?

Why do so many factors influence results?
  Low dimensionality measures.
Can we measure memory load with fewer controls?

Why do so many factors influence results?
   Low dimensionality measures.

Do the factors become separable in another space?
Can we measure memory load with fewer controls?

Why do so many factors influence results?
Low dimensionality measures.

Do the factors become separable in another space?
Let’s try using MEG.
WHAT IS MEG?
102 locations
HOW MIGHT MEG REFLECT LOAD?

Jensen et al., (2012)
HOW MIGHT MEG REFLECT LOAD?

Jensen et al., (2012)
Memory is a function of distributed processing
WHERE TO LOOK?

Memory is a function of distributed processing

Look for synchronized firing between sensors (brain regions)
WHERE TO LOOK?

Memory is a function of distributed processing

Look for synchronized firing between sensors (brain regions)

This study uses spectral coherence measurements.
SPECTRAL COHERENCE

\[ \text{coherence}(x, y) = \frac{E[S_{xy}]}{\sqrt{E[S_{xx}] \cdot E[S_{yy}]}}, \]

Amount of connectivity (synchronization) not caused by chance
SPECTRAL COHERENCE

\[
\text{coherence}(x, y) = \frac{E[S_{xy}]}{\sqrt{E[S_{xx}] \cdot E[S_{yy}]}} \quad \leftarrow \text{cross-correlation}
\]

\[
\quad \leftarrow \text{autocorrelations}
\]
SPECTRAL COHERENCE

\[ \text{coherence}(x, y) = \frac{\mathbb{E}[S_{xy}]}{\sqrt{\mathbb{E}[S_{xx}] \cdot \mathbb{E}[S_{yy}]}} \]

\[ \text{↔ cross-correlation} \]

\[ \text{↔ autocorrelations} \]

Amount of connectivity (synchronization) not caused by chance
SPECTRAL COHERENCE: PHASE SYNCHRONY

Fell & Axmacher (2011)
Collected 2 years ago at CMU
Collected 2 years ago at CMU

3 subjects
Collected 2 years ago at CMU

3 subjects

Heart of Darkness, ch. 2
12,342 words
80 (8 x 10) minutes

Synched with parallel audio recording and forced alignment
Collected 2 years ago at CMU

3 subjects

Heart of Darkness, ch. 2
12,342 words
80 (8 x 10) minutes
Synched with parallel audio recording and forced alignment

306-channel Elekta Neuromag, CMU
Movement/noise correction: SSP, SSS, tSSS
Band-pass filtered 0.01–50 Hz
Downsampled to 125 Hz
Visually scanned for muscle artifacts; none found
\[ d1 \text{ The cart broke.} \]
\[ d2 \text{ that the man bought} \]
The cart broke.

that the man bought

Depth annotations:
van Schijndel et al., (2013) parser
Remove words:

- in short or long sentences (<4 or >50 words)
- that follow a word at another depth
- that fail to parse

Partition data:

- Dev set: One third of corpus
- Test set: Two thirds of corpus
• Group by factor
• Compute coherence over subsets of 4 epochs
Coherence ($d_2 - d_1$)

Frequencies that matter

September 7, 2015 48 / 69
Coherence \( (d_2 - d_1) \)
POSSIBLE CONFOUNDS?

Sentence position

Unigram, Bigram, Trigram: COCA logprobs

PCFG surprisal: parser output
### Dev Results

<table>
<thead>
<tr>
<th>Factor</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigram</td>
<td>0.941</td>
</tr>
<tr>
<td>Bigram</td>
<td>0.257</td>
</tr>
<tr>
<td>Trigram</td>
<td>0.073</td>
</tr>
<tr>
<td>PCFG Surprisal</td>
<td>0.482</td>
</tr>
<tr>
<td>Sentence Position</td>
<td>0.031</td>
</tr>
<tr>
<td>Depth</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Depth 1 (40 items)       Depth 2 (1118 items)
### Test Results

<table>
<thead>
<tr>
<th>Factor</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigram</td>
<td>0.6480</td>
</tr>
<tr>
<td>Bigram</td>
<td>0.7762</td>
</tr>
<tr>
<td>Trigram</td>
<td>0.0264</td>
</tr>
<tr>
<td>PCFG Surprisal</td>
<td>0.3295</td>
</tr>
<tr>
<td>Sentence Position</td>
<td>0.4628</td>
</tr>
<tr>
<td>Depth</td>
<td>0.00002</td>
</tr>
</tbody>
</table>

Depth 1 (86 items)     Depth 2 (2142 items)
<table>
<thead>
<tr>
<th>Factor</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigram</td>
<td>0.6480</td>
</tr>
<tr>
<td>Bigram</td>
<td>0.7762</td>
</tr>
<tr>
<td><strong>Trigram</strong></td>
<td>0.0264</td>
</tr>
<tr>
<td>PCFG Surprisal</td>
<td>0.3295</td>
</tr>
<tr>
<td>Sentence Position</td>
<td>0.4628</td>
</tr>
<tr>
<td><strong>Depth</strong></td>
<td>0.00002</td>
</tr>
</tbody>
</table>

Bonferroni correction removes trigrams, but …
- Group by factor
- Compute coherence over subsets of 6 epochs
### TEST RESULTS: INCREASED RESOLUTION

<table>
<thead>
<tr>
<th>Factor</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trigram</td>
<td>0.3817</td>
</tr>
<tr>
<td>Depth</td>
<td>0.0046</td>
</tr>
</tbody>
</table>

- **Depth 1 (57 items)**
- **Depth 2 (1428 items)**
• Memory load is reflected in MEG connectivity
• Common confounds do not pose problems for oscillatory measures
CONCLUSIONS

- Cloze probabilities are insufficient as frequency control
- Hierarchic syntactic frequencies strongly influence processing
- Reading time studies need to use local *and* cumulative $n$-grams
- Oscillatory analyses could avoid control predictor explosion
• Stefan Frank, Matthew Traxler, Shari Speer, Roberto Zamparelli
• OSU Linguistics Targeted Investment for Excellence (2012-2013)
• National Science Foundation (DGE-1343012)
• University of Pittsburgh Medical Center MEG Seed Fund
• National Institutes of Health CRCNS (5R01HD075328-02)
• Cloze probabilities are insufficient as frequency control
• Hierarchic syntactic frequencies strongly influence processing
• Reading time studies need to use local \textit{and} cumulative $n$-grams
• Oscillatory analyses could avoid control predictor explosion
IMPROVED $n$-GRAM BASELINE

![N-gram Perplexity Graph]

- Perplexity on the y-axis
- $N$ on the x-axis
- Graph shows a decrease in perplexity as $N$ increases from 1 to 6.
FURTHER IMPROVED $n$-GRAM BASELINE

![N-gram Perplexity Chart]

- Perplexity vs. N
- Chart shows the decrease in perplexity with increasing N
- N-gram Perplexity

Frequencies that matter

September 7, 2015  61 / 69
FURTHER IMPROVED \(n\)-GRAM BASELINE

![Pruned vs. Unpruned N-gram Perplexity](image)

- **Pruned**: Red line
- **Unpruned**: Black line

- Perplexity on the y-axis
- \(N\)-gram order on the x-axis

- The graph shows the comparison between pruned and unpruned models for different \(N\)-gram orders.
How probable is each subtree?

Wall Street Journal (WSJ) section of the Penn Treebank:

(a)  
```
    VP-gNP
     |   
    VP-gNP VP-gNP
     |   |   
    TV Adv P NP
     |   |
    landed carefully behind
```
Transitive

(b)  
```
    VP-gNP
     |   
    VP PP-gNP
     |   |   
    VP Adv P t_i
     |   |
    IV carefully behind landed
```
Intransitive
How probable is each subtree?

Wall Street Journal (WSJ) section of the Penn Treebank:

(a) VP-gNP
   ├── VP-gNP
   │    └── Adv
   │        └── TV
   │             └── landed
   └── VP-gNP
       └── Adv
carefully
       └── behind

(b) VP-gNP
   ├── VP
   │    └── Adv
carefully
   │    └── behind
   └── PP-gNP
       └── t_i

Transitive 0.17
Intransitive 0.01
What is the probability of each interpretation?

\[
P(\text{syntactic configuration}) \cdot P(\text{generating the verb from that tree})
\]

\[
P(\text{Transitive}) = P(\text{VP-gNP} \rightarrow \text{VP-gNP PP}) \cdot P(\text{verb} | \text{TV}) \tag{1}
\]

\[
P(\text{Intransitive}) = P(\text{VP-gNP} \rightarrow \text{VP PP-gNP}) \cdot P(\text{verb} | \text{IV}) \tag{2}
\]
What is the probability of each interpretation?

\[
P(\text{syntactic configuration}) \cdot P(\text{generating the verb from that tree}) \cdot P(\text{subcat bias})/P(\text{preterminal prior})
\]

\[
P(\text{Transitive}) = P(\text{VP-gNP} \rightarrow \text{VP-gNP PP}) \cdot P(\text{verb} \mid \text{TV}) \tag{1}
\]

\[
P(\text{Intransitive}) = P(\text{VP-gNP} \rightarrow \text{VP PP-gNP}) \cdot P(\text{verb} \mid \text{IV}) \tag{2}
\]
What is the probability of each interpretation?

$$P(\text{syntactic configuration}) \cdot P(\text{subcat bias}) / P(\text{preterminal prior})$$

$$P(\text{Transitive}) = P(\text{VP-gNP} \rightarrow \text{VP-gNP} \text{ PP}) \cdot P(\text{verb} | \text{TV})$$

$$\propto P(\text{VP-gNP} \rightarrow \text{VP-gNP} \text{ PP}) \cdot \frac{P(\text{TV} | \text{verb})}{P(\text{TV})}$$

$$P(\text{Intransitive}) = P(\text{VP-gNP} \rightarrow \text{VP} \text{ PP-gNP}) \cdot P(\text{verb} | \text{IV})$$

$$\propto P(\text{VP-gNP} \rightarrow \text{VP} \text{ PP-gNP}) \cdot \frac{P(\text{IV} | \text{verb})}{P(\text{IV})}$$
What are the preterminal priors?

Relative prior probability from the WSJ:

$$P(TV):P(IV) = 2.6:1$$
What is the probability of each interpretation?

\[ P(\text{syntactic configuration}) \cdot P(\text{subcat bias}) / P(\text{preterminal prior}) \]

\[ P(\text{Transitive}) \propto P(\text{VP-gNP} \rightarrow \text{VP-gNP PP}) \cdot \frac{P(\text{TV} \mid \text{verb})}{P(\text{TV})} \]

\[ = 0.17 \cdot \frac{P(\text{TV} \mid \text{verb})}{2.6} \]  

\[ P(\text{Intransitive}) \propto P(\text{VP-gNP} \rightarrow \text{VP PP-gNP}) \cdot \frac{P(\text{IV} \mid \text{verb})}{P(\text{IV})} \]

\[ = 0.01 \cdot \frac{P(\text{IV} \mid \text{verb})}{1.0} \]
What is the probability of each interpretation?

\[
P(\text{syntactic configuration}) \cdot P(\text{subcat bias}) / P(\text{preterminal prior})
\]

\[
P(\text{Transitive}) \propto P(\text{VP-gNP} \rightarrow \text{VP-gNP PP}) \cdot \frac{P(\text{TV} | \text{verb})}{P(\text{TV})}
\]

\[
= 0.17 \cdot \frac{P(\text{TV} | \text{verb})}{2.6} = 0.065 \cdot P(\text{TV} | \text{verb}) \quad (1)
\]

\[
P(\text{Intransitive}) \propto P(\text{VP-gNP} \rightarrow \text{VP PP-gNP}) \cdot \frac{P(\text{IV} | \text{verb})}{P(\text{IV})}
\]

\[
= 0.01 \cdot \frac{P(\text{IV} | \text{verb})}{1.0} = 0.01 \cdot P(\text{IV} | \text{verb}) \quad (2)
\]
What is the probability of each interpretation?

\[ P(\text{syntactic configuration}) \cdot P(\text{subcat bias}) / P(\text{preterminal prior}) \]

\[ P(\text{Transitive}) \propto P(VP-gNP \rightarrow VP-gNP \text{ PP}) \cdot \frac{P(TV \mid \text{verb})}{P(TV)} \]

\[ = 0.17 \cdot \frac{P(TV \mid \text{verb})}{2.6} = 0.065 \cdot P(TV \mid \text{verb}) \quad (1) \]

\[ P(\text{Intransitive}) \propto P(VP-gNP \rightarrow VP \text{ PP-gNP}) \cdot \frac{P(IV \mid \text{verb})}{P(IV)} \]

\[ = 0.01 \cdot \frac{P(IV \mid \text{verb})}{1.0} = 0.01 \cdot P(IV \mid \text{verb}) \quad (2) \]

Pickering & Traxler (2003) experimentally determined subcat biases for a wide variety of verbs
(1) That’s the plane that the pilot landed carefully behind in the fog at the airport.
(2) That’s the truck that the pilot landed carefully behind in the fog at the airport.

Using Pickering & Traxler’s (2003) subcat bias data:
(1) That’s the plane that the pilot landed carefully behind in the fog at the airport.
(2) That’s the truck that the pilot landed carefully behind in the fog at the airport.

Using Pickering & Traxler’s (2003) subcat bias data:

\[
P(\text{Transitive} \mid \text{landed}) \propto 0.17 \cdot \frac{0.25}{2.6} = 0.016
\]

\[
P(\text{Intransitive} \mid \text{landed}) \propto 0.01 \cdot \frac{0.40}{1.0} = 0.004
\]