CONTROLLING FOR CONFOUNDING FACTORS IN ONLINE MEASURES OF SENTENCE COMPLEXITY

Marten van Schijndel

July 28, 2015

1Department of Linguistics, The Ohio State University
Occurrence frequencies have major influence on sentence processing. H0 demands that we then control for these factors in our studies. How do people try to account for frequencies?
Occurrence frequencies have major influence on sentence processing
Occurrence frequencies have major influence on sentence processing

$H_0$ demands that we then control for these factors in our studies
Occurrence frequencies have major influence on sentence processing

$H_0$ demands that we then control for these factors in our studies

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, ...

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

Confounds in complexity
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 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)
Nguyen et al. (2012)
(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) \[\text{VP-gNP} \rightarrow \text{VP-gNP} \rightarrow \text{TV} \rightarrow \text{landed} \]
(b) \[\text{VP-gNP} \rightarrow \text{VP} \rightarrow \text{IV} \rightarrow \text{landed} \]

Transitive

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.

(a) VP-gNP
   VP-gNP
   TV \( t_i \)
   landed

(b) VP-gNP
   VP-gNP
   PP
   \( t_i \)
   behind

Transitive

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:
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:

\[
\begin{align*}
P(\text{Transitive} \mid \text{landed}) &\propto 0.016 \\
P(\text{Intransitive} \mid \text{landed}) &\propto 0.004
\end{align*}
\]
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: \textit{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.

But how robust were their models?
Previous studies have debated whether humans use hierarchic syntax.

But how robust were their models?
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
- Hierarchic syntax not improved by accumulation
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
This work shows that:

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

Hierarchic syntax is still predictive over stronger baseline

Hierarchic syntax not improved by accumulation
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 \textbf{girl} ate ...

\textbf{FRANK & BOD (2011)}

Baseline:

- Sentence Position
- Word length
- \textit{N}-grams (Unigram, bigram)
The red apple that the \text{\underline{girl}} ate ...

**FRANK & BOD (2011)**

Baseline:
- Sentence Position
- Word length
- \textit{N}-grams (Unigram, bigram)
The red apple that the **girl** ate ...
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)
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
- Sequential helps over hierarchic
- ESN = ESN + PSG
- Hierarchic doesn’t help over sequential

van Schijndel
Confounds in complexity
**Frank & Bod (2011)**

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

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

**Outcome:**
- $\text{PSG} < \text{ESN} + \text{PSG}$
- $\text{ESN} = \text{ESN} + \text{PSG}$
<table>
<thead>
<tr>
<th>FRANK &amp; BOD (2011)</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>
<tr>
<td><strong>Test POS Predictors:</strong></td>
</tr>
<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
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
Fossum & Levy (2012)

Replicated Frank & Bod (2011):
PSG < ESN + PSG
ESN = ESN + PSG
Fossum & Levy (2012)

Replicated Frank & Bod (2011):

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

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

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

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

Better \textit{n}-gram baseline (more data) changes result:
PSG \equiv ESN + PSG \quad \text{Sequential doesn’t help over hierarchic}
ESN = ESN + PSG
Fossum & Levy (2012)

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

Better \(n\)-gram baseline (more data) changes result:
\[
\begin{align*}
\text{PSG} & = \text{ESN} + \text{PSG} \\
\text{ESN} & = \text{ESN} + \text{PSG}
\end{align*}
\]

Sequential doesn’t help over hierarchic

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
**BIGRAM EXAMPLE**

Reading time of *girl* after *red*

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

- **X**: bigram target
- **X**: bigram condition

- Fails to capture entire sequence;
- Conditions never generated;
- Probability of sequence is deficient.
**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;
Reading time of *girl* after *red*

The red apple that the _girl_ ate ...

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

Reading time of *girl* after *red*

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

- 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
IMPROVED \textit{n-gram baseline}

**Cumulative Bigram Example**

Reading time of \textit{girl} after \textit{red}:

The \textcolor{red}{1} \underline{red} \textcolor{red}{2} \underline{apple} \underline{that} \underline{the} \underline{girl} ate ...

\textbf{X}: bigram targets \quad \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-n-GRAMS PREDICT READING TIMES

First Pass and Go-Past
• Is hierarchic surprisal useful over the better baseline?
• 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
• Suggests previous findings were due to weaker $n$-gram baseline
• 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

Studies should use cumu-$n$-grams in their baselines
WHAT DOES THIS MEAN FOR OUR MODELS?

We need to carefully control for:

- Cloze probabilities
We need to carefully control for:

- Cloze probabilities
- $N$-gram frequencies (local and cumulative)
What does this mean for our models?

We need to carefully control for:

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

We need to carefully control for:

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

- Cloze probabilities
- $N$-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 without controlling for frequency effects?
Can we measure memory load without controlling for frequency effects?

Let’s try using MEG.
WHAT IS MEG?
WHAT IS MEG?

102 locations
How might MEG reflect load?

Jensen et al., (2012)
How might MEG reflect load?

Jensen et al., (2012)
WHERE TO LOOK?

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?
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

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

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

\[
\leftarrow \text{autocorrelations}
\]

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

Phase synchronization: phase lag = 0°

Phase synchronization: phase lag ≠ 0°

Fell & Axmacher (2011)

Nature Reviews | Neuroscience
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
The cart broke.

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 \((d_2 - d_1)\)

Frequency (Hz)

Time (s)
Coherence \((d_2 - d_1)\)
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)
<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
• 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 *and* cumulative *n*-grams
• Oscillatory analyses could avoid control predictor explosion