A Large-Scale Study of the Effects of Word Frequency and
Predictability in Naturalistic Reading

Cory Shain

NAACL 2019
Sam accused Harper of being touchy, which is like the pot calling the bottle kettle.
Sam accused Harper of being touchy, which is like the pot calling the kettle fast or slow.
Sam accused Harper of being touchy, which is like the pot calling the bottle kettle fast slow (Frequency effect).
Sam accused Harper of being touchy, which is like the pot calling the bottle kettle
Sam accused Harper of being touchy, which is like the pot calling the

bottle  
slow

kettle  
fast
Sam accused Harper of being touchy, which is like the pot calling the

**bottle**  **kettle**

*slow*  **fast**

**Predictability** effect
Both frequency and predictability effects have been shown by reading experiments.
Both *frequency* and *predictability* effects have been shown by reading experiments.

Do they arise from different processing mechanisms?
Are frequency and predictability different?

+ **Yes**
  + Predictability effects come from **anticipatory** mechanisms (context-dependent)
  + Frequency effects come from **retrieval** mechanisms (context-independent)
  + Seidenberg and McClelland 1989; Coltheart et al. 2001; Harm and Seidenberg 2004
  + **Prediction**: Independent frequency and predictability effects

+ **No**
  + Processing costs come from resource reallocation between competing interpretations, proportional to the information/surprisal of each word
  + Probability model subsumes lexical frequencies
  + Hale 2001; Norris 2005; Levy 2006; Rasmussen and Schuler 2018
  + **Prediction**: No independent frequency and predictability effects
Are frequency and predictability different?

Yes
- Predictability effects come from *anticipatory* mechanisms (context-dependent)
- Frequency effects come from *retrieval* mechanisms (context-independent)
- Seidenberg and McClelland 1989; Coltheart et al. 2001; Harm and Seidenberg 2004
- Prediction: Independent frequency and predictability effects

No
- Processing costs come from resource reallocation between competing interpretations, proportional to the information/surprisal of each word
- Probability model subsumes lexical frequencies
- Hale 2001; Norris 2000; Levy 2000; Halmoschond and Schuler 2018
- Prediction: No independent frequency and predictability effects
Are frequency and predictability different?

+ **Yes**
  
  - Predictability effects come from **anticipatory** mechanisms (context-dependent)
  
  - Frequency effects come from **retrieval** mechanisms (context-independent)

  - Seidenberg and McClelland 1989; Coltheart et al. 2001; Harm and Seidenberg 2004

  + **Prediction:** Independent frequency and predictability effects

+ **No**

  - Processing costs come from resource reallocation between competing interpretations, proportional to the information/surprisal of each word

  - Probability model subsumes lexical frequencies

  - Hale 2001; Norris 2005; Levy 2008; Harmmonsson and Schuler 2010

  + **Prediction:** No independent frequency and predictability effects
Are frequency and predictability different?

+ **Yes**
  - Predictability effects come from *anticipatory* mechanisms (context-dependent)
  - Frequency effects come from *retrieval* mechanisms (context-independent)
  - Seidenberg and McClelland 1989; Coltheart et al. 2001; Harm and Seidenberg 2004
  - **Prediction**: Independent frequency and predictability effects

+ **No**
  - Processing costs come from resource reallocation between competing interpretations, proportional to the information/surprisal of each word
  - Probability model subsumes lexical frequencies
  - Hale 2001; Norris 2009; Levy 2008; Rasmussen and Schuler 2018
  - **Prediction**: No independent frequency and predictability effects
Are frequency and predictability different?

+ **Yes**
  + Predictability effects come from *anticipatory* mechanisms (context-dependent)
  + Frequency effects come from *retrieval* mechanisms (context-independent)
  + Seidenberg and McClelland 1989; Coltheart et al. 2001; Harm and Seidenberg 2004
  + **Prediction:** Independent frequency and predictability effects

+ **No**
  + Processing costs come from resource reallocation between competing interpretations, proportional to the information/surprise of each word
  + Probability model subsumes lexical frequencies
  + Hale 2001; Norris 2005; Levy 2008; Rasmussen and Schuler 2018
  + **Prediction:** No independent frequency and predictability effects
Are frequency and predictability different?

**Yes**
- Predictability effects come from *anticipatory* mechanisms (context-dependent)
- Frequency effects come from *retrieval* mechanisms (context-independent)
- Seidenberg and McClelland 1989; Coltheart et al. 2001; Harm and Seidenberg 2004
- **Prediction:** Independent frequency and predictability effects

**No**
- Processing costs come from *resource reallocation* between competing interpretations, proportional to the information/surprisal of each word
- Probability model subsumes lexical frequencies
- Hale 2001; Norris 2006; Levy 2008; Rasmussen and Schuler 2018
- **Prediction:** No independent frequency and predictability effects
Are frequency and predictability different?

**Yes**
- Predictability effects come from **anticipatory** mechanisms (context-dependent)
- Frequency effects come from **retrieval** mechanisms (context-independent)
- Seidenberg and McClelland 1989; Coltheart et al. 2001; Harm and Seidenberg 2004
- **Prediction:** Independent frequency and predictability effects

**No**
- Processing costs come from **resource reallocation** between competing interpretations, proportional to the information/surprisal of each word
- Probability model subsumes lexical frequencies
- Hale 2001; Norris 2006; Levy 2008; Rasmussen and Schuler 2018
- **Prediction:** No independent frequency and predictability effects
Are frequency and predictability different?

+ Yes
  + Predictability effects come from **anticipatory** mechanisms (context-dependent)
  + Frequency effects come from **retrieval** mechanisms (context-independent)
  + Seidenberg and McClelland 1989; Coltheart et al. 2001; Harm and Seidenberg 2004
  + **Prediction:** Independent frequency and predictability effects

+ No
  + Processing costs come from **resource reallocation** between competing interpretations, proportional to the information/surprisal of each word
  + Probability model subsumes lexical frequencies
    + Hale 2001; Norris 2006; Levy 2008; Rasmussen and Schuler 2018
  + **Prediction:** No independent frequency and predictability effects
Are frequency and predictability different?

**Yes**
- Predictability effects come from **anticipatory** mechanisms (context-dependent)
- Frequency effects come from **retrieval** mechanisms (context-independent)
- Seidenberg and McClelland 1989; Coltheart et al. 2001; Harm and Seidenberg 2004
- **Prediction:** Independent frequency and predictability effects

**No**
- Processing costs come from **resource reallocation** between competing interpretations, proportional to the information/surprisal of each word
- Probability model subsumes lexical frequencies
- Hale 2001; Norris 2006; Levy 2008; Rasmussen and Schuler 2018
- **Prediction:** No independent frequency and predictability effects
Are frequency and predictability different?

+ **Yes**
  + Predictability effects come from *anticipatory* mechanisms (context-dependent)
  + Frequency effects come from *retrieval* mechanisms (context-independent)
  + Seidenberg and McClelland 1989; Coltheart et al. 2001; Harm and Seidenberg 2004
  + **Prediction:** Independent frequency and predictability effects

+ **No**
  + Processing costs come from *resource reallocation* between competing interpretations, proportional to the information/surprisal of each word
  + Probability model subsumes lexical frequencies
  + Hale 2001; Norris 2006; Levy 2008; Rasmussen and Schuler 2018
  + **Prediction:** No independent frequency and predictability effects
Are frequency and predictability different?

Although many studies have shown additive frequency and predictability effects (Staub 2015):

- Cloze estimates are course-grained (Smith and Levy 2013)
- Constructed stimuli can induce artifacts (Demberg and Keller 2008; Hasson and Honey 2012; Campbell and Tyler 2018)
Are frequency and predictability different?

Although many studies have shown additive frequency and predictability effects (Staub 2015):

- Cloze estimates are course-grained (Smith and Levy 2013)
- Constructed stimuli can induce artifacts (Demberg and Keller 2008; Hasson and Honey 2012; Campbell and Tyler 2018)
Are frequency and predictability different?

+ Although many studies have shown additive frequency and predictability effects (Staub 2015):
  + Cloze estimates are course-grained (Smith and Levy 2013)
  + Constructed stimuli can induce artifacts (Demberg and Keller 2008; Hasson and Honey 2012; Campbell and Tyler 2018)
This study

Are there distinct frequency/predictability effects in naturalistic sentence processing?
This study

+ **Challenge 1:** Collinearity
+ **Solution:** Large data
  - Natural Stories (self-paced reading) (Futrell et al. 2018)
  - Dundee (eye-tracking) (Kennedy et al. 2003)
  - UCL (eye-tracking) (Frank et al. 2013)
  - 1M+ data points
This study

Challenge 1: Collinearity
Solution: Large data
  - Natural Stories (self-paced reading) (Futrell et al. 2018)
  - Dundee (eye-tracking) (Kennedy et al. 2003)
  - UCL (eye-tracking) (Frank et al. 2013)
  - 1M+ data points
Challenge 1: Collinearity

Solution: Large data

- Natural Stories (self-paced reading) (Futrell et al. 2018)
- Dundee (eye-tracking) (Kennedy et al. 2003)
- UCL (eye-tracking) (Frank et al. 2013)
- 1M+ data points
This study

- **Challenge 1:** Collinearity
- **Solution:** Large data
  - Natural Stories (self-paced reading) (Futrell et al. 2018)
  - Dundee (eye-tracking) (Kennedy et al. 2003)
  - UCL (eye-tracking) (Frank et al. 2013)
  - 1M+ data points
This study

+ **Challenge 1:** Collinearity
+ **Solution:** Large data
  + Natural Stories (self-paced reading) (Futrell et al. 2018)
  + Dundee (eye-tracking) (Kennedy et al. 2003)
  + UCL (eye-tracking) (Frank et al. 2013)
  + 1M+ data points
Challenge 1: Collinearity

Solution: Large data

- Natural Stories (self-paced reading) (Futrell et al. 2018)
- Dundee (eye-tracking) (Kennedy et al. 2003)
- UCL (eye-tracking) (Frank et al. 2013)
- 1M+ data points
This study

+ **Challenge 2:** Temporal diffusion (Erlich and Rayner 1983)

+ **Solution:** Deconvolutional time series regression (Shain and Schuler 2018)
This study

- **Challenge 2:** Temporal diffusion (Erlich and Rayner 1983)
- **Solution:** Deconvolutional time series regression (Shain and Schuler 2018)
This study

Challenge 3: External validity

Solution: Non-parametric out-of-sample paired permutation test (cf. LRT)
Challenge 3: External validity

Solution: Non-parametric out-of-sample paired permutation test (cf. LRT)
This study

+ **Frequency**: Unigram logprob
+ **Predictability**: 5-gram surprisal
+ KenLM (Heafield et al. 2013) trained on Gigaword 3 (Graff et al. 2007)
+ By-subject random intercepts, slopes, and response shapes
+ Log-ms response
+ 50-50 train-test split
+ Ablative permutation tests
This study

- **Frequency**: Unigram logprob
- **Predictability**: 5-gram surprisal
- KenLM (Heafield et al. 2013) trained on Gigaword 3 (Graff et al. 2007)
- By-subject random intercepts, slopes, and response shapes
- Log-ms response
- 50-50 train-test split
- Ablative permutation tests
This study

- **Frequency:** Unigram logprob
- **Predictability:** 5-gram surprisal
- KenLM (Heafield et al. 2013) trained on Gigaword 3 (Graff et al. 2007)
- By-subject random intercepts, slopes, and response shapes
- Log-ms response
- 50-50 train-test split
- Ablative permutation tests
This study

+ **Frequency**: Unigram logprob
+ **Predictability**: 5-gram surprisal
+ KenLM (Heafield et al. 2013) trained on Gigaword 3 (Graff et al. 2007)
+ By-subject random intercepts, slopes, and response shapes
+ Log-ms response
+ 50-50 train-test split
+ Ablative permutation tests
This study

- **Frequency**: Unigram logprob
- **Predictability**: 5-gram surprisal
- KenLM (Heafield et al. 2013) trained on Gigaword 3 (Graff et al. 2007)
- By-subject random intercepts, slopes, and response shapes
- Log-ms response
- 50-50 train-test split
- Ablative permutation tests
This study

- **Frequency**: Unigram logprob
- **Predictability**: 5-gram surprisal
- KenLM (Heafield et al. 2013) trained on Gigaword 3 (Graff et al. 2007)
- By-subject random intercepts, slopes, and response shapes
- Log-ms response
- 50-50 train-test split
- Ablative permutation tests
This study

+ **Frequency:** Unigram logprob
+ **Predictability:** 5-gram surprisal
+ KenLM (Heafield et al. 2013) trained on Gigaword 3 (Graff et al. 2007)
+ By-subject random intercepts, slopes, and response shapes
+ Log-ms response
+ 50-50 train-test split
+ Ablative permutation tests
## Results: Effect sizes

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Effect estimate (log-ms)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unigram</td>
<td>5-gram</td>
<td></td>
</tr>
<tr>
<td>Natural Stories</td>
<td>-0.0018</td>
<td>0.0174</td>
<td></td>
</tr>
<tr>
<td>Dundee</td>
<td>-0.0067</td>
<td>0.0117</td>
<td></td>
</tr>
<tr>
<td>UCL</td>
<td>0.0005</td>
<td>0.0184</td>
<td></td>
</tr>
</tbody>
</table>
## Results: Effect sizes

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Effect estimate (log-ms)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Unigram</td>
<td>5-gram</td>
</tr>
<tr>
<td>Natural Stories</td>
<td>-0.0018</td>
<td>0.0174</td>
<td></td>
</tr>
<tr>
<td>Dundee</td>
<td>-0.0067</td>
<td>0.0117</td>
<td></td>
</tr>
<tr>
<td>UCL</td>
<td>0.0005</td>
<td>0.0184</td>
<td></td>
</tr>
</tbody>
</table>

**Larger-magnitude 5-gram effect**
## Results: Hypothesis test

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Pooled</th>
<th>Natural Stories</th>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-gram only vs. baseline</td>
<td>0.0001***</td>
<td>0.0001***</td>
<td>0.0001***</td>
</tr>
<tr>
<td>Unigram only vs. baseline</td>
<td>0.0001***</td>
<td>0.0001***</td>
<td>0.0001***</td>
</tr>
<tr>
<td>5-gram + Unigram vs. Unigram-only</td>
<td>0.0001***</td>
<td>0.0001***</td>
<td>0.0626</td>
</tr>
<tr>
<td>5-gram + Unigram vs. 5-gram-only</td>
<td>0.1515</td>
<td>0.1831</td>
<td>0.0105</td>
</tr>
</tbody>
</table>

No evidence of independent freq/pred effects.
## Results: Hypothesis test

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Pooled</th>
<th>Natural Stories</th>
<th>Dundee</th>
<th>UCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-gram only vs. baseline</td>
<td>0.0001***</td>
<td>0.0001***</td>
<td>0.0001***</td>
<td>0.0001***</td>
</tr>
<tr>
<td>Unigram only vs. baseline</td>
<td>0.0001***</td>
<td>0.0001***</td>
<td>0.0001***</td>
<td>0.0001***</td>
</tr>
<tr>
<td>5-gram + Unigram vs. Unigram-only</td>
<td>0.0001***</td>
<td>0.0001***</td>
<td>0.0626</td>
<td>0.0006***</td>
</tr>
<tr>
<td>5-gram + Unigram vs. 5-gram-only</td>
<td>0.1515</td>
<td>0.1831</td>
<td>0.0105</td>
<td>0.1491</td>
</tr>
</tbody>
</table>

No evidence of independent freq/pred effects
Conclusion

- No evidence of distinct, context-independent retrieval mechanism
- Findings disagree with previous experiments
  - Statistical vs. cloze predictability
  - Naturalistic vs. constructed stimuli
- Artificial stimuli/tasks engage problem-solving regions (Kaan and Swaab 2002; Novick et al. 2005; Blank and Fedorenko 2017)
- Comprehension-as-problem-solving may diminish influence of preceding words
- Frequency effects may still exist
  - Fail to reject null ≠ Accept null
  - At best, attenuated in naturalistic sentence processing
Conclusion

+ No evidence of distinct, context-independent retrieval mechanism
+ Findings disagree with previous experiments
  + Statistical vs. cloze predictability
  + Naturalistic vs. constructed stimuli
    Comprehension-as-problem-solving may diminish influence of preceding words.
+ Frequency effects may still exist
  + Fail to reject null ≠ Accept null
  + At best, attenuated in naturalistic sentence processing.
Conclusion

- No evidence of distinct, context-independent retrieval mechanism
- Findings disagree with previous experiments
  - Statistical vs. cloze predictability
  - Naturalistic vs. constructed stimuli
    - Artificial stimuli/tasks engage problem-solving regions (Kaan and Swaab 2002; Novick et al. 2005; Blank and Fedorenko 2017)
    - Comprehension-as-problem-solving may diminish influence of preceding words
- Frequency effects may still exist
  - Fail to reject null ≠ Accept null
  - At best, attenuated in naturalistic sentence processing
Conclusion

- No evidence of distinct, context-independent retrieval mechanism
- Findings disagree with previous experiments
  - Statistical vs. cloze predictability
  - Naturalistic vs. constructed stimuli
    - Artificial stimuli/tasks engage problem-solving regions (Kaan and Swaab 2002; Novick et al. 2005; Blank and Fedorenko 2017)
    - Comprehension-as-problem-solving may diminish influence of preceding words
- Frequency effects may still exist
  - Fail to reject null ≠ Accept null
  - At best, attenuated in naturalistic sentence processing
Conclusion

- No evidence of distinct, context-independent retrieval mechanism
- Findings disagree with previous experiments
  - Statistical vs. cloze predictability
  - Naturalistic vs. constructed stimuli
    - Artificial stimuli/tasks engage problem-solving regions (Kaan and Swaab 2002; Novick et al. 2005; Blank and Fedorenko 2017)
    - Comprehension-as-problem-solving may diminish influence of preceding words
- Frequency effects may still exist
  - Fail to reject null ≠ Accept null
  - At best, attenuated in naturalistic sentence processing
Conclusion

+ No evidence of distinct, context-independent retrieval mechanism
+ Findings disagree with previous experiments
  + Statistical vs. cloze predictability
  + Naturalistic vs. constructed stimuli
    + Artificial stimuli/tasks engage problem-solving regions (Kaan and Swaab 2002; Novick et al. 2005; Blank and Fedorenko 2017)
    + Comprehension-as-problem-solving may diminish influence of preceding words
+ Frequency effects may still exist
  + Fail to reject null ≠ Accept null
  + At best, attenuated in naturalistic sentence processing
No evidence of distinct, context-independent retrieval mechanism

Findings disagree with previous experiments

- Statistical vs. cloze predictability
- Naturalistic vs. constructed stimuli
  - Artificial stimuli/tasks engage problem-solving regions (Kaan and Swaab 2002; Novick et al. 2005; Blank and Fedorenko 2017)
  - Comprehension-as-problem-solving may diminish influence of preceding words

Frequency effects may still exist

- Fail to reject null ≠ Accept null
- At best, attenuated in naturalistic sentence processing
No evidence of distinct, context-independent retrieval mechanism

Findings disagree with previous experiments
- Statistical vs. cloze predictability
- Naturalistic vs. constructed stimuli
  - Artificial stimuli/tasks engage problem-solving regions (Kaan and Swaab 2002; Novick et al. 2005; Blank and Fedorenko 2017)
  - Comprehension-as-problem-solving may diminish influence of preceding words

Frequency effects may still exist
- Fail to reject null ≠ Accept null
  - At best, attenuated in naturalistic sentence processing
Conclusion

- No evidence of distinct, context-independent retrieval mechanism
- Findings disagree with previous experiments
  - Statistical vs. cloze predictability
  - Naturalistic vs. constructed stimuli
    - Artificial stimuli/tasks engage problem-solving regions (Kaan and Swaab 2002; Novick et al. 2005; Blank and Fedorenko 2017)
    - Comprehension-as-problem-solving may diminish influence of preceding words
- Frequency effects may still exist
  - Fail to reject null ≠ Accept null
  - At best, attenuated in naturalistic sentence processing
Thank you!

**Data preprocessing:**
https://github.com/modelblocks/modelblocks-release

**DTSR regression:**
https://github.com/coryshain/dtsr

**Acknowledgements:**
National Science Foundation grants #1551313 and #1816891
Anonymous NAACL 2019 reviewers
References


Frank, Stefan L et al. (2013). “Reading time data for evaluating broad-coverage models of English sentence processing”. In: Behavior Research Methods 45.4, pp. 1182–1190.
References


References


Baseline variables

- **ShiftedGamma IRF**
  
  \[ f(x; \alpha, \beta, \delta) = \frac{\beta^\alpha (x - \delta)^{\alpha - 1} e^{-\beta(x - \delta)}}{\Gamma(\alpha)} \]

- **Rate**: Deconvolutional intercept
- **Word length**: Word length in characters
- **Saccade length**: Length of last saccade in words
- **Previous was fixated**: Whether the previous word was fixated

- **Linear (Dirac Delta) IRF**
  
  - **Sentence position**: Index of word in sentence
  - **Trial**: Index of word in document
IRF estimates

(a) Natural Stories
(b) Dundee
(c) UCL
<table>
<thead>
<tr>
<th>Corpus</th>
<th>Length</th>
<th>Vocab</th>
<th>Token coverage</th>
<th>Type coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Stories</td>
<td>10256</td>
<td>3104</td>
<td>99.58%</td>
<td>98.65%</td>
</tr>
<tr>
<td>Dundee</td>
<td>51501</td>
<td>12871</td>
<td>99.26%</td>
<td>97.21%</td>
</tr>
<tr>
<td>UCL</td>
<td>4957</td>
<td>1576</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>