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

Cory Shain

NAACL 2019

bottle kettle

bottle **kettle**

fast *slow*

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 **kettle**

slow *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 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?

- **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

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

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

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)

This study

- **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

	Effect estimate (log-ms)	
Corpus	Unigram	5-gram
Natural Stories	-0.0018	0.0174
Dundee	-0.0067	0.0117
UCL	0.0005	0.0184

Results: Effect sizes

	Effect estimate (log-ms)	
Corpus	Unigram	5-gram
Natural Stories	-0.0018	0.0174
Dundee	-0.0067	0.0117
UCL	0.0005	0.0184

Larger-magnitude 5-gram effect

Results: Hypothesis test

Comparison	Pooled	Corpus Natural Stories	Dundee	UCL
5-gram only vs. baseline	**0.0001*** **	**0.0001*** **	**0.0001*** **	**0.0001*** **
Unigram only vs. baseline	**0.0001*** **	**0.0001*** **	**0.0001*** **	**0.0001*** **
5-gram + Unigram vs. Unigram-only	**0.0001*** **	**0.0001*** **	0.0626	**0.0006*** **
5-gram + Unigram vs. 5-gram-only	0.1515	0.1831	0.0105	0.1491

Results: Hypothesis test

Comparison	Pooled	Corpus: Natural Stories	Dundee	UCL
5-gram only vs. baseline	**0.0001*** **	**0.0001*** **	**0.0001*** **	**0.0001*** **
Unigram only vs. baseline	**0.0001*** **	**0.0001*** **	**0.0001*** **	**0.0001*** **
5-gram + Unigram vs. Unigram-only	**0.0001*** **	**0.0001*** **	0.0626	**0.0006*** **
5-gram + Unigram vs. 5-gram-only	0.1515	0.1831	0.0105	0.1491

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
- 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 drive influence of predictability
- 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 $\neq$ 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 $\neq$ 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 $\neq$ 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 $\neq$ 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

Blank, Idan and Evelina Fedorenko (2017). “Domain-general brain regions do not track linguistic input as closely as language-selective regions”. In: Journal of Neuroscience, pp. 3616–3642.

Campbell, Karen L and Lorraine K Tyler (2018). “Language-related domain-specific and domain-general systems in the human brain”. In: Current opinion in behavioral sciences 21, pp. 132–137.

Coltheart, Max et al. (2001). “DRC: a dual route cascaded model of visual word recognition and reading aloud.”. In: Psychological review 108.1, p. 204.

Demberg, Vera and Frank Keller (2008). “Data from eye-tracking corpora as evidence for theories of syntactic processing complexity”. In: Cognition 109.2, pp. 193–210.

Erlich, Kate and Keith Rayner (1983). “Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing”. In: Journal of Verbal Learning & Verbal Behavior 22, pp. 75–87.

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

Futrell, Richard et al. (2018). “The Natural Stories Corpus”. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation Ed. by Nicoletta Calzolari et al. Paris, France: European Language Resources Association (ELRA). ISBN: 979-10-95546-00-9.

Graff, David et al. (2007). English Gigaword Third Edition LDC2007T07. Philadelphia. URL: https://catalog.ldc.upenn.edu/LDC2007T07.

Hale, John (2001). “A probabilistic Earley parser as a psycholinguistic model”. In: Proceedings of the Second meeting of the North American Chapter of the Association for Com pp. 1–8. DOI: 10.3115/1073336.1073357. URL: http://www.aclweb.org/anthology/N01-1021http://portal.acm.org/citation.cfm?doid=1073336.1073357.

Harm, Michael W and Mark S Seidenberg (2004). “Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes.”. In: Psychological review 111.3, p. 662.

Hasson, Uri and Christopher J Honey (2012). “Future trends in Neuroimaging: Neural processes as expressed within real-life contexts”. In: NeuroImage 62.2, pp. 1272–1278.

References

Heafield, Kenneth et al. (2013). "Scalable modified Kneser-Ney language model estimation". In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, pp. 690–696.

Kaan, Edith and Tamara Y Swaab (2002). "The brain circuitry of syntactic comprehension". In: Trends in cognitive sciences 6.8, pp. 350–356.

Kennedy, Alan, James Pynte, and Robin Hill (2003). "The Dundee corpus". In: Proceedings of the 12th European conference on eye movement.

Levy, Roger (2008). "Expectation-based syntactic comprehension". In: Cognition 106.3, pp. 1126–1177.

Norris, Dennis (2006). "The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process.". In: Psychological review 113.2, p. 327.

Novick, Jared M, John C Trueswell, and Sharon L Thompson-Schill (2005). "Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension". In: Cognitive, Affective, \& Behavioral Neuroscience 5.3, pp. 263–281.

References

- Rasmussen, Nathan E and William Schuler (2018). "Left-Corner Parsing With Distributed Associative Memory Produces Surprisal and Locality Effects". In: Cognitive science 42, pp. 1009–1042.
- Seidenberg, Mark S and James L McClelland (1989). "A distributed, developmental model of word recognition and naming". In: Psychological review 96.4, p. 523.
- Shain, Cory and William Schuler (2018). "Deconvolutional time series regression: A technique for modeling temporally diffuse effects". In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.
- Smith, Nathaniel J and Roger Levy (2013). "The effect of word predictability on reading time is logarithmic". In: Cognition 128, pp. 302–319.
- Staub, Adrian (2015). "The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation". In: Language and Linguistics Compass 9.8, pp. 311–327.

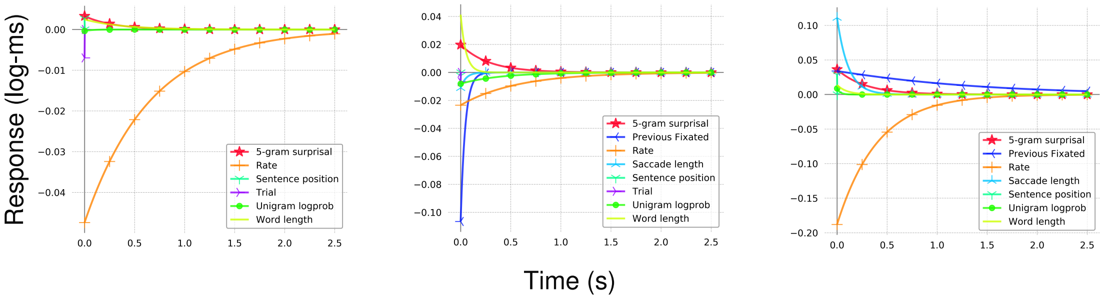
Baseline variables

- 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

Coverage

Corpus	Length	Vocab	Token coverage	Type coverage
Natural Stories	10256	3104	99.58%	98.65%
Dundee	51501	12871	99.26%	97.21%
UCL	4957	1576	100%	100%