Prediction- and Recall-Defined Online Complexity-Metrics

Marten van Schijndel
Department of Linguistics
The Ohio State University

November 5, 2012

Motivation

Observation isn't explanation

Current metrics of processing complexity (eg. surprisal variants [Hale, 2001, Levy, 2008, Roark et al., 2009] and UID [Levy and Jaeger, 2007]) are based on observation of complexity without providing an explanation for why it arises.

Goal: An Explanation

If a model based on current theories of the structure of domain-general working memory can predict processing complexity above the predictions of surprisal, it may provide a rationale for *why* humans have the language processing difficulties they do.

A Model of Prediction

- People use prediction (Cloze task, filled-gap effect)
- Processing difficulty may stem from incorrect predictions
- A model of prediction may predict processing difficulty

A Model of Prediction

- People use prediction (Cloze task, filled-gap effect)
- Processing difficulty may stem from incorrect predictions
- A model of prediction may predict processing difficulty

The professor would . . .

A Model of Prediction

- People use prediction (Cloze task, filled-gap effect)
- Processing difficulty may stem from incorrect predictions
- A model of prediction may predict processing difficulty

The professor would . . . (V, Neg)

A Model of Prediction

- People use prediction (Cloze task, filled-gap effect)
- Processing difficulty may stem from incorrect predictions
- A model of prediction may predict processing difficulty

The professor would . . . (V, Neg)
The professor would . . .

A Model of Prediction

- People use prediction (Cloze task, filled-gap effect)
- Processing difficulty may stem from incorrect predictions
- A model of prediction may predict processing difficulty

The professor would . . . (V, Neg)
The professor would . . . though

A Model of Prediction

- People use prediction (Cloze task, filled-gap effect)
- Processing difficulty may stem from incorrect predictions
- A model of prediction may predict processing difficulty

The professor would . . . (V, Neg)
The professor would . . . though Alice

A Model of Prediction

- People use prediction (Cloze task, filled-gap effect)
- Processing difficulty may stem from incorrect predictions
- A model of prediction may predict processing difficulty

The professor would . . . (V, Neg)
The professor would . . . though Alice advised

A Model of Prediction

- People use prediction (Cloze task, filled-gap effect)
- Processing difficulty may stem from incorrect predictions
- A model of prediction may predict processing difficulty

The professor would . . . (V, Neg)
The professor would . . . though Alice advised against it

A Model of Prediction

- People use prediction (Cloze task, filled-gap effect)
- Processing difficulty may stem from incorrect predictions
- A model of prediction may predict processing difficulty

The professor would . . . (V, Neg)
The professor would . . . though Alice advised against it (V, Neg)

A Model of Prediction

- People use prediction (Cloze task, filled-gap effect)
- Processing difficulty may stem from incorrect predictions
- A model of prediction may predict processing difficulty

The professor would . . . (V, Neg)
The professor would . . . though Alice advised against it (V, Neg)

Assumption: Parallel processing (competing hypotheses)

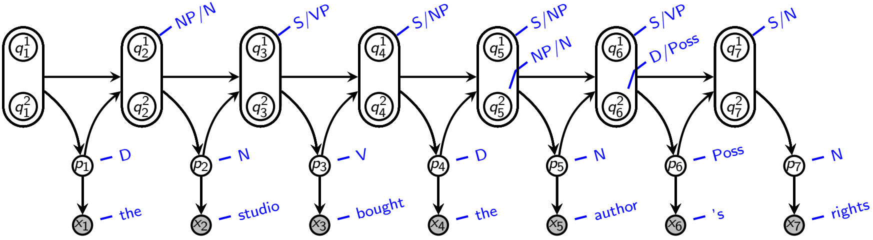
Cueing Predictions

The professor would . . . (V, Neg)
The professor would . . . though Alice advised against it (V, Neg)

- Memory research indicates that humans have different types of difficulty with different kinds of recall (short-term vs long-term)
- Sequential vs temporal cueing [Sederberg et al., 2008]
- Naturally lends itself to center-embeddeding

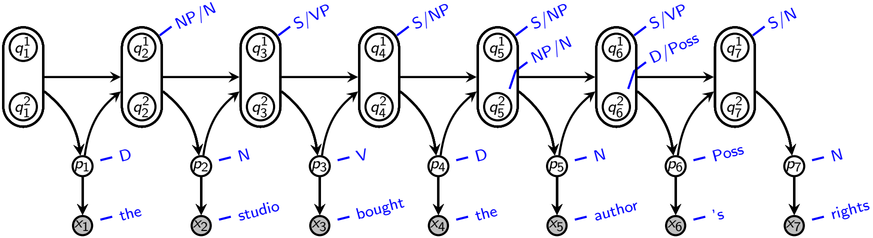
Cueing in Parsing

- Sequential cueing is captured via *active* and *awaited* components
- Temporal cueing is captured via tiers of embeddedness
- Grammar formalism is sensitive to embedding depth



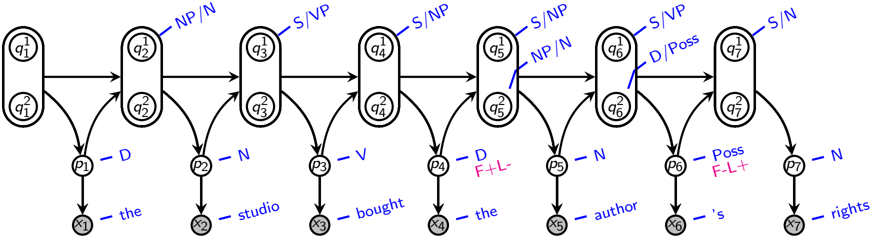
Parser Predictions

- F(irst): Predict the first element of a new tier
- L(ast): Predict that the last element of a tier was just seen
- F and L binary predictions made at each timestep metrics



Parser Predictions

- F(irst): Predict the first element of a new tier
- L(ast): Predict that the last element of a tier was just seen
- F and L binary predictions made at each timestep metrics



Proposed Complexity Metrics

Loosely correspond to Storage and Integration costs [Gibson, 2000]

- F+: Predict a new tier (incur a *storage* cost)
- DF+: F+ weighted by the tier number
- L+: Predict integration of a tier (incur an *integration* cost)
- DL+: L+ weighted by the tier number
- DistL+: L+ weighted by the length of the tier

Human Complexity

- Reading times provide a window into complexity
- Many different metrics (fixation duration, regression, etc)

People fixate longer on difficult words
People regress more after ambiguous words and difficult constructions

Human Complexity

- Reading times provide a window into complexity
- Many different metrics (fixation duration, regression, etc)

People fixate longer on difficult words
People regress more after ambiguous words and difficult constructions

Choice: First-Pass Fixation (Gaze Duration).

Training

- Parser and Lexicon: WSJ02-21 [Marcus et al., 1993]
  - 39,832 sentences
  - 950,028 words
- Ngrams: Brown [Francis and Kucera, 1979], WSJ02-21, BNC, Dundee[Kennedy et al., 2003]
  - 5,052,904 sentences
  - 87,302,312 words

Ngrams calculated using SRILM [Stolcke, 2002] with modified Kneser-Ney smoothing [Chen and Goodman, 1998]

Evaluation

- Dundee corpus [Kennedy et al., 2003]
  - 10 subjects
  - 2,388 sentences
  - 58,439 words
  - 260,124 subject/word pairs (first-pass fixations)
- Filtered Dundee corpus
  - 148,717 words
- Filtered Dundee corpus sans outliers (by subject)
  - 146,671 words

Exclusions: UNK-threshold 5, first and last of a line, first and last of a sentence, multiple capitals, words that contain a non-letter

Baseline Metrics

Fitting a linear mixed effects model

Derived from [Fossum and Levy, 2012] and [Frank and Bod, 2011]

- Number of characters
- Was previous word fixated?
- Unigram and Bigram probs
- Sentence position
- Will next word be fixated?
- Joint interactions

Plus

- Number of intervening words (non-significant)
- Cumulative Total Surprisal [Hale, 2001]

Simplest baseline is determined on development data before fitting test data with factors of interest (see Appendix)

Fixation durations are log-transformed prior to fitting to yield normal distributions (see Appendix)

Results

Metrics residualized from baseline on test set

Model	Improvement (p-value)	Model	p-value
F+	.047	F-	.00029
DF+	–	DF-	.0049
L+	.0014	L-	.020
DL+	–	DL-	–
DistL+	.0021	–	–

Metric improvement over baseline

Conclusion

An explanation

Some of the hypothesized metrics can predict reading times even over those predictions made by a strong baseline including a state-of-the-art complexity metric (cumulative total surprisal). This indicates that domain-general memory processes provide at least a partial account of *why* humans encounter difficulty during language processing.

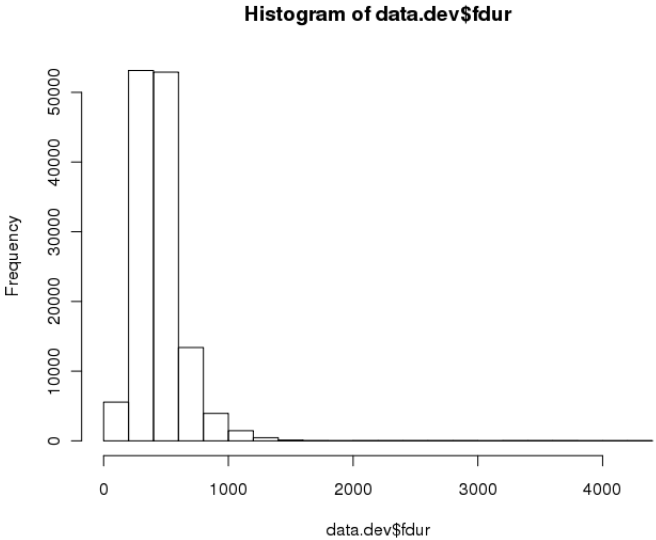
Fin

Thanks!

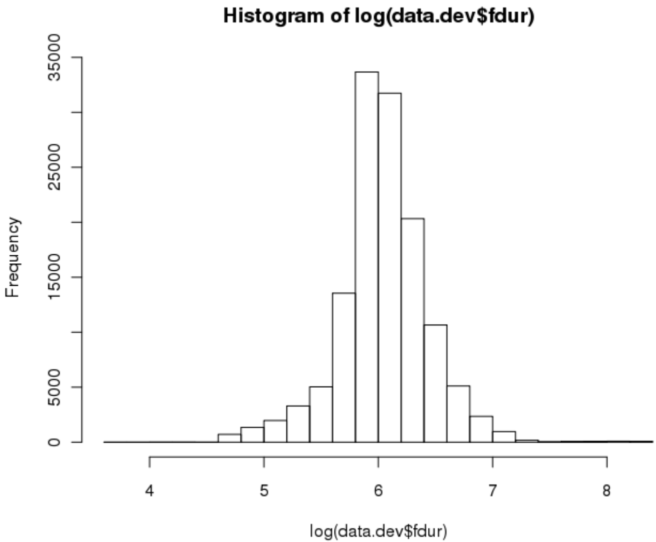
Especially to Kodi Weatherholtz and Rory Turnbull for their assistance with R-wrangling and working with linear mixed effect models!
Additional thanks due to William Schuler for advising on this project.
Any errors are my own

Questions?

Transforming the response variable



Transforming the response variable

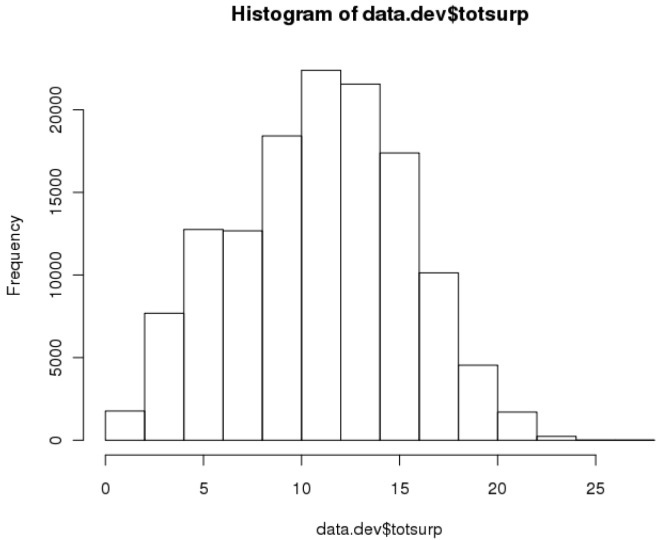


Transforming the response variable

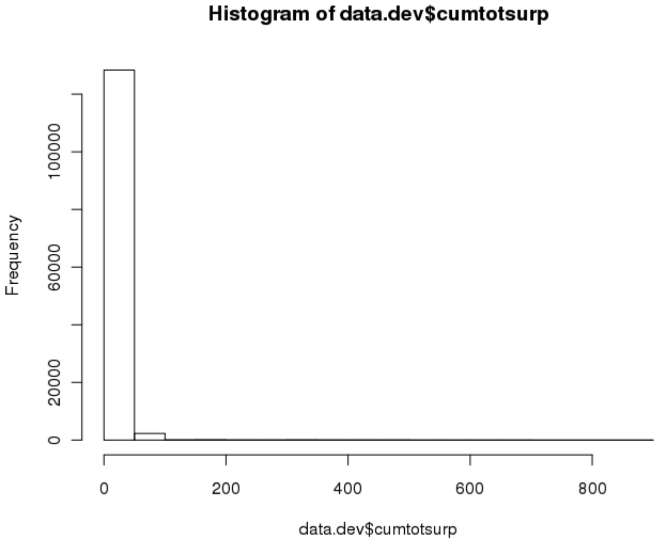
Model	LogLikelihood (dev)
fdur $\sim$ Baseline	-882,585
log(fdur) $\sim$ Baseline	-55,378

Improvement from transforming the response

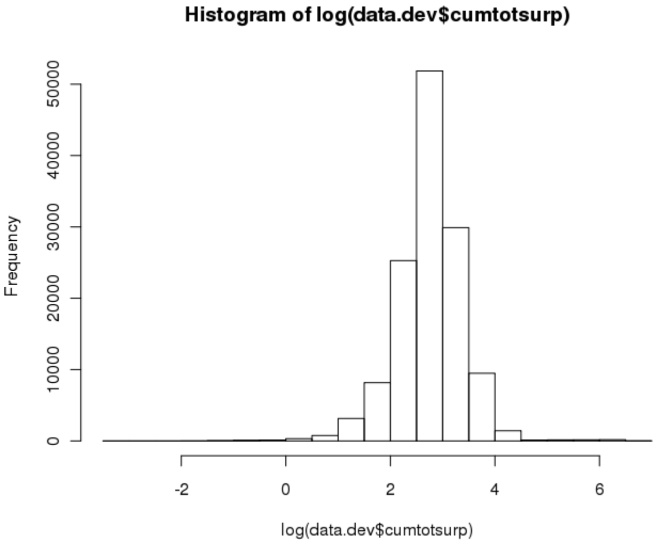
An underrepresented innovation



An underrepresented innovation



An underrepresented innovation



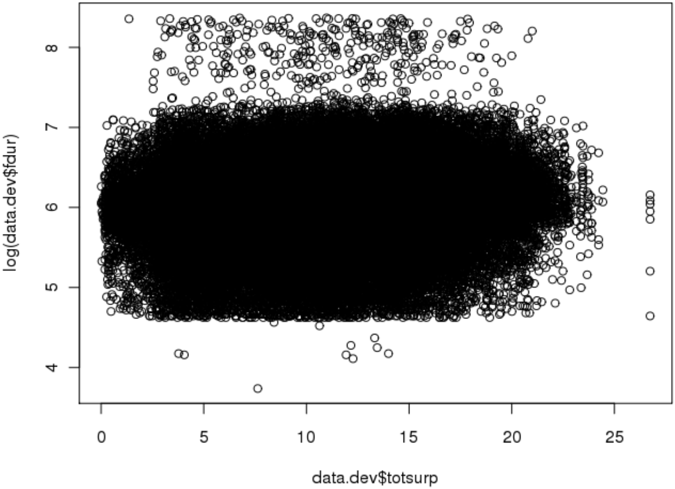
An underrepresented innovation

Model	LogLikelihood (dev)
Baseline + Total Surprisal	-56307
Baseline + log(Cum. Total Surprisal)	-55753
Baseline + Cum. Total Surprisal	-55378

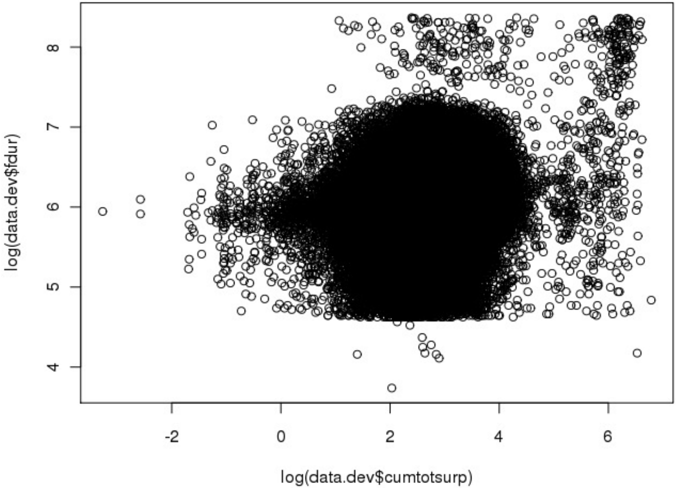
Improvement from accumulating metrics

Cumulative Total Surprisal seems to be the best

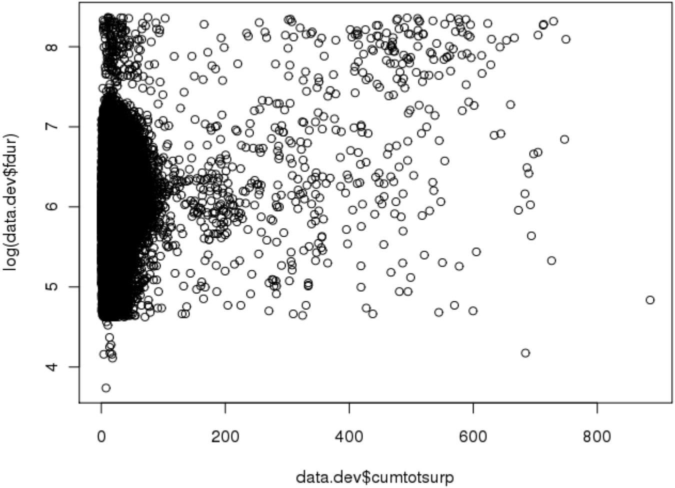
An underrepresented innovation



An underrepresented innovation



An underrepresented innovation



Finding the Simplest Baseline Model

1. Begin with all baseline effects thrown into model along with their joint interactions.
2. Reduce multicollinearity: Using Variance Inflation Factors (VIFs), remove largest contributor to multicollinearity until loglikelihood of model is negatively affected (interactions removed first)
3. Simplify model: Using t-scores, remove least significant factor until an ANOVA reveals a significant effect

Problems with multicollinearity

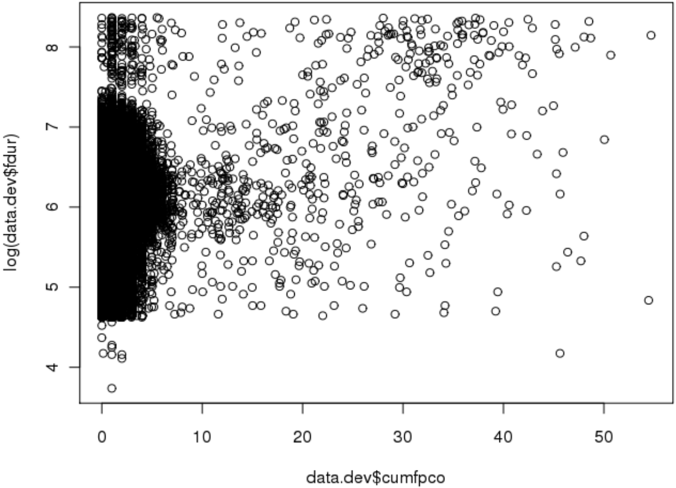
- Algorithms to determine coefficients fail or are inaccurate
- Results won't generalize to new populations
- Significance found will still be significant without collinearity but bias can lead to incorrect predictions on new data

Simplest Baseline Model

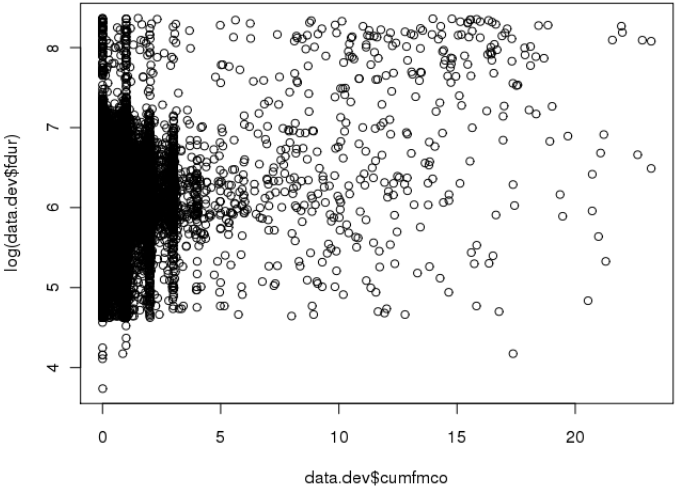
log(fdur)∼

- nchar
- sentpos
- previsfix
- nrchar:logwordprob
- sentpos:nextisfix
- sentpos:logfwprob
- nextisfix:cumtotsurp
- subject and item random intercepts
- logprob
- logfwprob
- cumtotsurp
- previsfix:logprob
- previsfix:logfwprob
- previsfix:cumtotsurp
- logprob:cumtotsurp
- logfwprob:cumtotsurp

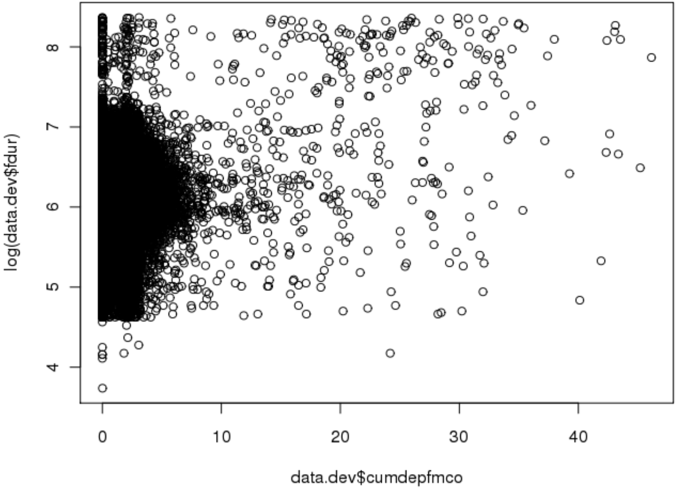
Graphs



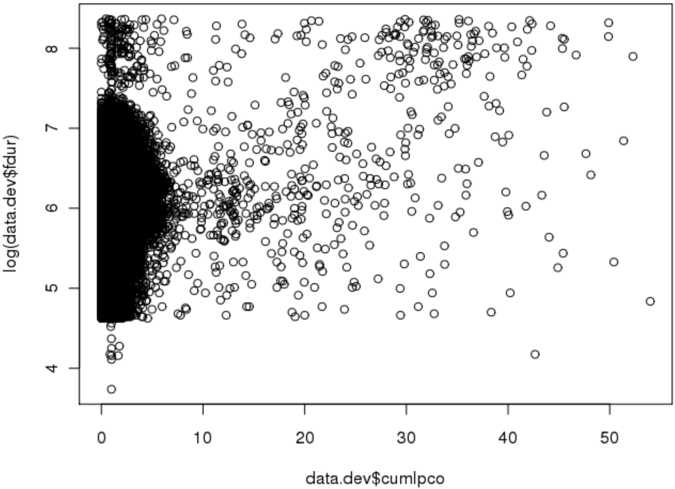
Graphs



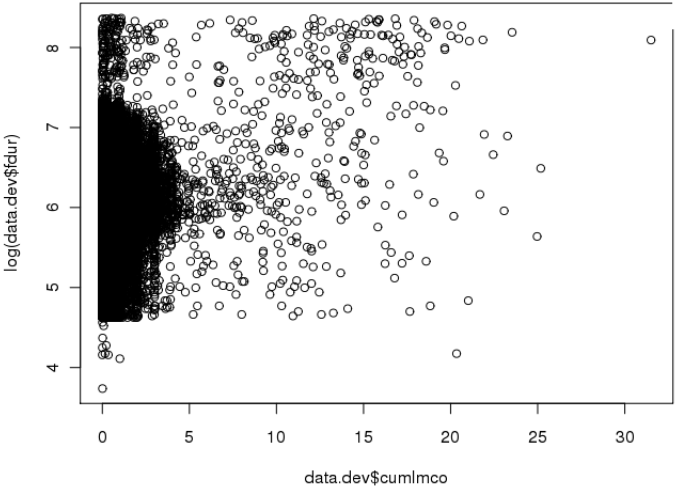
Graphs



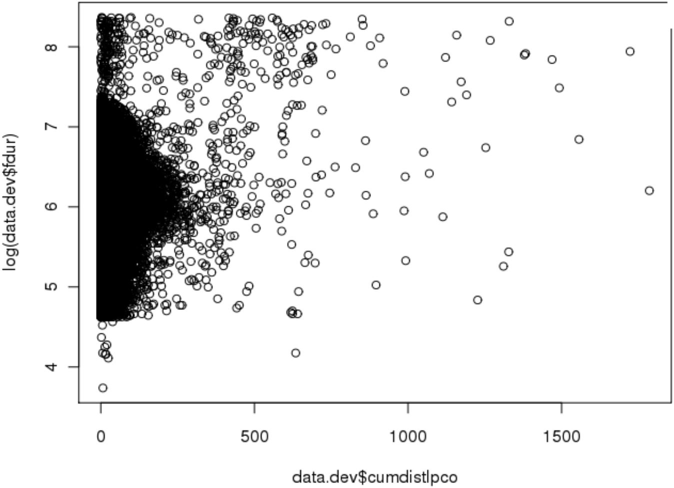
Graphs



Graphs



Graphs



Shhh. Secrets!

Look no farther! (Secrets...)

Shhh. Secrets!

Metrics residualized from baseline on test set

Model	Improvement (p-value)	Model	p-value
B+	–	B-	$2.61 * e^{-06}$
DistB+	.043	–	–

Metric improvement over baseline

Bibliography I

Chen, S. F. and Goodman, J. (1998).
An empirical study of smoothing techniques for language modeling.
Technical report, Harvard University.

Fossum, V. and Levy, R. (2012).
Sequential vs. hierarchical syntactic models of human incremental sentence processing.
In *Proceedings of CMCL-NAACL 2012*. Association for Computational Linguistics.

Francis, W. N. and Kucera, H. (1979).
The brown corpus: A standard corpus of present-day edited american english.

Bibliography II

Frank, S. and Bod, R. (2011).
Insensitivity of the human sentence-processing system to hierarchical structure.
Psychological Science.

Gibson, E. (2000).
The dependency locality theory: A distance-based theory of linguistic complexity.
In *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126.

Hale, J. (2001).
A probabilistic earley parser as a psycholinguistic model.
In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 159–166, Pittsburgh, PA.

Bibliography III

Kennedy, A., Pynte, J., and Hill, R. (2003).
The dundee corpus.
In *Proceedings of the 12th European conference on eye movement*.

Levy, R. (2008).
Expectation-based syntactic comprehension.
Cognition, 106(3):1126–1177.

Levy, R. and Jaeger, F. T. (2007).
Speakers optimize information density through syntactic reduction.
In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993).
Building a large annotated corpus of English: the Penn Treebank.
Computational Linguistics, 19(2):313–330.

Bibliography IV

Roark, B. (2001).
Probabilistic top-down parsing and language modeling.
Computational Linguistics, 27(2):249–276.

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009).
Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing.
Proceedings of the 2009 Conference on Empirical Methods in Natural Langauge Processing, pages 324–333.

Schuler, W. (2009).
Parsing with a bounded stack using a model-based right-corner transform.
In *Proceedings of NAACL/HLT 2009*, NAACL '09, pages 344–352, Boulder, Colorado. Association for Computational Linguistics.

Bibliography V

Schuler, W., AbdelRahman, S., Miller, T., and Schwartz, L. (2010).
Broad-coverage incremental parsing using human-like memory constraints.
Computational Linguistics, 36(1):1–30.

Sederberg, P. B., Howard, M. W., and Kahana, M. J. (2008).
A context-based theory of recency and contiguity in free recall.
Psychological Review, 115:893–912.

Stolcke, A. (2002).
Srilm – and extensible language modeling toolkit.
In *Seventh International Conference on Spoken Language Processing*.