Prediction- and Recall-Defined Online Complexity Metrics

Marten van Schijndel
Department of Linguistics
The Ohio State University

December 5, 2012

Motivation

Observation isn't explanation

Current metrics predict complexity with no cognitive explanation.

- Surprisal simply reflects corpus statistics.
- Entropy reduction and UID reflect interpreted corpus statistics.

Goal: An Explanation

- Can current theories of working memory predict difficulty over extant complexity metrics?
- Provide a rationale for *why* humans have certain difficulties

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)

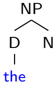
Assumption: Parallel processing (competing hypotheses)

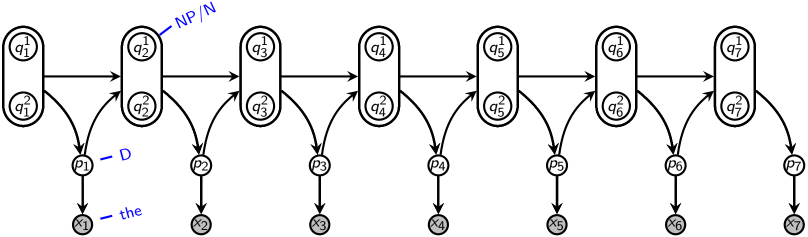
Cueing Predictions

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

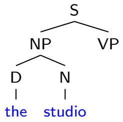
- Sequential (skilled, content-based) cueing [Botvinick, 2007]
- Temporal (context-based) cueing [Howard and Kahana, 2002]
- Naturally lends itself to center-embedding

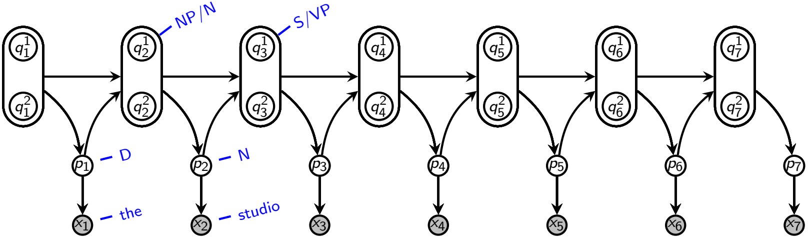
Practice Parse #1364



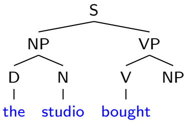


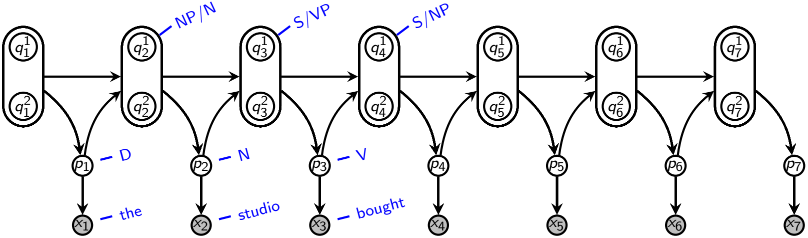
Practice Parse #1364



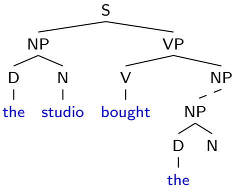


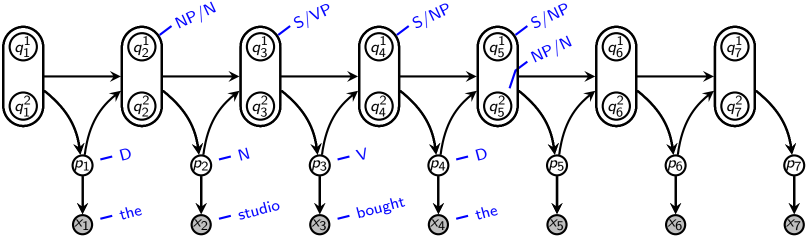
Practice Parse #1364



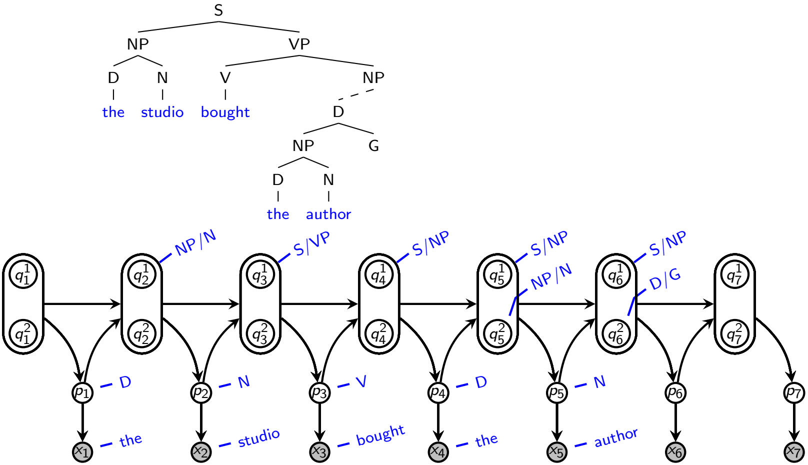


Practice Parse #1364

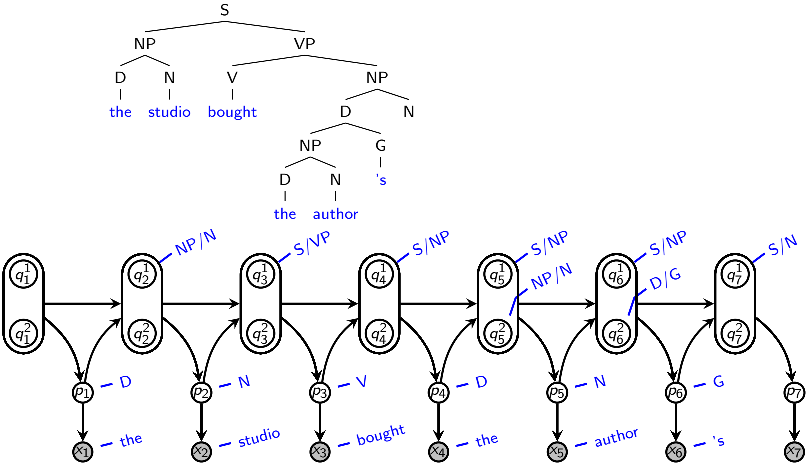




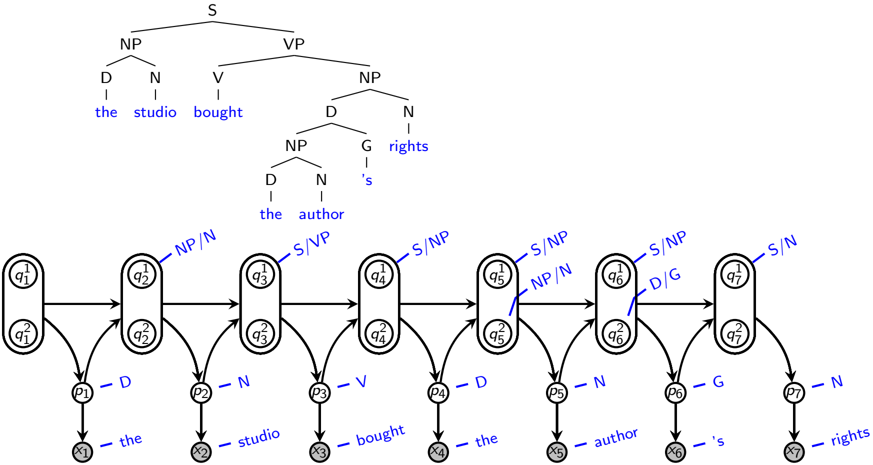
Practice Parse #1364



Practice Parse #1364

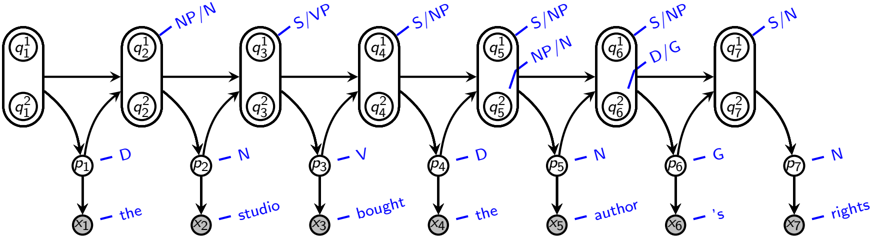


Practice Parse #1364



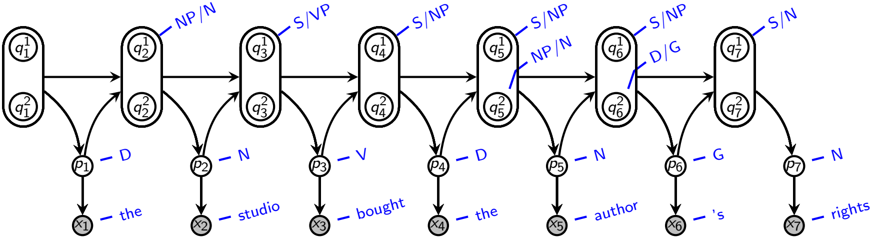
Cueing in Parsing

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



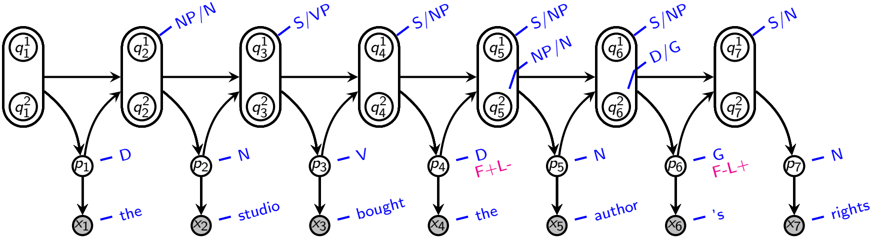
Parser Predictions

- 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)
- DepF+: F+ weighted by the tier number
- L+: Predict integration of a tier (incur an *integration* cost)
- DepL+: L+ weighted by the tier number

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: Go-Past 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 (go-past durations)
- Filtered Dundee corpus
  - 154,168 words

Exclusions: UNK-threshold 5, first and last of a line, fixations skipping an entire line (track/attention loss)

Baseline Metrics

Fitting a linear mixed effects model

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

- Number of characters
- Previous (next) word fixated?
- Unigram and Bigram probs
- Sentence position
- Joint interactions

Plus

- Spillover Predictors
- Number of intervening words
- Cum. Total Surprisal [Hale, 2001]
- Cum. Entropy Reduction [Hale, 2003]

Durations are log-transformed prior to fitting to yield more normal distributions

Results

Metrics residualized from baseline

Model	t-score	p-value
F–L–	3.13	.0017
F+L–	2.76	.0058
F–L+	-3.16	.0016
F+L+	–	–

Model	t-score	p-value
F+	–	–
DepF+	–	–
L+	-3.68	.0002
DepL+	-4.47	$8 \cdot 10^{-6}$

Model	t-score	p-value
DepF+L–	–	–
DepF–L+	-3.81	.0001
DepF+L+	–	–

Significance of Improvement over Baseline

Discussion

Corroborates

- Antilocality in ACT-R [Vasishth and Lewis, 2006]
- Embedding difference [Wu et al., 2010]

Possible Explanations

- Processing 'momentum' [Just and Varma, 2007]
- Increased resting activation

Conclusion

An explanation

- Some proposed metrics can predict reading times even over a strong baseline
- Indicates that domain-general memory processes provide at least a partial account of *why* language processing difficulties occur.

Plus

- Suggests antilocality effects present in English, too.

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?

Results: The Villains

Metrics residualized from baseline (w/o complexity) (w/FL)

Model	t-score	p-value
Totsurp	–	$< 2.2 \cdot 10^{-16}$
Totsurp$_R$	13.82	$< 2.2 \cdot 10^{-16}$
Lexsurp	–	$< 2.2 \cdot 10^{-16}$
Lexsurp$_R$	13.26	$< 2.2 \cdot 10^{-16}$
Synsurp	–	$1 \cdot 10^{-6}$
Synsurp$_R$	3.21	.001
EntRed	–	–
EntRed$_R$	–	–

Model	t-score	p-value
Totsurp	–	$< 2.2 \cdot 10^{-16}$
Totsurp$_R$	10.89	$< 2.2 \cdot 10^{-16}$
Lexsurp	–	$< 2.2 \cdot 10^{-16}$
Lexsurp$_R$	11.41	$< 2.2 \cdot 10^{-16}$
Synsurp	–	–
Synsurp$_R$	–	–
EntRed	–	.04
EntRed$_R$	–	.32

Significance of Improvement over Baseline

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

Bibliography I

Botvinick, M. (2007).
Multilevel structure in behavior and in the brain: a computational model of Fuster's hierarchy.
Philosophical Transactions of the Royal Society, Series B: Biological Sciences, 362:1615–1626.

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 2012*. Association for Computational Linguistics.

Bibliography II

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

Frank, S. (Forthcoming).
Uncertainty reduction as a measure of cognitive load in sentence comprehension.
Topics in Cognitive Science.

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

Bibliography III

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, Cambridge, MA. MIT Press.

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.

Hale, J. (2003).
Grammar, Uncertainty and Sentence Processing.
PhD thesis, Cognitive Science, The Johns Hopkins University.

Bibliography IV

Howard, M. W. and Kahana, M. J. (2002).
A distributed representation of temporal context.
Journal of Mathematical Psychology, 45:269–299.

Just, M. A. and Varma, S. (2007).
The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition.
Cognitive, Affective, & Behavioral Neuroscience, 7:153–191.

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

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 V

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

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.

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.

Bibliography VI

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

Vasishth, S. and Lewis, R. L. (2006).
Argument-head distance and processing complexity: Explaining both locality and antilocality effects.
Language, 82(4):767–794.

Wu, S., Bachrach, A., Cardenas, C., and Schuler, W. (2010).
Complexity metrics in an incremental right-corner parser.
In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 1189–1198.