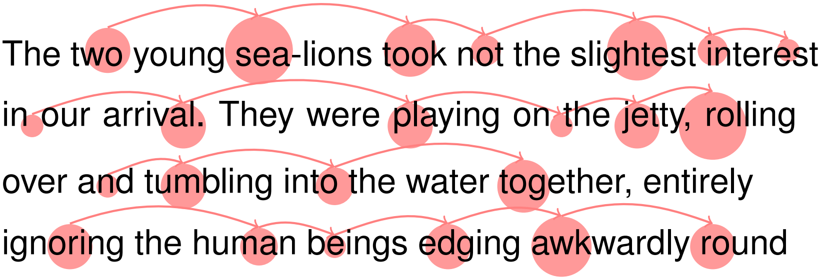
Modeling Human Reading with Neural Attention

Michael Hahn	Frank Keller
Stanford University	University of Edinburgh
mhahn2@stanford.edu	keller@inf.ed.ac.uk

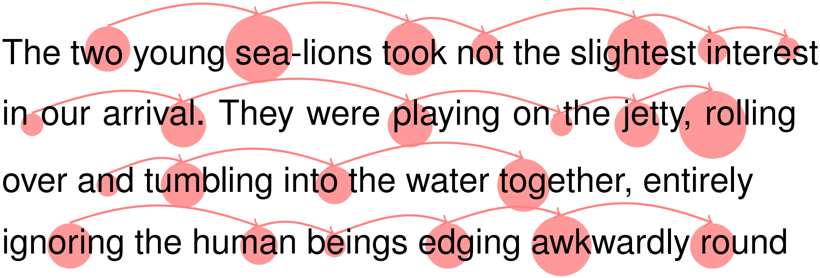
EMNLP 2016

Eye Movements in Human Reading



adapted from the Dundee corpus [Kennedy and Pynte, 2005]

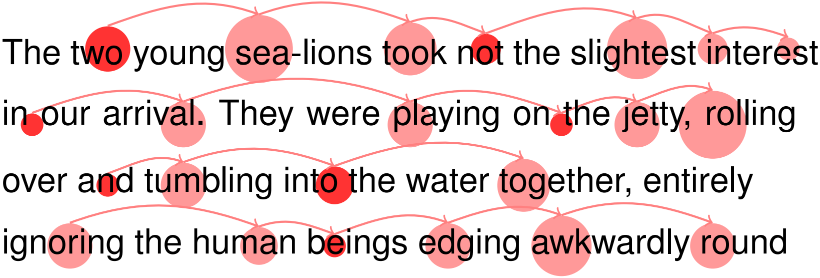
Eye Movements in Human Reading



adapted from the Dundee corpus [Kennedy and Pynte, 2005]

- **Fixations** static
- **Saccades** take 20–40 ms, no information obtained from text

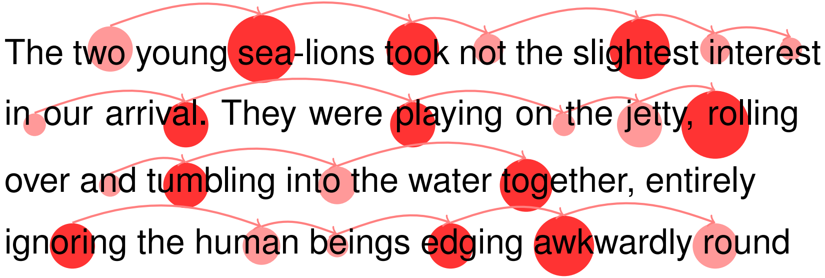
Eye Movements in Human Reading



adapted from the Dundee corpus [Kennedy and Pynte, 2005]

- **Fixations** static
- **Saccades** take 20–40 ms, no information obtained from text
- **Fixation times** vary from ≈ **100 ms** to ≈ 300ms

Eye Movements in Human Reading



adapted from the Dundee corpus [Kennedy and Pynte, 2005]

- **Fixations** static
- **Saccades** take 20–40 ms, no information obtained from text
- **Fixation times** vary from $\approx$ 100 ms to $\approx$ **300ms**

Eye Movements in Human Reading

The two young sea-lions took not the slightest interest
in our arrival. They were playing on the jetty, rolling
over and tumbling into the water together, entirely
ignoring the human beings edging awkwardly round

adapted from the Dundee corpus [Kennedy and Pynte, 2005]

- **Fixations** static
- **Saccades** take 20–40 ms, no information obtained from text
- **Fixation times** vary from $\approx$ 100 ms to $\approx$ 300ms
- $\approx 40\%$ of words are **skipped**

Computational Models I

1. models of saccade generation in cognitive psychology
    - EZ-Reader [Reichle et al., 1998, 2003, 2009]
    - SWIFT [Engbert et al., 2002, 2005]
    - Bayesian inference [Bicknell and Levy, 2010]
2. machine learning models trained on eye-tracking data [Nilsson and Nivre, 2009, 2010, Hara et al., 2012, Matthies and Søgaard, 2013]

Computational Models I

1. models of saccade generation in cognitive psychology
   - EZ-Reader [Reichle et al., 1998, 2003, 2009]
   - SWIFT [Engbert et al., 2002, 2005]
   - Bayesian inference [Bicknell and Levy, 2010]
2. machine learning models trained on eye-tracking data [Nilsson and Nivre, 2009, 2010, Hara et al., 2012, Matthies and Søgaard, 2013]

These models...

- involve theoretical assumptions about human eye-movements, or
- require selection of relevant eye-movement features, and
- estimate parameters from eye-tracking corpora

Computational Models II: Surprisal

$$\text{Surprisal}(w_i|\boldsymbol{w}_{1\ldots i-1}) = -\log P(w_i|\boldsymbol{w}_{1\ldots i-1}) \tag{1}$$

- measures predictability of word in context
- computed by language model

Computational Models II: Surprisal

$$\text{Surprisal}(w_i|\boldsymbol{w}_{1\ldots i-1}) = -\log P(w_i|\boldsymbol{w}_{1\ldots i-1}) \tag{1}$$

- measures predictability of word in context
- computed by language model
- correlates with word-by-word reading times [Hale, 2001, McDonald and Shillcock, 2003a,b, Levy, 2008, Demberg and Keller, 2008, Frank and Bod, 2011, Smith and Levy, 2013]
- but cannot explain...
  - reverse saccades
  - re-fixations
  - spillover
  - skipping
    $\approx$ 40% of words are skipped

Tradeoff Hypthesis

Goal

Build unsupervised models jointly accounting for **reading times** and **skipping**

Tradeoff Hypthesis

Goal

Build unsupervised models jointly accounting for **reading times** and **skipping**

- reading is recent innovation in evolutionary terms
- humans learn it without access to other people's eye-movements

Tradeoff Hypthesis

Goal

Build unsupervised models jointly accounting for **reading times** and **skipping**

- reading is recent innovation in evolutionary terms
- humans learn it without access to other people's eye-movements

Hypothesis

Human reading optimizes a tradeoff between

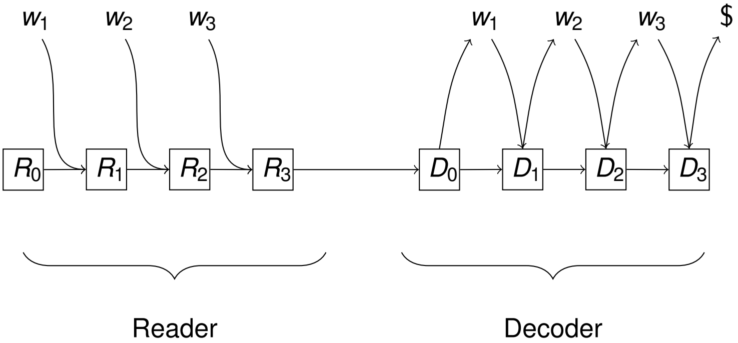
- **Precision** of language understanding:
  Encode the input so that it can be reconstructed accurately
- **Economy** of attention:
  Fixate as few words as possible

Tradeoff Hypothesis

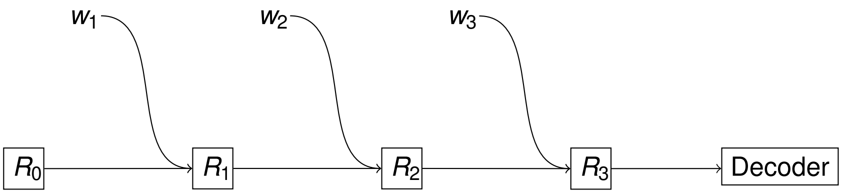
Approach: NEAT (NEural Attention Tradeoff)

1. develop generic architecture integrating
   - neural language modeling
   - attention mechanism
2. train end-to-end to optimize tradeoff between precision and economy
3. evaluate on human eyetracking corpus

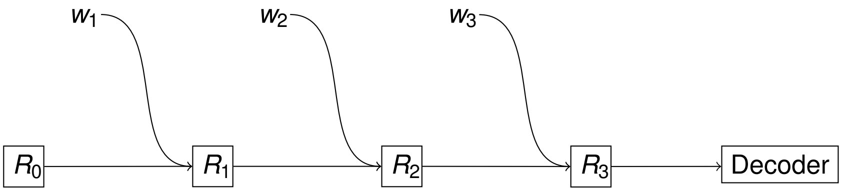
Architecture I: Recurrent Autoencoder



Architecture II: Real-Time Predictions

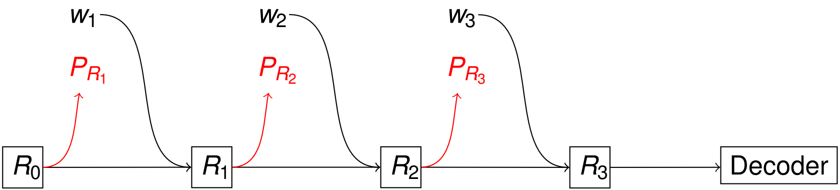


Architecture II: Real-Time Predictions



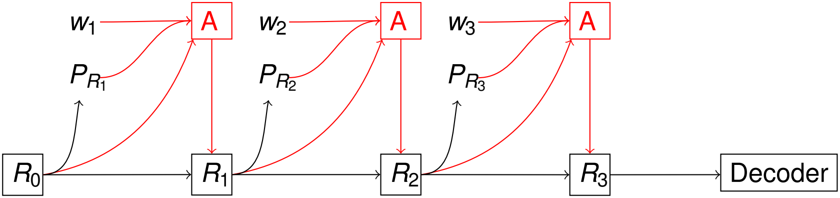
- Humans constantly make predictions about the upcoming input

Architecture II: Real-Time Predictions



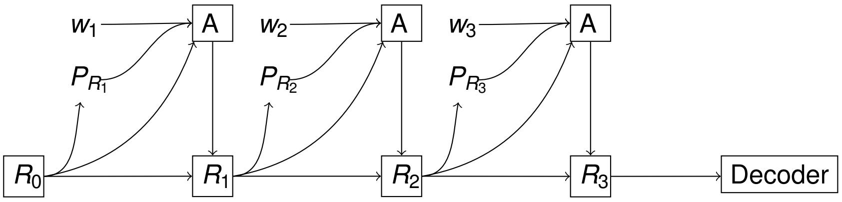
- Humans constantly make predictions about the upcoming input
- Reader outpus probability distribution P_R over the lexicon at each time step
- Describes which words are likely to come next

Architecture III: Skipping



- Attention module **shows word** to R or **skips** it

Architecture III: Skipping



- Attention module **shows word** to R or **skips** it
- A computes a probability + draws a sample $\omega \in \{\text{READ}, \text{SKIP}\}$
- R receives special 'SKIPPED' vector when skipping

Implementing the Tradeoff Hypothesis

Training Objective

Solve prediction and reconstruction with minimal attention:

$$\arg_\theta \min \{ \mathrm{E}_{\boldsymbol{w},\boldsymbol{\omega}} [L(\boldsymbol{\omega} | \boldsymbol{w}, \theta) + \alpha \cdot \| \boldsymbol{\omega} \|_{\ell_1}] \}$$

Loss on Prediction + Reconstruction

# of fixated words

Implementing the Tradeoff Hypothesis

Training Objective

Solve prediction and reconstruction with minimal attention:

$$\arg_\theta \min \{ \mathrm{E}_{\boldsymbol{w},\boldsymbol{\omega}} [L(\boldsymbol{\omega} | \boldsymbol{w}, \theta) + \alpha \cdot ||\boldsymbol{\omega}||_{\ell_1}] \}$$

Loss on Prediction + Reconstruction

\# of fixated words

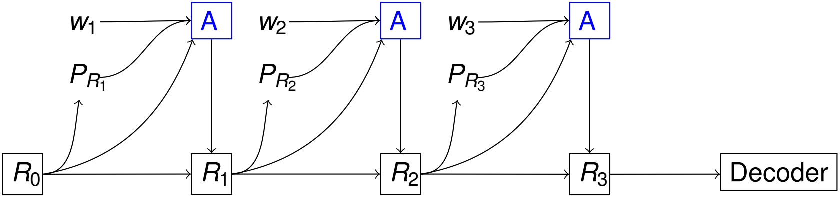
- $\boldsymbol{w}$ is word sequence drawn from corpus
- $\boldsymbol{\omega}$ sampled from attention module A
- $\alpha > 0$: encourages NEAT to attend to as few words as possible

Implementation and Training

- Implementation
  - one-layer LSTM network with 1,000 memory cells
  - attention network: one-layer feedforward network
- optimized by SGD + REINFORCE policy gradient method [Williams, 1992]
- trained on corpus of newstext [Hermann et al., 2015]
  - 195,462 articles from Daily Mail
  - $\approx$ 200 million tokens
- Input data split into sequences of 50 tokens

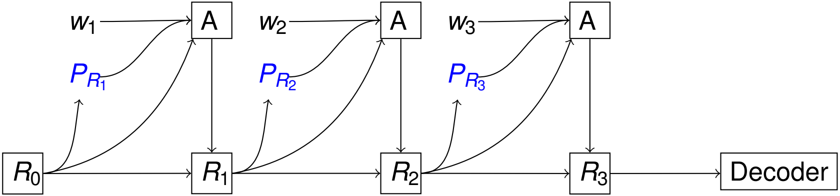
NEAT as a Model of Reading

- **Attention module** models fixations and skips
- **NEAT surprisal** models reading times of fixated words



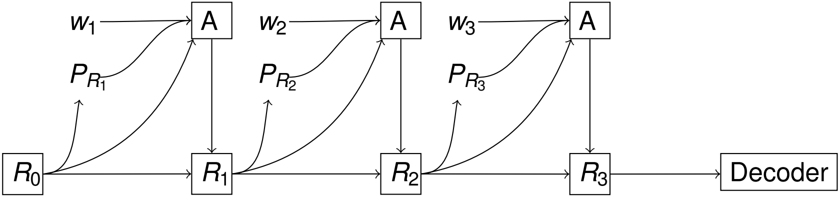
NEAT as a Model of Reading

- **Attention module** models fixations and skips
- **NEAT surprisal** models reading times of fixated words



NEAT as a Model of Reading

- **Attention module** models fixations and skips
- **NEAT surprisal** models reading times of fixated words



The only ingredients are

- architecture
- objective
- unlabeled corpus

No eye-tracking data, lexicon, grammar, ... needed.

Evaluation Setup

- English section of the Dundee corpus [Kennedy and Pynte, 2005]
  - 20 texts from *The Independent*
  - annotated with eye-movement data from ten English native speakers who were asked to answer questions after each text.
- split into development (1–3) and test set (4–20)
- Size: 78,300 tokens (dev); 281,911 tokens (test)
- exclude from the evaluation words at the beginning or end of lines, outliers, cases of track loss, out-of-vocabulary words
- Fixation rate: 62.1% (dev), 61.3% (test)

Intrinsic Evaluation: Prediction and Reconstruction

	Perplexity		Fix. Rate
	Prediction	Reconstruction	
NEAT	180	4.5	60.4%
$\omega \sim Bin(0.62)$	333	56	62.1%
Word Length	230	40	62.1%
Word Freq.	219	39	62.1%
Full Surprisal	211	34	62.1%
Human	218	39	61.3%
$\omega \equiv 1$	107	1.6	100%

- For Word Length, Word Frequency, Full Surprisal, we take threshold predictions matching the fixation rate of the development set.

Intrinsic Evaluation: Prediction and Reconstruction

	Perplexity		Fix. Rate
	Prediction	Reconstruction	
NEAT	180	4.5	60.4%
$\omega \sim Bin(0.62)$	333	56	62.1%
Word Length	230	40	62.1%
Word Freq.	219	39	62.1%
Full Surprisal	211	34	62.1%
Human	218	39	61.3%
$\omega \equiv 1$	107	1.6	100%

- For Word Length, Word Frequency, Full Surprisal, we take threshold predictions matching the fixation rate of the development set.

Evaluating Reading Times: Linear Mixed Models

$$\text{FirstPassDuration} = \beta_0 + \sum_{i \in \text{Predictors}} \beta_i x_i + \sum_{j \in \text{RandomEffects}} \gamma_j y_j + \varepsilon$$

Evaluating Reading Times: Linear Mixed Models

$$\text{FirstPassDuration} = \beta_0 + \sum_{i \in \text{Predictors}} \beta_i x_i + \sum_{j \in \text{RandomEffects}} \gamma_j y_j + \varepsilon$$

	β	SE	t
(Intercept)	247.4	7.1	34.7*
Word Length	12.9	0.2	60.6*
Previous Word Freq.	−5.3	0.3	−18.3*
Prev. Word Fixated	−24.7	0.8	−30.6*
Obj. Landing Pos.	−8.1	0.2	−41.3*
Word Pos. in Sent.	−0.1	0.03	−3.0*
Log Word Freq.	−1.6	0.2	−7.7*
Launch Distance	−0.005	0.01	−0.4
Residualized NEAT Surprisal	2.8	0.1	23.7*

Baseline Predictors

Evaluating Reading Times: Linear Mixed Models

$$\text{FirstPassDuration} = \beta_0 + \sum_{i \in \text{Predictors}} \beta_i x_i + \sum_{j \in \text{RandomEffects}} \gamma_j y_j + \varepsilon$$

	β	SE	t	
(Intercept)	247.4	7.1	34.7*	
Word Length	12.9	0.2	60.6*	Baseline Predictors
Previous Word Freq.	-5.3	0.3	-18.3*	
Prev. Word Fixated	-24.7	0.8	-30.6*	
Obj. Landing Pos.	-8.1	0.2	-41.3*	
Word Pos. in Sent.	-0.1	0.03	-3.0*	
Log Word Freq.	-1.6	0.2	-7.7*	
Launch Distance	-0.005	0.01	-0.4	
Residualized NEAT Surprisal	2.8	0.1	23.7*	

Evaluating Reading Times: Linear Mixed Models

$$\text{FirstPassDuration} = \beta_0 + \sum_{i \in \text{Predictors}} \beta_i x_i + \sum_{j \in \text{RandomEffects}} \gamma_j y_j + \varepsilon$$

	β	SE	t	
(Intercept)	247.4	7.1	34.7*	
Word Length	12.9	0.2	60.6*	Baseline Predictors
Previous Word Freq.	-5.3	0.3	-18.3*	
Prev. Word Fixated	-24.7	0.8	-30.6*	
Obj. Landing Pos.	-8.1	0.2	-41.3*	
Word Pos. in Sent.	-0.1	0.03	-3.0*	
Log Word Freq.	-1.6	0.2	-7.7*	
Launch Distance	-0.005	0.01	-0.4	
Residualized NEAT Surprisal	2.8	0.1	23.7*	

- NEAT surprisal captures more than word length, frequency, ...
- even though it only has access to 60.4% of the words

Evaluating Reading Times: Deviance

- Assume we have models M_1, M_2 for the same data
- They assign likelihoods $P_1 = P(\text{Data}|M_1)$, $P_2 = P(\text{Data}|M_2)$
- **Deviance**

$$2 \times \log \frac{P_2}{P_1}$$

Evaluating Reading Times: Deviance

- Assume we have models M_1, M_2 for the same data
- They assign likelihoods $P_1 = P(\text{Data}|M_1)$, $P_2 = P(\text{Data}|M_2)$
- **Deviance**

$$2 \times \log \frac{P_2}{P_1}$$

- Here:
  M_1: Model containing only baseline predictors
  M_2: Model including surprisal

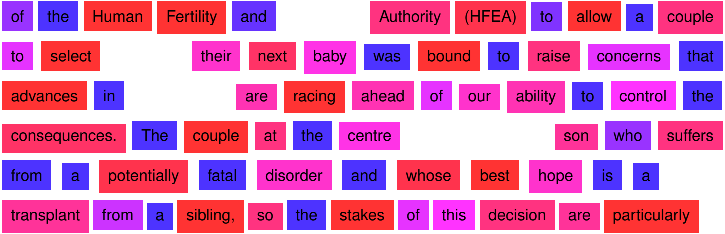
Full surprisal	$\boldsymbol{\omega} \equiv 1$	980
NEAT surprisal	$\boldsymbol{\omega} \equiv P_A(\boldsymbol{w})$	867
Random surprisal	$\boldsymbol{\omega} \equiv Binom(0.604)$	832

Evaluating Fixations I: Heatmaps

Evaluating Fixations I: Heatmaps

HUMAN

MODEL



Evaluating Fixations II: Accuracy

	Acc	$F1_{\text{fix}}$	$F1_{\text{skip}}$
NEAT	63.7	70.4	53.0
Lower and Upper Bounds			
Random Baseline	52.6	62.1	37.9
Intersubject Agreement	69.5	76.6	53.6
Feature-Based Models			
Nilsson and Nivre [2009]	69.5	75.2	62.6
Matthies and Søgaard [2013]	69.9	72.3	66.1
Word Frequency	67.9	74.0	58.3
Word Length	68.4	77.1	49.0

Evaluating Fixations II: Accuracy

	Acc	$F1_{fix}$	$F1_{skip}$
NEAT	63.7	70.4	53.0
Lower and Upper Bounds			
Random Baseline	52.6	62.1	37.9
Intersubject Agreement	69.5	76.6	53.6
Feature-Based Models			
Nilsson and Nivre [2009]	69.5	75.2	62.6
Matthies and Søgaard [2013]	69.9	72.3	66.1
Word Frequency	67.9	74.0	58.3
Word Length	68.4	77.1	49.0

- NEAT outperforms random baseline

Evaluating Fixations II: Accuracy

	Acc	$F1_{fix}$	$F1_{skip}$
NEAT	63.7	70.4	53.0
Lower and Upper Bounds			
Random Baseline	52.6	62.1	37.9
Intersubject Agreement	69.5	76.6	53.6
Feature-Based Models			
Nilsson and Nivre [2009]	69.5	75.2	62.6
Matthies and Søgaard [2013]	69.9	72.3	66.1
Word Frequency	67.9	74.0	58.3
Word Length	68.4	77.1	49.0

- NEAT outperforms random baseline
- supervised models at upper limit

Evaluating Fixations II: Accuracy

	Acc	$F1_{\text{fix}}$	$F1_{\text{skip}}$
NEAT	63.7	70.4	53.0
Lower and Upper Bounds			
Random Baseline	52.6	62.1	37.9
Intersubject Agreement	69.5	76.6	53.6
Feature-Based Models			
Nilsson and Nivre [2009]	69.5	75.2	62.6
Matthies and Søgaard [2013]	69.9	72.3	66.1
Word Frequency	67.9	74.0	58.3
Word Length	68.4	77.1	49.0

- NEAT outperforms random baseline
- supervised models at upper limit
- bulk of data explained by word length/frequency predictors

Fixations of Successive Words

- Humans more likely to fixate a word when the previous word was skipped

$$P(\omega_i = \text{READ} | \omega_{i-1} = \text{READ}) < P(\omega_i = \text{READ})$$

Fixations of Successive Words

- Humans more likely to fixate a word when the previous word was skipped

$$P(\omega_i = \text{READ} | \omega_{i-1} = \text{READ}) < P(\omega_i = \text{READ})$$

- Ratio:

Setting	$\frac{P(\omega_i = \text{READ} \mid \omega_{i-1} = \text{READ})}{P(\omega_i = \text{READ})}$
NEAT	0.81
Human	0.85
Word Frequency	0.91
Random	1.0

Fixations of Successive Words

- Humans more likely to fixate a word when the previous word was skipped

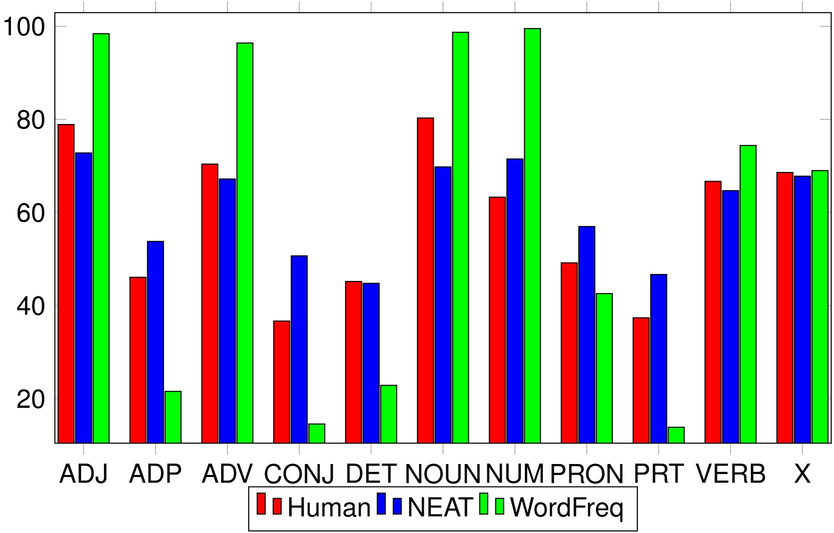
$$P(\omega_i = \text{READ} | \omega_{i-1} = \text{READ}) < P(\omega_i = \text{READ})$$

- Ratio:

Setting	$\frac{P(\omega_i = \text{READ} \mid \omega_{i-1} = \text{READ})}{P(\omega_i = \text{READ})}$
NEAT	0.81
Human	0.85
Word Frequency	0.91
Random	1.0

- Mixed models show effect beyond word frequency

Fixation Rates by POS Categories



Conclusion

- unsupervised model of reading predicting **reading times** and **skipping**
- based on tradeoff between
  precision of understanding $\Leftrightarrow$ **economy** of attention
- trained end-to-end without linguistic knowledge, eyetracking data, or feature extraction
- Experiments on the Dundee corpus
  - provides accurate predictions for human skipping behavior
  - predicts reading times, while only accessing 60.4% of the words
  - known qualitative properties of skipping emerge, without specifying relevant features in advance

References I

K. Bicknell and R. Levy. A rational model of eye movement control in reading. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1168–1178. Association for Computational Linguistics, 2010. URL http://dl.acm.org/citation.cfm?id=1858800.

V. Demberg and F. Keller. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210, 2008. URL http://www.sciencedirect.com/science/article/pii/S0010027708001741.

R. Engbert, A. Longtin, and R. Kliegl. A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision research*, 42(5):621–636, 2002. URL http://www.sciencedirect.com/science/article/pii/S0042698901003017.

R. Engbert, A. Nuthmann, E. M. Richter, and R. Kliegl. SWIFT: A Dynamical Model of Saccade Generation During Reading. *Psychological Review*, 112(4):777–813, 2005. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.112.4.777.

S. Frank and R. Bod. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22: 829–834, 2011.

J. Hale. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of NAACL*, volume 2, pages 159–166, 2001.

T. Hara, D. M. Y. Kano, and A. Aizawa. Predicting word fixations in text with a CRF model for capturing general reading strategies among readers. In *Proceedings of the First Workshop on Eye-tracking and Natural Language Processing*, pages 55–70, 2012. URL http://anthology.aclweb.org/W/W12/W12-49.pdf#page=65.

K. M. Hermann, T. Kočiskỳ, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. *arXiv preprint arXiv:1506.03340*, 2015. URL http://arxiv.org/abs/1506.03340.

A. Kennedy and J. Pynte. Parafoveal-on-foveal effects in normal reading. *Vision Research*, 45(2):153–168, January 2005. URL http://linkinghub.elsevier.com/retrieve/pii/S0042698904003979.

R. Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, March 2008. URL http://linkinghub.elsevier.com/retrieve/pii/S0010027707001436.

F. Matthies and A. Søgaard. With Blinkers on: Robust Prediction of Eye Movements across Readers. In *EMNLP*, pages 803–807, 2013. URL http://www.aclweb.org/website/old_anthology/D/D13/D13-1075.pdf.

S. A. McDonald and R. C. Shillcock. Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, 14(6):648–652, November 2003a.

References II

S. A. McDonald and R. C. Shillcock. Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research*, 43(16):1735–1751, July 2003b. URL http://www.sciencedirect.com/science/article/pii/S0042698903002372.

M. Nilsson and J. Nivre. Learning where to look: Modeling eye movements in reading. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 93–101. Association for Computational Linguistics, 2009. URL http://dl.acm.org/citation.cfm?id=1596392.

M. Nilsson and J. Nivre. Towards a data-driven model of eye movement control in reading. In *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics*, pages 63–71. Association for Computational Linguistics, 2010. URL http://dl.acm.org/citation.cfm?id=1870073.

E. D. Reichle, A. Pollatsek, D. L. Fisher, and K. Rayner. Toward a model of eye movement control in reading. *Psychological Review*, 105(1):125–157, January 1998.

E. D. Reichle, K. Rayner, and A. Pollatsek. The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(04):445–476, 2003. URL http://journals.cambridge.org/abstract_S0140525X03000104.

E. D. Reichle, T. Warren, and K. McConnell. Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, 16(1):1–21, February 2009. URL http://www.springerlink.com/index/10.3758/PBR.16.1.1.

N. J. Smith and R. Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319, September 2013. URL http://linkinghub.elsevier.com/retrieve/pii/S0010027713000413.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4): 229–256, 1992. URL http://link.springer.com/article/10.1007/BF00992696.

Correlations with Known Predictors

	Human	NEAT
Restricted Surprisal	0.465	0.762
Full Surprisal	0.512	0.720
Log Word Freq.	−0.608	−0.760
Word Length	0.663	0.521