Addressing surprisal deficiencies in reading time models

Marten van Schijndel William Schuler

December 11, 2016

Department of Linguistics, The Ohio State University

Overview

- Surprisal (PCFG, *N*-gram) is a way to estimate text complexity

Overview

- Surprisal (PCFG, *N*-gram) is a way to estimate text complexity
- Experienced complexity is reflected in reading speed

- Surprisal (PCFG, *N*-gram) is a way to estimate text complexity
- Experienced complexity is reflected in reading speed

Claim:

Current surprisal models inadequately estimate reading complexity

Overview

- Surprisal (PCFG, *N*-gram) is a way to estimate text complexity
- Experienced complexity is reflected in reading speed

Claim:
Current surprisal models inadequately estimate reading complexity

This work:
A simple tweak to fix surprisal

Reading complexity is estimated based on region ending

1 2

The red apple that the $\boxed{\text{girl}}$ ate …

Reading complexity is estimated based on region ending

The red apple that the $\boxed{\text{girl}}$ ate ...

w_1 w_2 w_3 w_4 w_5 w_6

Reading model of 'girl':
sentence position

Reading complexity is estimated based on region ending

$$\text{The red apple that the } \underbrace{\overbrace{\boxed{\text{girl}}}^{\text{4 chars}}}_{w_6} \text{ ate} \ldots$$

Reading model of ‘girl’:
sentence position, word length

$$\overbrace{\boxed{\text{girl}}}^{\text{4 chars}}$$

The red apple that <u>the</u> $\underset{w_6}{\boxed{\text{girl}}}$ ate ...

Reading model of 'girl':
sentence position, word length, P(girl|the)

Reading complexity is estimated based on region ending

The $\overset{1}{\text{red}}$ apple that the $\overset{2}{\boxed{\text{girl}}}$ ate ...

↑
important

Reading model of ‘girl’:
sentence position, word length, P(girl|the)

Reading complexity is estimated based on region ending

$$\text{The } \overset{1}{\text{red}}\ \underbrace{\text{apple that the}}_{\substack{\uparrow \\ ?}}\ \overset{2}{\boxed{\text{girl}}}\ \text{ate} \ldots$$

Reading model of ‘girl’:
sentence position, word length, P(girl|the)

Reading complexity is estimated based on region ending

$$\text{The red}\overset{1}{\ }\underbrace{\text{apple that the}}\ \overset{2}{\boxed{\text{girl}}}\ \text{ate} \ldots$$

$$\uparrow\ ?\ \downarrow$$

$$\text{The red}\overset{1}{\ }\overbrace{\text{canoe that the}}\ \overset{2}{\boxed{\text{girl}}}\ \text{ate} \ldots$$

Reading model of ‘girl’:
sentence position, word length, P(girl|the)

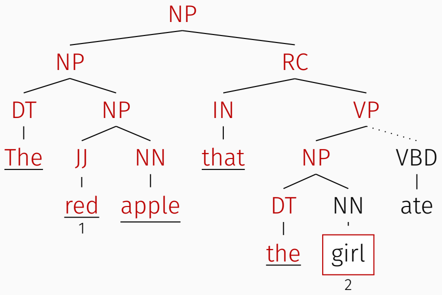
This study: *n*-gram and PCFG surprisal

This study: *n*-gram and PCFG surprisal

The red apple that the $\boxed{\text{girl}}$ ate ...

$$N\text{-gram-surp}(\text{girl}) = -\log \mathrm{P}(\text{girl} \mid \text{the})$$

This study: *n*-gram and PCFG surprisal



$$\text{PCFG-surp(girl)} = -\log \mathrm{P}(T_6 = \text{girl} \mid T_1 \ldots T_5 = \text{The} \ldots \text{the})$$

Accumulated surprisal fixes the theoretical problem

Cumulative *N*-gram Surprisal

The red[1] apple that the girl[2] ate ...

Cumulative N-gram Surprisal

The $\overset{1}{\underline{\text{red}}}$ $\boxed{\text{apple}}$ that the $\overset{2}{\text{girl}}$ ate …

$$\text{cumu-}n\text{-gram}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}+1}^{f_t} -\log \mathrm{P}(w_i \mid w_{i-n} \ldots w_{i-1})$$

Cumulative N-gram Surprisal

The red^{1} $\boxed{\underline{\text{apple}}}$ $\boxed{\text{that}}$ the girl^{2} ate ...

$$\text{cumu-}n\text{-gram}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}+1}^{f_t} -\log \mathsf{P}(w_i \mid w_{i-n} \ldots w_{i-1})$$

Cumulative N-gram Surprisal

The red[1] $\boxed{\text{apple}}$ $\boxed{\underline{\text{that}}}$ $\boxed{\text{the}}$ girl[2] ate ...

$$\text{cumu-}n\text{-gram}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}+1}^{f_t} -\log \mathrm{P}(w_i \mid w_{i-n} \ldots w_{i-1})$$

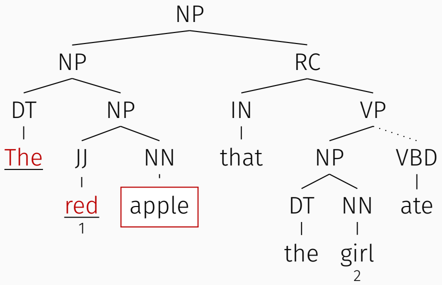
Accumulated surprisal fixes the theoretical problem

Cumulative *N*-gram Surprisal

The red[1] $\boxed{\text{apple}}$ $\boxed{\text{that}}$ $\boxed{\underline{\text{the}}}$ $\boxed{\text{girl}}$[2] ate ...

$$\text{cumu-}n\text{-gram}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}+1}^{f_t} -\log \mathrm{P}(w_i \mid w_{i-n} \ldots w_{i-1})$$

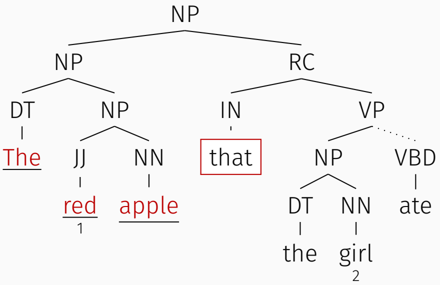
Cumulative PCFG Surprisal



$$\text{Cumu-PCFG}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}}^{f_t} -\log \mathrm{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

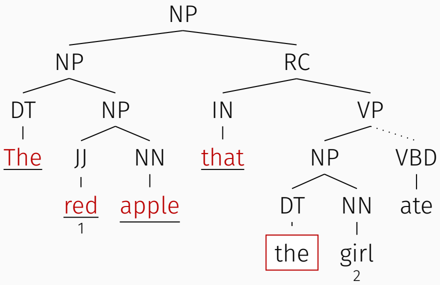
Accumulated surprisal fixes the theoretical problem

Cumulative PCFG Surprisal



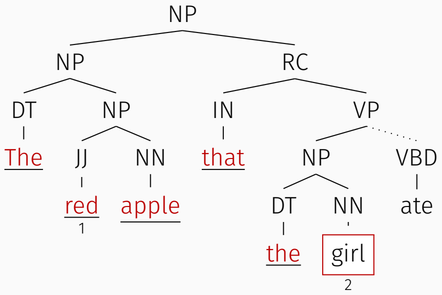
$$\text{Cumu-PCFG}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}}^{f_t} -\log \mathrm{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

Cumulative PCFG Surprisal



$$\text{Cumu-PCFG}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}}^{f_t} -\log \mathrm{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

Cumulative PCFG Surprisal



$$\text{Cumu-PCFG}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}}^{f_t} -\log \mathrm{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

How well does this fix work?

N-gram surprisal

- 5-grams
- Trained on Gigaword 3.0 (Graff and Cieri, 2003)
- Computed with KenLM (Heafield et al., 2013)

How well does this fix work?

N-gram surprisal

- 5-grams
- Trained on Gigaword 3.0 (Graff and Cieri, 2003)
- Computed with KenLM (Heafield et al., 2013)

PCFG surprisal

- Nguyen et al., (2012) Generalized Categorial Grammar
- Trained on WSJ 02-21 (Marcus et al., 1993)
- Computed with van Schijndel et al., (2013) parser

How well does this fix work?

University College London (UCL) Corpus (Frank et al., 2013)

- 43 subjects
- reading short sentences from online novels
- frequent comprehension questions

Baseline mixed effects model

Fixed Factors

- sentence position
- word length
- region length
- whether the previous word was fixated

How well does this fix work?

Baseline mixed effects model

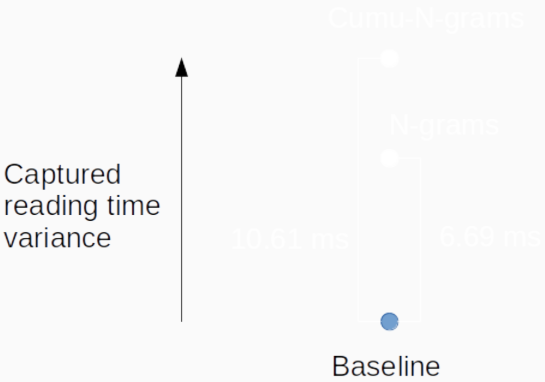
Fixed Factors

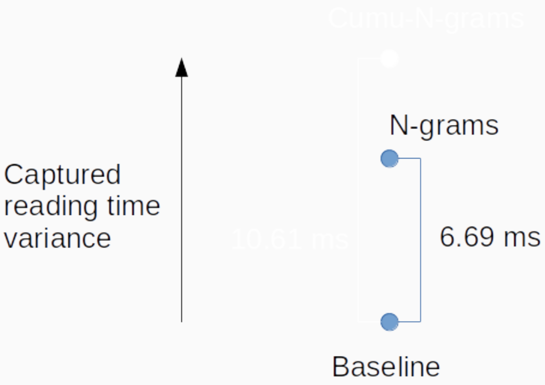
- sentence position
- word length
- region length
- whether the previous word was fixated

Random Factors

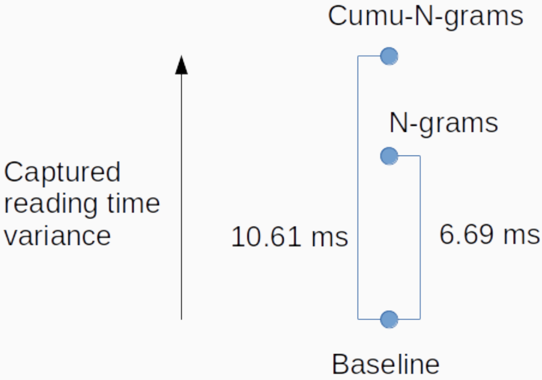
- All fixed factors as by-subject random slopes
- Item, subject and subject$\times$sentence intercepts

Accumulation improves *n*-gram surprisal




N-grams
Captured reading time variance
6.69 ms
Baseline

Accumulation improves n-gram surprisal



Accumulation does not help PCFG surprisal

After adding cumulative *n*-gram surprisal to model:

Accumulation does not help PCFG surprisal

After adding cumulative *n*-gram surprisal to model:

- PCFG surprisal is not useful ($p > 0.05$)

Accumulation does not help PCFG surprisal

After adding cumulative *n*-gram surprisal to model:

- PCFG surprisal is not useful ($p > 0.05$)
- Cumulative PCFG surprisal is not useful ($p > 0.05$)

What does accumulation model?

Subsequent regression

The red[1] apple that the girl ate …

Subsequent regression

The red[1] apple that the girl[2] ate ...

Subsequent regression

The red[1] apple[3] that the girl[2] ate …

Possible accumulation influences

Subsequent regression

1 3 4 2

The red apple that the girl ate …

Possible accumulation influences

Subsequent regression

1 3 4 2 5
The red apple that the girl ate …

Parafovial processing

1

The red apple that the girl ate ...

Parafovial processing

Th(e red apple that t)he girl ate …

Parafovial processing

Th(e red[1] apple that t)he girl[2] ate …

Prediction (entropy)

The red[1] apple that the girl ate …

Possible accumulation influences

Prediction (entropy)

The red[1] (apple that the girl) ate …

Prediction (entropy)

The red[1] (apple that the girl[2]) ate …

Accumulation alternative: Successor surprisal

Cumulative surprisal only handles subsequent regression

Accumulation alternative: Successor surprisal

Cumulative surprisal only handles subsequent regression

Parafovial: Th(e $\overset{1}{\text{red}}$ apple that t)he $\overset{2}{\text{girl}}$ ate …

Prediction: The $\overset{1}{\text{red}}$ $\underbrace{\text{(apple that the}}_{\text{accumulated}}$ $\overset{2}{\text{girl}}$) ate …

Accumulation alternative: Successor surprisal

Cumulative surprisal only handles subsequent regression

Parafovial: Th(e $\overset{1}{\text{red}}$ apple that t)he $\overset{2}{\text{girl}}$ ate …

Prediction: The $\overset{1}{\text{red}}$ $\underbrace{\text{(apple that the}}_{\text{accumulated}}$ $\overset{2}{\text{girl}}$) ate …

Other accumulation mechanisms presuppose earlier accumulation

Successor effects influence reading times

Upcoming material influences reading times

Upcoming material influences reading times

- Orthographic effects
  (Pynte, Kennedy, & Ducrot, 2004; Angele, Tran, & Rayner, 2013)

Successor effects influence reading times

Upcoming material influences reading times

- Orthographic effects
  (Pynte, Kennedy, & Ducrot, 2004; Angele, Tran, & Rayner, 2013)
- Lexical effects
  (Kliegl et al., 2006; Li et al., 2014; Angele et al., 2015)

Successor N-grams

The red[1] apple that the girl[2] ate …

$$\text{future-}n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(w_i \mid w_{i-n} \ldots w_{i-1})$$

The $\underset{}{\overset{1}{\underline{\text{red}}}}$ $\boxed{\text{apple}}$ that the $\overset{2}{\text{girl}}$ ate ...

$$\text{future-}n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(w_i \mid w_{i-n} \ldots w_{i-1})$$

Successor N-grams

The red $\overset{1}{}$ $\boxed{\underline{\text{apple}}}$ $\boxed{\text{that}}$ the $\overset{2}{\text{girl}}$ ate …

$$\text{future-}n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(w_i \mid w_{i-n} \ldots w_{i-1})$$

Successor *N*-grams

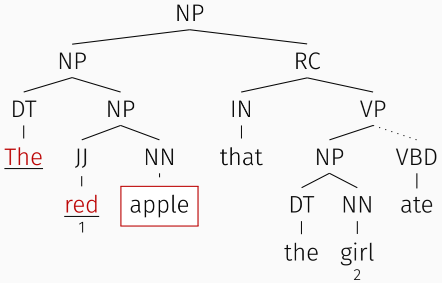
The red[1] $\boxed{\text{apple}}$ $\boxed{\underline{\text{that}}}$ $\boxed{\text{the}}$ girl[2] ate ...

$$\text{future-}n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(w_i \mid w_{i-n} \ldots w_{i-1})$$

Successor N-grams

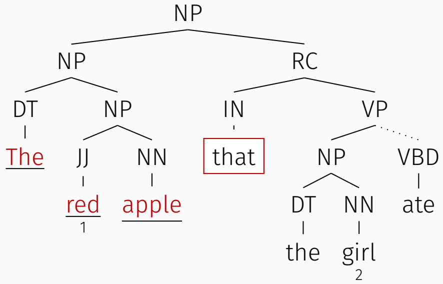
The red[1] $\boxed{\text{apple}}$ $\boxed{\text{that}}$ $\boxed{\underline{\text{the}}}$ $\boxed{\text{girl}}$[2] ate ...

$$\text{future-}n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(w_i \mid w_{i-n} \dots w_{i-1})$$



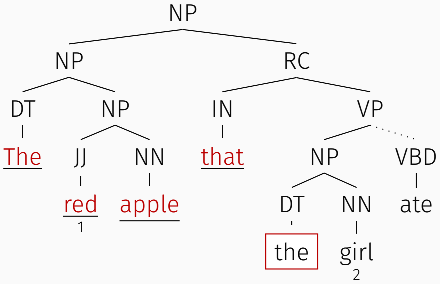
$$\text{Future-PCFG}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(T_i = w_i \mid \; T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

Successor PCFG Surprisal

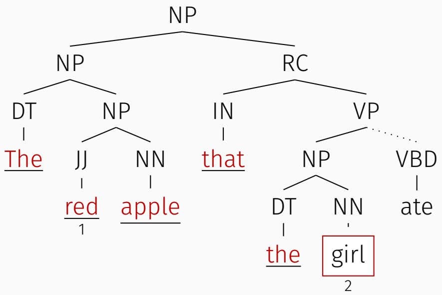


$$\text{Future-PCFG}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

Successor PCFG Surprisal

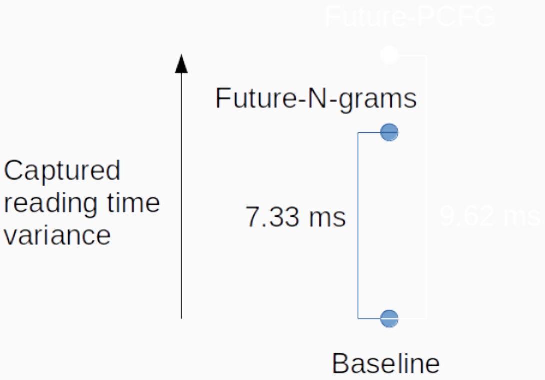


$$\text{Future-PCFG}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

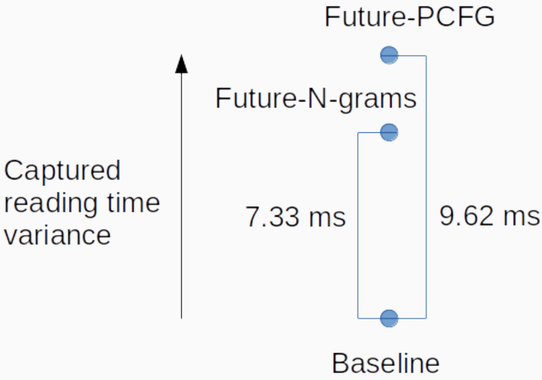


$$\text{Future-PCFG}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

SUCCESSOR *N*-GRAMS WORK



Successor PCFG works better



Conclusion: Accumulate surprisal!

- *N*-gram surprisal should be accumulated to predict reading times

Conclusion: Accumulate surprisal!

- *N*-gram surprisal should be accumulated to predict reading times
- *N*-gram surprisal accumulates post-saccade
- PCFG surprisal accumulates pre-saccade

Conclusion: Accumulate surprisal!

- *N*-gram surprisal should be accumulated to predict reading times
- *N*-gram surprisal accumulates post-saccade
- PCFG surprisal accumulates pre-saccade
  - Likely approximates entropy ($\rho = 0.2$)

Conclusion: Accumulate surprisal!

- *N*-gram surprisal should be accumulated to predict reading times
- *N*-gram surprisal accumulates post-saccade
- PCFG surprisal accumulates pre-saccade
    - Likely approximates entropy ($\rho = 0.2$)
    - Evidence that entropy directly predicts RTs

Conclusion: Accumulate surprisal!

- *N*-gram surprisal should be accumulated to predict reading times
- *N*-gram surprisal accumulates post-saccade
- PCFG surprisal accumulates pre-saccade
  - Likely approximates entropy ($\rho = 0.2$)
  - Evidence that entropy directly predicts RTs
  - Evidence of syntactic successor effects

Thanks! Questions?

Thanks to:

- Stefan Frank
- National Science Foundation (DGE-1343012)

Successor *n*-grams work

Successor *n*-grams are most predictive for 2 future words ($p < 0.001$)

Successor *n*-grams work

Successor *n*-grams are most predictive for 2 future words ($p < 0.001$)

6% of UCL saccades (n=3500) $>$2 words