Addressing surprisal deficiencies in reading time models

Marten van Schijndel William Schuler

December 11, 2016

Department of Linguistics, The Ohio State University

Overview

- Surprisal is a way to estimate text complexity

Overview

- Surprisal is a way to estimate text complexity
- Experienced complexity is reflected in reading speed

- Surprisal is a way to estimate text complexity
- Experienced complexity is reflected in reading speed

Claim:
Current surprisal models inadequately estimate reading complexity

- Surprisal is a way to estimate text complexity
- Experienced complexity is reflected in reading speed

Claim:
Current surprisal models inadequately estimate reading complexity

Consequence:
Other reading time predictors may get too much credit

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

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

Reading model of 'girl':
sentence position, word length

Reading complexity is estimated based on region ending

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

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

1 2

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

↑

important

Reading model of 'girl':

sentence position, word length, P(girl|the)

Reading complexity is estimated based on region ending

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

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

Reading complexity is estimated based on region ending

1 2

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

↑
?
↓

1 2

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

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

Surprisal: probability of observation given context

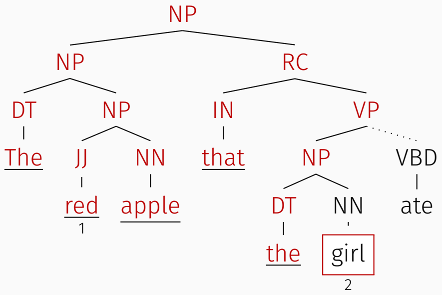
This study: *n*-gram and PCFG surprisal

This study: *n*-gram and PCFG surprisal

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

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

This study: *n*-gram and PCFG surprisal



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

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 ...

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

Cumulative *N*-gram Surprisal

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

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

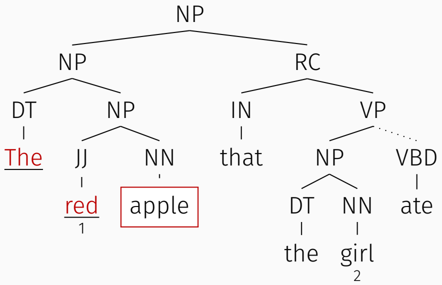
Cumulative *N*-gram Surprisal

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

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

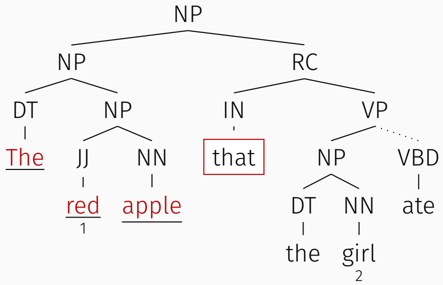
Accumulated surprisal fixes the theoretical problem

Cumulative PCFG Surprisal



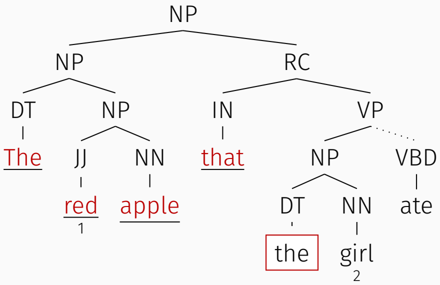
$$\text{Cumu-PCFG}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}}^{f_t} \sum_{T \in \text{Trees}} -\log \mathrm{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

Cumulative PCFG Surprisal



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

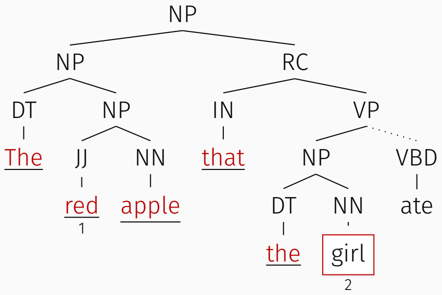
Cumulative PCFG Surprisal



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

Accumulated surprisal fixes the theoretical problem

Cumulative PCFG Surprisal



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

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

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

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

- 43 subjects
- reading online novels
- frequent comprehension questions

How well does this fix work?

Baseline mixed effects model

Fixed Factors

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

Baseline mixed effects model

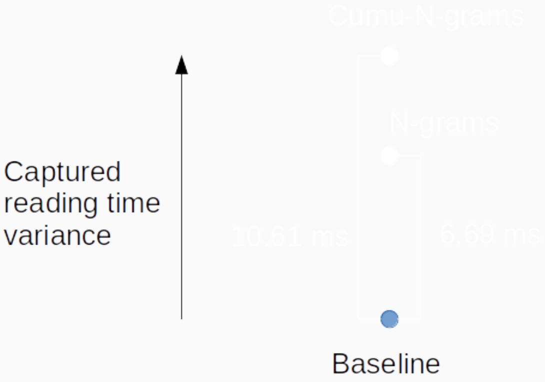
Fixed Factors

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

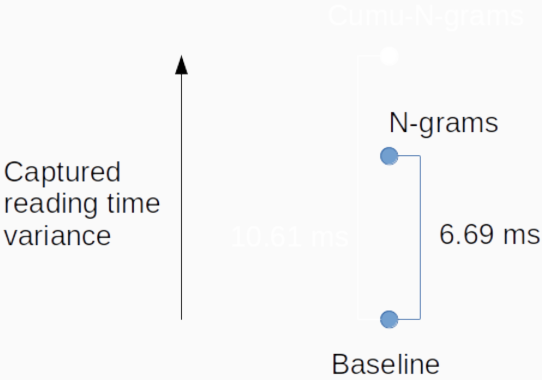
Random Factors

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

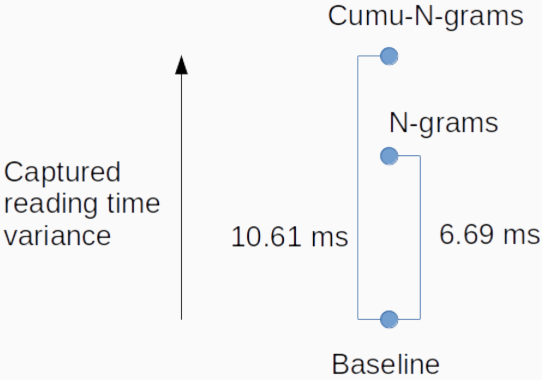
Accumulation improves *n*-gram surprisal



Accumulation improves *n*-gram surprisal



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 ($p > 0.05$)

Parafovial processing

1

The red apple that the girl ate …

Possible accumulation influences

Parafovial processing

1

Th(e red apple that t)he girl ate …

Possible accumulation influences

Parafovial processing

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

Prediction

1
The red apple that the girl ate …

Prediction

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

Possible accumulation influences

Prediction

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

Subsequent regression

1
The red apple that the girl ate …

Subsequent regression

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

Possible accumulation influences

Subsequent regression

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

Possible accumulation influences

Subsequent regression

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

Subsequent regression

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

Accumulation alternative: Successor surprisal

Cumulative PCFG surprisal only handles subsequent regression

Cumulative PCFG 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 PCFG surprisal only handles subsequent regression

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

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

Other accumulation mechanisms presuppose earlier accumulation

Upcoming material influences reading times

Upcoming material influences reading times

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

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 …

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

Successor N-grams

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

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

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

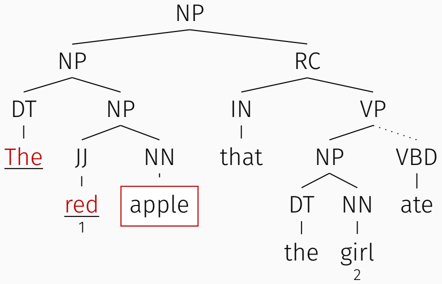
$$\text{future-}n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(w_i \mid w_{i-n} \ldots w_{i-1})$$

Successor N-grams

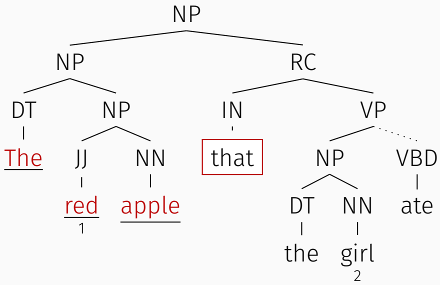
The red[1] $\boxed{\text{apple}}$ $\boxed{\text{that}}$ $\boxed{\underline{\text{the}}}$ $\boxed{\text{girl}}$[2] ate ...

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

Successor PCFG Surprisal

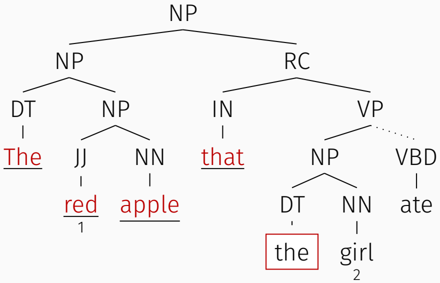


$$\text{Future-PCFG}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} \sum_{T \in \text{Trees}} -\log \mathrm{P}(T_i = w_i \mid T_1 \dots T_{i-1} = w_1 \dots w_{i-1})$$



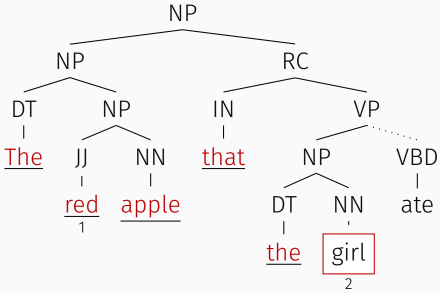
$$\text{Future-PCFG}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} \sum_{T \in \text{Trees}} -\log \mathrm{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

Successor PCFG Surprisal



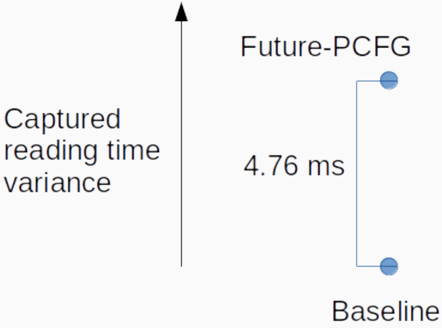
$$\text{Future-PCFG}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} \sum_{T \in \text{Trees}} -\log \mathrm{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

Successor PCFG Surprisal

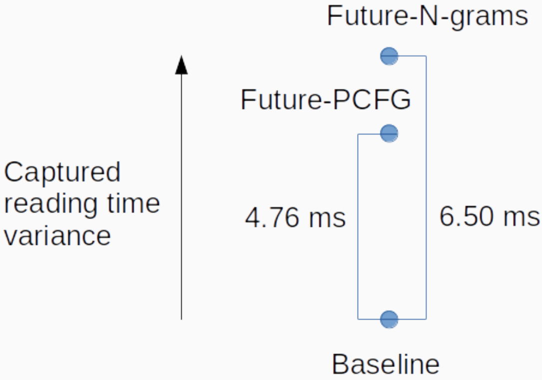


$$\text{Future-PCFG}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} \sum_{T \in \text{Trees}} -\log \mathrm{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

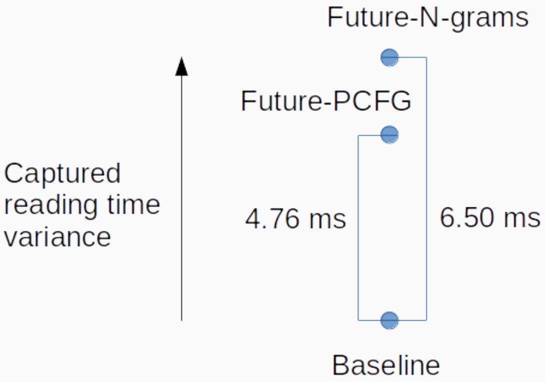
Successor PCFG works



Successor *n*-grams work better



SUCCESSOR N-GRAMS WORK BETTER



PCFG surprisal may require a richer grammar

Successor *n*-grams have limited influence

Successor *n*-grams have limited influence

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

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

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

Conclusion

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

Conclusion

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

Conclusion

- *N*-gram surprisal should be accumulated to predict reading times
- *N*-gram surprisal accumulates pre- and post-saccade
  - Pre-saccade *n*-grams are limited

Conclusion

- *N*-gram surprisal should be accumulated to predict reading times
- *N*-gram surprisal accumulates pre- and post-saccade
  - Pre-saccade *n*-grams are limited
- PCFG surprisal does not accumulate

Conclusion

- *N*-gram surprisal should be accumulated to predict reading times
- *N*-gram surprisal accumulates pre- and post-saccade
  - Pre-saccade *n*-grams are limited
- PCFG surprisal does not accumulate
  - PCFG surprisal still predictive on Dundee

Conclusion

- *N*-gram surprisal should be accumulated to predict reading times
- *N*-gram surprisal accumulates pre- and post-saccade
  - Pre-saccade *n*-grams are limited
- PCFG surprisal does not accumulate
  - PCFG surprisal still predictive on Dundee
  - UCL corpus may not be syntactically complex enough

Conclusion

- *N*-gram surprisal should be accumulated to predict reading times
- *N*-gram surprisal accumulates pre- and post-saccade
  - Pre-saccade *n*-grams are limited
- PCFG surprisal does not accumulate
  - PCFG surprisal still predictive on Dundee
  - UCL corpus may not be syntactically complex enough
  - PCFG surprisal may need a richer grammar

Thanks! Questions?

Thanks to:

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

Model	N-gram vs Cumu-N-gram		
	β	Log-Likelihood	AIC
Baseline		−12702	25476
Base+Basic	0.035	−12689*	25451
Base+Cumulative	0.055	−12683*	25440
Base+Both		−12683*	25442

Base random: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram
Base fixed: sentpos, wlen, rlen, prevfix

Significance for the Base+Both model applies to improvement over the Base+Basic model.

Future Surprisal Results

Model	Future-N-grams vs Future-PCFG		
	β	Log-Likelihood	AIC
Baseline		-12276	24642
Base+Future-N-grams	0.034	-12259*	24610
Base+Future-PCFG	0.025	-12266*	24624
Base+Both		-12259*	24612

Base random: sentpos, wlen, rlen, prevfix, cumu-5-gram, future-5-grams, future-PCFG
Base fixed: sentpos, wlen, rlen, prevfix, cumu-5-gram

Significance for the Base+Both model applies to improvement over the Base+Future-PCFG model.