The statistics of the unseen influence reading times

Marten van Schijndel

September 8, 2017

Department of Cognitive Science, Johns Hopkins University
(Department of Linguistics, The Ohio State University)

1. The frequencies of skipped material affect linguistic processing
2. Upcoming frequencies affect linguistic processing

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:
Shows that material skipped by saccades slows reading
Presents a simple way for surprisal to address that complexity

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

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

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

Reading complexity is estimated based on region ending

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

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 ...

$\uparrow$

important

Reading model of 'girl':

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

Reading complexity is estimated based on region ending

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

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

Reading complexity is estimated based on region ending

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

$$\overset{\uparrow}{\underset{\downarrow}{?}}$$

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

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

Surprisal: probability of observation given context

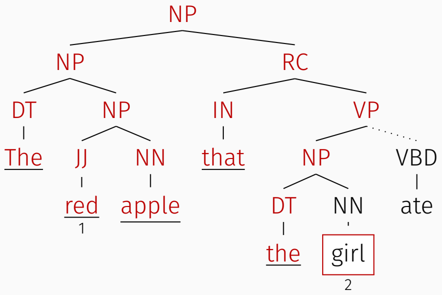
This study: *n*-gram and PCFG surprisal

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



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

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 ...

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

Cumulative *N*-gram Surprisal

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

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

Accumulated surprisal fixes the theoretical problem

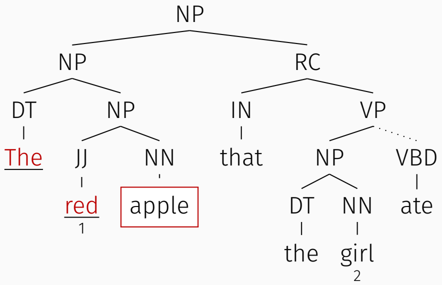
Cumulative N-gram Surprisal

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

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

Accumulated surprisal fixes the theoretical problem

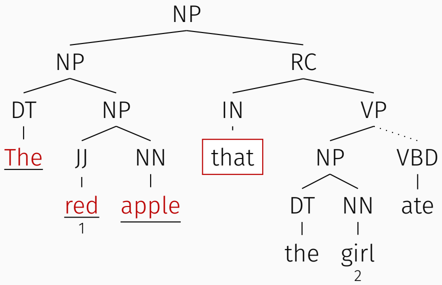
Cumulative PCFG Surprisal



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

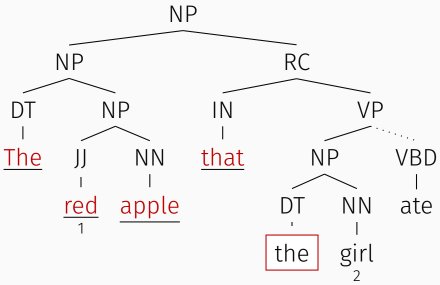
Accumulated surprisal fixes the theoretical problem

Cumulative PCFG Surprisal



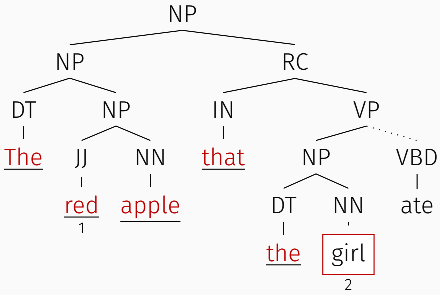
$$\text{Cumu-PCFG}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}}^{f_t} -\log \text{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

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



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

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

How well does this fix work?

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

- 43 subjects
- reading 361 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

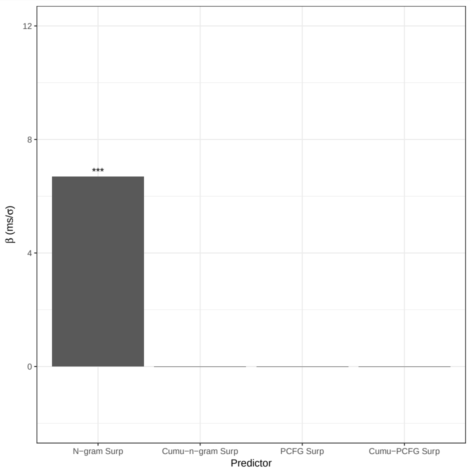
Fixed Factors

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

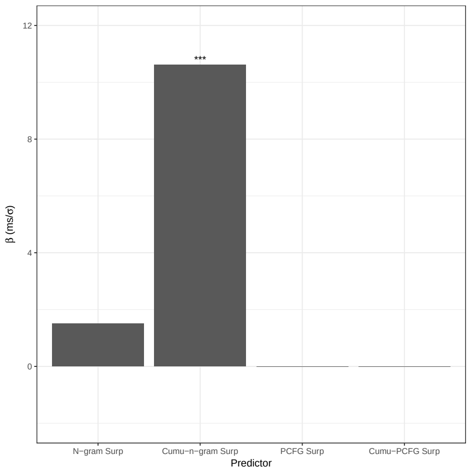
Random Factors

- All fixed factors as by-subject random slopes
- Item, subject and subject×sentence intercepts

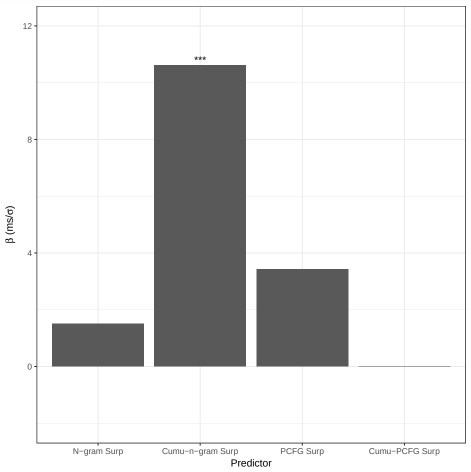
Accumulation improves *n*-gram surprisal



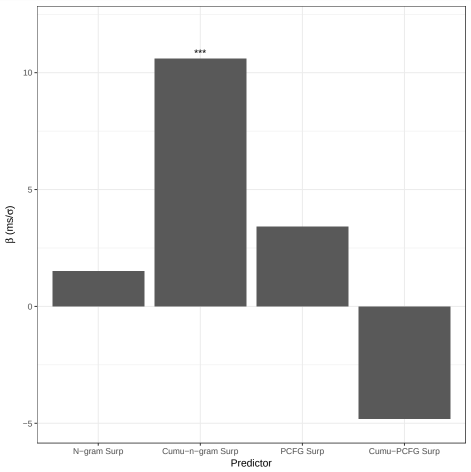
Accumulation improves *n*-gram surprisal



Accumulation improves *n*-gram surprisal



Accumulation does not help PCFG surprisal



What does accumulation model?

Possible accumulation influences

Subsequent regression

1
The red 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 …

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 ...

Inference

1
The red apple that the girl ate …

Inference

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

Possible accumulation influences

Inference

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

Parafovial processing

The red[1] apple that the girl ate …

Parafovial processing

Th(e red apple that t)he girl ate …

1

Parafovial processing

Th(e red apple that t)he girl ate …

Prediction (entropy)

1
The red apple that the girl ate …

Prediction (entropy)

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

Possible accumulation influences

Prediction (entropy)

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

Accumulation alternative: Successor surprisal

Cumulative surprisal handles regression and inference

Accumulation alternative: Successor surprisal

Cumulative surprisal handles regression and inference

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 handles regression and inference

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

How much influence does upcoming material have?

Successor effects influence reading times

Upcoming material influences reading times

Successor effects influence 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)

Upcoming material affects processing

Angele et al. (2015)

A ¦ ch*ild ¦ XXXXXXX ¦ the ¦ fish

Upcoming material affects processing

Angele et al. (2015)

A	child*	XXXXXXX	the	fish
A	child	annoyed*	XXX	fish

Upcoming material affects processing

Angele et al. (2015)

A	child*	XXXXXXX	the	fish
A	child	annoyed*	XXX	fish
A	child	annoyed	the*	XXXX

Upcoming material affects processing

Angele et al. (2015)

A	child*	XXXXXXX	the	fish
A	child	annoyed*	XXX	fish
A	child	annoyed	the*	XXXX

Lexical frequency of the upcoming masked word affects processing

Upcoming material affects processing

Angele et al. (2015)

A	*child	XXXXXXX	the	fish
A	child	ann*oyed	XXX	fish
A	child	annoyed	*the	XXXX

Lexical frequency of the upcoming masked word affects processing

Hypothesis: Effect is due to uncertainty over continuations

Angele et al. (2015)

A	child*	XXXXXXX	the	fish
A	child	annoyed*	XXX	fish
A	child	annoyed	the*	XXXX

Lexical frequency of the upcoming masked word affects processing

Hypothesis: Effect is due to uncertainty over continuations

Problem: Uncertainty is expensive to calculate

Entropy measures uncertainty

Shannon (1948)

$$H(X) \stackrel{def}{=} -\sum_{x \in X} \mathrm{P}(x) \log \mathrm{P}(x) \tag{1}$$

Entropy measures uncertainty

Shannon (1948)

$$H(X) \stackrel{def}{=} -\sum_{x \in X} \mathrm{P}(x) \log \mathrm{P}(x) \tag{1}$$

Roark et al. (2009) distinguishes two kinds of entropy (over words and preterminals)

$$\mathrm{Lex}H(w_{1..i-1}) \stackrel{def}{=} -\sum_{w_i \in V} \mathrm{P}_G(w_i \mid w_{1..i-1}) \log \mathrm{P}_G(w_i \mid w_{1..i-1}) \tag{2}$$

$$\mathrm{Syn}H(w_{1..i-1}) \stackrel{def}{=} -\sum_{p_i \in G} \mathrm{P}_G(p_i \mid w_{1..i-1}) \log \mathrm{P}_G(p_i \mid w_{1..i-1}) \tag{3}$$

Roark et al. (2009) showed

- Syn$H$ predicts self-paced reading times
- Lex$H$ is not predictive of SPR times

Roark et al. (2009) showed

- Syn*H* predicts self-paced reading times
- Lex*H* is not predictive of SPR times
  (No Angele et al., 2015, effect)

Entropy measures uncertainty

Roark et al. (2009) showed

- Syn$H$ predicts self-paced reading times
- Lex$H$ is not predictive of SPR times
  (No Angele et al., 2015, effect)

But

- Small training corpus (V is poor)
- Small test corpus:
  $\sim$ 200 sentences, $\sim$ 4000 words, 23 subjects

Test data in this work

Natural Stories self-paced reading corpus (Futrell et al., in prep)

- 181 subjects
- 10 narrative texts
- 485 sentences (10256 words)
- Each text followed by 6 comprehension questions
- Events removed if $<$100 ms or $>$3000 ms

Parsed using Roark (2001) parser

Fitted with *lmer*

Spaces were masked

_ _

Spaces were masked

```
A -----------------------
```

Spaces were masked

```
- child -----------------
```

Spaces were masked

```
------- annoyed ---------
```

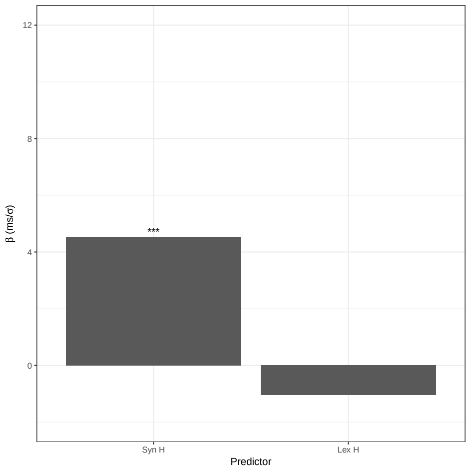
Spaces were masked

```
---------------- the -----
```

Spaces were masked

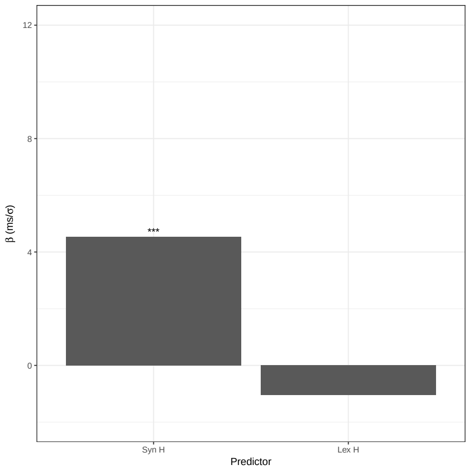
```
-------------------- fish.
```

Syntactic entropy predicts RTs



Replication of Roark et al. (2009)

Syntactic entropy predicts RTs



Replication of Roark et al. (2009)

But Angele et al. (2015) found a *lexical* frequency effect

Can we make Lex$H$ more tractable?

$$S_G(w_i, w_{1..i-1}) \stackrel{def}{=} -\log \mathrm{P}_G(w_i \mid w_{1..i-1}) \tag{4}$$

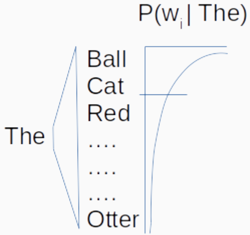
$$\begin{aligned}
\mathrm{Lex}H_G(w_{1..i-1}) &\stackrel{def}{=} \sum_{w_i \in V} -\mathrm{P}_G(w_i \mid w_{1..i-1}) \log \mathrm{P}_G(w_i \mid w_{1..i-1}) && (5)\\
&= \sum_{w_i \in V} \mathrm{P}_G(w_i \mid w_{1..i-1})\, S_G(w_i, w_{1..i-1}) && (6)\\
&= \mathrm{E}[S_G(w_i, w_{1..i-1})] && (7)
\end{aligned}$$

Can we make Lex$H$ more tractable?

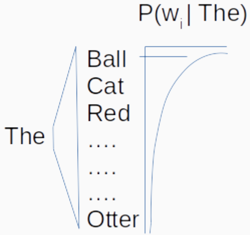
$$S_G(w_i, w_{1..i-1}) \stackrel{def}{=} -\log \mathrm{P}_G(w_i \mid w_{1..i-1}) \tag{4}$$

$$\begin{aligned}
\mathrm{Lex}H_G(w_{1..i-1}) &\stackrel{def}{=} \sum_{w_i \in V} -\mathrm{P}_G(w_i \mid w_{1..i-1}) \log \mathrm{P}_G(w_i \mid w_{1..i-1}) && (5)\\
&= \sum_{w_i \in V} \mathrm{P}_G(w_i \mid w_{1..i-1})\, S_G(w_i, w_{1..i-1}) && (6)\\
&= \mathrm{E}[S_G(w_i, w_{1..i-1})] && (7)
\end{aligned}$$

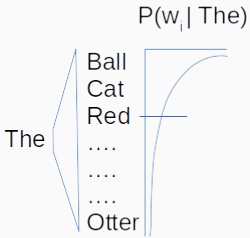
We can use a corpus instead of explicitly computing the expectation


P(w_i | The)
The
Ball
Cat
Red
....
....
....
Otter

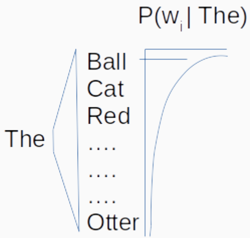
Surprisal approximates entropy in the aggregate



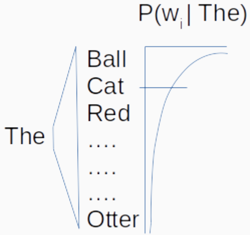
Surprisal approximates entropy in the aggregate



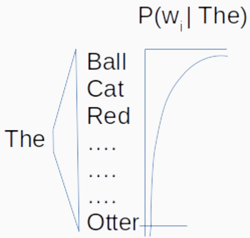
Surprisal approximates entropy in the aggregate

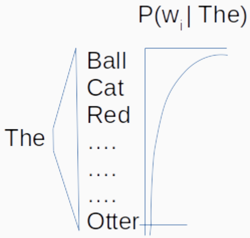


Surprisal approximates entropy in the aggregate

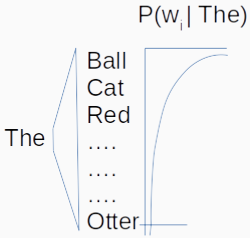


Surprisal approximates entropy in the aggregate





Ex: The boy annoyed the fish.



We can treat large corpora as our samplers.

Possible entropy approximations

We can try:

- Future Roark surprisal
  (same distribution as SynH)

Possible entropy approximations

We can try:

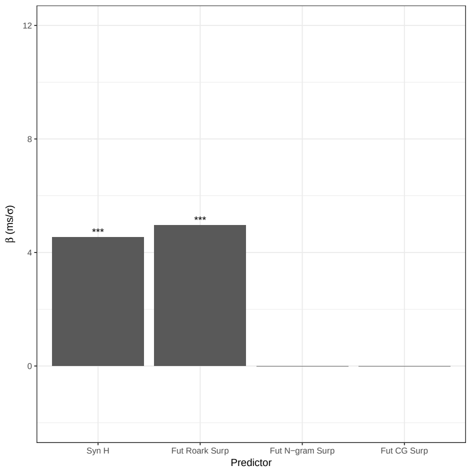
- Future Roark surprisal
  (same distribution as SynH)
- Future 5-gram Surprisal
  (similar to what Angele et al., observed)

Possible entropy approximations

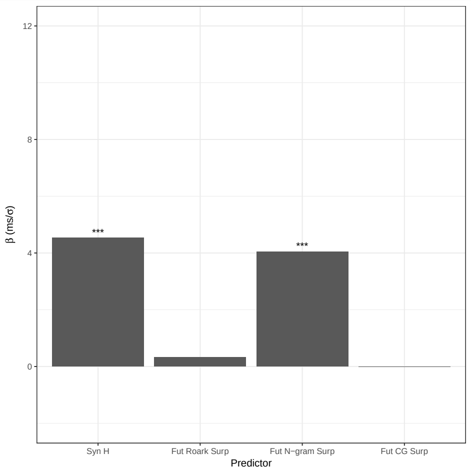
We can try:

- Future Roark surprisal
  (same distribution as SynH)
- Future 5-gram Surprisal
  (similar to what Angele et al., observed)
- Future categorial grammar surprisal
  (tests how specific syntactic prediction is)

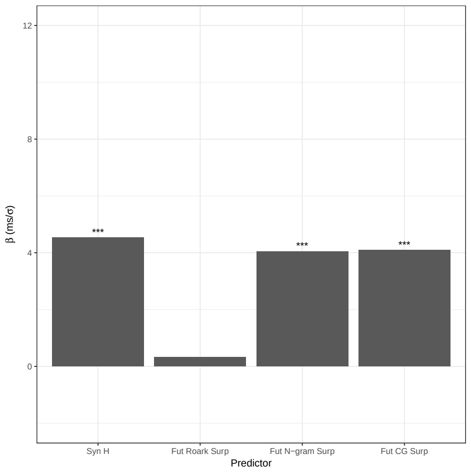
Future surprisal predicts RTs



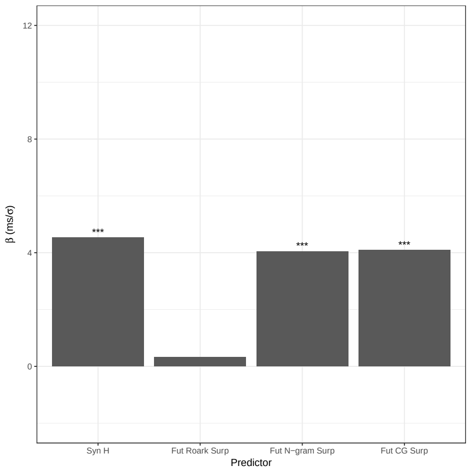
Uncertainty over both words and syntax



Uncertainty over both words and syntax



Uncertainty over both words and syntax



Support for Angele et al. hypothesis

Why does this pre-slowing occur?

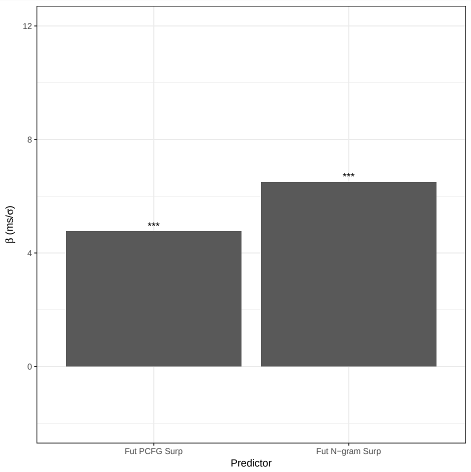
- Better encoding of w_i to help with w_{i+1}

Why does this pre-slowing occur?

- Better encoding of w_i to help with w_{i+1}
- A kind of Uniform Information Density (UID; Jaeger, 2010)
  - Optimizes per-millisecond informativity

Can this approximation method be used with accumulation?
(eye-tracking)

Accumulated future surprisal works



SUCCESSOR n-GRAMS HAVE LIMITED INFLUENCE

Successor n-grams have limited influence

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

Successor *n*-grams have limited influence

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

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

Successor *n*-grams have limited influence

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

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

Successor *n*-grams are most predictive for 1 SPR word ($p < 0.001$)

Conclusions

- Skipped Material in eye-tracking
  - *N*-gram surprisal should be accumulated to predict reading times

Conclusions

- Skipped Material in eye-tracking
  - *N*-gram surprisal should be accumulated to predict reading times
  - PCFG surprisal does not accumulate

Conclusions

- Skipped Material in eye-tracking
  - *N*-gram surprisal should be accumulated to predict reading times
  - PCFG surprisal does not accumulate
- Upcoming Material

Conclusions

- Skipped Material in eye-tracking
  - *N*-gram surprisal should be accumulated to predict reading times
  - PCFG surprisal does not accumulate
- Upcoming Material
  - Uncertainty about upcoming words slows processing

- Skipped Material in eye-tracking
  - *N*-gram surprisal should be accumulated to predict reading times
  - PCFG surprisal does not accumulate
- Upcoming Material
  - Uncertainty about upcoming words slows processing
  - That influence can be detected prior to any expectation violation

- Skipped Material in eye-tracking
  - *N*-gram surprisal should be accumulated to predict reading times
  - PCFG surprisal does not accumulate
- Upcoming Material
  - Uncertainty about upcoming words slows processing
  - That influence can be detected prior to any expectation violation
  - Future surprisal can efficiently approximate that uncertainty

Conclusions

- Skipped Material in eye-tracking
  - *N*-gram surprisal should be accumulated to predict reading times
  - PCFG surprisal does not accumulate
- Upcoming Material
  - Uncertainty about upcoming words slows processing
  - That influence can be detected prior to any expectation violation
  - Future surprisal can efficiently approximate that uncertainty
  - Syntactic uncertainty is fine-grained

Thanks! Questions?

This work was done with William Schuler

Thanks to:

- Stefan Frank, Klinton Bicknell
- The reviewers for their very helpful comments
- National Science Foundation (DGE-1343012)

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

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

Successor N-grams

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

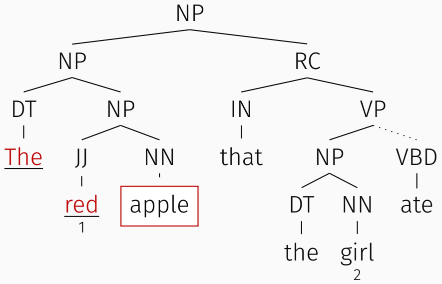
$$\text{future-}n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(w_i \mid w_{i-n} \dots w_{i-1})$$

Successor N-grams

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

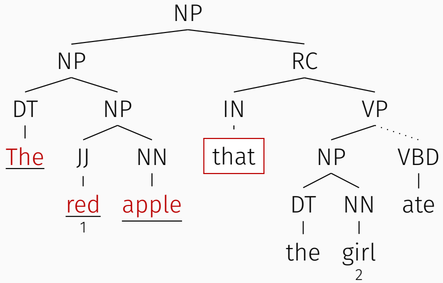
$$\text{future-}n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(w_i \mid w_{i-n} \ldots w_{i-1})$$

Successor PCFG Surprisal



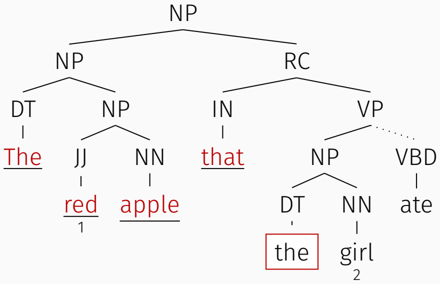
$$\text{Future-PCFG}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

Successor PCFG Surprisal



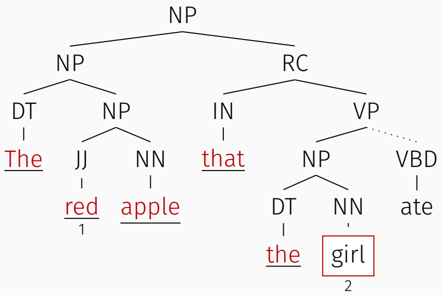
$$\text{Future-PCFG}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log \mathrm{P}(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})$$

Successor PCFG Surprisal



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

Successor PCFG Surprisal



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