

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

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

- 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

- 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

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

The red apple that the girl ate ...

w_1 w_2 w_3 w_4 w_5 w_6

Reading model of 'girl':
sentence position

The red apple that the girl ate ...

4 chars
w₆

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

The red apple that the girl ate ...

4 chars
w₆

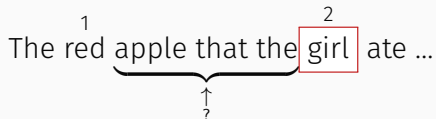
Reading model of 'girl':
sentence position, word length, $P(\text{girl}|\text{the})$

The red apple that the ¹girl² ate ...
↑
important

Reading model of 'girl':
sentence position, word length, $P(\text{girl}|\text{the})$

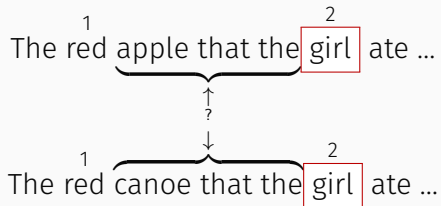
READING COMPLEXITY IS ESTIMATED BASED ON REGION ENDING

The ¹red apple that the ²girl ate ...



Reading model of 'girl':
sentence position, word length, $P(\text{girl}|\text{the})$

READING COMPLEXITY IS ESTIMATED BASED ON REGION ENDING



Reading model of 'girl':
sentence position, word length, $P(\text{girl}|\text{the})$

This study: n -gram and PCFG surprisal

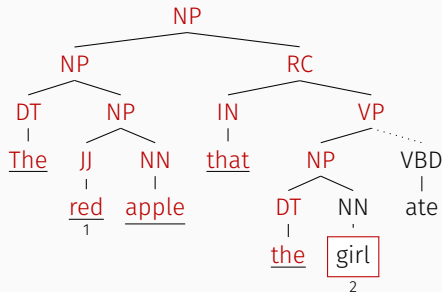
This study: n -gram and PCFG surprisal

The red apple that the girl ate ...

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

SURPRISAL: PROBABILITY OF OBSERVATION GIVEN CONTEXT

This study: n -gram and PCFG surprisal



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

Cumulative N -gram Surprisal

The red¹ apple that the girl² ate ...

Cumulative N -gram Surprisal

The red¹ apple that the girl² ate ...

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

Cumulative N -gram Surprisal

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

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

Cumulative N -gram Surprisal

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

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

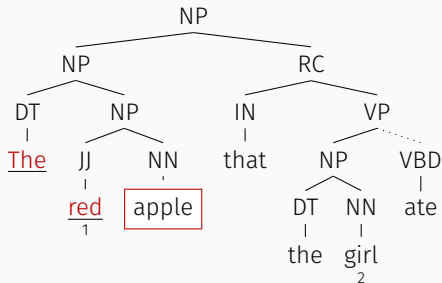
Cumulative N -gram Surprisal

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

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

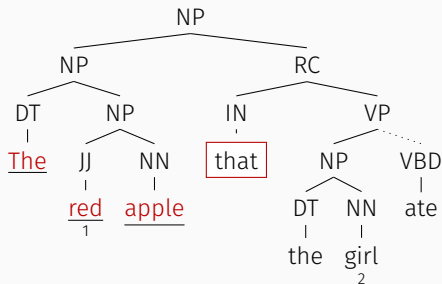
ACCUMULATED SURPRISAL FIXES THE THEORETICAL PROBLEM

Cumulative PCFG Surprisal



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

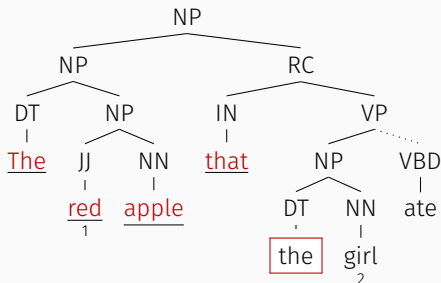
Cumulative PCFG Surprisal



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

ACCUMULATED SURPRISEL FIXES THE THEORETICAL PROBLEM

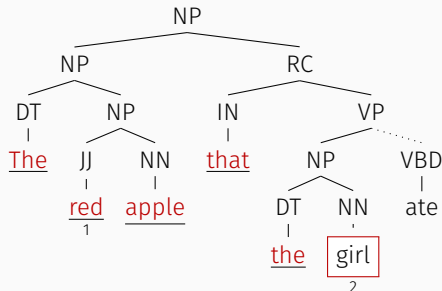
Cumulative PCFG Surprisal



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

ACCUMULATED SURPRISAL FIXES THE THEORETICAL PROBLEM

Cumulative PCFG Surprisal



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

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 361 short sentences from 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

HOW WELL DOES THIS FIX WORK?

Baseline mixed effects model

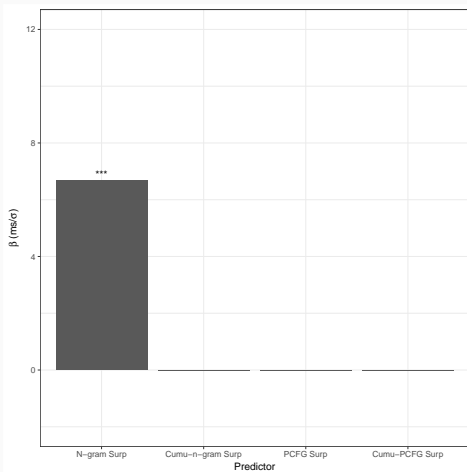
Fixed Factors

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

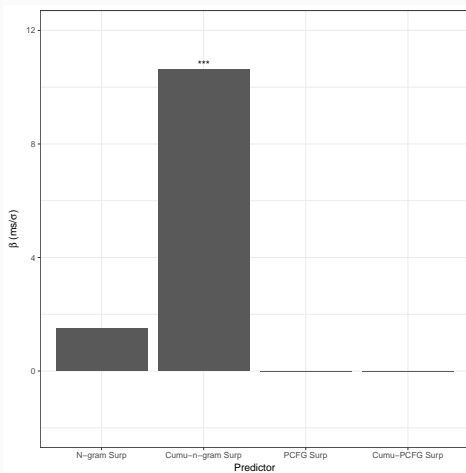
Random Factors

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

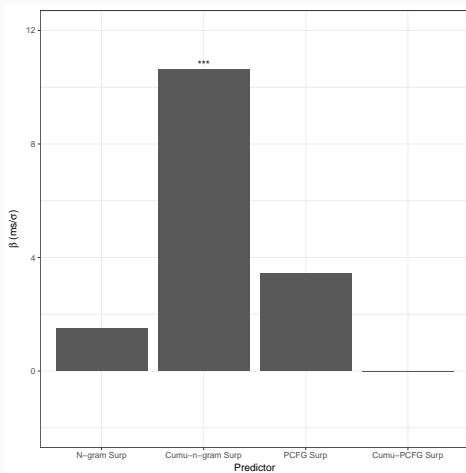
ACCUMULATION IMPROVES N-GRAM SURPRISAL



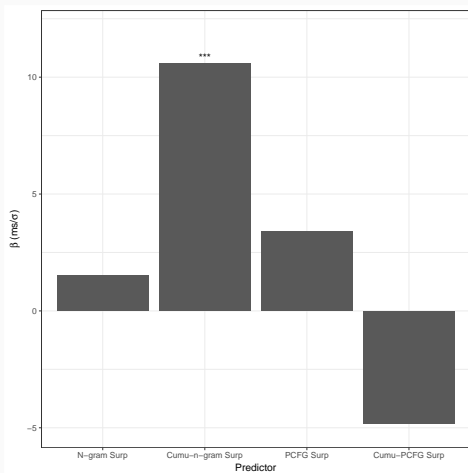
ACCUMULATION IMPROVES N-GRAM SURPRISAL



ACCUMULATION IMPROVES N-GRAM SURPRISAL



ACCUMULATION DOES NOT HELP PCFG SURPRISAL



What does accumulation model?

Subsequent regression

¹
The red apple that the girl ate ...

Subsequent regression

The red¹ apple that the girl² ate ...

Subsequent regression

The ¹red ³apple that the ²girl ate ...

Subsequent regression

¹ ³ ⁴ ²
The red apple that the girl ate ...

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 red¹ apple that the girl² ate ...

Inference

The red¹ (apple that the girl²) ate ...

Parafoveal processing

¹
The red apple that the girl ate ...

Parafoveal processing

Th(e¹ red apple that t)he girl ate ...

Parafoveal processing

Th(e¹ red apple that t)he² girl ate ...

Prediction (entropy)

The red¹ apple that the girl ate ...

Prediction (entropy)

The red¹ (apple that the girl) ate ...

Prediction (entropy)

The red¹ (apple that the girl²) ate ...

Cumulative surprisal handles regression and inference

Cumulative surprisal handles regression and inference

Parafoveal: Th(e red ¹ apple that t)he ² girl ate ...

Prediction: The red ¹ (apple that the ² girl) ate ...
accumulated

Cumulative surprisal handles regression and inference

Parafoveal: Th(e red ¹ apple that t)he ² girl ate ...

Prediction: The red ¹ (apple that the ² girl) ate ...
accumulated

Other accumulation mechanisms presuppose earlier accumulation

How much influence does upcoming material have?

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)

Angele et al. (2015)

A | child | XXXXXXX | the | fish

Angele et al. (2015)

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

Angele et al. (2015)

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

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

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

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

Shannon (1948)

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

Shannon (1948)

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

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

$$\text{Lex}H(w_{1..i-1}) \stackrel{\text{def}}{=} - \sum_{w_i \in V} P_G(w_i | w_{1..i-1}) \log P_G(w_i | w_{1..i-1}) \quad (2)$$

$$\text{Syn}H(w_{1..i-1}) \stackrel{\text{def}}{=} - \sum_{p_i \in G} P_G(p_i | w_{1..i-1}) \log P_G(p_i | w_{1..i-1}) \quad (3)$$

Roark et al. (2009) showed

- $SynH$ predicts self-paced reading times
- $LexH$ is not predictive of SPR times

Roark et al. (2009) showed

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

Roark et al. (2009) showed

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

But

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

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



A -----

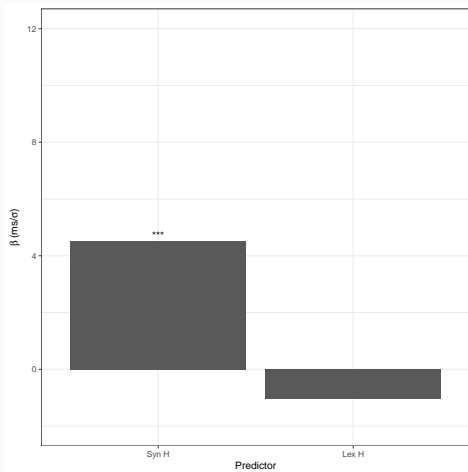
- child -----

----- annoyed -----

----- the -----

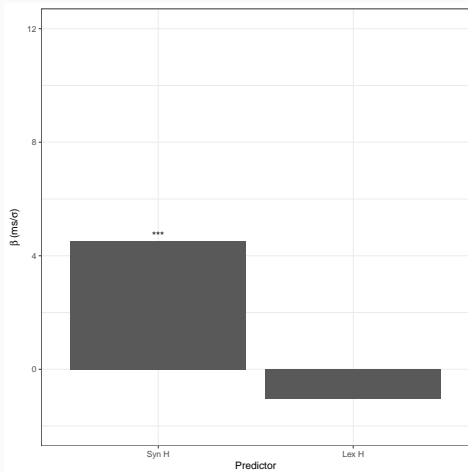
----- fish.

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?

$$S_G(w_i, w_{1..i-1}) \stackrel{\text{def}}{=} -\log P_G(w_i | w_{1..i-1}) \quad (4)$$

$$\text{Lex}H_G(w_{1..i-1}) \stackrel{\text{def}}{=} \sum_{w_i \in V} -P_G(w_i | w_{1..i-1}) \log P_G(w_i | w_{1..i-1}) \quad (5)$$

$$= \sum_{w_i \in V} P_G(w_i | w_{1..i-1}) S_G(w_i, w_{1..i-1}) \quad (6)$$

$$= E[S_G(w_i, w_{1..i-1})] \quad (7)$$

CAN WE MAKE LEXH MORE TRACTABLE?

$$S_G(w_i, w_{1..i-1}) \stackrel{\text{def}}{=} -\log P_G(w_i | w_{1..i-1}) \quad (4)$$

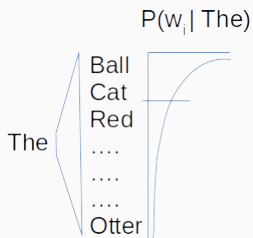
$$\text{Lex}H_G(w_{1..i-1}) \stackrel{\text{def}}{=} \sum_{w_i \in V} -P_G(w_i | w_{1..i-1}) \log P_G(w_i | w_{1..i-1}) \quad (5)$$

$$= \sum_{w_i \in V} P_G(w_i | w_{1..i-1}) S_G(w_i, w_{1..i-1}) \quad (6)$$

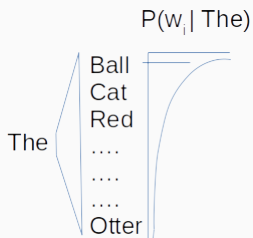
$$= E[S_G(w_i, w_{1..i-1})] \quad (7)$$

We can use a corpus instead of explicitly computing the expectation

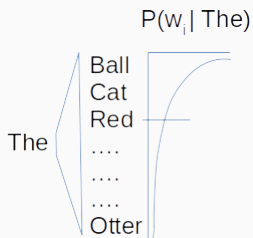
ENTROPY GIVES MEAN SURPRISAL



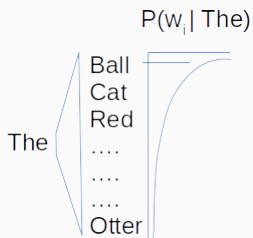
SURPRISAL APPROXIMATES ENTROPY IN THE AGGREGATE



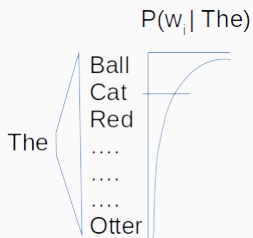
SURPRISAL APPROXIMATES ENTROPY IN THE AGGREGATE



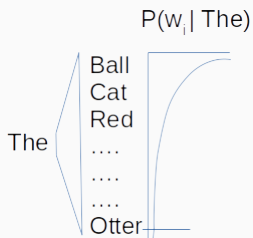
SURPRISAL APPROXIMATES ENTROPY IN THE AGGREGATE



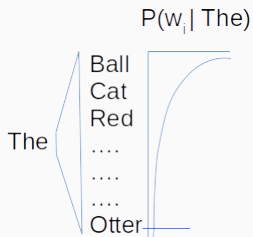
SURPRISAL APPROXIMATES ENTROPY IN THE AGGREGATE



SURPRISAL APPROXIMATES ENTROPY IN THE AGGREGATE

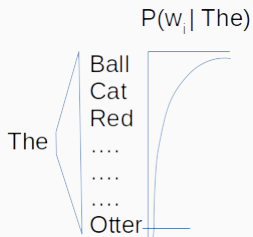


SURPRISAL APPROXIMATES ENTROPY IN THE AGGREGATE



Ex: The boy annoyed the fish.

SURPRISAL APPROXIMATES ENTROPY IN THE AGGREGATE



We can treat large corpora as our samplers.

We can try:

- Future Roark surprisal
(same distribution as SynH)

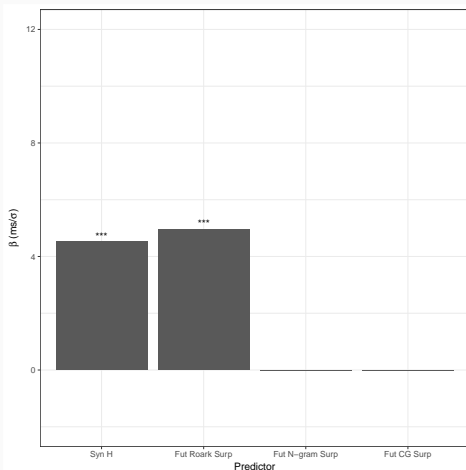
We can try:

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

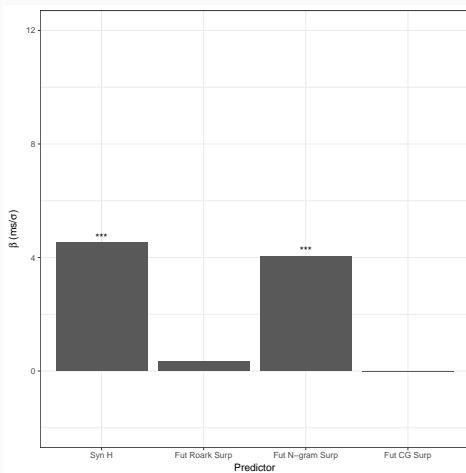
We can try:

- 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)

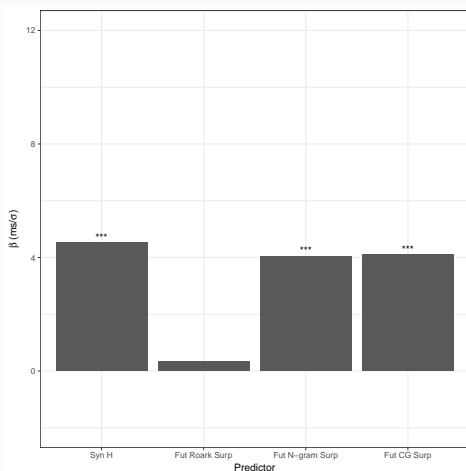
FUTURE SURPRISAL PREDICTS RTs



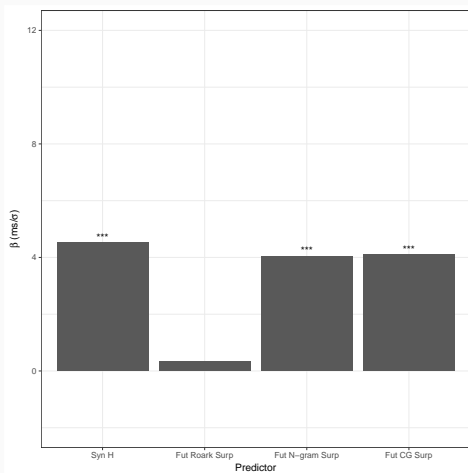
UNCERTAINTY OVER BOTH WORDS AND SYNTAX



UNCERTAINTY OVER BOTH WORDS AND SYNTAX



UNCERTAINTY OVER BOTH WORDS AND SYNTAX



Support for Angele et al. hypothesis

WHY DOES THIS PRE-SLOWING OCCUR?

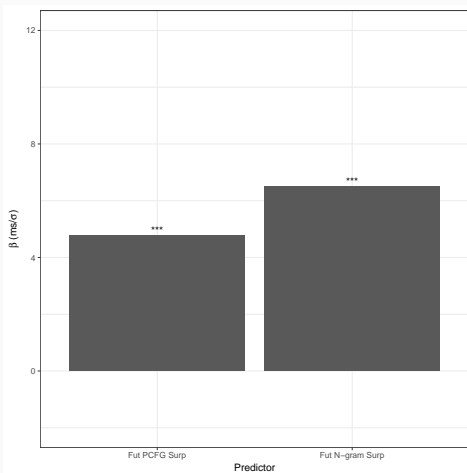
- Better encoding of w_i to help with w_{i+1}

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 are most predictive for 2 future ET words ($p < 0.001$)

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 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 ($p < 0.001$)

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

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

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

- 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

- 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

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)

The red ¹apple that the girl ²ate ...

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

The ¹red apple that the girl ²ate ...

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

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

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

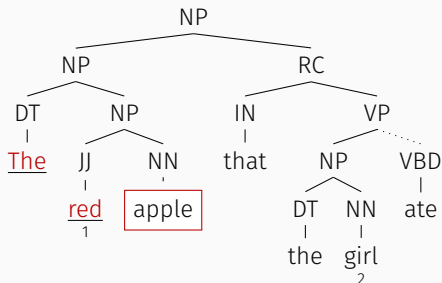
The red ¹ apple that the ² girl ate ...

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

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

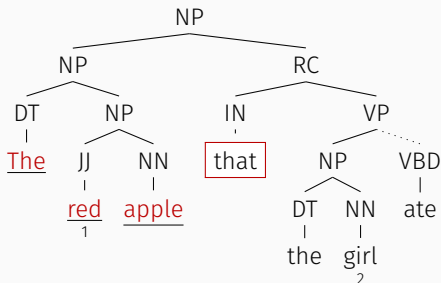
$$\text{future-}n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log P(w_i \mid w_{i-n} \dots w_{i-1})$$

SUCCESSOR PCFG SURPRISAL



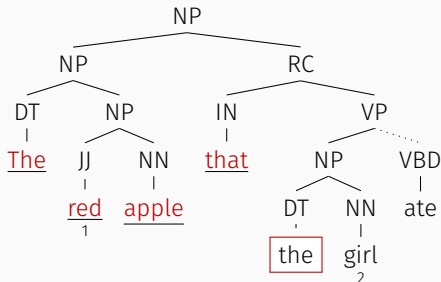
$$\text{Future-PCFG}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log P(T_i = w_i \mid T_1 \dots T_{i-1} = w_1 \dots w_{i-1})$$

SUCCESSOR PCFG SURPRISAL



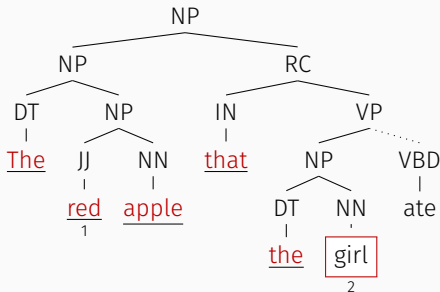
$$\text{Future-PCFG}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log P(T_i = w_i \mid T_1 \dots T_{i-1} = w_1 \dots w_{i-1})$$

SUCCESSOR PCFG SURPRISAL



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

SUCCESSOR PCFG SURPRISAL



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