HIERARCHIC SYNTAX IMPROVES READING TIME PREDICTION

Marten van Schijndel and William Schuler
Department of Linguistics
The Ohio State University

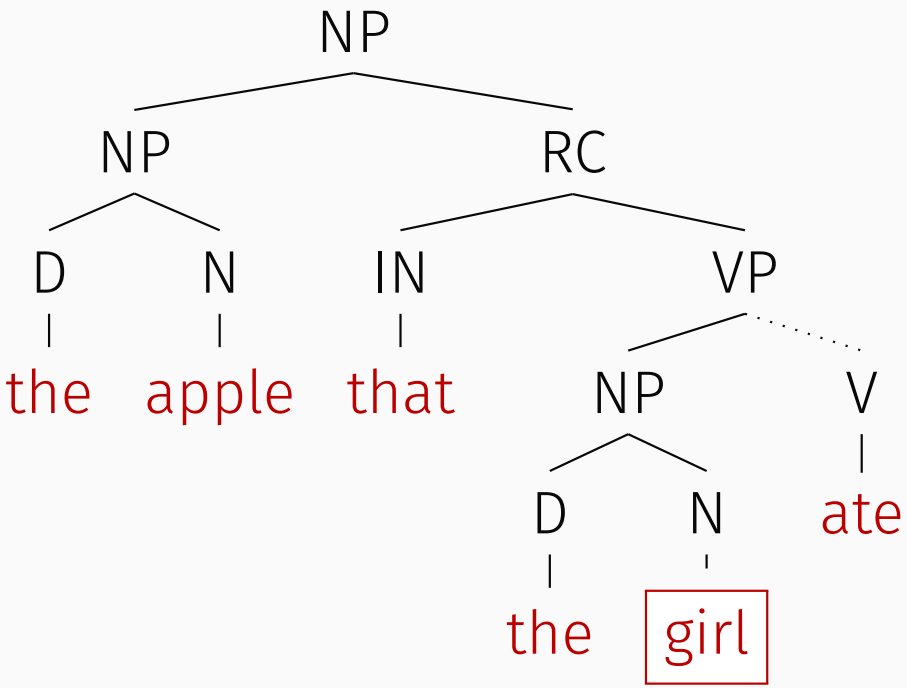
June 3, 2015

OVERVIEW

Previous studies have debated whether humans use hierarchic syntax

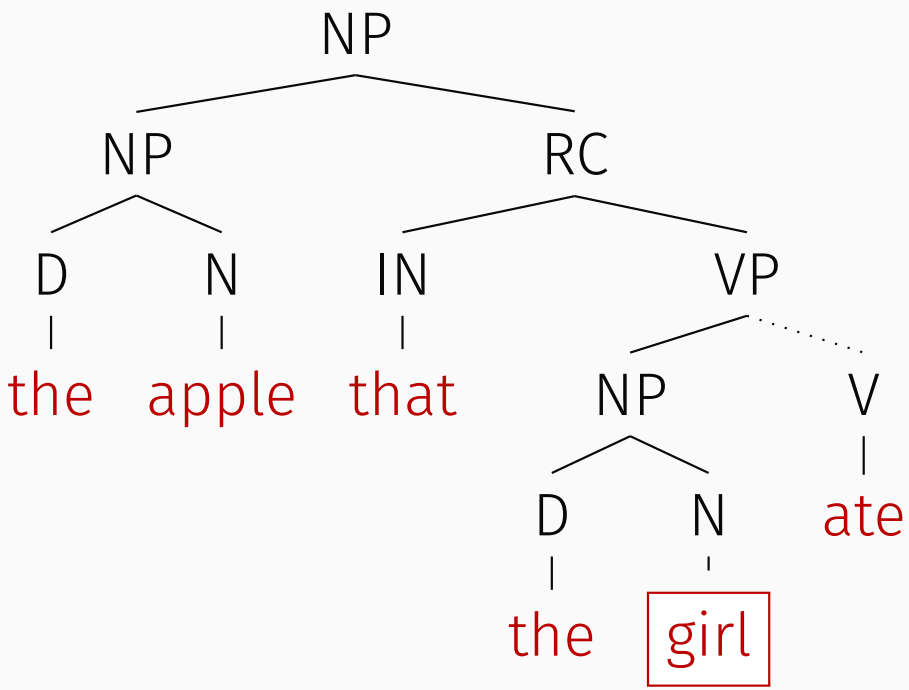
OVERVIEW

Previous studies have debated whether humans use hierarchic syntax



OVERVIEW

Previous studies have debated whether humans use hierarchic syntax



But standard baseline predictors may be deficient

OVERVIEW

This work shows that:

OVERVIEW

This work shows that:

Baselines can be greatly improved (accumulation)

OVERVIEW

This work shows that:

Baselines can be greatly improved (accumulation)

Hierarchic syntax is still predictive over stronger baseline

OVERVIEW

This work shows that:

Baselines can be greatly improved (accumulation)

Hierarchic syntax is still predictive over stronger baseline

Hierarchic syntax not improved by accumulation

OVERVIEW

This work shows that:

Baselines can be greatly improved (accumulation)

Hierarchic syntax is still predictive over stronger baseline

Hierarchic syntax not improved by accumulation

Long distance dependencies independently improve model

Hierarchic syntax in reading?

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

Frank & Bod (2011)

Hierarchic syntax in reading?

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

Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Hierarchic syntax in reading?

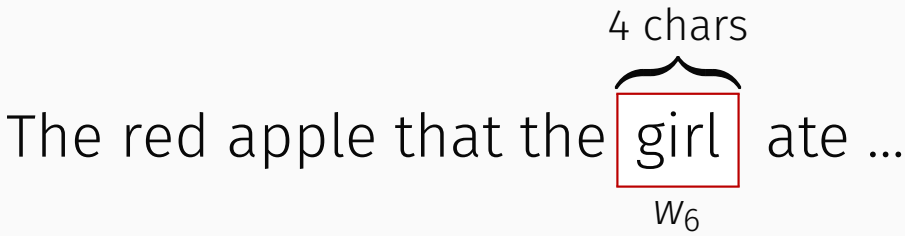
The red apple that the girl ate ...
w_1 w_2 w_3 w_4 w_5 w_6

Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Hierarchic syntax in reading?

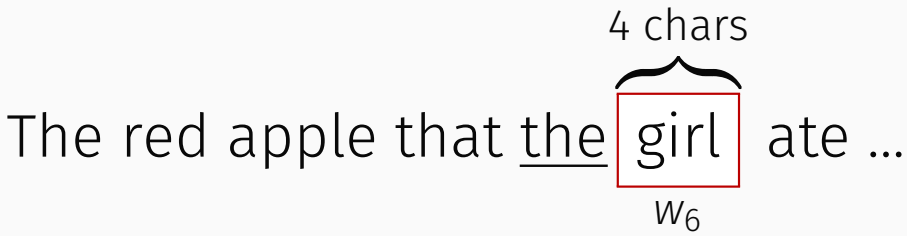


Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Hierarchic syntax in reading?

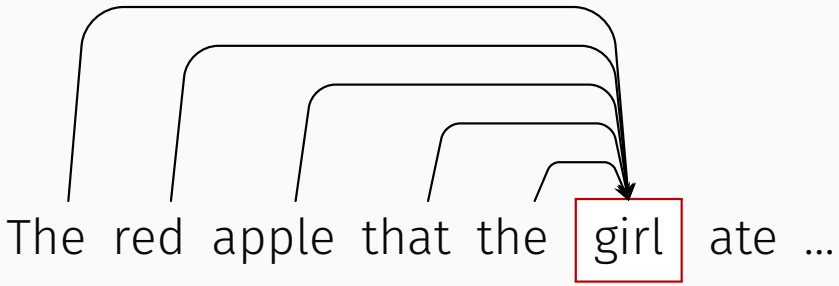


Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Hierarchic syntax in reading?



Frank & Bod (2011)

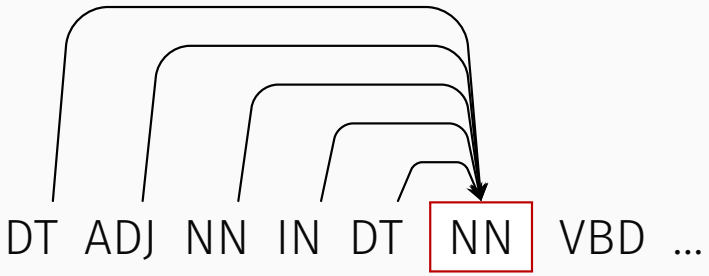
Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

Hierarchic syntax in reading?



Frank & Bod (2011)

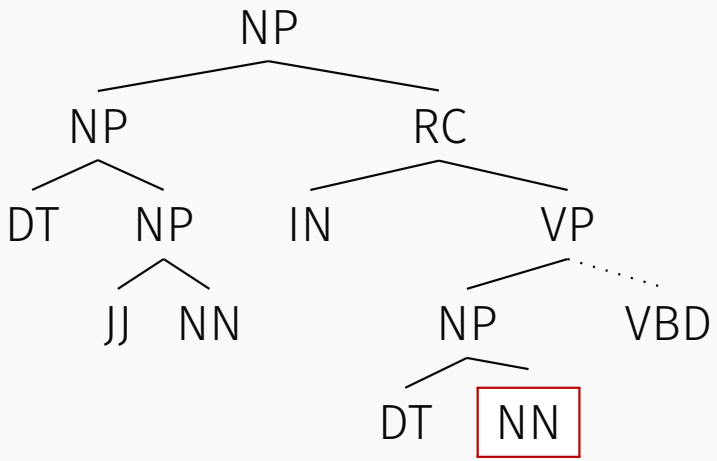
Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

Hierarchic syntax in reading?



Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

Hierarchic syntax in reading?

Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

Hierarchic syntax in reading?

Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

Outcome:

PSG $<$ ESN + PSG

ESN $=$ ESN + PSG

Hierarchic syntax in reading?

Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

Outcome:

PSG $<$ ESN + PSG Sequential helps over hierarchic

ESN $=$ ESN + PSG

Hierarchic syntax in reading?

Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

Outcome:

PSG $<$ ESN + PSG

ESN $=$ ESN + PSG Hierarchic doesn't help over sequential

Fossum & Levy (2012)

Replicated Frank & Bod (2011):

PSG $<$ ESN + PSG

ESN $=$ ESN + PSG

Fossum & Levy (2012)

Replicated Frank & Bod (2011):

PSG $<$ ESN + PSG

ESN $=$ ESN + PSG

Better *n*-gram baseline (more data) changes result:

PSG $=$ ESN + PSG

ESN $=$ ESN + PSG

FOSSUM & LEVY (2012)

Replicated Frank & Bod (2011):

PSG $<$ ESN + PSG

ESN $=$ ESN + PSG

Better *n*-gram baseline (more data) changes result:

PSG $=$ ESN + PSG Sequential doesn't help over hierarchic

ESN $=$ ESN + PSG

FOSSUM & LEVY (2012)

Replicated Frank & Bod (2011):

PSG $<$ ESN + PSG

ESN $=$ ESN + PSG

Better *n*-gram baseline (more data) changes result:

PSG $=$ ESN + PSG Sequential doesn't help over hierarchic

ESN $=$ ESN + PSG

Also: lexicalized syntax improves PSG fit

IMPROVED n-GRAM BASELINE

Previous reading time studies:

- Unigrams/Bigrams/Trigrams
  Trained on WSJ, Dundee, BNC

IMPROVED n-GRAM BASELINE

Previous reading time studies:

- Unigrams/Bigrams/Trigrams
  Trained on WSJ, Dundee, BNC
- Only from region boundaries

BIGRAM EXAMPLE

Reading time of *girl* after *red*

The red apple that the girl ate ...

region

X: bigram target X: bigram condition

BIGRAM EXAMPLE

Reading time of *girl* after *red*

The red apple that the girl ate ...

region

X: bigram target X: bigram condition

- Fails to capture entire sequence;

Improved n-gram baseline

Bigram Example

Reading time of *girl* after *red*

The red^{1} apple that the girl^{2} ate ...

region

X: bigram target X: bigram condition

- Fails to capture entire sequence;
- Conditions never generated;

BIGRAM EXAMPLE

Reading time of *girl* after *red*

The red apple that the girl ate ...

region

X: bigram target X: bigram condition

- Fails to capture entire sequence;
- Conditions never generated;
- Probability of sequence is deficient

CUMULATIVE BIGRAM EXAMPLE

Reading time of *girl* after *red*:

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

$\boxed{\text{X}}$: bigram targets $\underline{\text{X}}$: bigram conditions

CUMULATIVE BIGRAM EXAMPLE

Reading time of *girl* after *red*:

The ^{1}red apple that the ^{2}girl ate ...

X: bigram targets X: bigram conditions

- Captures entire sequence;
- Well-formed sequence probability;
- Reflects processing that must be done by humans

IMPROVED n-GRAM BASELINE

Previous reading time studies:

- Unigrams/Bigrams/Trigrams
- Trained on WSJ, Dundee, BNC
- Only from region boundaries

IMPROVED n-GRAM BASELINE

Previous reading time studies:

- Unigrams/Bigrams/Trigrams
- Trained on WSJ, Dundee, BNC
- Only from region boundaries

This study:

- 5-grams (w/ backoff)
- Trained on Gigaword 4.0
- Cumulative and Non-cumulative

EVALUATION

Dundee Corpus (Kennedy et al., 2003)

- 10 subjects
- 2,388 sentences
- 58,439 words
- 194,882 first pass durations
- 193,709 go-past durations

Exclusions:

- Unknown words (5 tokens)
- First and last of a line
- Regions larger than 4 words (track loss)

CUMU-n-GRAMS PREDICT READING TIMES

Baseline:

Fixed Effects

- Sentence Position
- Word length
- Region Length
- Preceding word fixated?

Random Effects

- Item/Subject Intercepts
- By Subject Slopes:
  - All Fixed Effects
  - N-grams (5-grams)
  - N-grams (Cumu-5-grams)

CUMU-n-GRAMS PREDICT READING TIMES

Baseline:

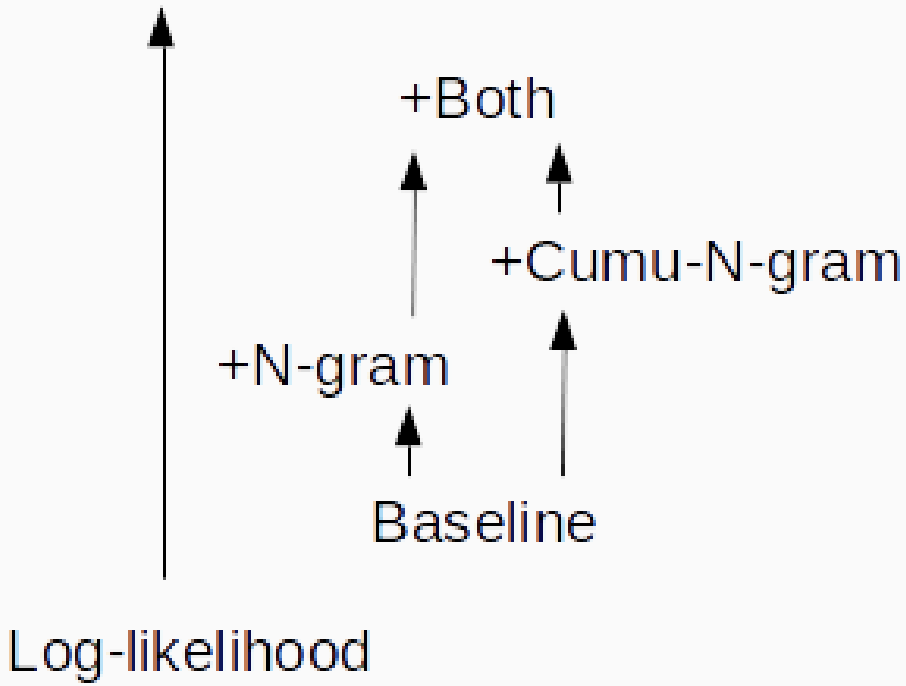
Fixed Effects

- Sentence Position
- Word length
- Region Length
- Preceding word fixated?

Random Effects

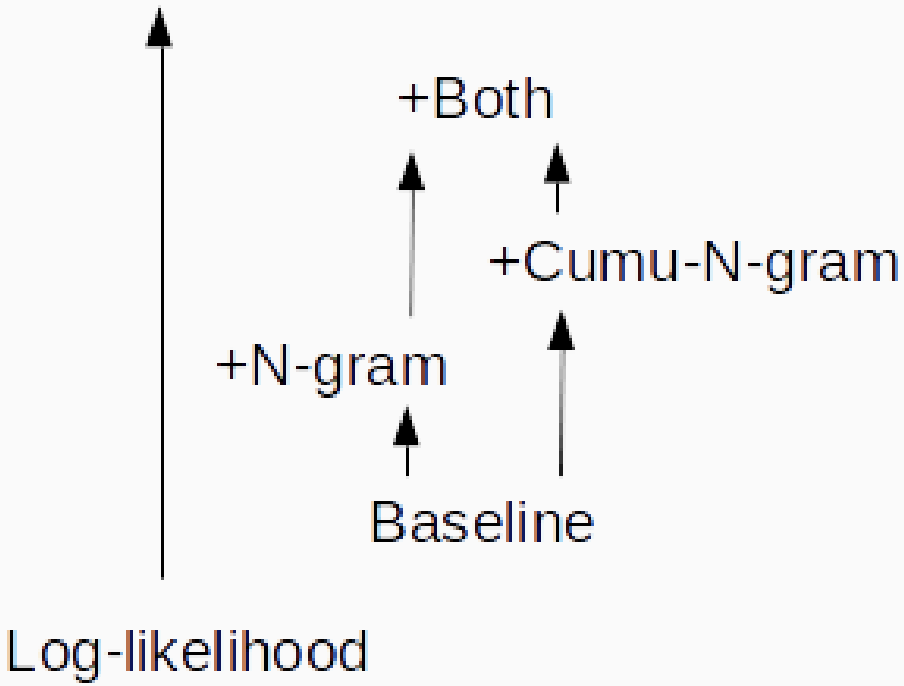
- Item/Subject Intercepts
- By Subject Slopes:
  - All Fixed Effects
  - N-grams (5-grams) $\leftarrow$
  - N-grams (Cumu-5-grams) $\leftarrow$

CUMU-n-GRAMS PREDICT READING TIMES

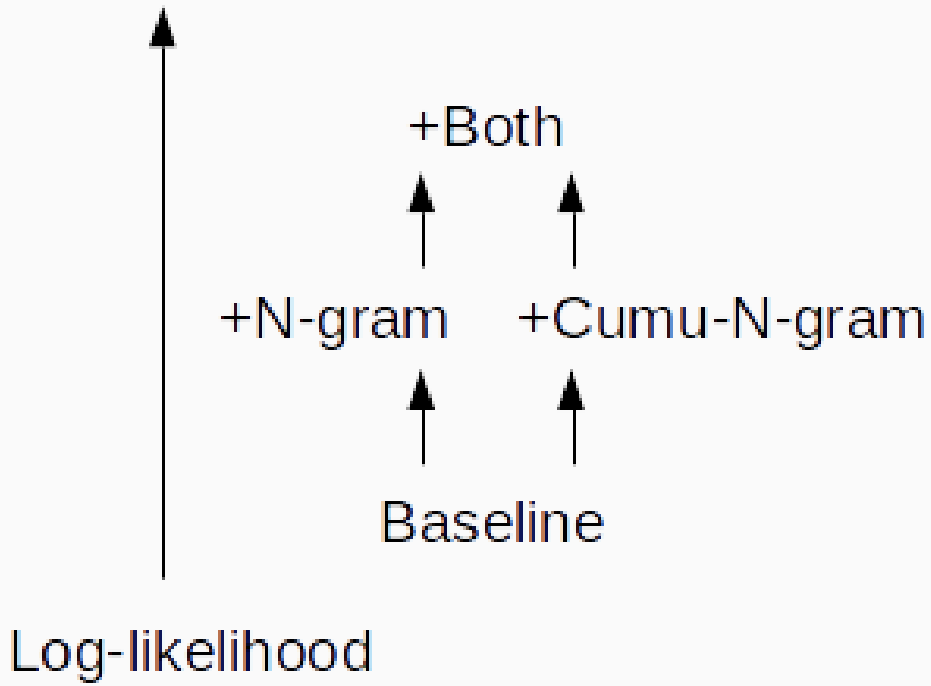


First Pass

CUMU-n-GRAMS PREDICT READING TIMES



First Pass



Go-Past

FOLLOW-UP QUESTIONS

FOLLOW-UP QUESTIONS

- Is hierarchic surprisal useful over the better baseline?

FOLLOW-UP QUESTIONS

- Is hierarchic surprisal useful over the better baseline?
- If so, can it be similarly improved through accumulation?

FOLLOW-UP QUESTIONS

- Is hierarchic surprisal useful over the better baseline?
- If so, can it be similarly improved through accumulation?
  van Schijndel & Schuler (2013) found it could over weaker baselines

Grammar:
Berkeley parser, WSJ, 5 split-merge cycles (Petrov & Klein 2007)

Hierarchic surprisal predicts reading times

Baseline:

Fixed Effects

- Same as before
- *N*-grams (5-grams)
- *N*-grams (Cumu-5-grams)

HIERARCHIC SURPRISAL PREDICTS READING TIMES

Baseline:

Fixed Effects

- Same as before
- *N*-grams (5-grams)
- *N*-grams (Cumu-5-grams)

Random Effects

- Same as before
- By Subject Slopes:
  - Hierarchic surprisal
  - Cumu-Hierarchic surprisal

Hierarchic surprisal predicts reading times

Baseline:

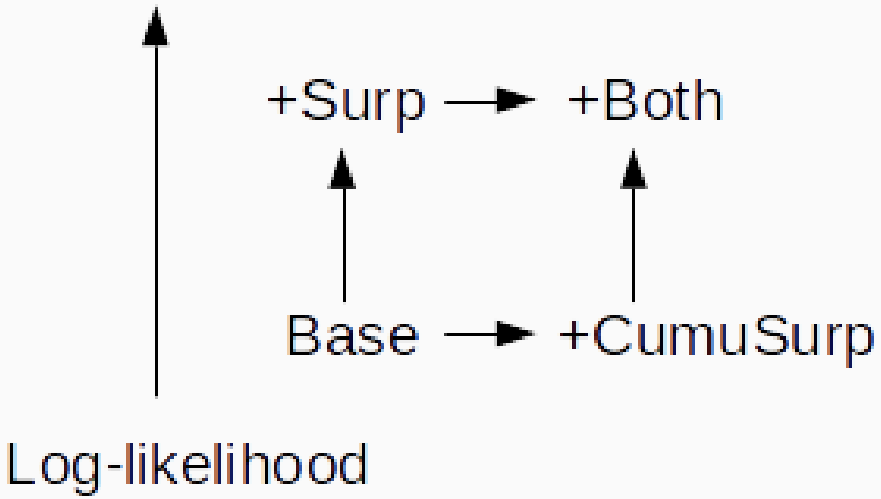
Fixed Effects

- Same as before
- *N*-grams (5-grams)
- *N*-grams (Cumu-5-grams)

Random Effects

- Same as before
- By Subject Slopes:
  - Hierarchic surprisal ←
  - Cumu-Hierarchic surprisal ←

HIERARCHIC SURPRISAL PREDICTS READING TIMES



First Pass and Go-Past

Cumulative surprisal doesn't help?!

- Suggests previous findings were due to weaker *n*-gram baseline

Cumulative surprisal doesn't help?!

- Suggests previous findings were due to weaker *n*-gram baseline
- Suggests only local PCFG surprisal affects reading times

Cumulative surprisal doesn't help?!

- Suggests previous findings were due to weaker *n*-gram baseline
- Suggests only local PCFG surprisal affects reading times

But... long-distance dependencies should affect reading times!

Cumulative surprisal doesn't help?!

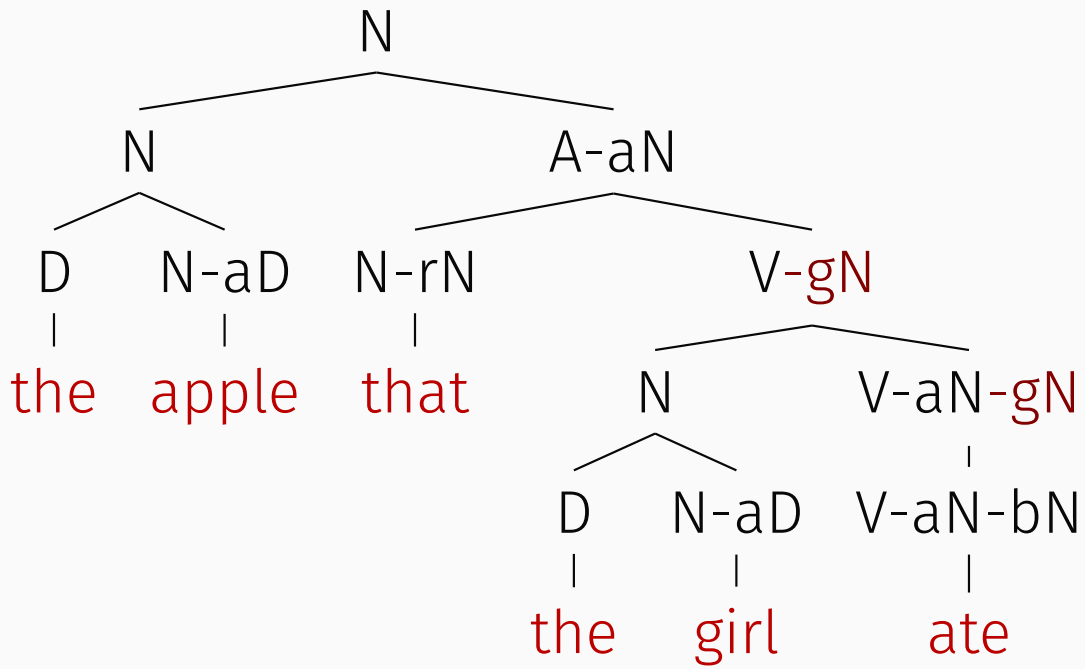
- Suggests previous findings were due to weaker *n*-gram baseline
- Suggests only local PCFG surprisal affects reading times

But... long-distance dependencies should affect reading times!

Let's try a PCFG that tracks long-distance deps

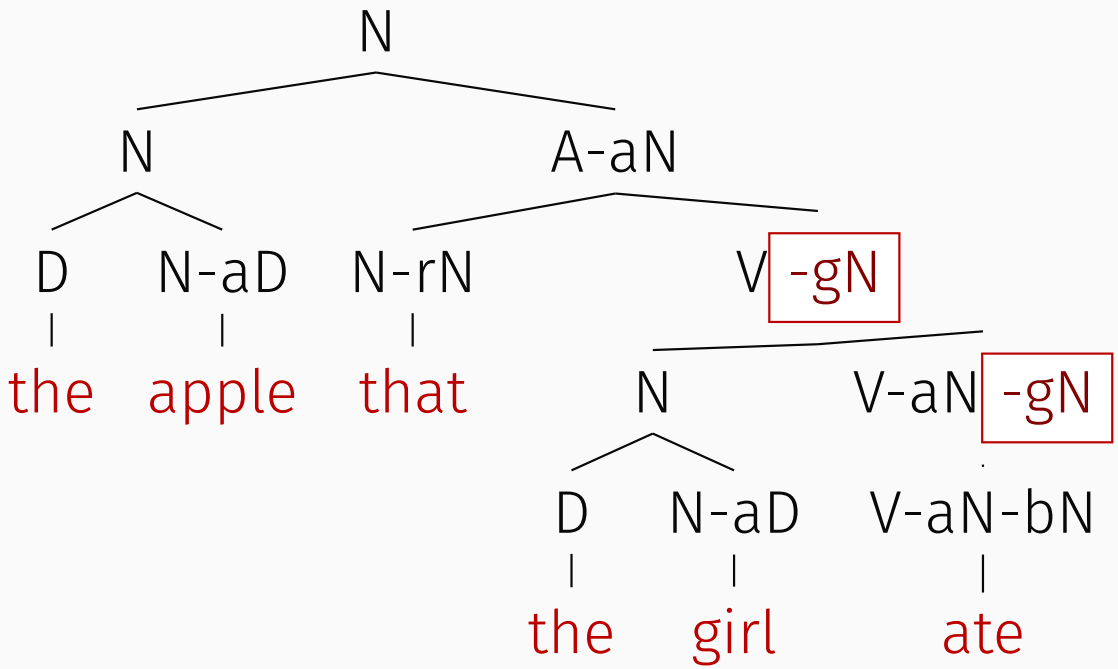
GENERALIZED CATEGORIAL GRAMMAR (GCG)

Nguyen et al. (2012)



GENERALIZED CATEGORIAL GRAMMAR (GCG)

Nguyen et al. (2012)



Long-distance surprisal predicts reading times

Baseline:

Fixed Effects

- Same as before

Random Effects

- Same as before
- By Subject Slopes:
  - Hierarchic PTB surprisal
  - Hierarchic GCG surprisal

Long-distance surprisal predicts reading times

Baseline:

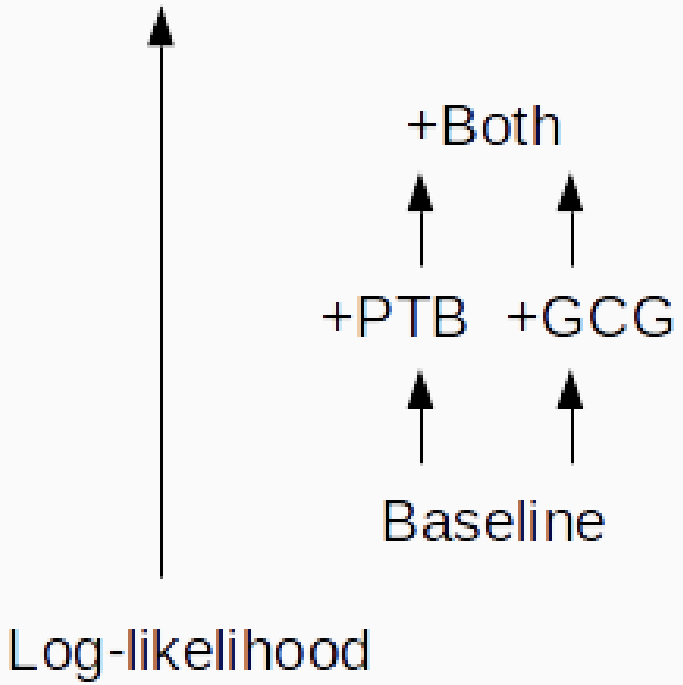
Fixed Effects

- Same as before

Random Effects

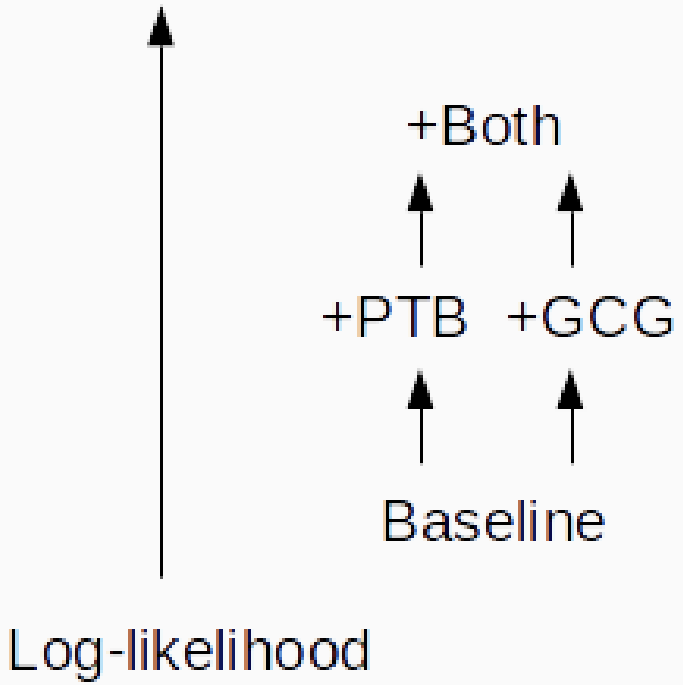
- Same as before
- By Subject Slopes:
  - Hierarchic PTB surprisal ←
  - Hierarchic GCG surprisal ←

LONG-DISTANCE SURPRISAL PREDICTS READING TIMES



First Pass and Go-Past

LONG-DISTANCE SURPRISAL PREDICTS READING TIMES



First Pass and Go-Past

Both help independently

Hierarchic syntax predicts reading times over strong linear baseline

DISCUSSION AND CONCLUSIONS

Hierarchic syntax predicts reading times over strong linear baseline

Long-distance dependencies *do* affect reading times

DISCUSSION AND CONCLUSIONS

Hierarchic syntax predicts reading times over strong linear baseline

Long-distance dependencies *do* affect reading times

Studies should use cumu-*n*-grams in their baselines

Compare to Echo State Networks

FUTURE WORK

Compare to Echo State Networks

Test anticipatory accumulation

QUESTIONS?

Thanks to:

- Stefan Frank
- Attendees of CUNY 2015
- National Science Foundation (DGE-1343012)

CUMU-n-GRAMS PREDICT READING TIMES

First Pass Evaluation (Log-Likelihood):

Base -1212399	
Base+N-gram -1212396 ($p < 0.05$)	**Base+Cumu-n-gram** -1212392 ($p < 0.01$)
Base+Both -1212387 ($p < 0.01$)	**Base+Both** -1212387 ($p < 0.01$)

Comparable with go-past durations

Go-Past Evaluation (Log-Likelihood):

Base -1261582	
Base+*N*-gram -1261577 ($p < 0.01$)	**Base+Cumu-*n*-gram** -1261576 ($p < 0.01$)
Base+Both -1261570 ($p < 0.01$)	**Base+Both** -1261570 ($p < 0.01$)

HIERARCHIC SURPRISAL PREDICTS READING TIMES

First Pass Evaluation (Log-Likelihood):

Base −1212260	
Base+Surp −1212253 ($p < 0.01$)	Base+CumuSurp −1212259
Base+Both −1212253	**Base+Both** −1212253 ($p < 0.01$)

Comparable with go-past durations

Go-Past Evaluation (Log-Likelihood):

Base −1261488	
Base+Surp −1261481 ($p < 0.01$)	Base+CumuSurp −1261487
Base+Both −1261481	**Base+Both** −1261481 ($p < 0.01$)

Long-distance surprisal predicts reading times

First Pass Evaluation (Log-Likelihood):

Base -1212242	
Base+PTB -1212239 ($p < 0.01$)	**Base+GCG** -1212239 ($p < 0.05$)
Base+Both -1212235 ($p < 0.05$)	**Base+Both** -1212235 ($p < 0.01$)

Both help independently

PCFG surprisal helps more with go-past durations

Go-Past Evaluation (Log-Likelihood):

Base -1261474	
Base+PTB -1261468 ($p < 0.01$)	**Base+GCG** -1261470 ($p < 0.01$)
Base+Both -1261465 ($p < 0.01$)	**Base+Both** -1261465 ($p < 0.01$)

Again, both help independently.

FIXED EFFECT COEFFICIENTS FOR BASE+PTB+GCG

Predictor	First Pass		Go-Past	
	coef	t value	coef	t value
sentpos	−2.47	−3.59	−2.82	−3.38
wlen	25.90	8.67	28.98	9.97
prevfix	−30.16	−7.81	−37.42	−11.49
n-gram	−2.39	−1.81	−6.70	−3.36
cumu-*n*-gram	−14.69	−7.36	−11.68	−5.01
rlen	−5.67	−1.31	−12.51	−2.59
surp-GCG	4.97	2.87	5.74	2.73
surp-PTB	4.20	3.23	4.85	3.29