# HIERARCHIC SYNTAX IMPROVES READING TIME PREDICTION

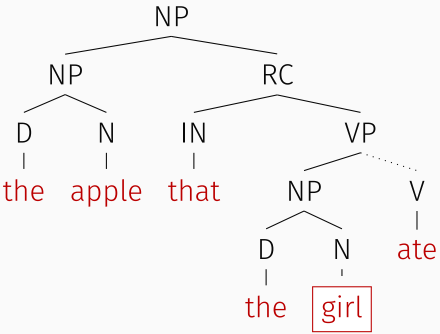
Marten van Schijndel and William Schuler
Department of Linguistics
The Ohio State University

June 3, 2015

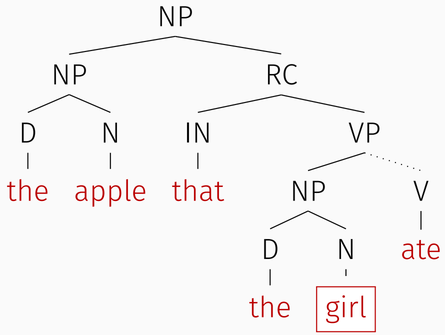
Previous studies have debated whether humans use hierarchic syntax

## OVERVIEW

Previous studies have debated whether humans use hierarchic syntax



Previous studies have debated whether humans use hierarchic syntax




But standard baseline predictors may be deficient

This work shows that:

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

This work shows that:
Baselines can be greatly improved (accumulation)

Hierarchic syntax is still predictive over stronger baseline

Hierarchic syntax not improved by accumulation

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

1 2

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

## Frank & Bod (2011)

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

## Frank & Bod (2011)

Baseline:

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

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

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

## Frank & Bod (2011)

Baseline:

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

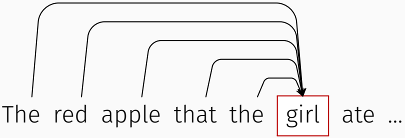
$$\text{The red apple that } \underline{\text{the}}\ \underbrace{\overbrace{\boxed{\text{girl}}}^{\text{4 chars}}}_{w_6}\ \text{ate ...}$$

## Frank & Bod (2011)

Baseline:

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

# HIERARCHIC SYNTAX IN READING?




## Frank & Bod (2011)

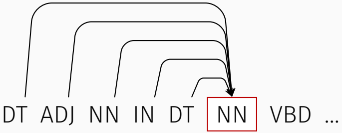
Baseline:

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

Test POS Predictors:

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

# HIERARCHIC SYNTAX IN READING?



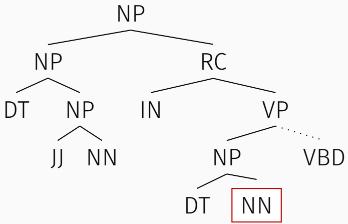

## Frank & Bod (2011)

Baseline:

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

Test POS Predictors:

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




## Frank & Bod (2011)

Baseline:

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

Test POS Predictors:

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

## Frank & Bod (2011)

Baseline:

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

Test POS Predictors:

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

## Frank & Bod (2011)

Baseline:

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

Test POS Predictors:

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

Outcome:

$\text{PSG} < \text{ESN} + \text{PSG}$

$\text{ESN} = \text{ESN} + \text{PSG}$

## 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

## 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 $\boxed{=}$ 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 $\boxed{=}$ 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 $\boxed{=}$ 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

Previous reading time studies:

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

## Bigram Example

Reading time of *girl* after *red*

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

$\boxed{X}$: bigram target $\underline{X}$: bigram condition

## Bigram Example

Reading time of *girl* after *red*

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

$\boxed{X}$: bigram target    $\underline{X}$: bigram condition

- Fails to capture entire sequence;

### Bigram Example

Reading time of *girl* after *red*

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

$\boxed{\text{X}}$: bigram target  $\underline{\text{X}}$: bigram condition

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

## Bigram Example

Reading time of *girl* after *red*

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

$\boxed{\text{X}}$: bigram target    $\underline{\text{X}}$: 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 red[1] $\boxed{\text{apple}}$ $\boxed{\text{that}}$ $\boxed{\text{the}}$ $\boxed{\text{girl}}$[2] ate ...

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

## Cumulative Bigram Example

Reading time of *girl* after *red*:

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

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

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

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

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)

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)

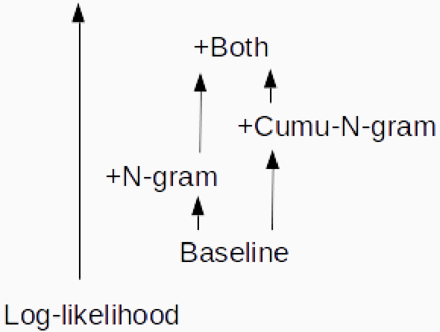
Baseline:

Fixed Effects

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

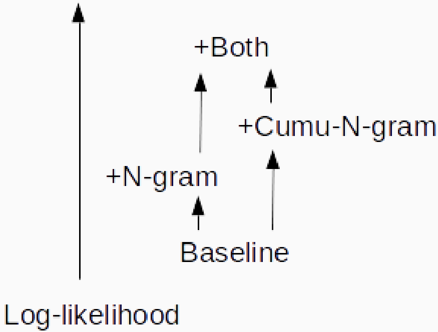
Random Effects

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

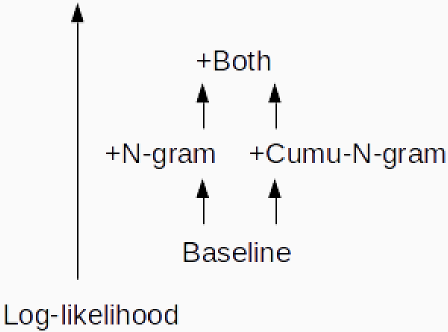



First Pass

# CUMU-$n$-GRAMS PREDICT READING TIMES




First Pass




Go-Past

## 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)

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

Baseline:

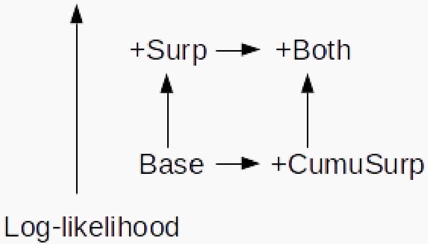
Fixed Effects

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

Random Effects

- Same as before
- By Subject Slopes:
  - Hierarchic surprisal $\leftarrow$
  - Cumu-Hierarchic surprisal $\leftarrow$

# 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

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

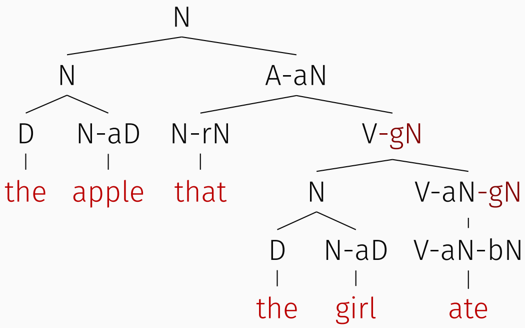
But… long-distance dependencies should affect reading times!

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

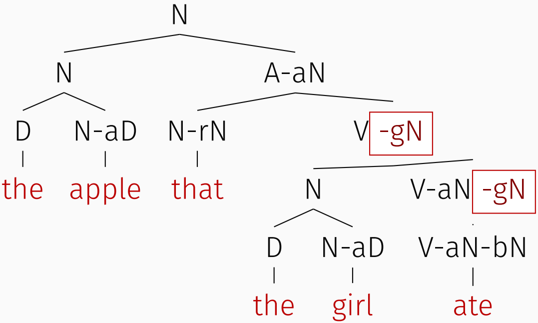
But... long-distance dependencies should affect reading times!

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

Nguyen et al. (2012)



Nguyen et al. (2012)



Baseline:

Fixed Effects

- Same as before

Random Effects

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

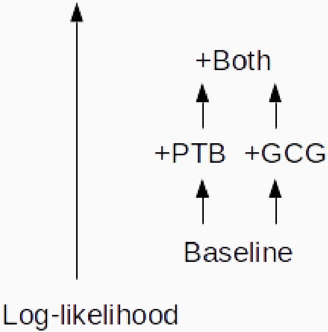
Baseline:

Fixed Effects

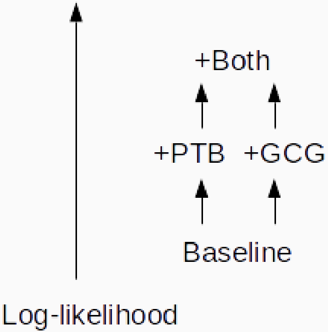
- Same as before

Random Effects

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




First Pass and Go-Past




First Pass and Go-Past

Both help independently

Hierarchic syntax predicts reading times over strong linear baseline

Hierarchic syntax predicts reading times over strong linear baseline

Long-distance dependencies *do* affect reading times

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

# FUTURE WORK

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)

First Pass Evaluation (Log-Likelihood):

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

Comparable with go-past durations

Go-Past Evaluation (Log-Likelihood):

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

First Pass Evaluation (Log-Likelihood):

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

Comparable with go-past durations

Go-Past Evaluation (Log-Likelihood):

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

First Pass Evaluation (Log-Likelihood):

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

Both help independently

PCFG surprisal helps more with go-past durations

Go-Past Evaluation (Log-Likelihood):

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

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 |