# Character-based Surprisal as a Model of Reading Difficulty in the Presence of Errors

Michael Hahn
Stanford

Frank Keller
University of Edinburgh

Yonatan Bisk
University of Washington

Yonatan Belinkov
Harvard & MIT

# Human Reading is...

- Effortless and Fast: ~ 250 words per minute (Rayner, White, Johnson, & Liversedge, 2006)

# Human Reading is...

- Effortless and Fast: ~ 250 words per minute (Rayner, White, Johnson, & Liversedge, 2006)
- Adaptive and task-dependent (Kaakinen & Hyönä, 2010; Schotter et al. 2014; Hahn & Keller, 2018)

# Human Reading is...

- Effortless and Fast: ~ 250 words per minute (Rayner, White, Johnson, & Liversedge, 2006)
- Adaptive and task-dependent (Kaakinen & Hyönä, 2010; Schotter et al. 2014; Hahn & Keller, 2018)
- Robust:
  - We often encounter errors (hand-written notes, emails, text messages, and social media posts)
  - Intuitively: easy to cope with, often go unnoticed

# Human Reading is...

- Effortless and Fast: ~ 250 words per minute (Rayner, White, Johnson, & Liversedge, 2006)
- Adaptive and task-dependent (Kaakinen & Hyönä, 2010; Schotter et al. 2014; Hahn & Keller, 2018)
- Robust:
  - We often encounter errors (hand-written notes, emails, text messages, and social media posts)
  - Intuitively: easy to cope with, often go unnoticed



Source: https://www.grammarly.com/blog/autocorrect-text-fails/

# Human Reading is...

- Effortless and Fast: ~ 250 words per minute (Rayner, White, Johnson, & Liversedge, 2006)
- Adaptive and task-dependent (Kaakinen & Hyönä, 2010; Schotter et al. 2014; Hahn & Keller, 2018)
- Robust:
  - We often encounter errors (hand-written notes, emails, text messages, and social media posts)
  - Intuitively: easy to cope with, often go unnoticed

**Aim of this paper:**

1. Experimentally investigate reading in the face of errors
2. Propose simple model to account for results

# Types of Errors

- Focus on errors that change the form of a word

# Types of Errors

- Focus on errors that change the form of a word
  - **letter transposition**

# Types of Errors

- Focus on errors that change the form of a word
  - **letter transposition**

# Types of Errors

- Focus on errors that change the form of a word
  - letter transposition
  - **misspellings**


- Typically, writer didn’t know standard spelling
- Typically conforms to phonotactics

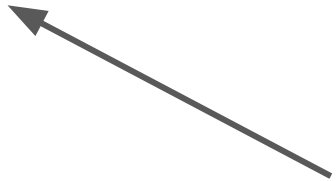
# Types of Errors

- Focus on errors that change the form of a word
  - letter transposition
  - misspellings

- We don’t study semantic, syntactic, … errors.

# Types of Errors

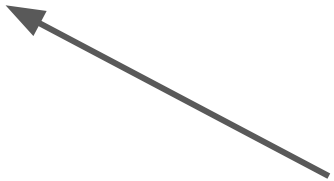
- Focus on errors that change the form of a word
  - **letter transposition**
  - misspellings



Known to cause reading difficulty... (Rayner et al., 2006; Johnson et al., 2007; White et al. 2008)

# Types of Errors

- Focus on errors that change the form of a word
  - **letter transposition**
  - misspellings

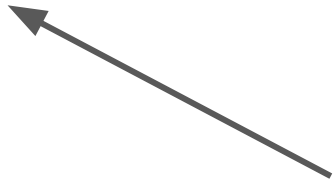


Known to cause reading difficulty... (Rayner et al., 2006; Johnson et al., 2007; White et al. 2008)

… but artificial and rare

# Types of Errors

- Focus on errors that change the form of a word
  - letter transposition
  - **misspellings**



Known to cause reading difficulty... (Rayner et al., 2006; Johnson et al., 2007; White et al. 2008)

… but artificial and rare

# Types of Errors

- Focus on errors that change the form of a word
  - letter transposition
  - misspellings

Known to cause reading difficulty... (Rayner et al., 2006; Johnson et al., 2007; White et al. 2008)

… but artificial and rare

Prediction: Misspellings will cause less difficulty than transpositions.

# Eye-Tracking Experiment

Q: How is human reading affected by errors in the input?

# Eye-Tracking Experiment

Q: How is human reading affected by errors in the input?

**Predictions:**

1. Transpositions more difficult than misspellings

    - Transpositions create rare / phonotactically invalid letter sequences.

innocetn vs inocent

# Eye-Tracking Experiment

Q: How is human reading affected by errors in the input?

**Predictions:**

1. Transpositions more difficult than misspellings
2. Higher error rates increase difficulty on all words
    - Errors degrade the context available for processing other words.

# Eye-Tracking Experiment

- 20 newspaper texts from the DeepMind QA corpus (Hermann et al., 2015)
- length: min 149, max 805, mean 323 words
- balanced selection of topics
- +2 practice texts

# Eye-Tracking Experiment

- 20 newspaper texts from the DeepMind QA corpus (Hermann et al., 2015)
- length: min 149, max 805, mean 323 words
- balanced selection of topics
- +2 practice texts

- Introduced errors automatically (Belinkov and Bisk, 2018)
  - transpositions
  - misspellings from corpus of human edits (Geertzen et al., 2014)
- Error rates: 10% or 50% erroneous words

Sabra Dipping Co. is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. Food and Drug Administration said Wednesday. The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported. The potential for contamination was

**Question:** A random sample from a _____ store tested positive for Listeria monocytogenes.

**Answers:** (1) Michigan (2) Washington (3) Ohio (4) Georgia

Misspellings, 10% error rate

Sabra Dipping Co. is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. Food and Drag Administration said Wednesday. Ihe nationwide recall is voluntary. So far, NO illnes caused by the hummus have been reported. The potential for cotamination was

Misspellings, 10% error rate

Sabra Dipping Co. is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. Food and Drag Administration said Wednesday. Ihe nationwide recall is voluntary. So far, NO illnes caused by the hummus have been reported. The potential for cotamination was

Misspellings, 50% error rate

Sabra Dipping Co. is recalling 30,000 casses off hummus dur por possibe cotamination wift Listeria, DE u.s Food ang Drag Administation sayed Wednesday.
them nationwide recall is voluntary. Soo far, NO illnes caused bye the hummus heve been reported.
THe potential fpr contamination wass discovered

Misspellings, 50% error rate

Sabra Dipping Co. is recalling 30,000 casses off hummus dur por possibe cotamination wift Listeria, DE u.s Food ang Drag Administation sayed Wednesday.
them nationwide recall is voluntary. Soo far, NO illnes caused bye the hummus heve been reported.
THe potential fpr contamination wass discovered

Transpositions, 10% error rate

Sabra Dipping Co. is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. Food and Drgu Administration said Wednesday.
The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported.
The potential for contaminatino was discovered

Transpositions, 50% error rate

Sarba Dipping Co. si recallign 30,000 caess fo humums ude ot possible ocntamination with Litseria, teh U.S. Food and Durg Administration said Wednesdya.
Teh nationwide ercall is voluntary. So afr, no illnesses caused yb teh hummsu hvae been reported.
Teh ptoential for contaminatino wsa discovered

# Eye-Tracking Experiment: Design

- 4 versions for each text
- Within participants:
  - all participants read all texts
  - each of them in 1 of 4 versions
- 16 participants
- Random order of texts per participant

| | Error Rate | |
|---|---|---|
| | 10% | 50% |
| Transpositions | 5 texts | 5 texts |
| Misspellings | 5 texts | 5 texts |

# Predictors

1. **ErrorType**: mispelling or transposition?
2. **ErrorRate**: 10% or 50% erroneous words overall?

# Predictors

1. **ErrorType**: mispelling or transposition?
2. **ErrorRate**: 10% or 50% erroneous words overall?
3. **Error**: current word correct or erroneous?
4. **WordLength**: Length of the word in characters.
5. **LastFix**: Was the preceding word fixated? (controls for preview effects.)

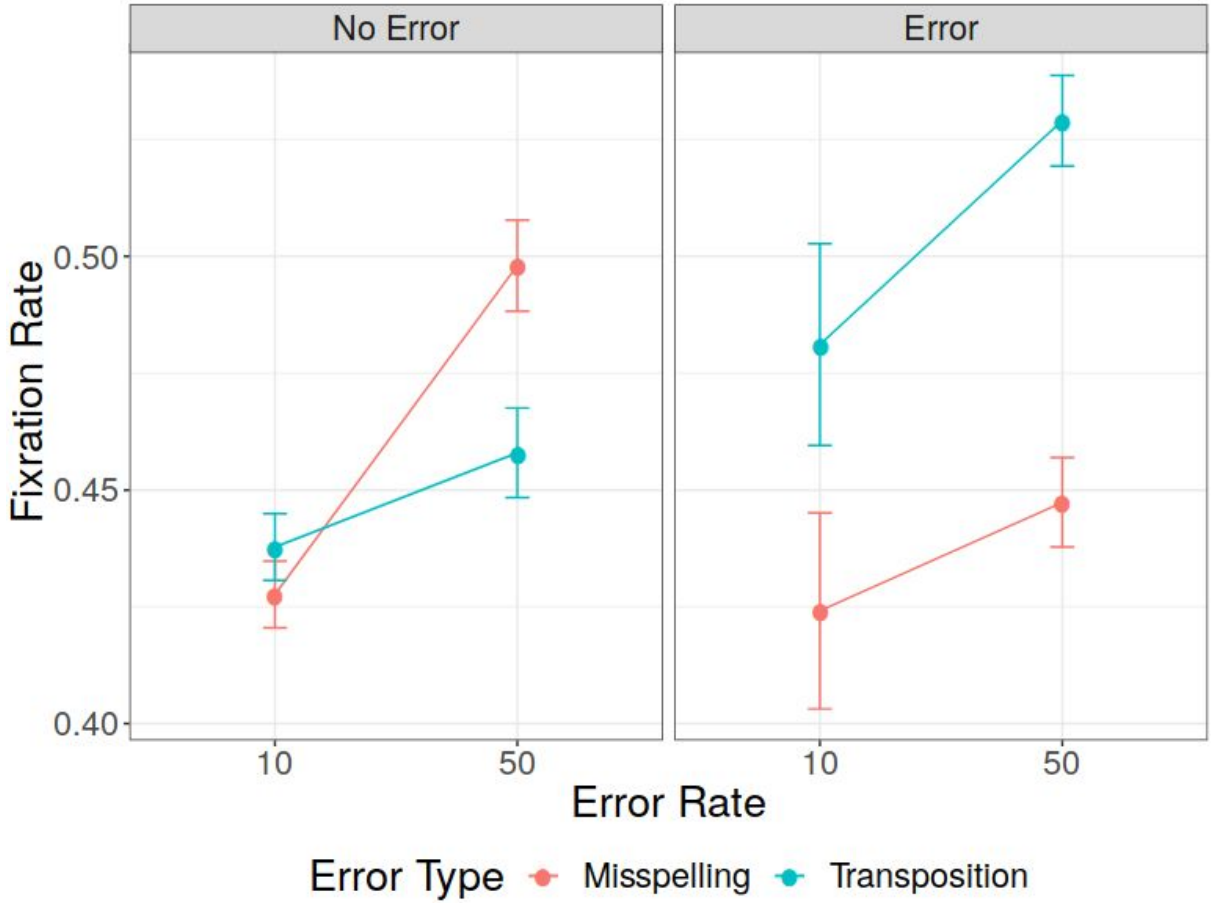
| | First Pass | | Fixation Rate | |
|---|---|---|---|---|
| (Intercept) | 248.41 | (6.34)*** | −0.16 | (0.12) |
| ErrType | 1.41 | (1.32) | 0.08 | (0.02)*** |
| ErrRate | 7.20 | (1.60)*** | 0.16 | (0.02)*** |
| Error | 23.77 | (4.12)*** | 0.21 | (0.07)*** |
| WLength | 22.18 | (2.02)*** | 0.83 | (0.04)*** |
| LastFix | 3.10 | (4.18) | 0.22 | (0.18) |
| ErrRate × LastFix | 6.71 | (2.77)* | 0.16 | (0.04)*** |
| Error × LastFix | | — | 0.26 | (0.10)** |
| WLength × LastFix | | — | 0.74 | (0.10)*** |

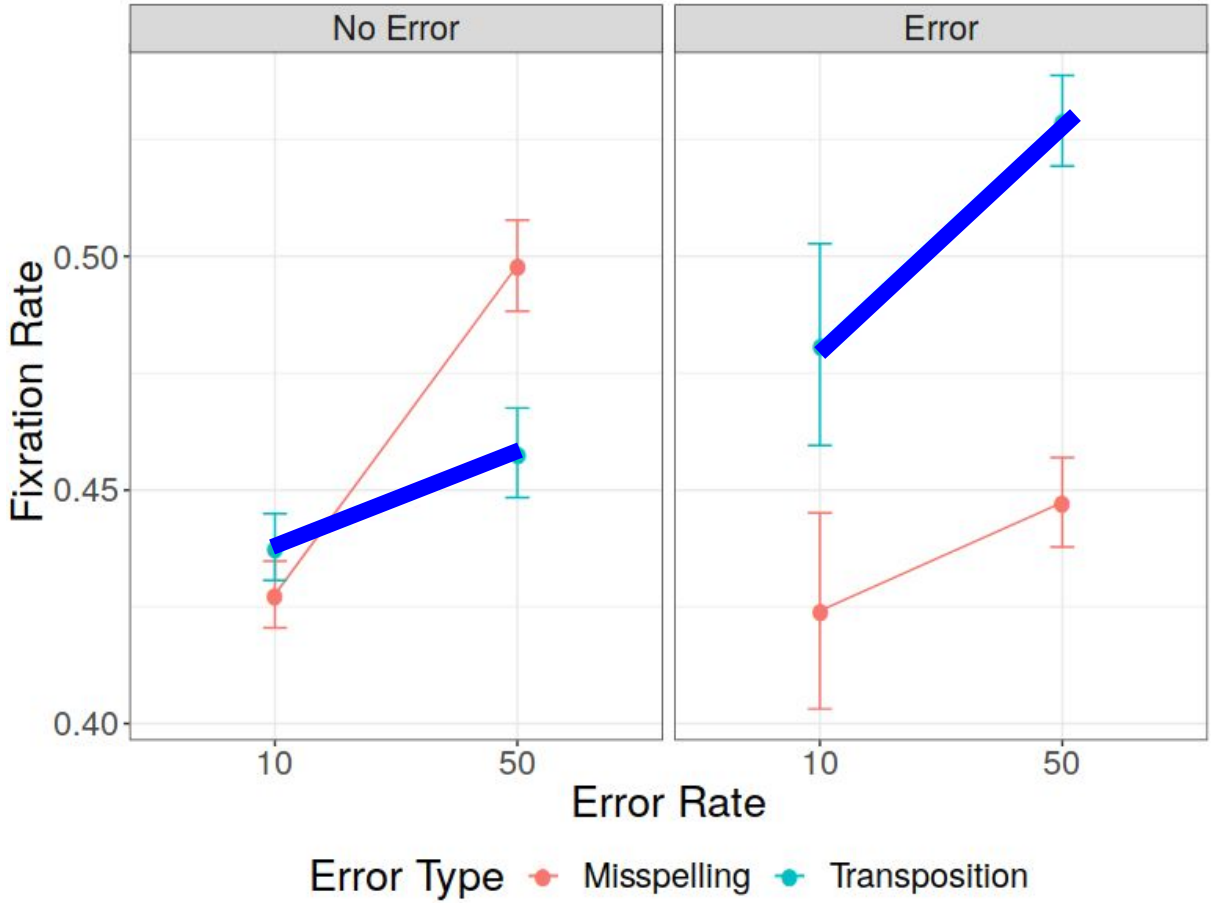
$Pr(\beta < 0)$: *** < 0.001, ** < 0.01, * < 0.05

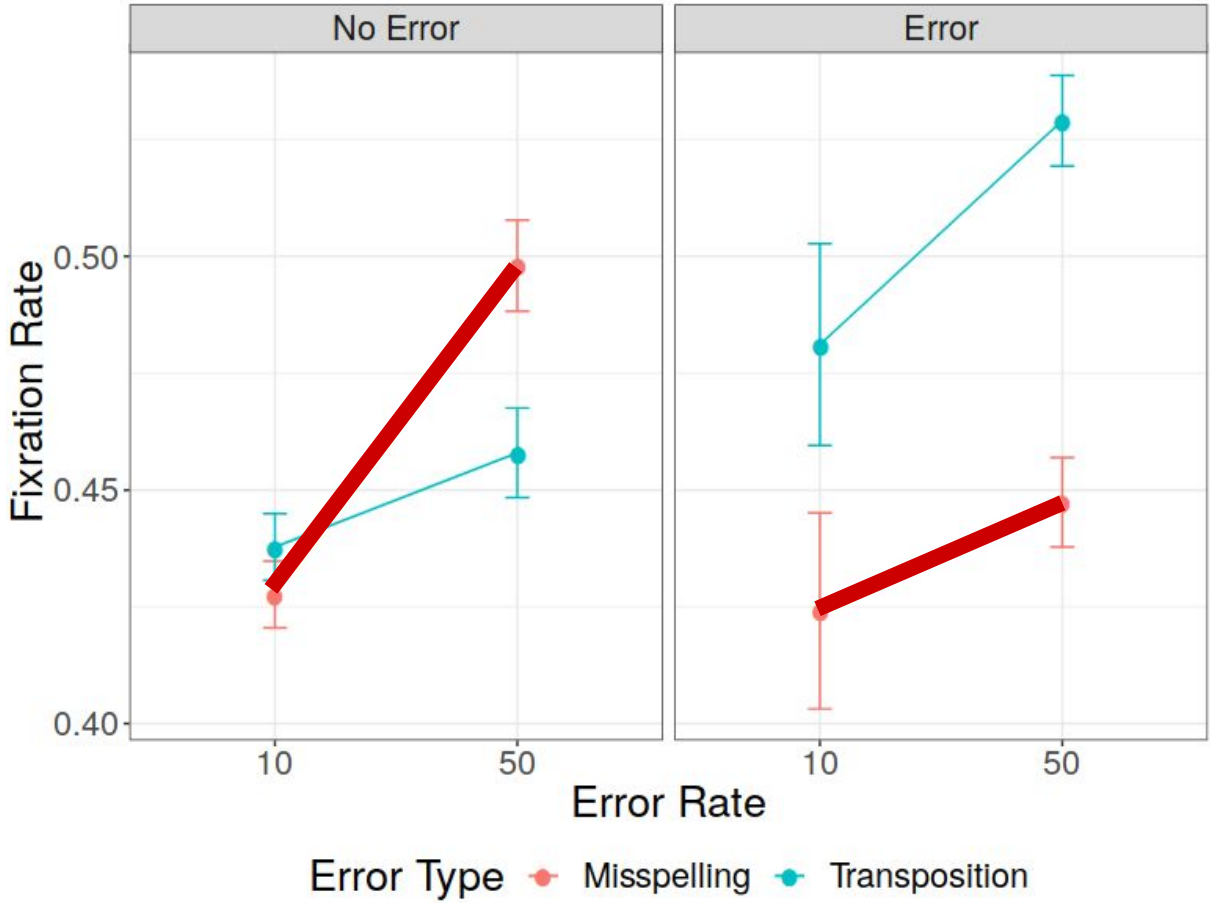
Transpositions increase fixations

| | First Pass | Fixation Rate |
|---|---|---|
| (Intercept) | 248.41 (6.34)*** | −0.16 (0.12) |
| ErrType | 1.41 (1.32) | 0.08 (0.02)*** |
| ErrRate | 7.20 (1.60)*** | 0.16 (0.02)*** |
| Error | 23.77 (4.12)*** | 0.21 (0.07)*** |
| WLength | 22.18 (2.02)*** | 0.83 (0.04)*** |
| LastFix | 3.10 (4.18) | 0.22 (0.18) |
| ErrRate × LastFix | 6.71 (2.77)* | 0.16 (0.04)*** |
| Error × LastFix | — | 0.26 (0.10)** |
| WLength × LastFix | — | 0.74 (0.10)*** |

$Pr(\beta < 0)$: *** < 0.001, ** < 0.01, * < 0.05


No Error
Error
Fixration Rate
0.50
0.45
0.40
10
50
10
50
Error Rate
Error Type
Misspelling
Transposition

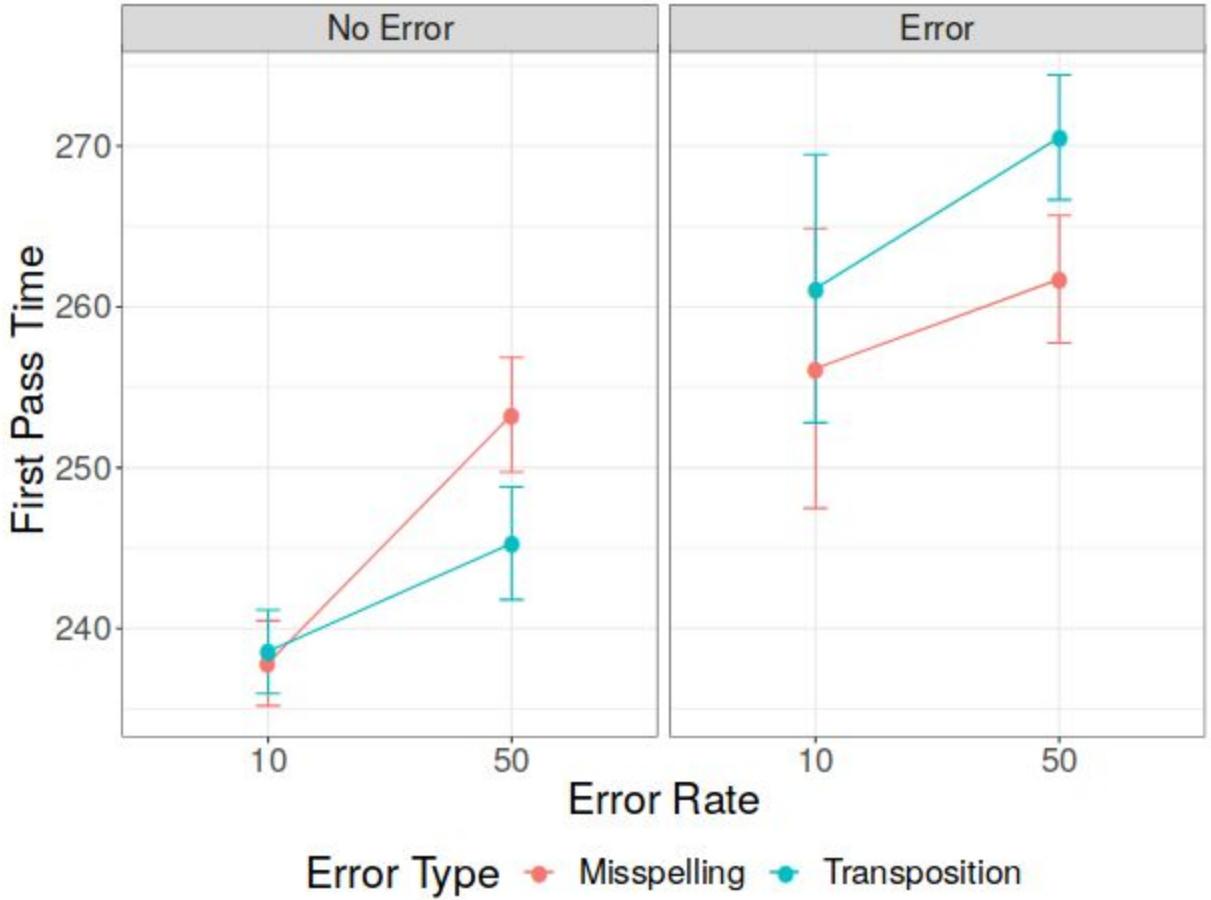

No Error
Error
Fixration Rate
0.50
0.45
0.40
10
50
10
50
Error Rate
Error Type
Misspelling
Transposition

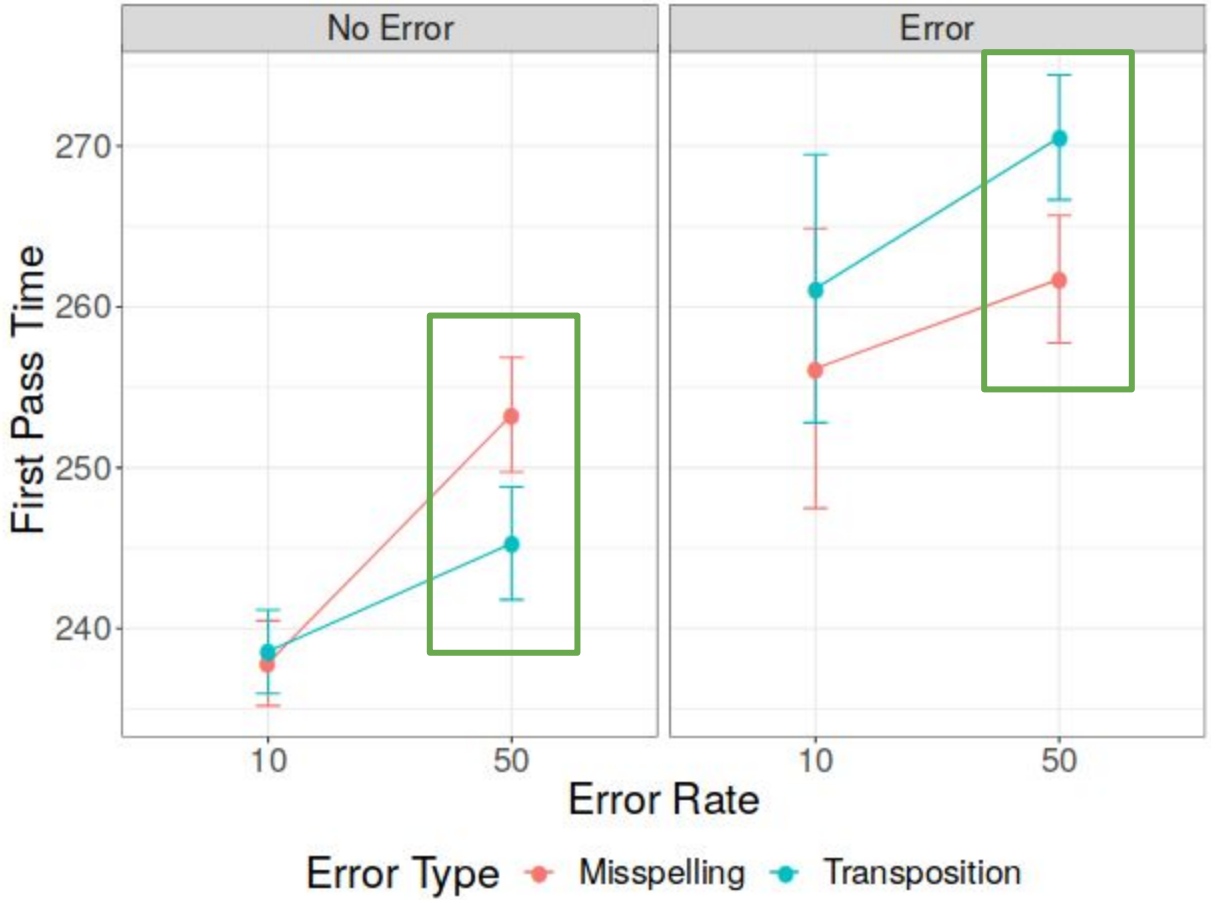

No Error
Error
Fixration Rate
0.50
0.45
0.40
10
50
10
50
Error Rate
Error Type
Misspelling
Transposition

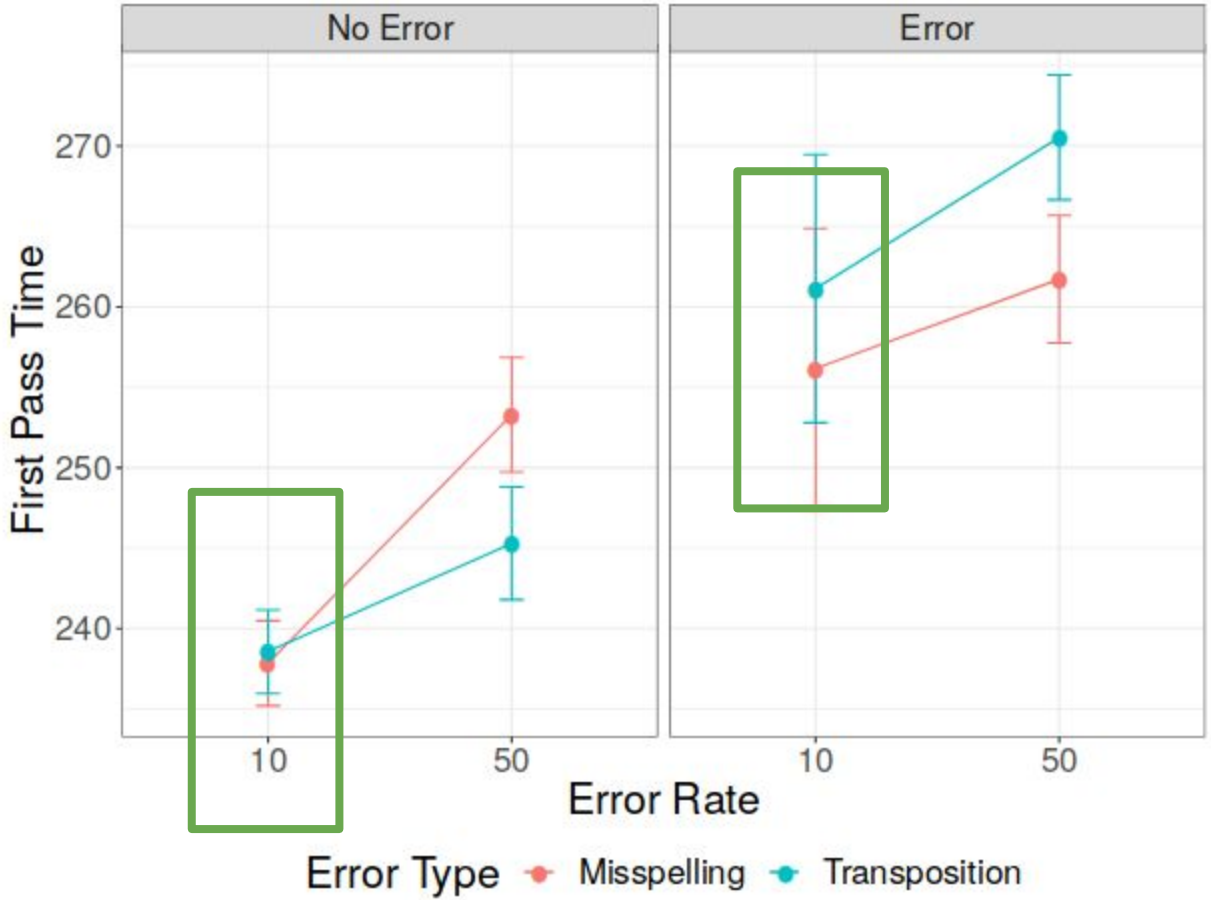
Error rate

| | First Pass | Fixation Rate |
|---|---|---|
| (Intercept) | 248.41 (6.34)*** | −0.16 (0.12) |
| ErrType | 1.41 (1.32) | 0.08 (0.02)*** |
| ErrRate | 7.20 (1.60)*** | 0.16 (0.02)*** |
| Error | 23.77 (4.12)*** | 0.21 (0.07)*** |
| WLength | 22.18 (2.02)*** | 0.83 (0.04)*** |
| LastFix | 3.10 (4.18) | 0.22 (0.18) |
| ErrRate × LastFix | 6.71 (2.77)* | 0.16 (0.04)*** |
| Error × LastFix | — | 0.26 (0.10)** |
| WLength × LastFix | — | 0.74 (0.10)*** |

$Pr(\beta < 0)$: *** < 0.001, ** < 0.01, * < 0.05


No Error
Error
First Pass Time
270
260
250
240
10
50
10
50
Error Rate
Error Type
Misspelling
Transposition

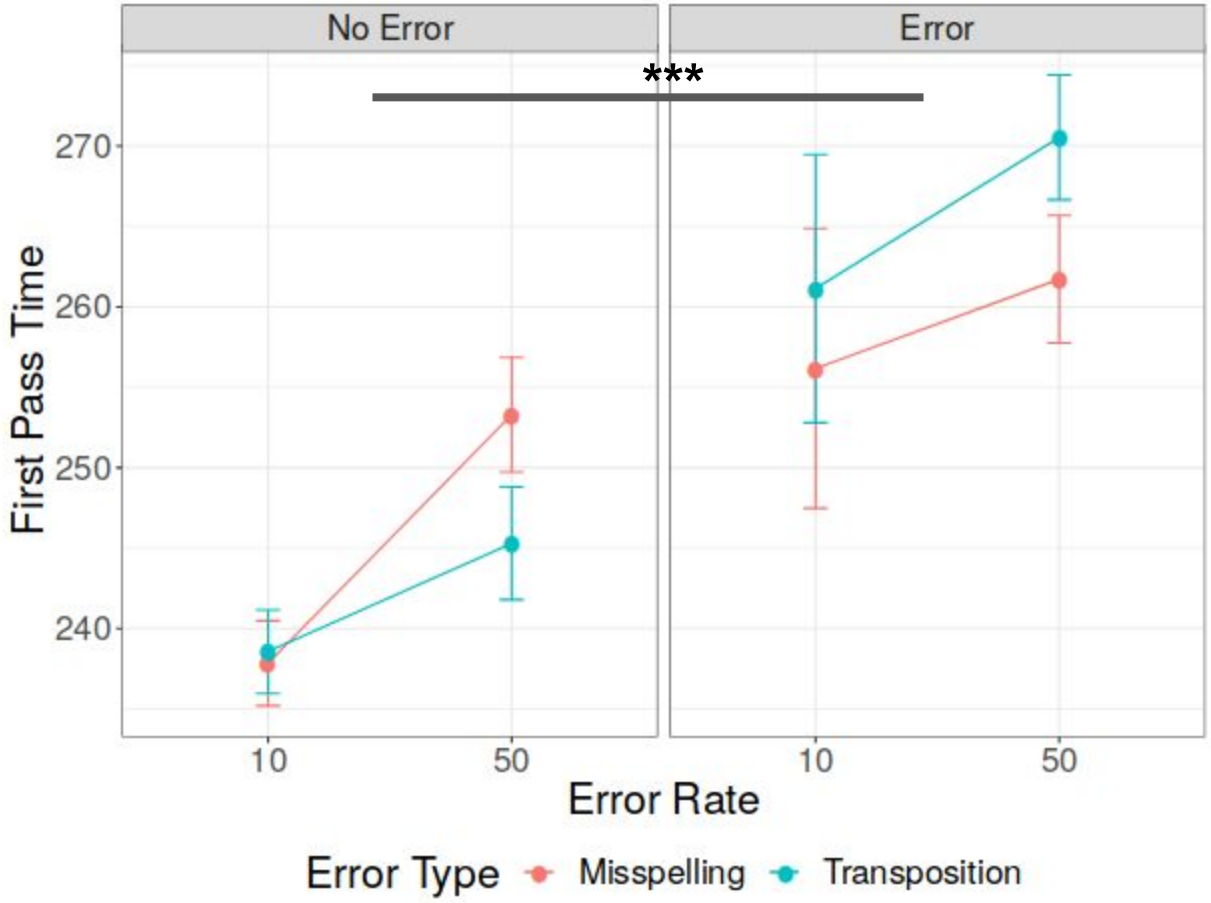

No Error
Error
First Pass Time
270
260
250
240
10
50
10
50
Error Rate
Error Type
Misspelling
Transposition


No Error
Error
First Pass Time
270
260
250
240
10
50
10
50
Error Rate
Error Type
Misspelling
Transposition

Erroneous words

| | First Pass | | Fixation Rate | |
|---|---|---|---|---|
| (Intercept) | 248.41 | (6.34)*** | −0.16 | (0.12) |
| ErrType | 1.41 | (1.32) | 0.08 | (0.02)*** |
| ErrRate | 7.20 | (1.60)*** | 0.16 | (0.02)*** |
| Error | 23.77 | (4.12)*** | 0.21 | (0.07)*** |
| WLength | 22.18 | (2.02)*** | 0.83 | (0.04)*** |
| LastFix | 3.10 | (4.18) | 0.22 | (0.18) |
| ErrRate × LastFix | 6.71 | (2.77)* | 0.16 | (0.04)*** |
| Error × LastFix | | — | 0.26 | (0.10)** |
| WLength × LastFix | | — | 0.74 | (0.10)*** |

$Pr(\beta < 0)$: *** < 0.001, ** < 0.01, * < 0.05


No Error
Error
***
First Pass Time
270
260
250
240
10
50
10
50
Error Rate
Error Type
Misspelling
Transposition

**Erroneous words** more **likely to be read** when **preview available**

| | First Pass | Fixation Rate |
|---|---|---|
| (Intercept) | 248.41 (6.34)*** | −0.16 (0.12) |
| ErrType | 1.41 (1.32) | 0.08 (0.02)*** |
| ErrRate | 7.20 (1.60)*** | 0.16 (0.02)*** |
| Error | 23.77 (4.12)*** | 0.21 (0.07)*** |
| WLength | 22.18 (2.02)*** | 0.83 (0.04)*** |
| LastFix | 3.10 (4.18) | 0.22 (0.18) |
| ErrRate × LastFix | 6.71 (2.77)* | 0.16 (0.04)*** |
| Error × LastFix | — | 0.26 (0.10)** |
| WLength × LastFix | — | 0.74 (0.10)*** |

$Pr(\beta < 0)$: *** < 0.001, ** < 0.01, * < 0.05

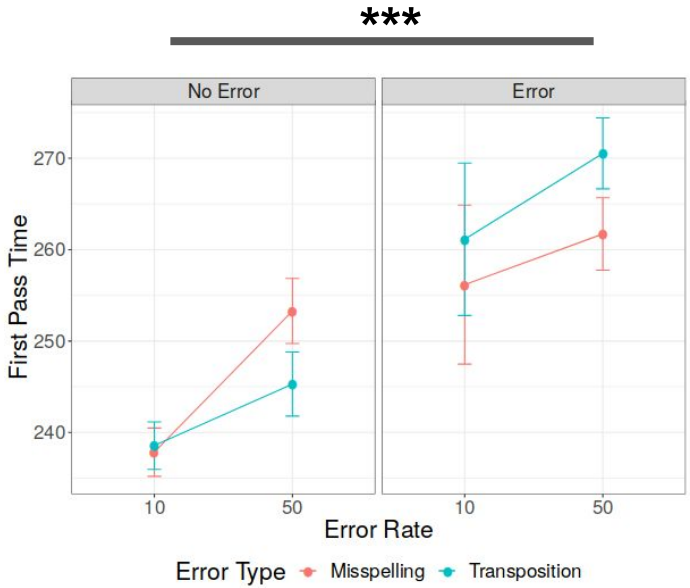
**Preview** seems to **increase effects** (for Fixations)

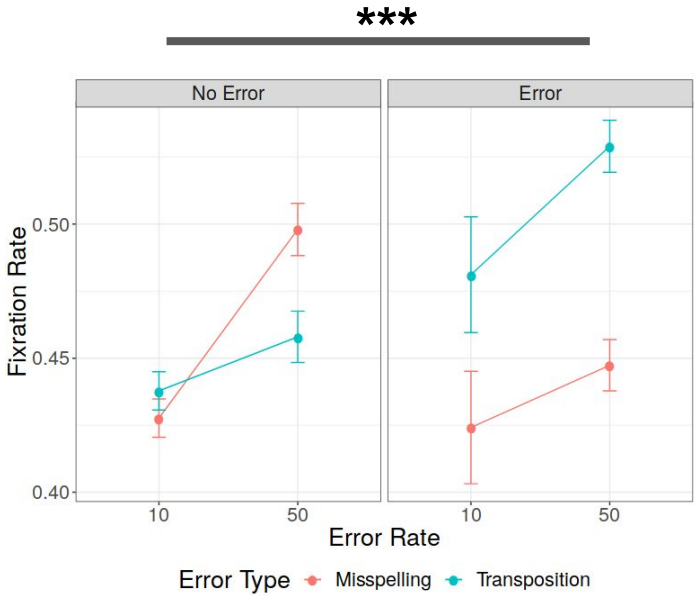
| | First Pass | | Fixation Rate | |
|---|---|---|---|---|
| (Intercept) | 248.41 | (6.34)*** | −0.16 | (0.12) |
| ErrType | 1.41 | (1.32) | 0.08 | (0.02)*** |
| ErrRate | 7.20 | (1.60)*** | 0.16 | (0.02)*** |
| Error | 23.77 | (4.12)*** | 0.21 | (0.07)*** |
| WLength | 22.18 | (2.02)*** | 0.83 | (0.04)*** |
| LastFix | 3.10 | (4.18) | 0.22 | (0.18) |
| ErrRate × LastFix | 6.71 | (2.77)* | 0.16 | (0.04)*** |
| Error × LastFix | | — | 0.26 | (0.10)** |
| WLength × LastFix | | — | 0.74 | (0.10)*** |

$Pr(\beta < 0)$: *** < 0.001, ** < 0.01, * < 0.05

# Experimental Results

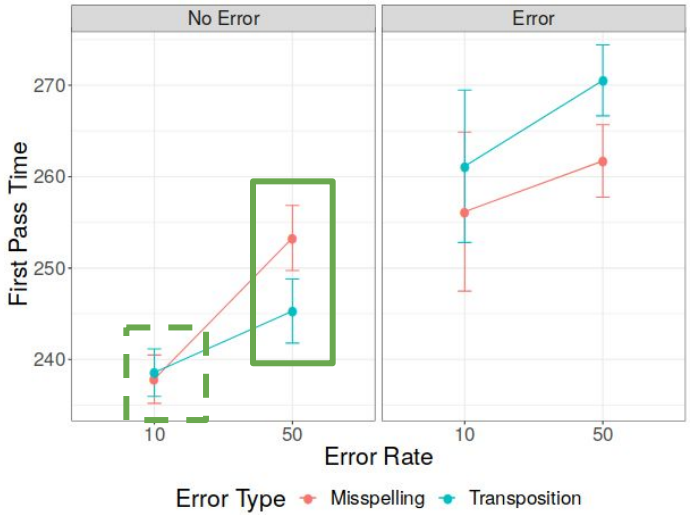
1. Erroneous words read longer & more likely to be fixated

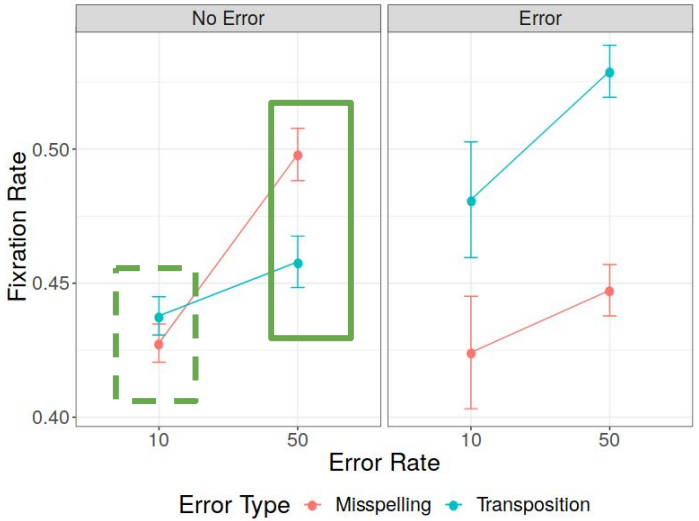


# Experimental Results

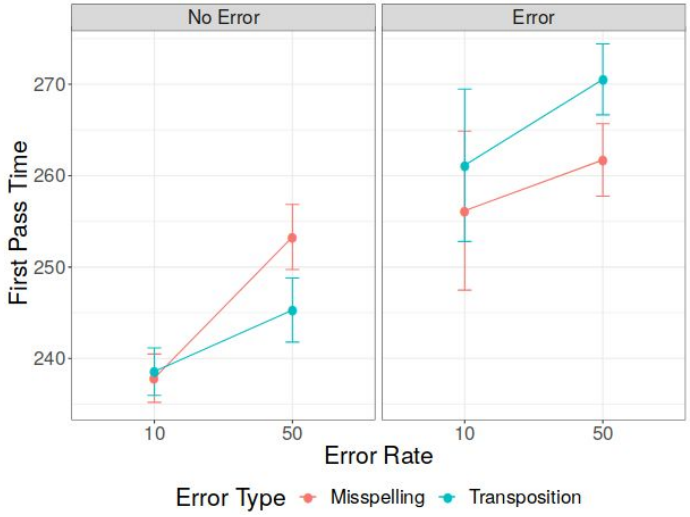
1. Erroneous words read longer & more likely to be fixated
2. High error rate ⇒ increased reading times & fixations, *even on correct words*

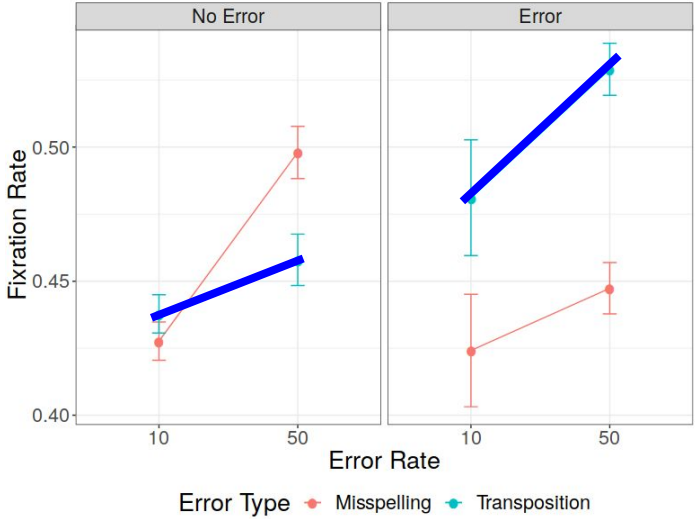


# Experimental Results

1. Erroneous words read longer & more likely to be fixated
2. High error rate ⇒ increased reading times & fixations, *even on correct words*
3. Transpositions increase fixation rate compared to misspellings



# Experimental Results

1. Erroneous words read longer & more likely to be fixated
2. High error rate ⇒ increased reading times & fixations, *even on correct words*
3. Transpositions increase fixation rate compared to misspellings
4. Whether the previous word is fixated or not modulates effect of error and error rate

# Surprisal Model

Most models of reading do not explicitly deal with errors.

Models using lexicon for word lookup cannot deal with errors without further assumptions.

# Surprisal Model

Most models of reading do not explicitly deal with errors

Models using lexicon for word lookup cannot deal with errors without further assumptions

**Example:** Surprisal model of processing difficulty (Hale, 2003; Levy, 2008)

- forced to treat all error words as out of vocabulary items
- cannot distinguish between error types

# Surprisal Model

Most models of reading do not explicitly deal with errors

Models using lexicon for word lookup cannot deal with errors without further assumptions

Idea: We need more fine-grained surprisal, computing expectations in terms of characters, not words:

- **inocent** more surprising than **innocent**,
- but not as surprising as completely unfamiliar string

# Character-Based Surprisal Model

Character-based neural language model (LSTM, Hochreiter & Schmidhuber, 1997)

- assigns probabilities to any sequence of characters
- ⇒ can compute surprisal even for words never seen in training data

# Character-Based Surprisal Model

Character-based neural language model (LSTM, Hochreiter & Schmidhuber, 1997)

- assigns probabilities to any sequence of characters
- ⇒ can compute surprisal even for words never seen in training data

Setup:

- trained on the DeepMind QA corpus
- create 7 models to control for random weight initialization
- use resulting model to compute surprisal on the 20 texts, in each condition

Using the Product Rule of Probability:

$$-\log P(x_t \ldots x_{t+T} | x_1 \ldots x_{t-1}) = \sum_{i=t}^{t+T} -\log P(x_i | x_1 \ldots x_{i-1})$$

Surprisal of a Word = Sum of Character Surprisals

-log P(innocent |they are) =

-log P(innocent |they are) = - log P(i|they are )

$$-\log P(\text{innocent} \mid \text{they are}) = -\log P(\text{i} \mid \text{they are}) - \log P(\text{n} \mid \text{they are i})$$

$$-\log P(\text{innocent} \mid \text{they are}) = -\log P(\text{i} \mid \text{they are}) - \log P(\text{n} \mid \text{they are i}) - \log P(\text{n} \mid \text{they are in})$$

$$-\log P(\text{innocent} \mid \text{they are}) = -\log P(\text{i} \mid \text{they are }) - \log P(\text{n} \mid \text{they are i}) - \log P(\text{n} \mid \text{they are in}) \dots - \log P(\text{n} \mid \text{they are innoce}) - \log P(\text{t} \mid \text{they are innocen})$$

# Predictions

1. Transpositions more surprising than misspellings:

    e.g., innocetn contains the rare character sequence tn

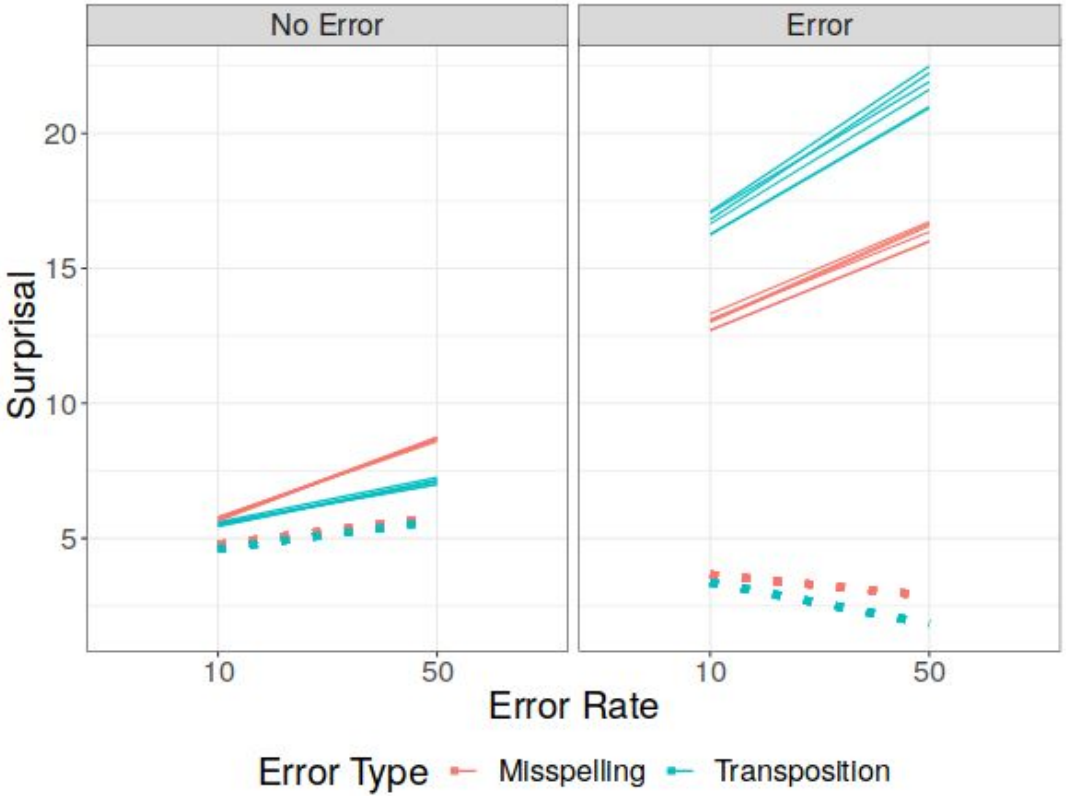
# Predictions

1. Transpositions more surprising than misspellings:

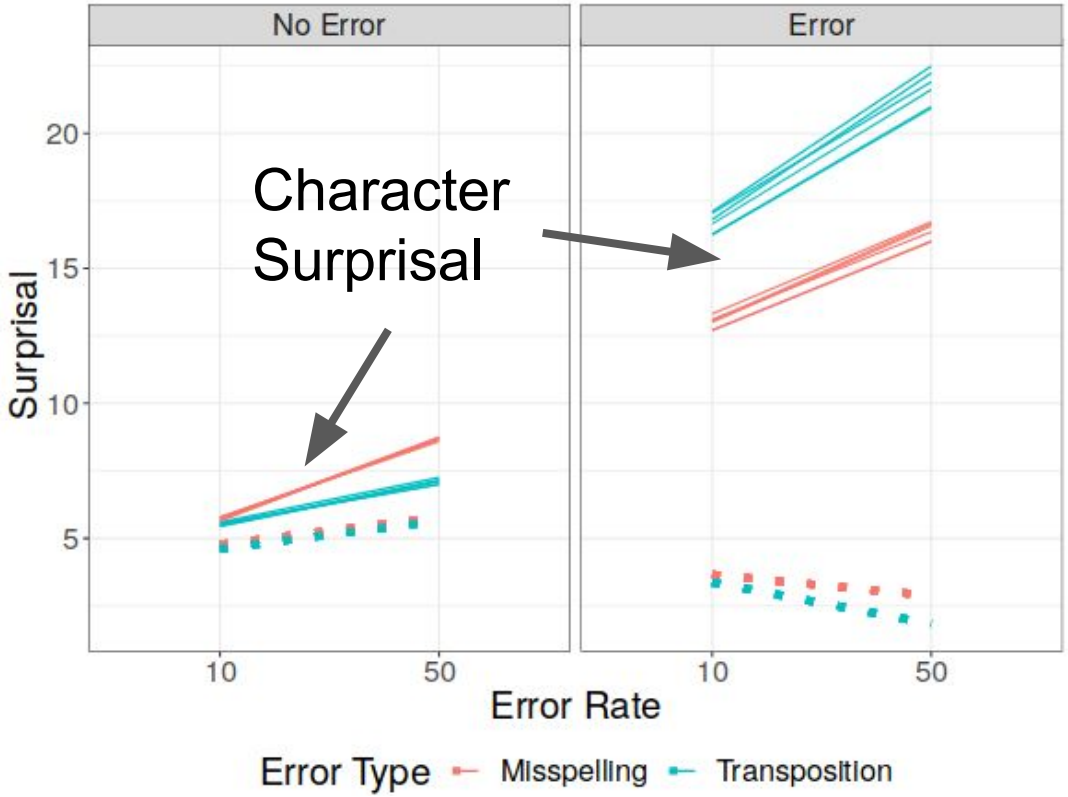
   e.g., innocetn contains the rare character sequence tn

2. High error rates degrade context ⇒ make all words harder to predict

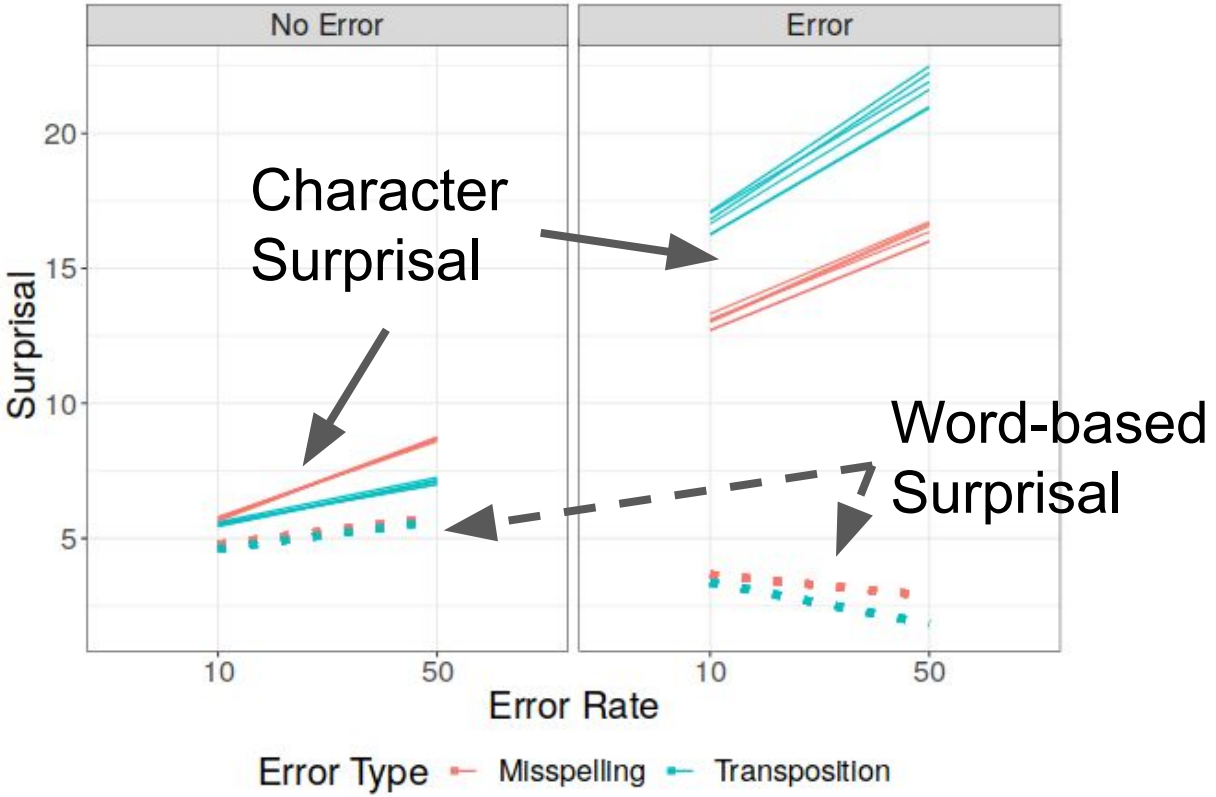
# Results


No Error
Error
Surprisal
20
15
10
5
10
50
10
50
Error Rate
Error Type
Misspelling
Transposition

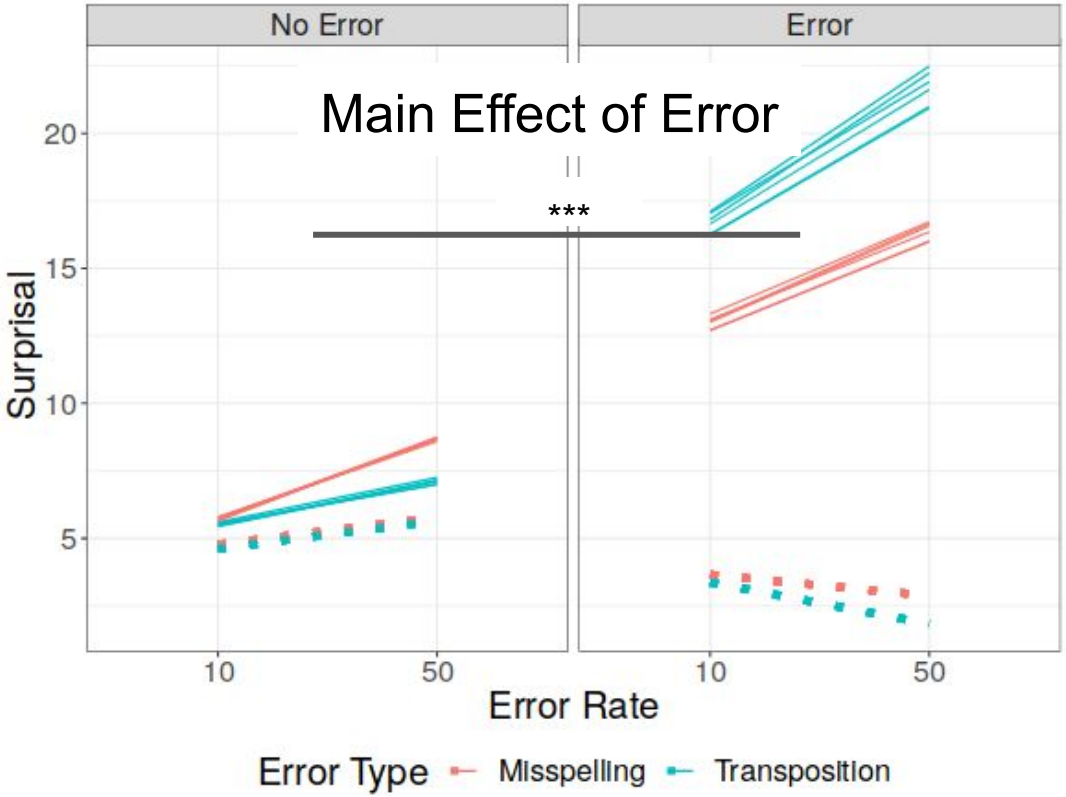
# Results


No Error
Error
Character Surprisal
Surprisal
20
15
10
5
10
50
10
50
Error Rate
Error Type
Misspelling
Transposition

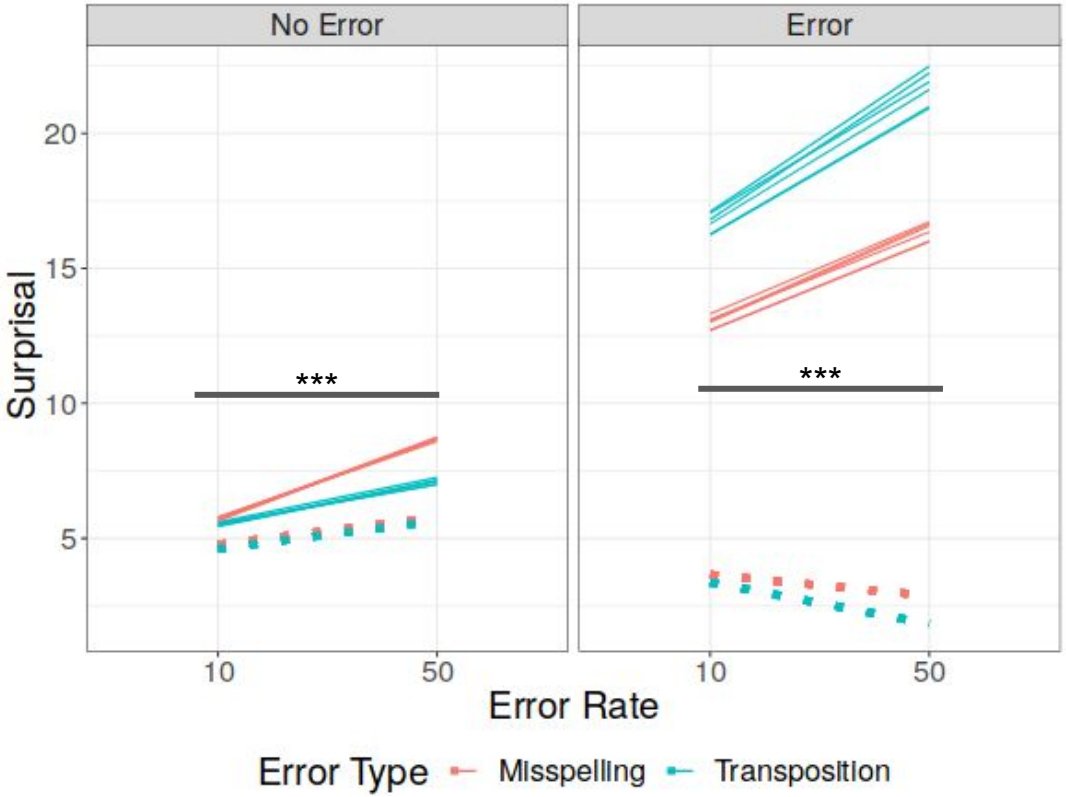
# Results


No Error
Error
Surprisal
20
15
10
5
Character
Surprisal
Word-based
Surprisal
10
50
10
50
Error Rate
Error Type
Misspelling
Transposition

# Results

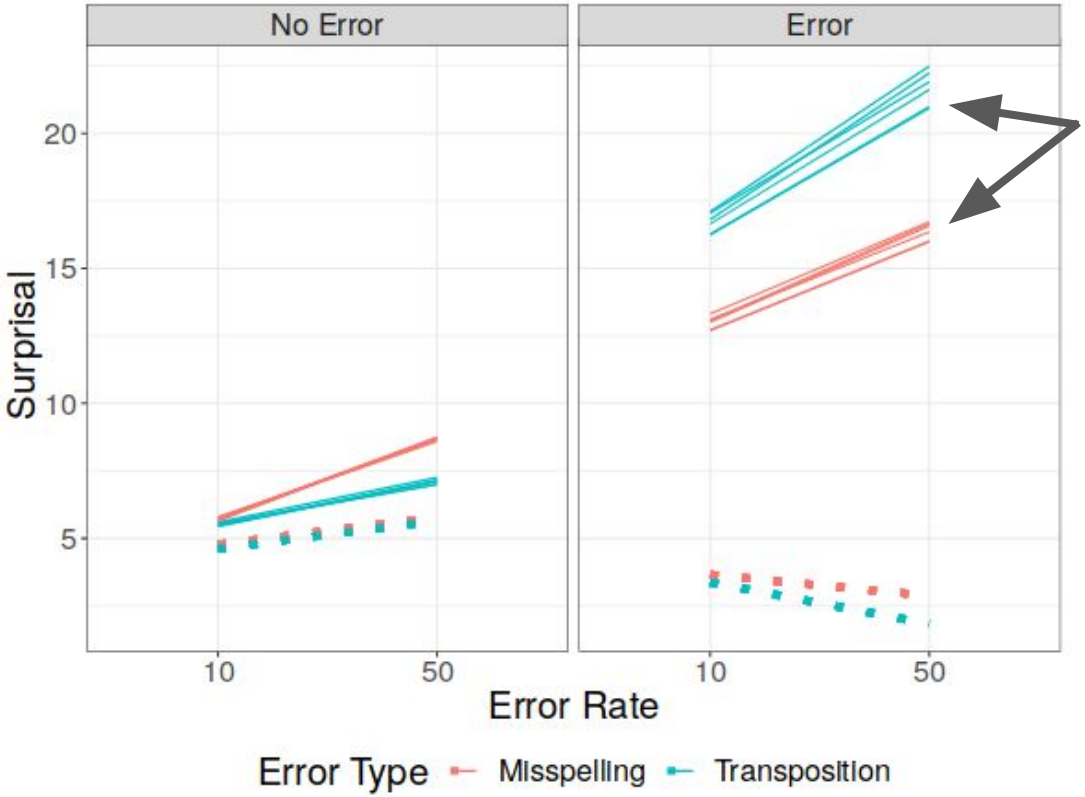

No Error
Error
Main Effect of Error
***
Surprisal
20
15
10
5
10
50
10
50
Error Rate
Error Type
Misspelling
Transposition

# Results



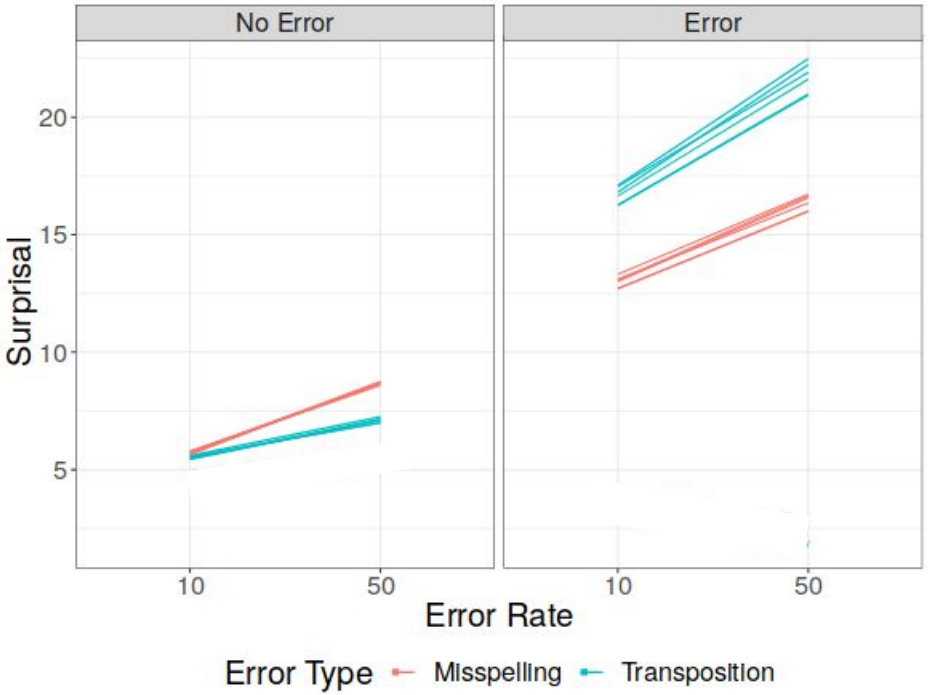

**Higher error rates** make all words more surprising

# Results

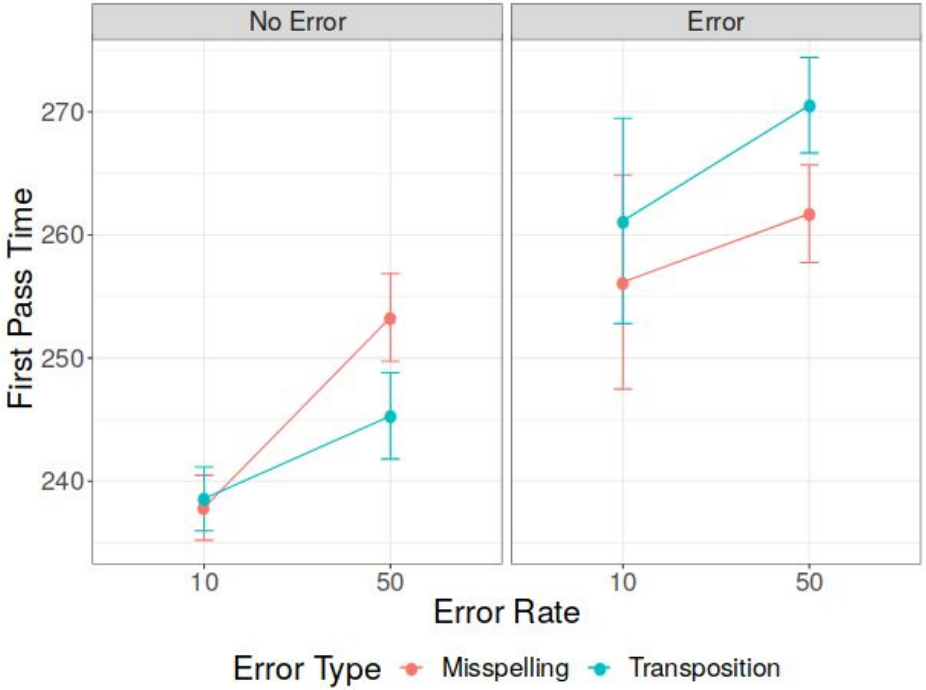



Transpositions cause higher surprisal than misspellings

## Surprisal Model




## First Pass Times



# Predicting Reading Measures

Baseline predictors unrelated to error manipulation

| | First Pass | Fixation Rate |
|---|---|---|
| (Intercept) | 248.73 (5.55)*** | −0.15 (0.09) |
| WLENGTH | 22.22 (0.79)*** | 0.75 (0.01)*** |
| LASTFIX | 2.65 (1.34) | 0.22 (0.02)*** |
| WLENGTH × LASTFIX | — | 0.60 (0.19)*** |
| RESIDCHARSURP-ORACLE | 9.89 (0.78)*** | 0.09 (0.01)*** |
| RESIDCHARSURP | 13.82 (0.66)*** | 0.14 (0.01)*** |
| ΔAIC | −273.88 | −205.83 |
| ΔBIC | −273.88 | −205.83 |

$Pr(\beta < 0)$: *** < 0.001, ** < 0.01, * < 0.05

# Predicting Reading Measures

| | First Pass | Fixation Rate |
|---|---|---|
| (Intercept) | 248.73 (5.55)*** | −0.15 (0.09) |
| WLENGTH | 22.22 (0.79)*** | 0.75 (0.01)*** |
| LASTFIX | 2.65 (1.34) | 0.22 (0.02)*** |
| WLENGTH × LASTFIX | — | 0.60 (0.19)*** |
| RESIDCHARSURP-ORACLE | 9.89 (0.78)*** | 0.09 (0.01)*** |
| RESIDCHARSURP | 13.82 (0.66)*** | 0.14 (0.01)*** |
| ΔAIC | −273.88 | −205.83 |
| ΔBIC | −273.88 | −205.83 |

$Pr(\beta < 0)$: *** < 0.001, ** < 0.01, * < 0.05

Character Surprisal

# Predicting Reading Measures

Baseline surprisal: using *corrected* words

| | First Pass | Fixation Rate |
|---|---|---|
| (Intercept) | 248.73 (5.55)*** | −0.15 (0.09) |
| WLENGTH | 22.22 (0.79)*** | 0.75 (0.01)*** |
| LASTFIX | 2.65 (1.34) | 0.22 (0.02)*** |
| WLENGTH × LASTFIX | — | 0.60 (0.19)*** |
| RESIDCHARSURP-ORACLE | 9.89 (0.78)*** | 0.09 (0.01)*** |
| RESIDCHARSURP | 13.82 (0.66)*** | 0.14 (0.01)*** |
| ΔAIC | −273.88 | −205.83 |
| ΔBIC | −273.88 | −205.83 |

$Pr(\beta < 0)$: *** < 0.001, ** < 0.01, * < 0.05

# Predicting Reading Measures

| | First Pass | Fixation Rate |
|---|---|---|
| (Intercept) | 248.73 (5.55)*** | −0.15 (0.09) |
| WLENGTH | 22.22 (0.79)*** | 0.75 (0.01)*** |
| LASTFIX | 2.65 (1.34) | 0.22 (0.02)*** |
| WLENGTH × LASTFIX | — | 0.60 (0.19)*** |
| RESIDCHARSURP-ORACLE | 9.89 (0.78)*** | 0.09 (0.01)*** |
| RESIDCHARSURP | 13.82 (0.66)*** | 0.14 (0.01)*** |
| ΔAIC | −273.88 | −205.83 |
| ΔBIC | −273.88 | −205.83 |

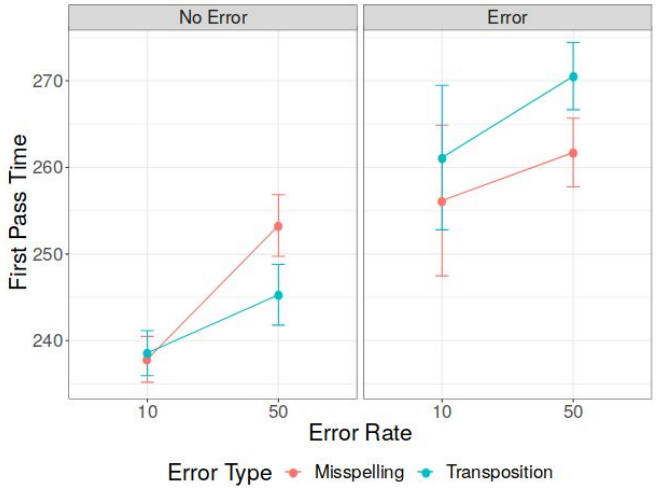
$Pr(\beta < 0)$: *** < 0.001, ** < 0.01, * < 0.05

Character Surprisal improves model fit

# Conclusion

1. Investigated reading in the face of errors (transpositions & misspellings)

- transpositions cause more reading difficulty than misspellings
- High error rate makes all words are harder to read, even the ones without errors

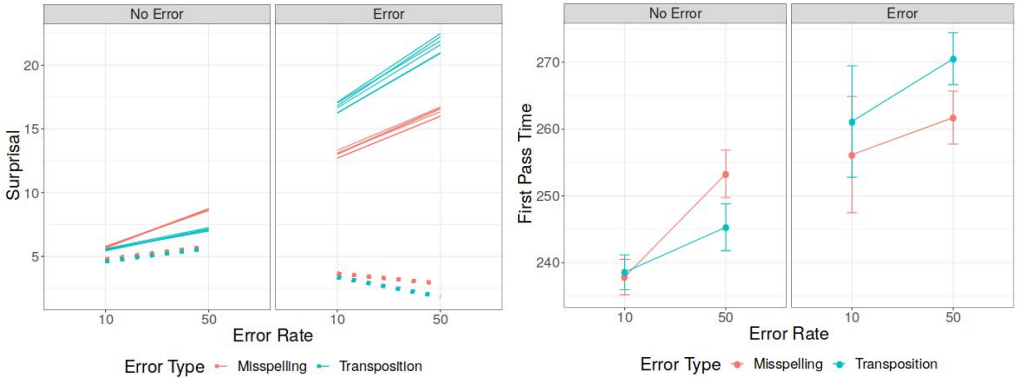


# Conclusion

1. Investigated reading in the face of errors (transpositions & misspellings)

- transpositions cause more reading difficulty than misspellings
- High error rate makes all words are harder to read, even the ones without errors

2. Character-based surprisal explains results.



# Conclusion

1. Investigated reading in the face of errors (transpositions & misspellings)

- transpositions cause more reading difficulty than misspellings
- High error rate makes all words are harder to read, even the ones without errors

2. Character-based surprisal explains results.

3. Future work: Integrate character-based surprisal with existing neural models of human reading (Hahn & Keller, 2018), to model effects of landing position, preview, ....

Thanks!