# A-maze: Easier measurement of incremental processing

Veronica Boyce

3 June 2020

# Why measure reading time?

# Why measure reading time?

Linguistic and psycholinguistic theories make predictions about processing difficulty.

# Why measure reading time?

Linguistic and psycholinguistic theories make predictions about processing difficulty.

Examples of increased difficulty

# Why measure reading time?

Linguistic and psycholinguistic theories make predictions about processing difficulty.

Examples of increased difficulty

- Constructions not in the grammar

# Why measure reading time?

Linguistic and psycholinguistic theories make predictions about processing difficulty.

Examples of increased difficulty

- Constructions not in the grammar
- Lexical items that are harder to retrieve

# Why measure reading time?

Linguistic and psycholinguistic theories make predictions about processing difficulty.

Examples of increased difficulty

- Constructions not in the grammar
- Lexical items that are harder to retrieve
- Words that force reparsing or reanalysis

# Why measure reading time?

Linguistic and psycholinguistic theories make predictions about processing difficulty.

Examples of increased difficulty

- Constructions not in the grammar
- Lexical items that are harder to retrieve
- Words that force reparsing or reanalysis

We assume that harder processing manifests in longer reading/reaction time (RT).

# Why measure reading time?

Linguistic and psycholinguistic theories make predictions about processing difficulty.

Examples of increased difficulty

- Constructions not in the grammar
- Lexical items that are harder to retrieve
- Words that force reparsing or reanalysis

We assume that harder processing manifests in longer reading/reaction time (RT).

RT patterns may be phenomena that theories need to explain.

# Two common methods

# Two common methods

## Eye-tracking

# Two common methods

## Eye-tracking

- Expensive
- Hard to analyse

# Two common methods

## Eye-tracking

- Expensive
- Hard to analyse

## Self-paced reading

The- - - - - - - - - - -

# Two common methods

## Eye-tracking

- Expensive
- Hard to analyse

## Self-paced reading

- - - cat - - - - - - - -

# Two common methods

## Eye-tracking

- Expensive
- Hard to analyse

## Self-paced reading

- - - - - - drank - - - -

# Two common methods

## Eye-tracking

- Expensive
- Hard to analyse

## Self-paced reading

- - - - - - - - - - -milk.

# Two common methods

## Eye-tracking

- Expensive
- Hard to analyse

## Self-paced reading

- - - - - - - - - - -milk.

- Lots of spillover
- Messy data

# A third option: Maze

# Maze Task

The x-x-x

e i

# Maze Task

The x-x-x

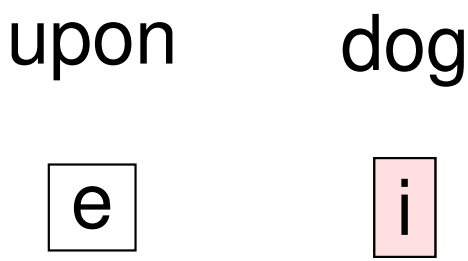
e i

# Maze Task

upon dog

e i

# Maze Task

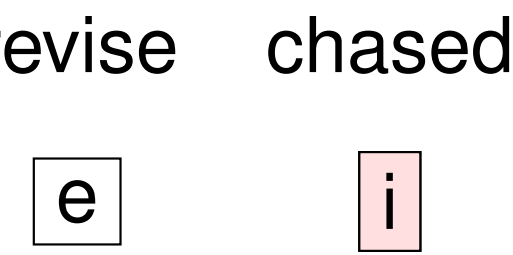


# Maze Task

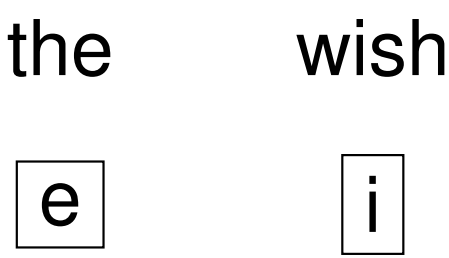
revise chased

e i

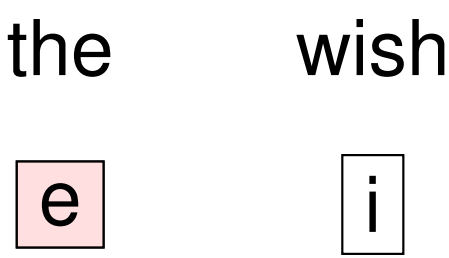
# Maze Task



# Maze Task



# Maze Task



# Maze Task

mitigate. squirrel.

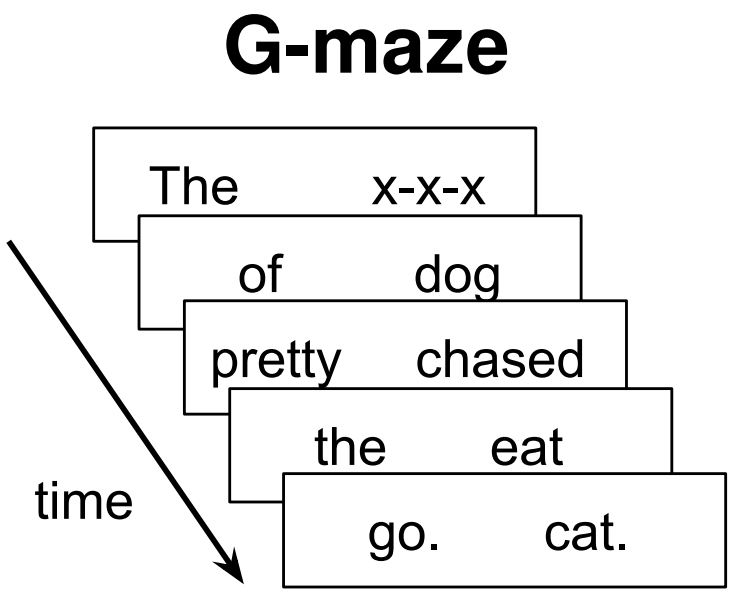
e i

# Maze Task

mitigate. squirrel.

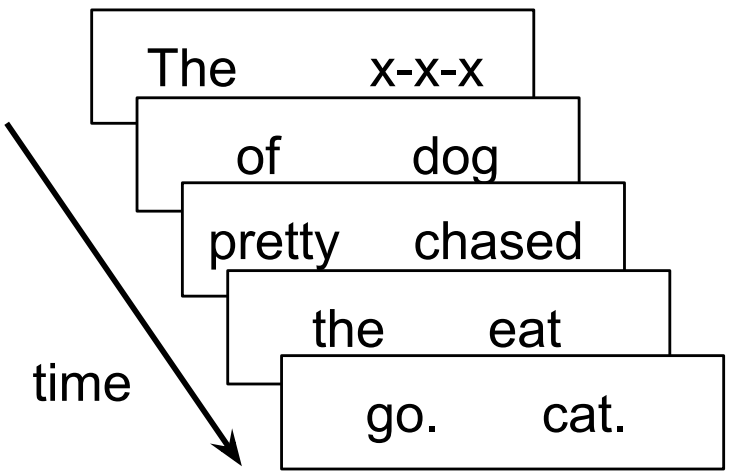
e i

# A third option: Maze

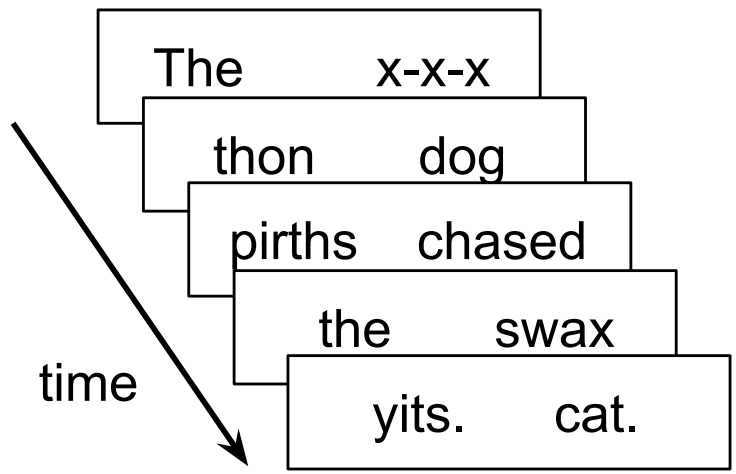


# A third option: Maze

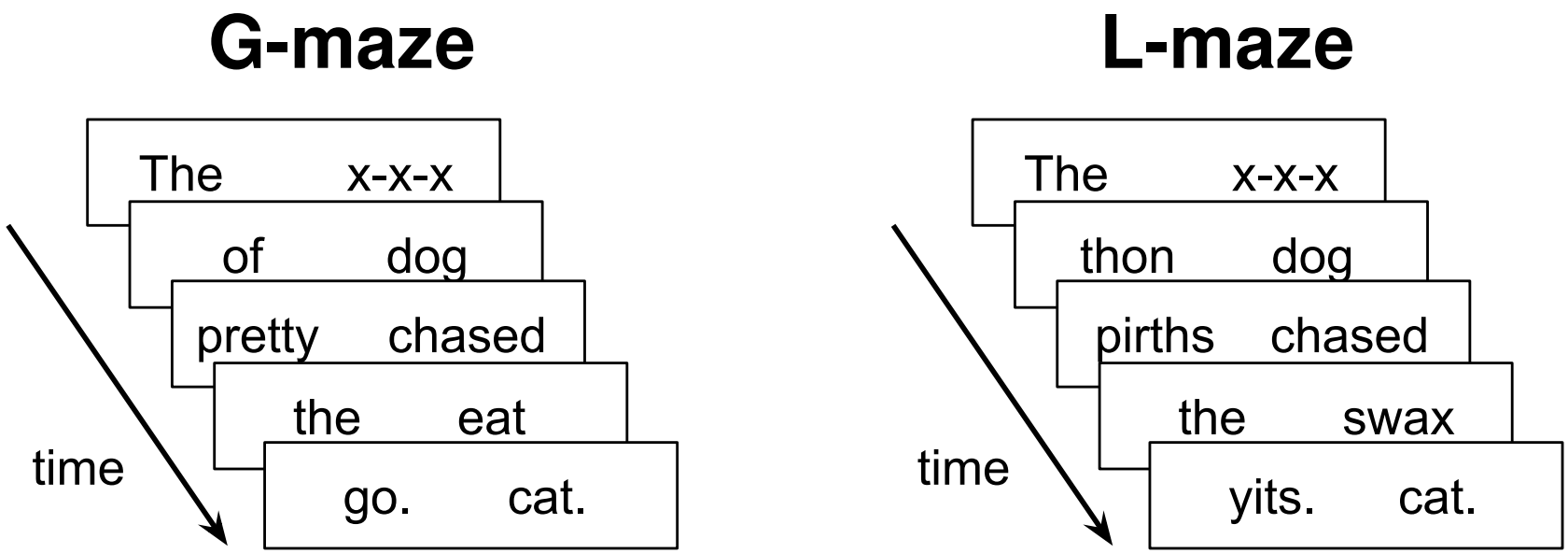
## G-maze




## L-maze



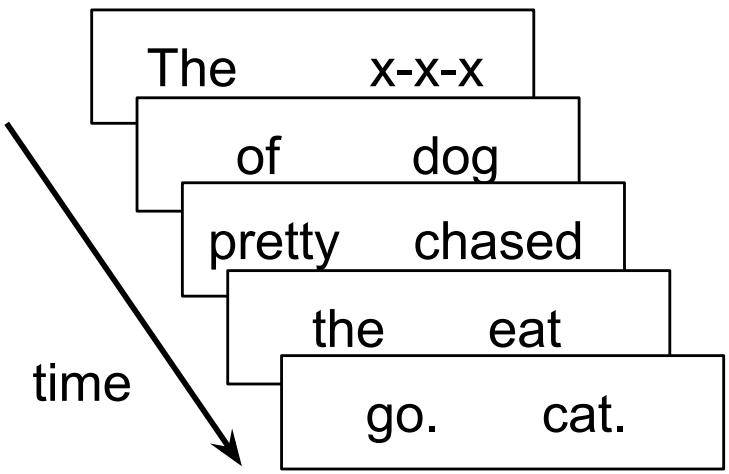
# A third option: Maze



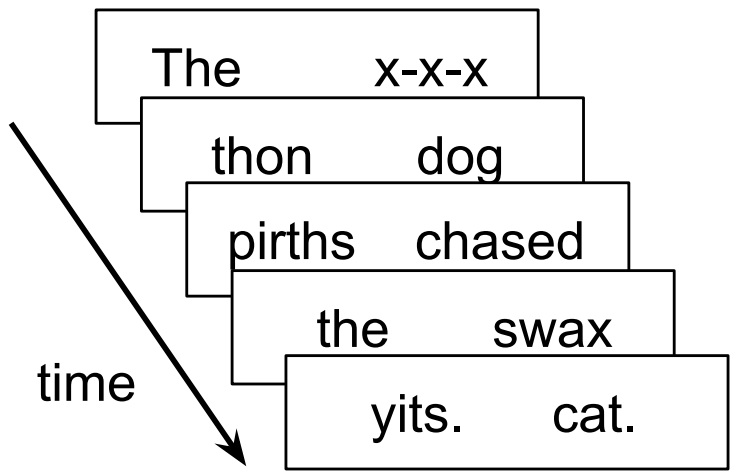

Sentence ends when a mistake is made.

# A third option: Maze

## G-maze




## L-maze




Sentence ends when a mistake is made.
Central claim: forces extremely incremental processing (no spillover)

(Forster et al. 2009; Witzel et al. 2012)

# Web Implementation

Maze well suited to run on the web

- Implement in Ibex

# Web Implementation

Maze well suited to run on the web

- Implement in Ibex

Test by replicating Witzel et al. (2012)

- Witzel et al (2012): Comparison of eye-tracking, SPR, L-maze, G-maze (all in-lab)
- Got materials and data from Witzel
- We run SPR, L-maze, and G-maze on MTurk

# Materials

**Relative Clause**

*Low:* The son of the lady who politely introduced herself was popular at the party.

*High:* The son of the lady who politely introduced himself was popular at the party.

**Adverb Clause**

*Low:* James will fix the car he drove yesterday, but he will need some help.

*High:* James will fix the car he drove tomorrow, but he will need some help.

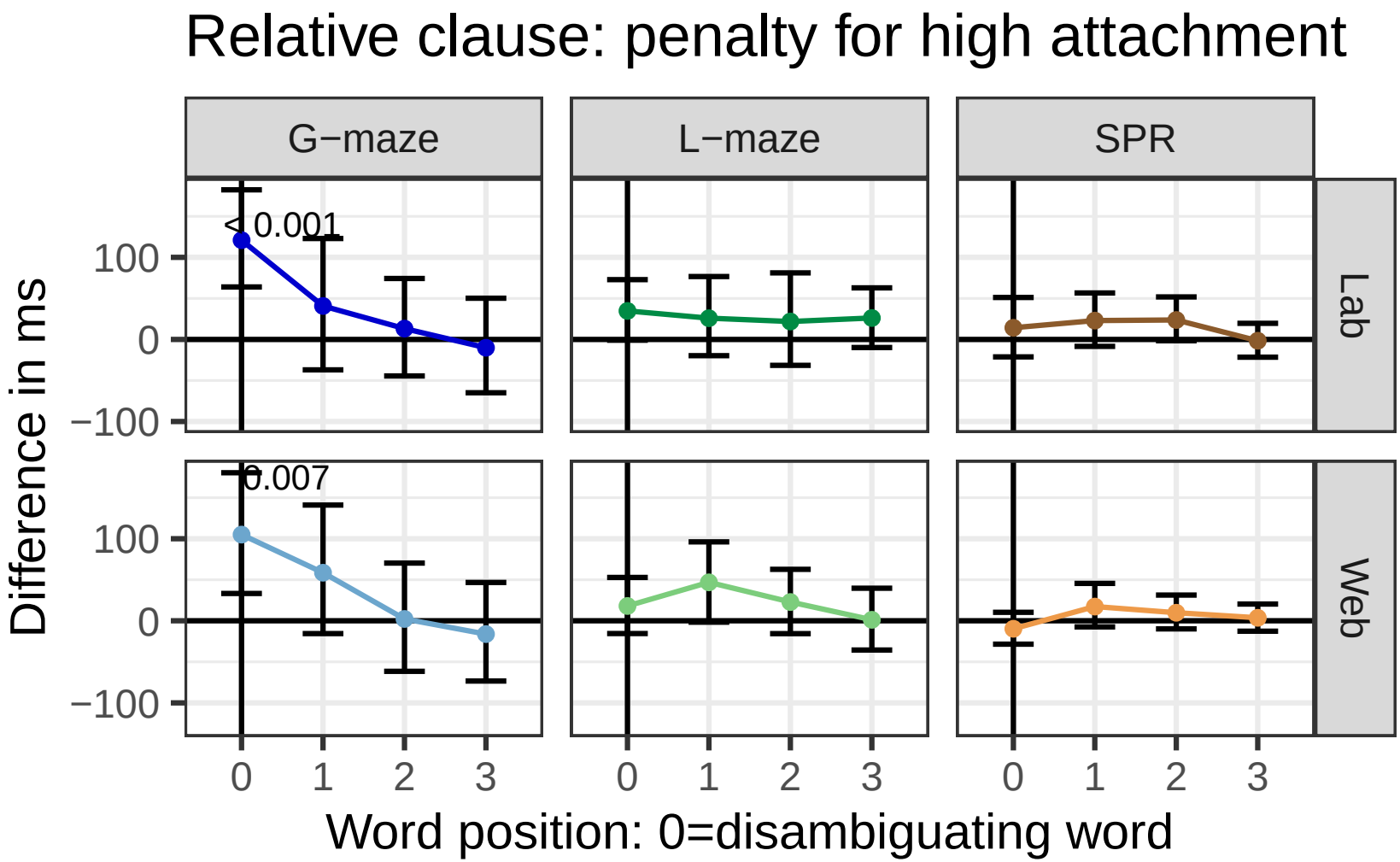
**Sentence v Noun Phrase conjunction**

*Comma:* The swimmer disappointed her coach, and her mother tried to console her.

*No comma:* The swimmer disappointed her coach and her mother tried to console her.

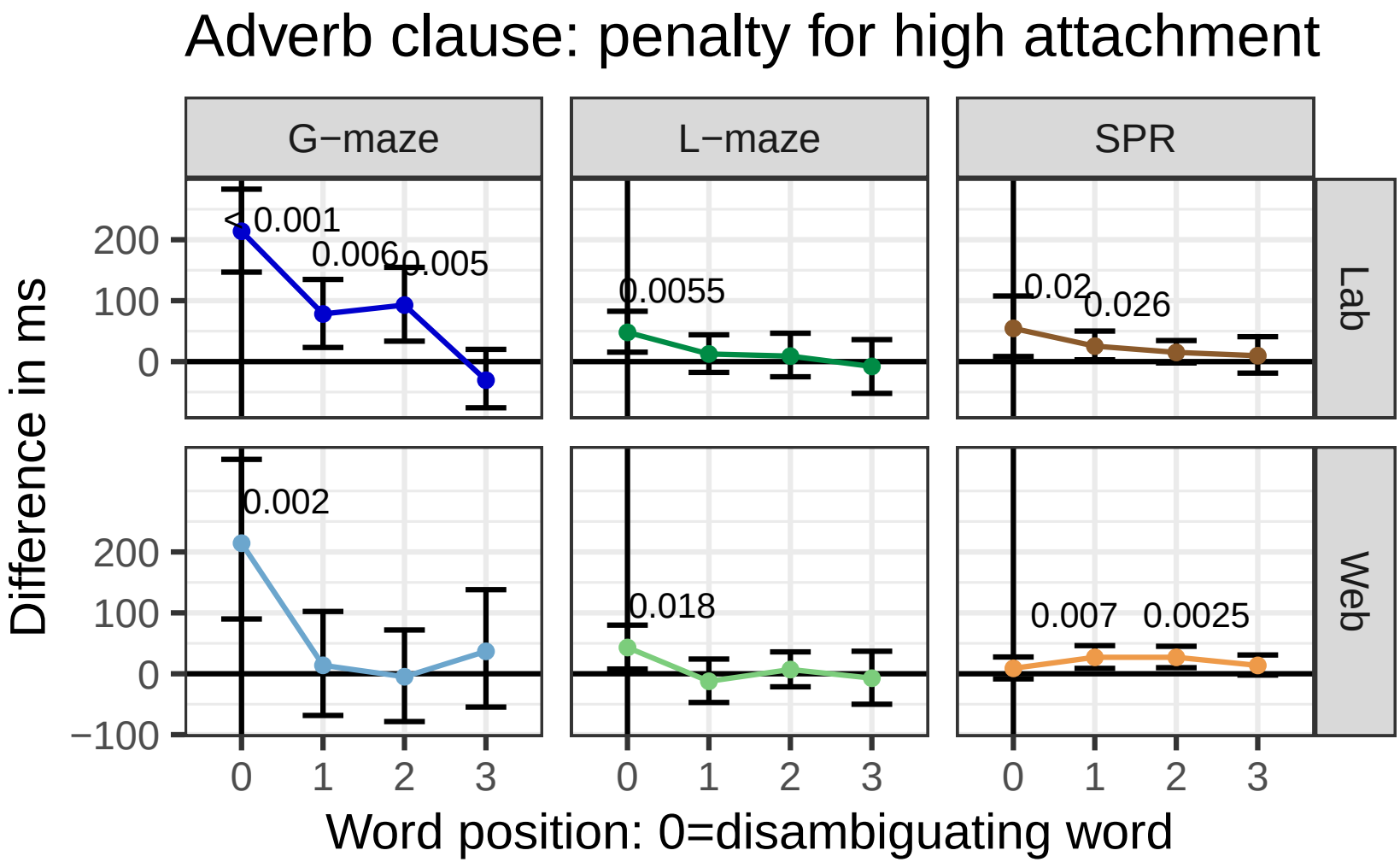
# Results

The son of the lady who politely introduced herself / himself was popular at the party.



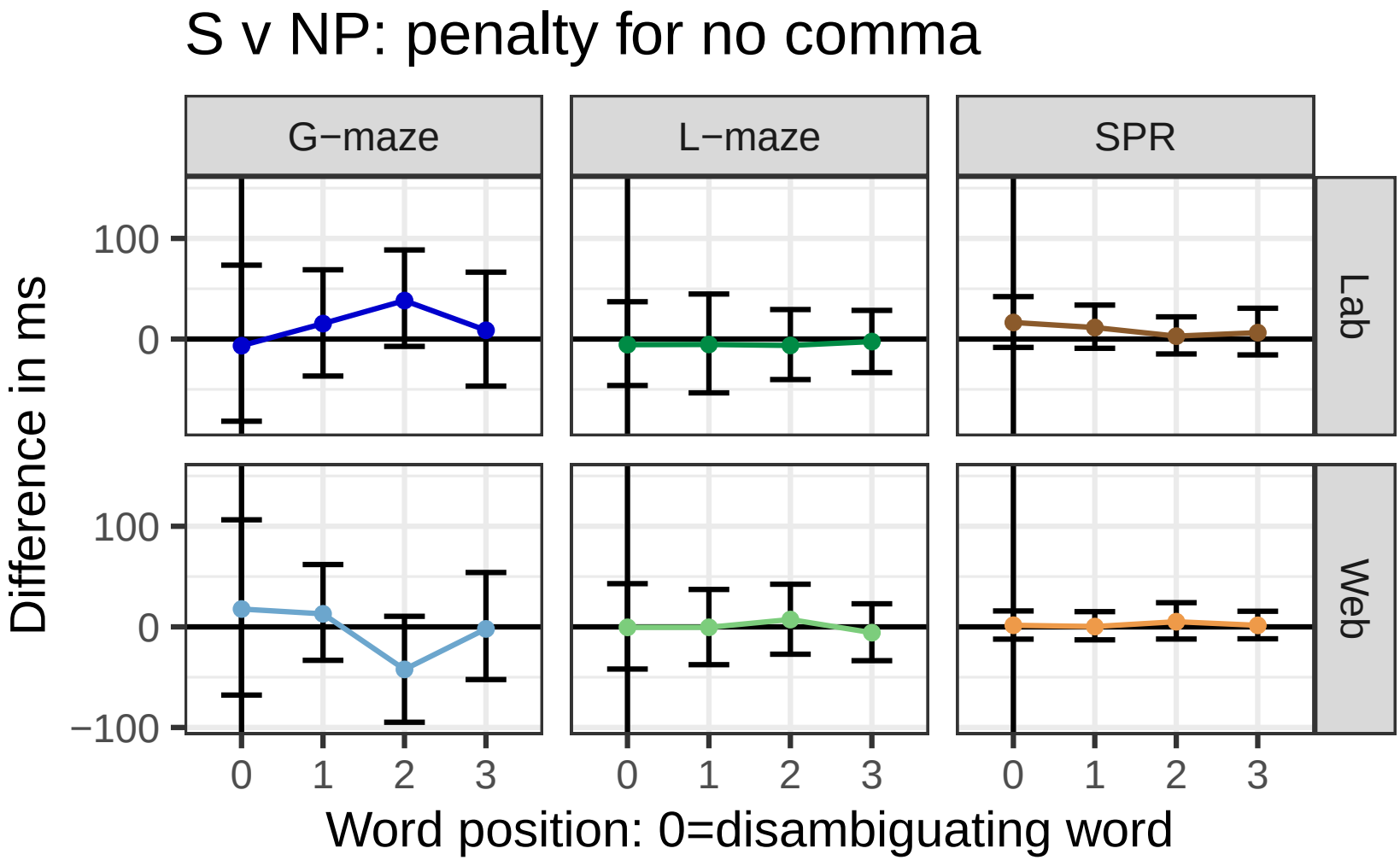
# Results

James will fix the car he drove yesterday / tomorrow, but he will need some help.



# Results

The swimmer disappointed her coach, and her mother tried / tried to console her.



# Interim Summary

**Good news:** G-maze works well over the web (better than L-maze or SPR)

# Interim Summary

**Good news:** G-maze works well over the web (better than L-maze or SPR)

**Bad news:** I do not want to write G-maze materials.

# Meanwhile in Natural Language Processing

# Meanwhile in Natural Language Processing

Language models (LMs)

- Trained on large corpora to predict the next word
- Given a partial sentence, return probabilities of the next word

# Meanwhile in Natural Language Processing

Language models (LMs)

- Trained on large corpora to predict the next word
- Given a partial sentence, return probabilities of the next word

Surprisal: negative log probability

- 2 bits of surprisal = 1/4
- 10 bits of surprisal $\approx$ 1/1000
- +1 surprisal = half as likely

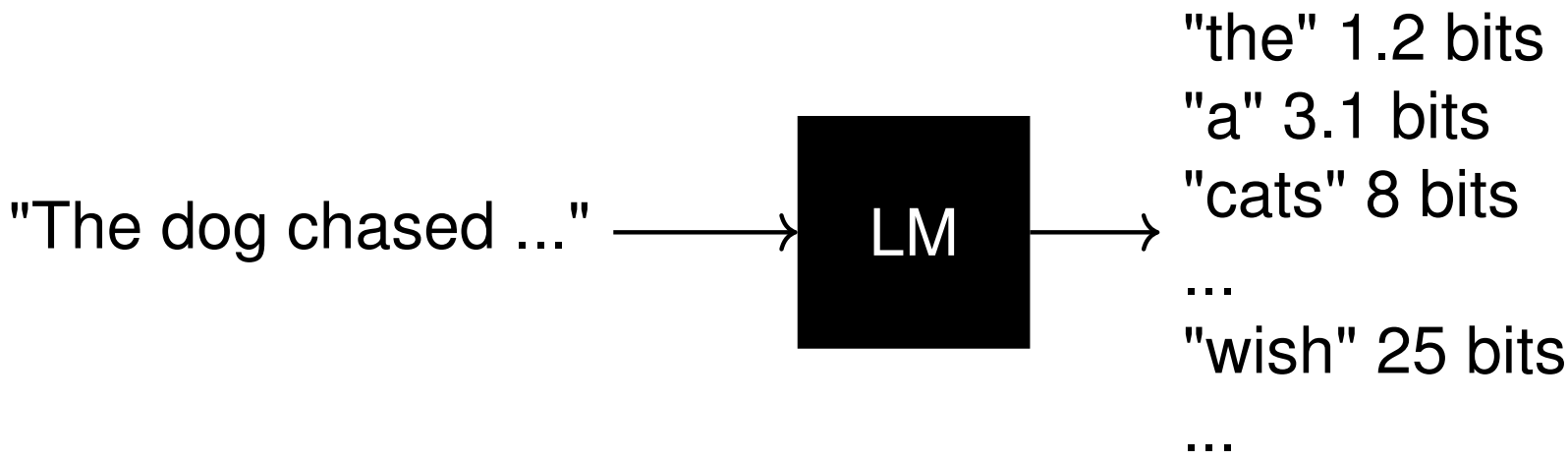
# Meanwhile in Natural Language Processing

Language models (LMs)

- Trained on large corpora to predict the next word
- Given a partial sentence, return probabilities of the next word

Surprisal: negative log probability

- 2 bits of surprisal = 1/4
- 10 bits of surprisal $\approx$ 1/1000
- +1 surprisal = half as likely



# Can we use LMs to choose distractors?

Use high surprisal according to LM as a proxy for bad in context

# Can we use LMs to choose distractors?

Use high surprisal according to LM as a proxy for bad in context

- Model the target sentence word by word

# Can we use LMs to choose distractors?

Use high surprisal according to LM as a proxy for bad in context

- Model the target sentence word by word
- At each position, choose a high surprisal word

# Can we use LMs to choose distractors?

Use high surprisal according to LM as a proxy for bad in context

- Model the target sentence word by word
- At each position, choose a high surprisal word

Want quality control on distractors

# Can we use LMs to choose distractors?

Use high surprisal according to LM as a proxy for bad in context

- Model the target sentence word by word
- At each position, choose a high surprisal word

Want quality control on distractors

- Restrict to a list of possible distractors

# Can we use LMs to choose distractors?

Use high surprisal according to LM as a proxy for bad in context

- Model the target sentence word by word
- At each position, choose a high surprisal word

Want quality control on distractors

- Restrict to a list of possible distractors
- Only consider distractors of same length, frequency as target word

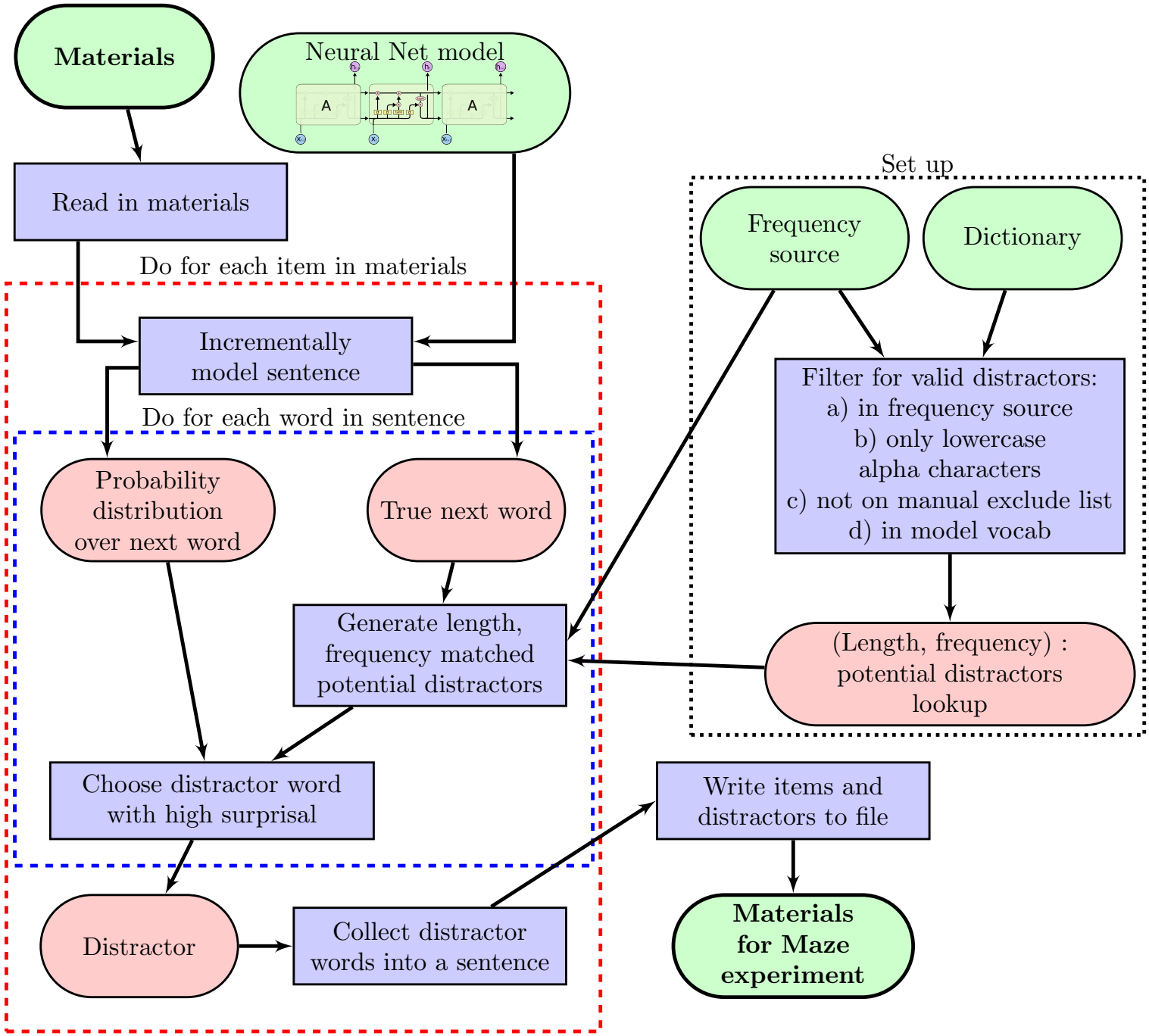
# Can we use LMs to choose distractors?

Use high surprisal according to LM as a proxy for bad in context

- Model the target sentence word by word
- At each position, choose a high surprisal word

Want quality control on distractors

- Restrict to a list of possible distractors
- Only consider distractors of same length, frequency as target word
- Check distractors until we find one with high surprisal


Materials
Neural Net model
A
A
Read in materials
Set up
Frequency source
Dictionary
Do for each item in materials
Incrementally model sentence
Filter for valid distractors:
a) in frequency source
b) only lowercase alpha characters
c) not on manual exclude list
d) in model vocab
Do for each word in sentence
Probability distribution over next word
True next word
Generate length, frequency matched potential distractors
(Length, frequency) : potential distractors lookup
Choose distractor word with high surprisal
Write items and distractors to file
Distractor
Collect distractor words into a sentence
Materials for Maze experiment

# Does A-maze work?

# Does A-maze work?

Test it on materials from Witzel et al. (2012)

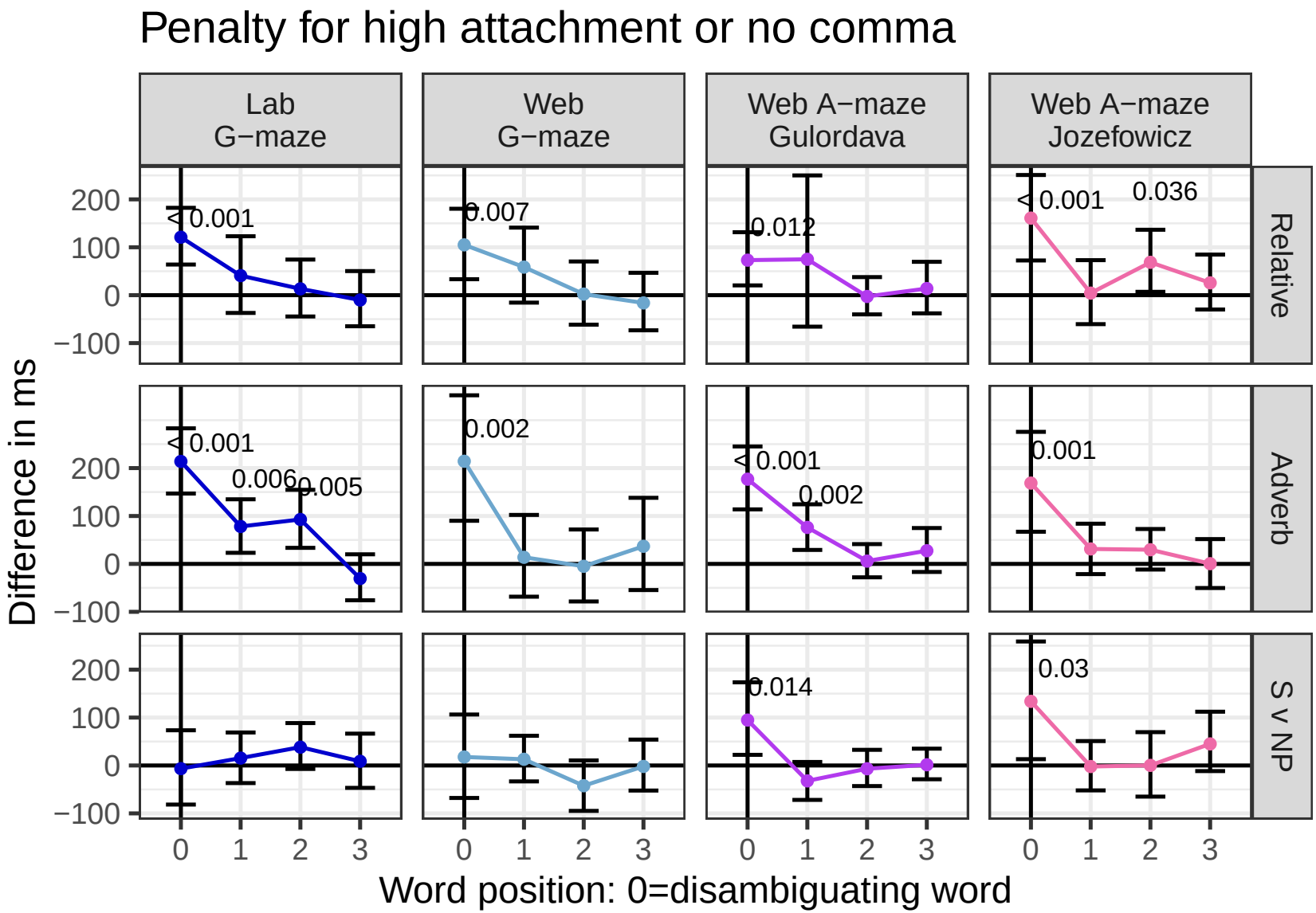
# Does A-maze work?

Test it on materials from Witzel et al. (2012)

Try with two pre-trained LSTM models (Gulordava 2018, Jozefowicz 2016)

# Does A-maze work? Yes.

# Does A-maze work? Yes.

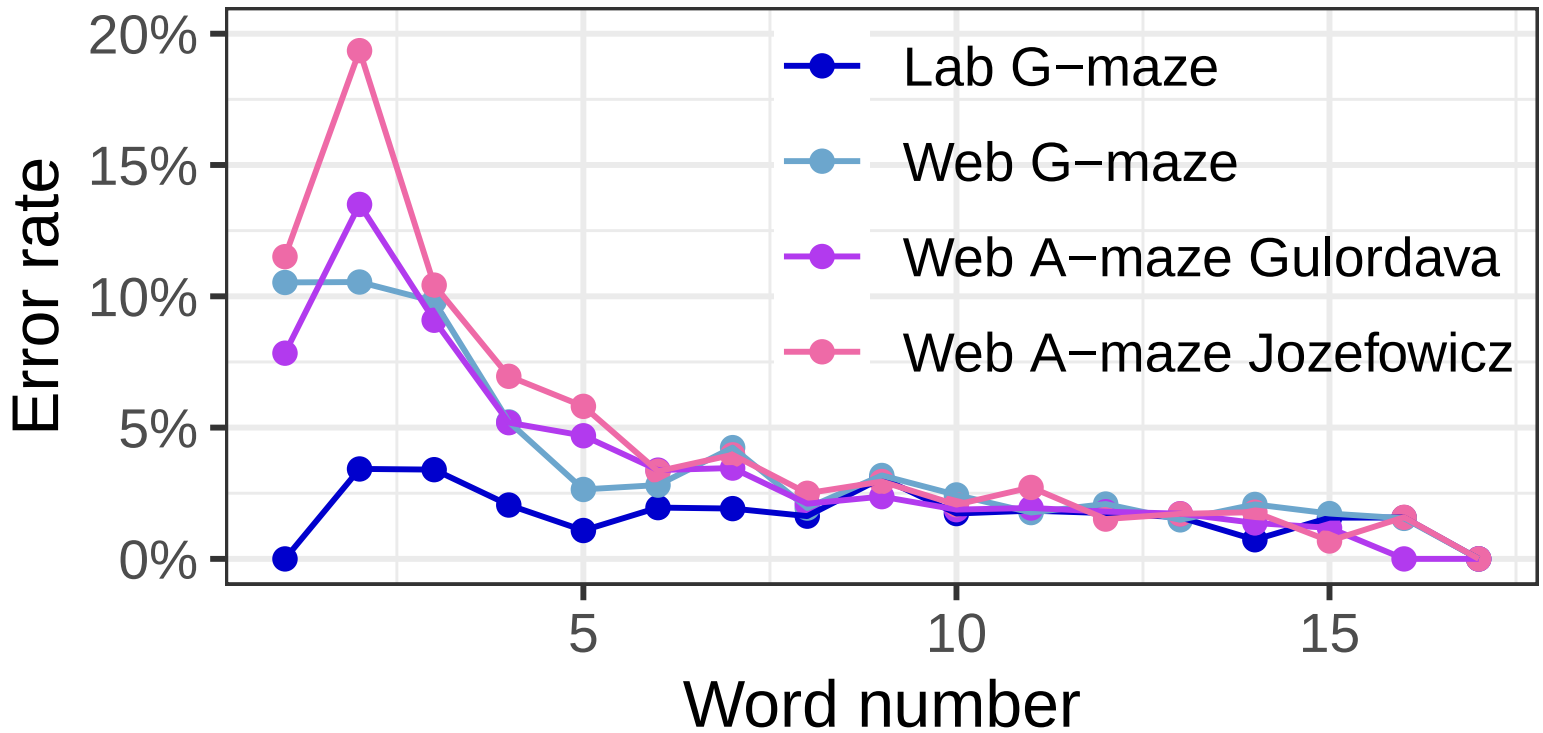


# Caveats

A lot of mistakes on the second word of the sentence ("The dog ...")

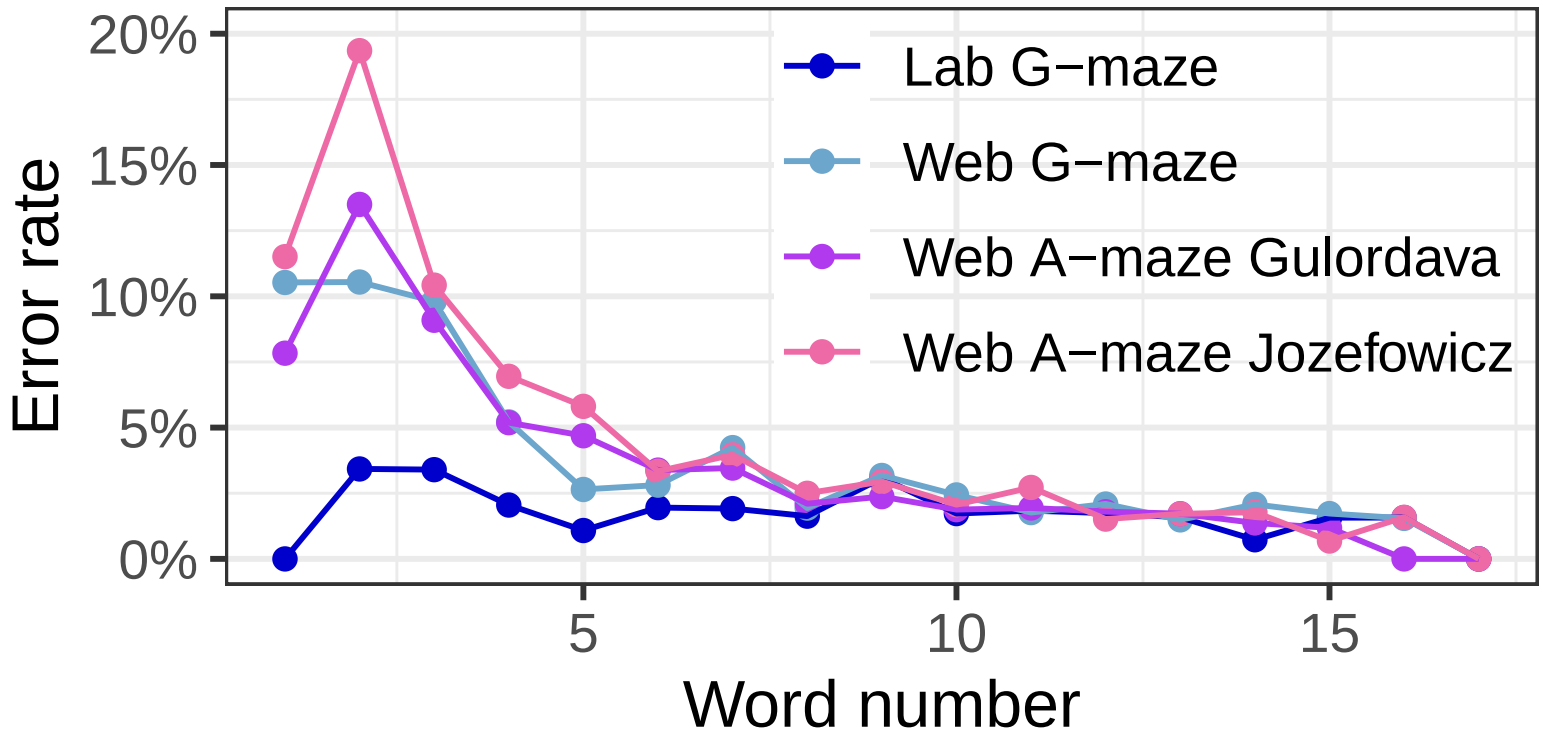
# Caveats

A lot of mistakes on the second word of the sentence ("The dog ...")



# Caveats

A lot of mistakes on the second word of the sentence ("The dog ...")




Bad participants or bad distractors?
Either way, lots of data loss.

# Caveats

## Definitely some bad distractors

| Prefix | Correct | Distractor | Error Rate |
|---|---|---|---|
| Gulordava | | | |
| The | niece | cooks | 44% |
| The swimmer | disappointed | propositions | 30% |
| The | semester | steroids | 29% |
| Jozefowicz | | | |
| The | husband | authors | 46% |
| Jim | listened | survived | 43% |
| The | uncle | roads | 42% |
| The | knight | saints | 40% |

# Potential improvements

Option 1: Choose better distractors.

# Potential improvements

Option 1: Choose better distractors.

- Tweak thresholds of distractor selection.

# Potential improvements

Option 1: Choose better distractors.

- Tweak thresholds of distractor selection.
- Could search for optimal thresholds.

# Potential improvements

Option 1: Choose better distractors.

- Tweak thresholds of distractor selection.
- Could search for optimal thresholds.
- Could manually check and fix distractors.

# Potential improvements

Option 2: Be forgiving.

# Potential improvements

Option 2: Be forgiving.

- Just don’t terminate on mistakes.

# Potential improvements

Option 2: Be forgiving.

- Just don't terminate on mistakes.
- Instead have participant correct mistake and continue.

# Maze with Error Correction

The x-x-x

e i

# Maze with Error Correction

The x-x-x

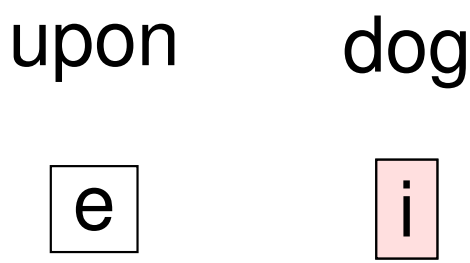
e i

# Maze with Error Correction

upon dog

e i

# Maze with Error Correction

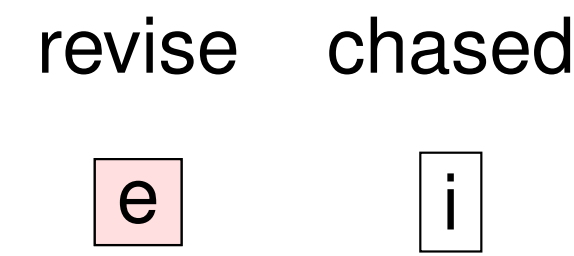


# Maze with Error Correction

revise chased

e i

# Maze with Error Correction



# Maze with Error Correction

revise chased

e i

Incorrect. Please try again.

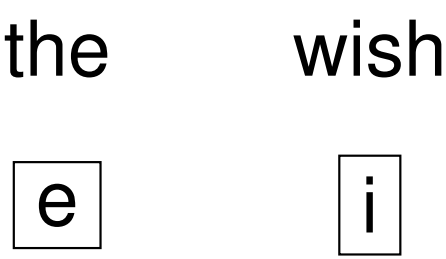
# Maze with Error Correction

revise chased

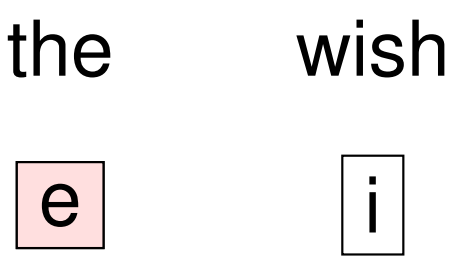
e i

Incorrect. Please try again.

# Maze with Error Correction



# Maze with Error Correction



# Maze with Error Correction

mitigate. squirrel.

e i

# Maze with Error Correction

mitigate. squirrel.

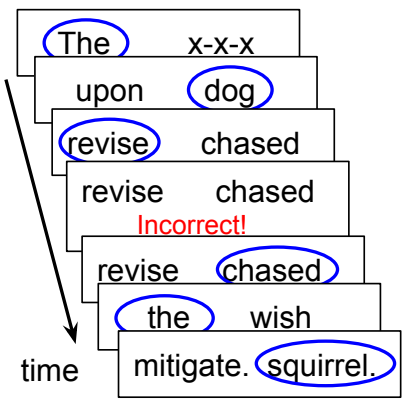
e i

# Maze with Error Correction

The “solution” to our problems!

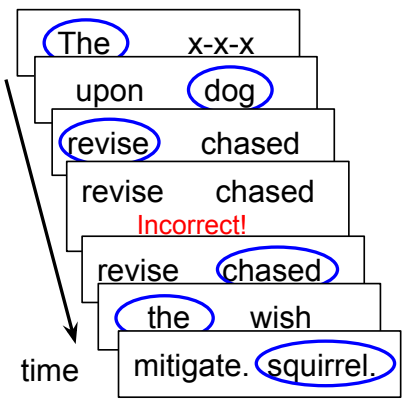
# Maze with Error Correction

The “solution” to our problems!



# Maze with Error Correction

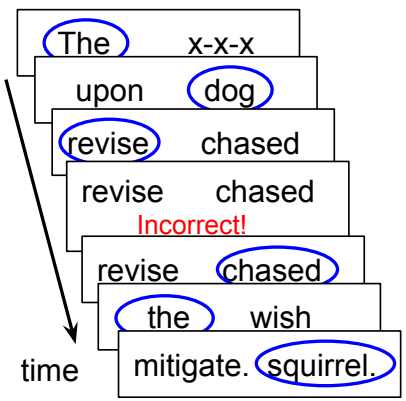
The “solution” to our problems!




- Easily implemented as a toggle in Ibex

# Maze with Error Correction

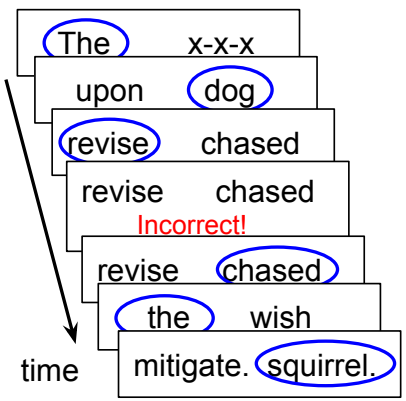
The “solution” to our problems!




- Easily implemented as a toggle in Ibex
- Have all the data

# Maze with Error Correction

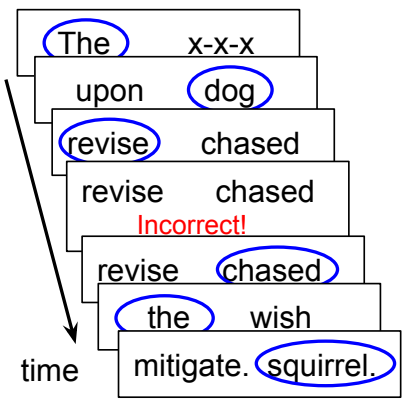
The “solution” to our problems!




- Easily implemented as a toggle in Ibex
- Have all the data
- Identify bad participants v bad distractors

# Maze with Error Correction

The “solution” to our problems!




- Easily implemented as a toggle in Ibex
- Have all the data
- Identify bad participants v bad distractors

Side effect: Can now run multi-sentence items!

# Natural Stories

Natural stories corpus (Futrell et al. 2017)

- 10 stories, each about 1000 words
- Some unusual constructions, but read fluently
- 6 comprehension questions per story

# Natural Stories

Tulip mania was a period in the Dutch Golden Age during which contract prices for bulbs of the recently introduced tulip reached extraordinarily high levels and then suddenly collapsed. At the peak of tulip mania in February sixteen thirty-seven, tulip contracts sold for more than ten times the annual income of a skilled craftsman. It is generally considered the first recorded economic bubble. [...]

Q: When did tulip mania reach its peak?
A: 1630's 1730's

# Experiment Plan

Experiment questions:

- Will people do this task?
- Can participants comprehend the stories?

# Experiment Plan

Experiment questions:

- Will people do this task?
- Can participants comprehend the stories?

Checks on RT data

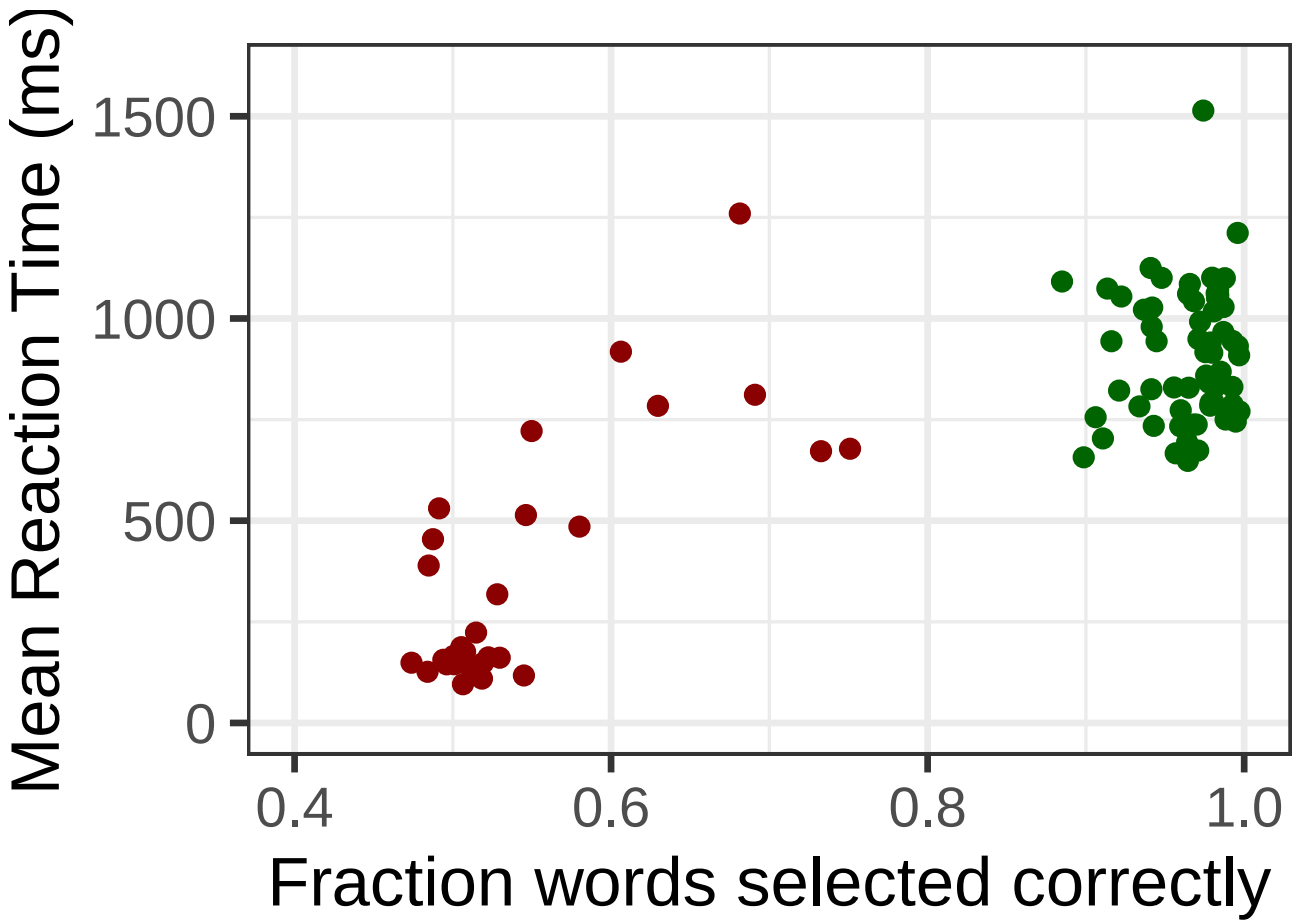
- Check if RT is linear in surprisal (Smith & Levy 2013)
- Check for known effects of word length, frequency
- Check for spillover effects

# Participant accuracy

100 participants each read 1 story

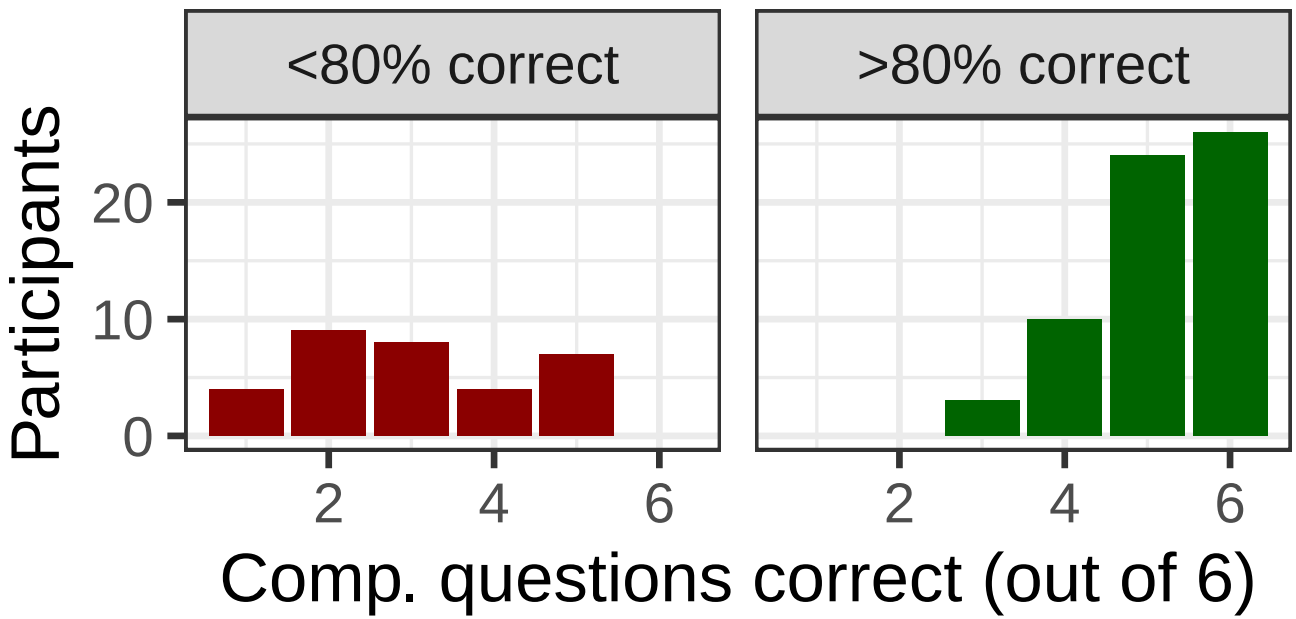
# Participant accuracy

100 participants each read 1 story



# Comprehension questions

# Comprehension questions



# Surprisal Effects

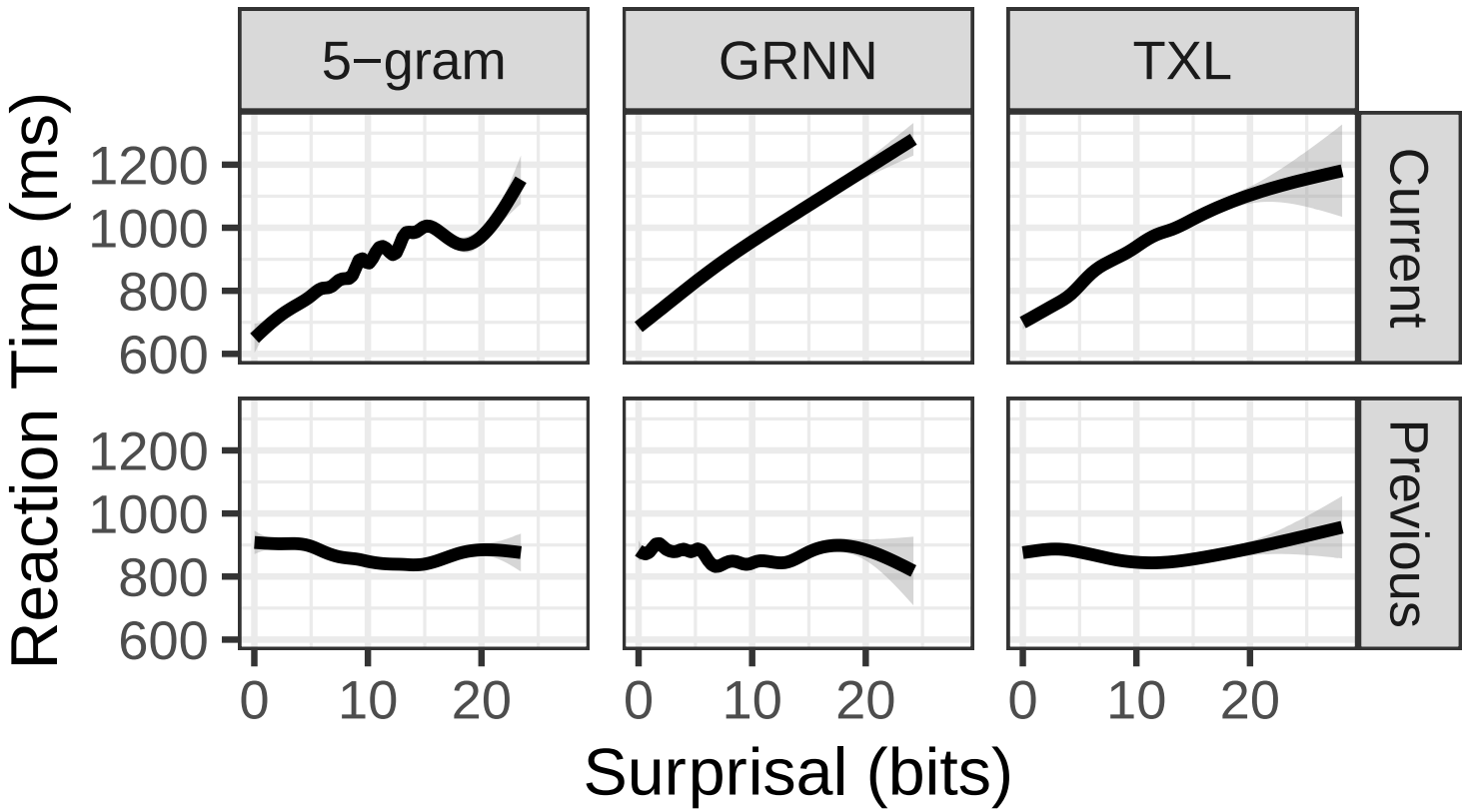
Use 3 LMs to estimate surprisal: smoothed 5-gram, Gulordava RNN, Transformer-XL (Gulordava et al. 2018, Dai et al. 2019)

# Surprisal Effects

Use 3 LMs to estimate surprisal: smoothed 5-gram, Gulordava RNN, Transformer-XL (Gulordava et al. 2018, Dai et al. 2019)
Fit GAMs on pre-error data.

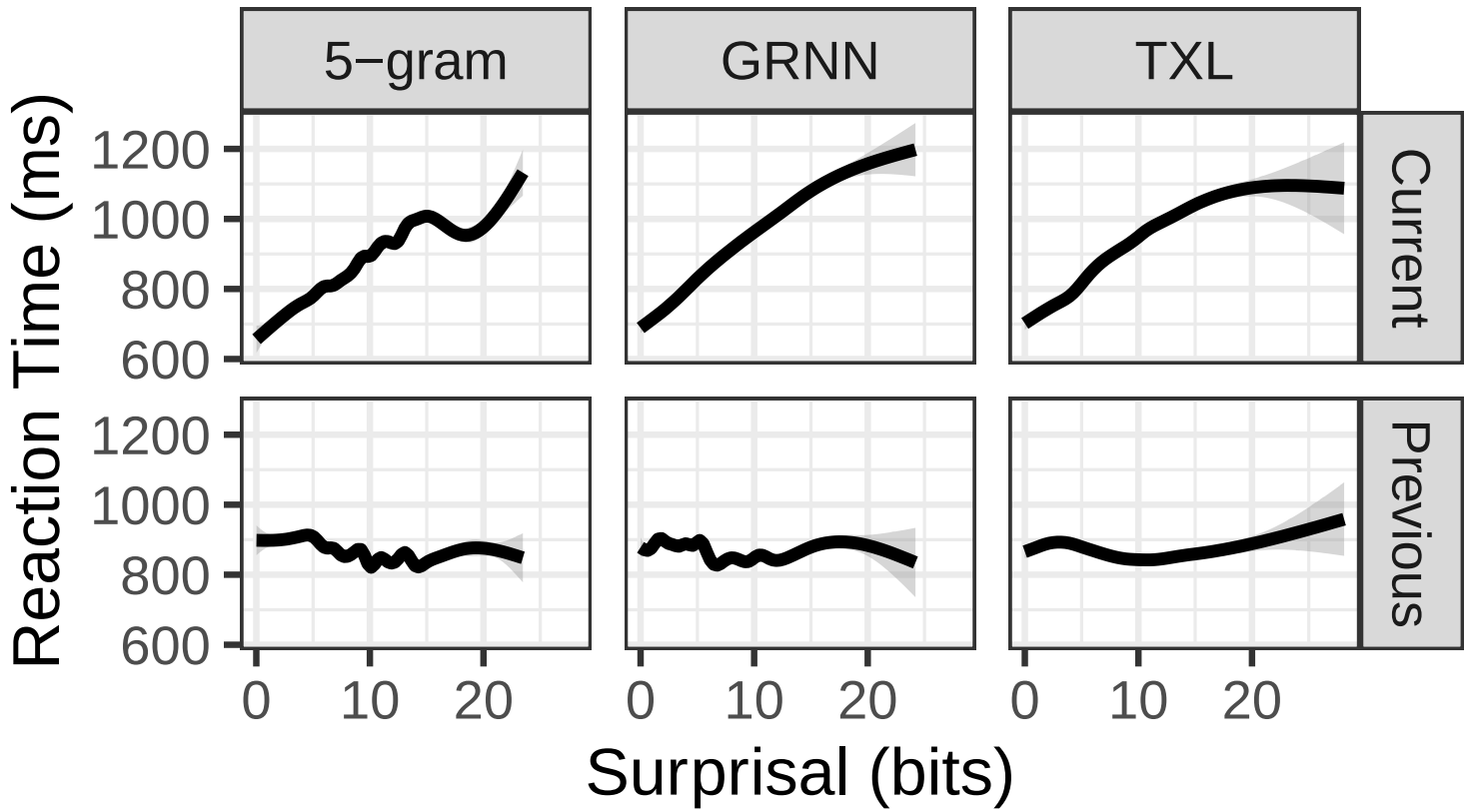
- Limit to single-token words.
- Fit to both current and past word surprisal.

# Surprisal Effects



# What about post-mistake data?

Exclude data from mistakes or the two words after a mistake.



# Summary

# Summary

- People will read in the Maze task for 15-20 minutes

# Summary

- People will read in the Maze task for 15-20 minutes
- It's possible to comprehend during Maze

# Summary

- People will read in the Maze task for 15-20 minutes
- It's possible to comprehend during Maze
- Distractors are generally good enough

# Summary

- People will read in the Maze task for 15-20 minutes
- It's possible to comprehend during Maze
- Distractors are generally good enough
- Find expected RT patterns

# Summary

- People will read in the Maze task for 15-20 minutes
- It's possible to comprehend during Maze
- Distractors are generally good enough
- Find expected RT patterns
- Very little spillover

# Use A-maze!

Easy to use!

# Use A-maze!

Easy to use!

- Runs on command line

# Use A-maze!

Easy to use!

- Runs on command line
- Match distractors across minimal pair sentences

# Use A-maze!

Easy to use!

- Runs on command line
- Match distractors across minimal pair sentences
- Customize surprisal thresholds, vocabulary lists

# Use A-maze!

Easy to use!

- Runs on command line
- Match distractors across minimal pair sentences
- Customize surprisal thresholds, vocabulary lists
- Can output pre-formatted for Ibex

# Use A-maze!

# Use A-maze!

Contribute to A-maze:

- Written in Python 3

# Use A-maze!

Contribute to A-maze:

- Written in Python 3
- Interface with other language models

# Use A-maze!

Contribute to A-maze:

- Written in Python 3
- Interface with other language models
- Add more frequency sources

# Use A-maze!

Contribute to A-maze:

- Written in Python 3
- Interface with other language models
- Add more frequency sources
- Extend to non-English languages

## Documentation: vboyce.github.io/Maze

with links to the following:

- A-maze code: github.com/vboyce/Maze
- Web-maze code: github.com/vboyce/Ibex-with-Maze
- Sample task: syntaxgym.org:666
- Paper: psyarxiv.com/b7nqd/

# Matching distractors

If unspecified: Match by position

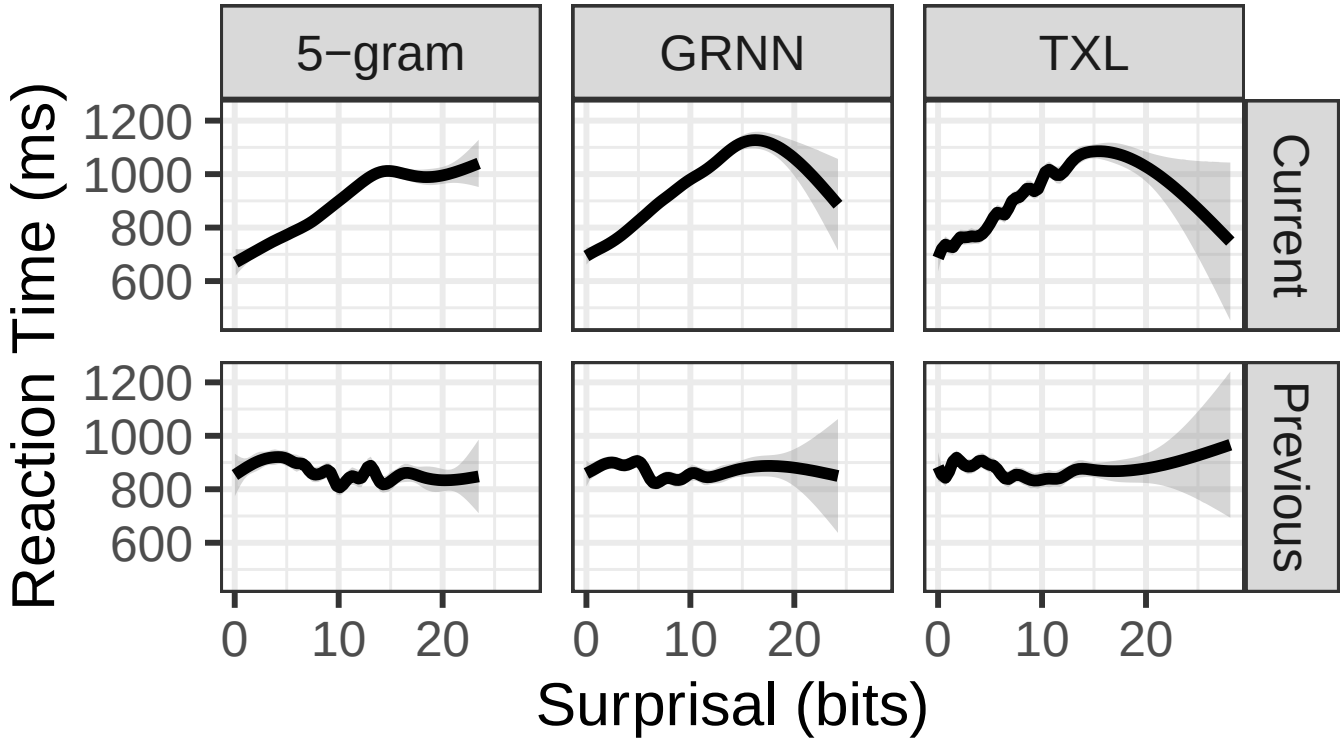
- The son of the lady who politely introduced herself / himself was popular at the party.

Can specify labels for each word to pair (within item)

- The cat who the dog scared hid in a box.
  pre-1 pre-2 who art noun verb main-verb post-1 post-2 post-3
- The dog who scared the cat sniffed around the couch.
  pre-1 pre-2 who verb art noun main-verb post-1 post-2 post-3

# Post-mistake only

Exclude data from mistakes or the two words after a mistake.



# Regression coefficients

| | 5-gram | | | GRNN | | | TXL | |
|---|---|---|---|---|---|---|---|---|
| | Est | CI | *p* | Est | CI | *p* | Est | CI |
| Intercept | 865.3 | [829.9, 902.9] | 0.00 | 871.1 | [837.9, 905.3] | 0.00 | 870.8 | [832.5, 907.8] |
| Surprisal | 11.7 | [9.3, 14.1] | 0.00 | 23.7 | [21, 26.5] | 0.00 | 18.5 | [16.1, 21.1] |
| Frequency | -2.9 | [-6.3, 0.5] | 0.10 | 2.9 | [-0.2, 6] | 0.06 | 0.4 | [-2.7, 3.5] |
| Length | 20.5 | [15.4, 25.6] | 0.00 | 18.5 | [13.3, 23.7] | 0.00 | 21.4 | [16.2, 26.6] |
| Surprisal:Length | -2.0 | [-3, -1] | 0.00 | -1.8 | [-2.7, -0.9] | 0.00 | -1.4 | [-2.2, -0.6] |
| Freq:Length | -1.0 | [-2.5, 0.4] | 0.16 | -0.1 | [-1.2, 1] | 0.82 | 0.2 | [-0.9, 1.2] |
| Past Surprisal | 1.6 | [-0.5, 3.6] | 0.14 | 2.7 | [0.8, 4.5] | 0.00 | 0.8 | [-0.9, 2.5] |
| Past Freq | 2.6 | [-0.1, 5.4] | 0.06 | 1.9 | [-0.2, 4.2] | 0.08 | 1.2 | [-1.1, 3.6] |
| Past Length | -4.8 | [-9, -0.1] | 0.04 | -6.6 | [-10.9, -2.1] | 0.00 | -5.2 | [-9.3, -0.7] |
| Past Surp:Length | -0.2 | [-1.2, 0.8] | 0.72 | -0.9 | [-1.7, -0.2] | 0.01 | -0.6 | [-1.3, 0.2] |
| Past Freq:Length | -1.0 | [-2.3, 0.3] | 0.15 | -1.8 | [-2.9, -0.8] | 0.00 | -1.5 | [-2.6, -0.5] |