A-maze: Easier measurement of incremental processing

Veronica Boyce

26 April 2021

Plan

Plan

- Maze task
- A-maze
- Study 1: methods comparison
- Variant of A-maze
- Study 2: test on Natural Stories Corpus

Plan

mit computational psycholinguistics lab

- Maze task
- A-maze
- Study 1: methods comparison
- Variant of A-maze
- Study 2: test on Natural Stories Corpus

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

- 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

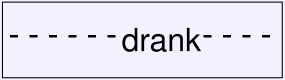
- - - cat - - - - - - - -

Two common methods

Eye-tracking

- Expensive
- Hard to analyse

Self-paced reading

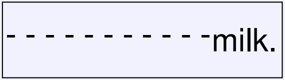


Two common methods

Eye-tracking

- Expensive
- Hard to analyse

Self-paced reading

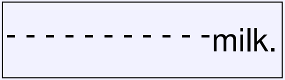


Two common methods

Eye-tracking

- Expensive
- Hard to analyse

Self-paced reading



- Lots of spillover
- Messy data

Two common methods

Eye-tracking

- Expensive
- Hard to analyse

Self-paced reading

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

- Lots of spillover
- Messy data

Different methods have different trade-offs

An alternative: Maze

The x-x-x

An alternative: Maze

The x-x-x

An alternative: Maze

upon dog

An alternative: Maze

upon dog

An alternative: Maze

revise   chased

An alternative: Maze

revise chased

An alternative: Maze

the wish

An alternative: Maze

the wish

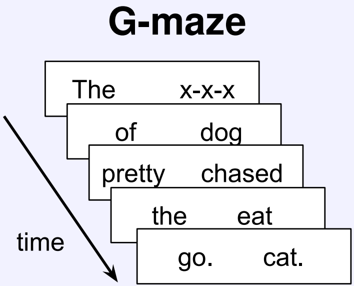
An alternative: Maze

mitigate. squirrel.

An alternative: Maze

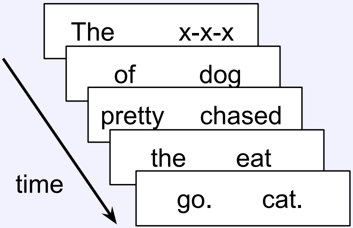
mitigate. squirrel.

A third option: Maze

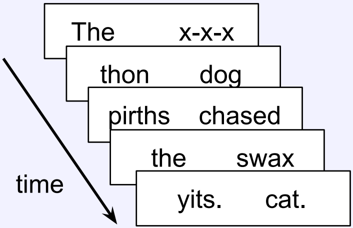


A third option: Maze

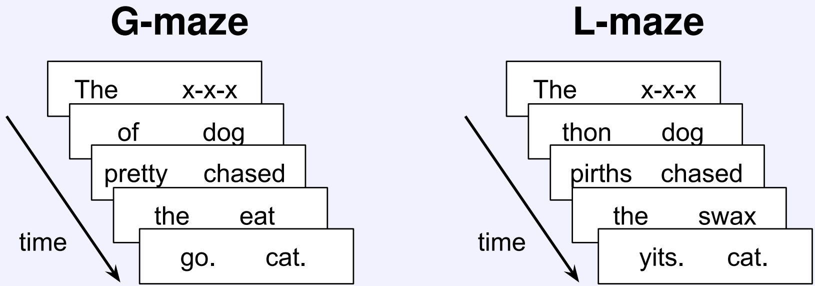
G-maze



L-maze

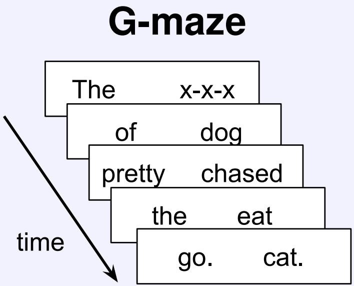


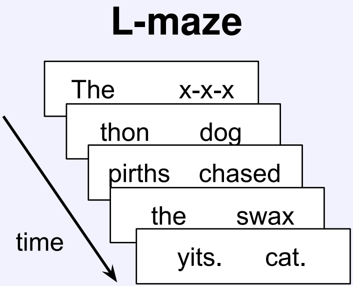
A third option: Maze



Sentence ends when a mistake is made.

A third option: Maze





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

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

Maze Made Easy

Can we use Maze instead of web SPR?

Maze Made Easy

Can we use Maze instead of web SPR?

Needs some tweaks:

Maze Made Easy

Can we use Maze instead of web SPR?

Needs some tweaks:

- Run on web

Maze Made Easy

Can we use Maze instead of web SPR?

Needs some tweaks:

- Run on web
- Easily generate distractors

Maze Made Easy

Can we use Maze instead of web SPR?

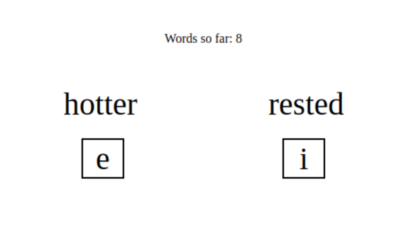
Needs some tweaks:

- Run on web
- Easily generate distractors
- Work for multi-sentence items

Run on web

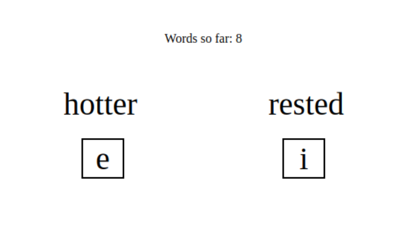
Run on web

Wrote an Ibex module



Run on web

Wrote an Ibex module



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.

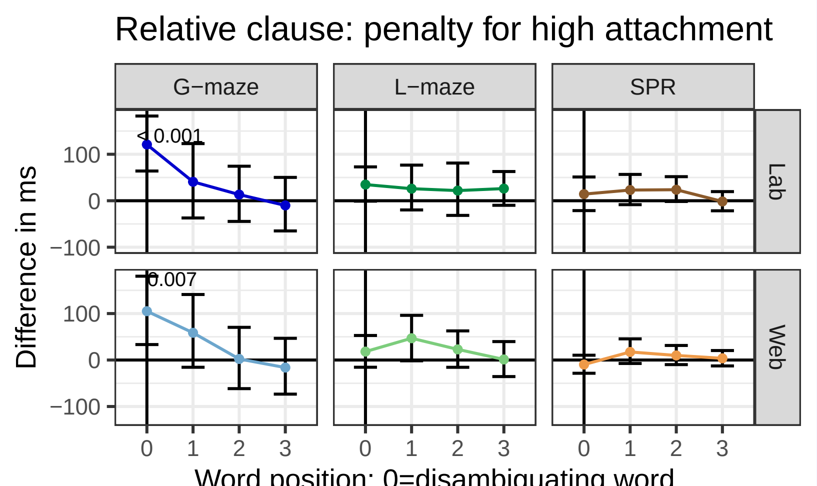
Sentence v Noun Phrase conjunction

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.

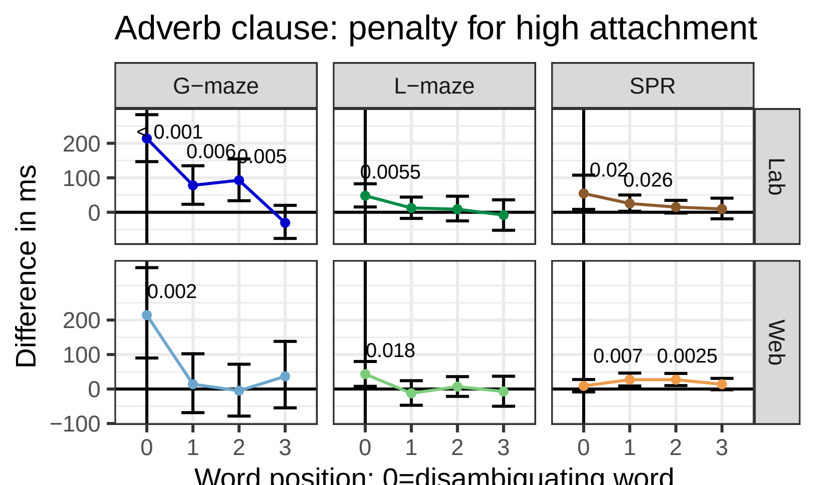
Results

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



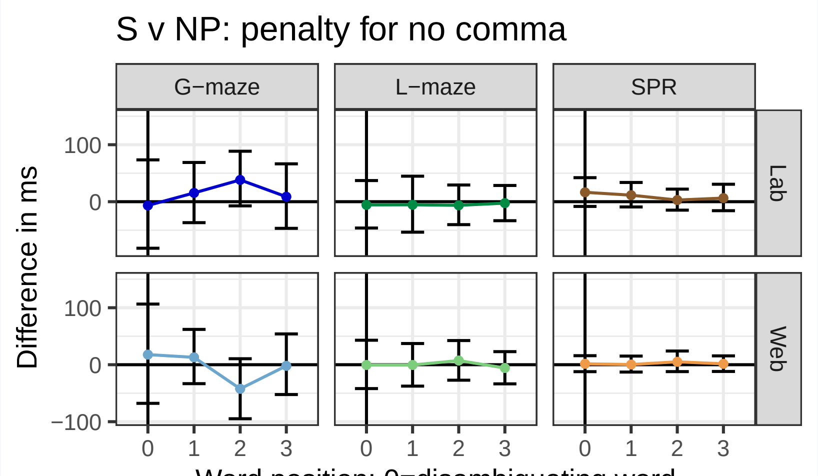
Results

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.



Maze Made Easy

Can we use Maze instead of web SPR?

Needs some tweaks:

- Run on web ✓
- Easily generate distractors
- Work for multi-sentence items

Generating distractors

Generating distractors

Goal: Find a word that can't continue a partial sentence

- Ex. *The dog chased*

Generating distractors

Goal: Find a word that can't continue a partial sentence

- Ex. *The dog chased*
- Tedious (and hard!) to do by hand

Generating distractors

Goal: Find a word that can't continue a partial sentence

- Ex. *The dog chased*
- Tedious (and hard!) to do by hand

What makes something an unacceptable continuation?

Generating distractors

Goal: Find a word that can't continue a partial sentence

- Ex. *The dog chased*
- Tedious (and hard!) to do by hand

What makes something an unacceptable continuation?

- Ungrammatical

Generating distractors

Goal: Find a word that can't continue a partial sentence

- Ex. *The dog chased*
- Tedious (and hard!) to do by hand

What makes something an unacceptable continuation?

- Ungrammatical
- ...or otherwise really unlikely

Generating distractors

Goal: Find a word that can't continue a partial sentence

- Ex. *The dog chased*
- Tedious (and hard!) to do by hand

What makes something an unacceptable continuation?

- Ungrammatical
- ...or otherwise really unlikely
- $\approx$ high surprisal

Generating distractors

Goal: Find a word that can't continue a partial sentence

- Ex. *The dog chased*
- Tedious (and hard!) to do by hand

What makes something an unacceptable continuation?

- Ungrammatical
- ...or otherwise really unlikely
- $\approx$ high surprisal

Can we use Neural Language Models?

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

Surprisal: negative log probability

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

"The dog chased ..." $\longrightarrow$ LM $\longrightarrow$

"the" 1.2 bits
"a" 3.1 bits
"cats" 8 bits
...
"wish" 25 bits
...

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

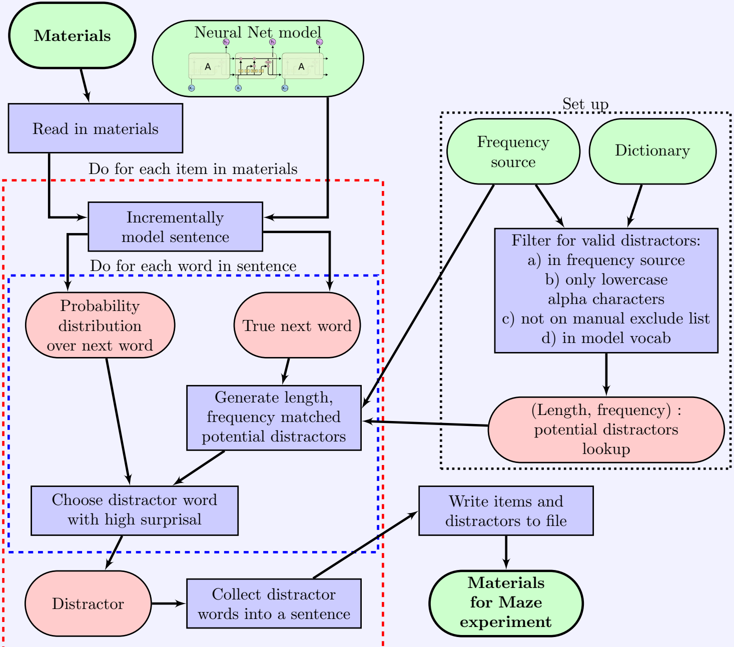
Can we use LMs to choose distractors?

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

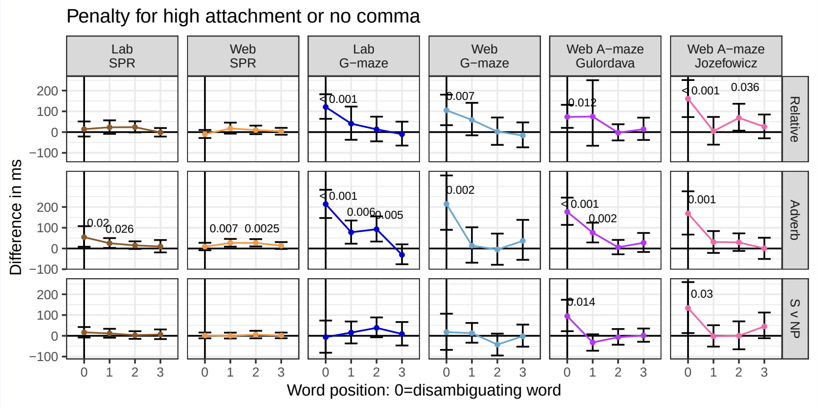
- Model the target sentence word by word
- At each position, choose a high surprisal word

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 it work?



Error bars: 95% CI

Does it work?

Does it work?

Yes, at least well enough.

Does it work?

Yes, at least well enough.

- Caveat: Sometimes generates plausible distractors.

Does it work?

Yes, at least well enough.

- Caveat: Sometimes generates plausible distractors.
- Sloggett et al (2020) also found A-maze results comparable with G-maze

Maze Made Easy

Can we use Maze instead of web SPR?

Needs some tweaks:

- Run on web ✓
- Easily generate distractors ✓
- Work for multi-sentence items

Long items

Want to run multi-sentence items.

Long items

Want to run multi-sentence items.
Problem: Errors terminate sentences.

Long items

Want to run multi-sentence items.
Problem: Errors terminate sentences.

- Treat whole story as a unit:

Long items

Want to run multi-sentence items.
Problem: Errors terminate sentences.

- Treat whole story as a unit: Few participants make it to the end.

Long items

Want to run multi-sentence items.
Problem: Errors terminate sentences.

- Treat whole story as a unit: Few participants make it to the end.
- Treat each sentence as a unit:

Long items

Want to run multi-sentence items.
Problem: Errors terminate sentences.

- Treat whole story as a unit: Few participants make it to the end.
- Treat each sentence as a unit: Some participants miss key context.

Long items

Want to run multi-sentence items.
Problem: Errors terminate sentences.

- Treat whole story as a unit: Few participants make it to the end.
- Treat each sentence as a unit: Some participants miss key context.

What if after an error, participants corrected errors and the sentence continued?

Maze with Error Correction

The x-x-x

Maze with Error Correction

The x-x-x

Maze with Error Correction

upon dog

Maze with Error Correction

upon dog

Maze with Error Correction

revise chased

Maze with Error Correction

revise chased

Maze with Error Correction

revise chased

Incorrect. Please try again.

Maze with Error Correction

revise chased

Incorrect. Please try again.

Maze with Error Correction

the wish

Maze with Error Correction

the wish

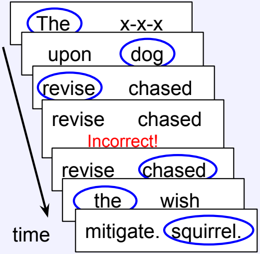
Maze with Error Correction

mitigate. squirrel.

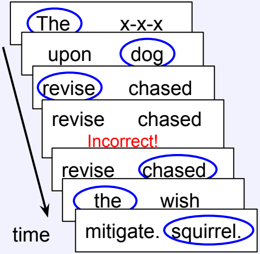
Maze with Error Correction

mitigate. squirrel.

Maze with Error Correction

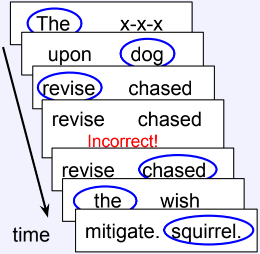


Maze with Error Correction



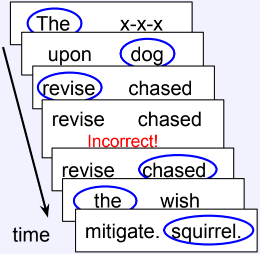
- Can be toggled in Ibex Maze

Maze with Error Correction



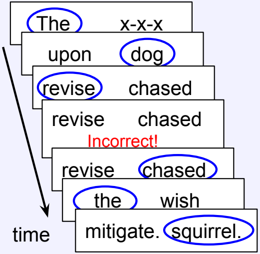
- Can be toggled in Ibex Maze
- Long materials feasible

Maze with Error Correction



- Can be toggled in Ibex Maze
- Long materials feasible
- Have all the data

Maze with Error Correction



- Can be toggled in Ibex Maze
- Long materials feasible
- Have all the data
- Compensates for bad distractors

Maze Made Easy

Can we use Maze instead of web SPR?

Needs some tweaks:

- Run on web ✓
- Easily generate distractors ✓
- Work for multi-sentence items ✓ ?

Various open questions to address

Various open questions to address

- Will people read long texts in Maze?

Various open questions to address

- Will people read long texts in Maze?
- Will they comprehend what they read?

Various open questions to address

- Will people read long texts in Maze?
- Will they comprehend what they read?
- Does error correction Maze work?

Various open questions to address

- Will people read long texts in Maze?
- Will they comprehend what they read?
- Does error correction Maze work?
- Do we get predictability effects?

Natural Stories

Natural stories corpus (Futrell et al. 2017)

Natural Stories

Natural stories corpus (Futrell et al. 2017)

- 10 stories, each about 1000 words

Natural Stories

Natural stories corpus (Futrell et al. 2017)

- 10 stories, each about 1000 words
- 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?

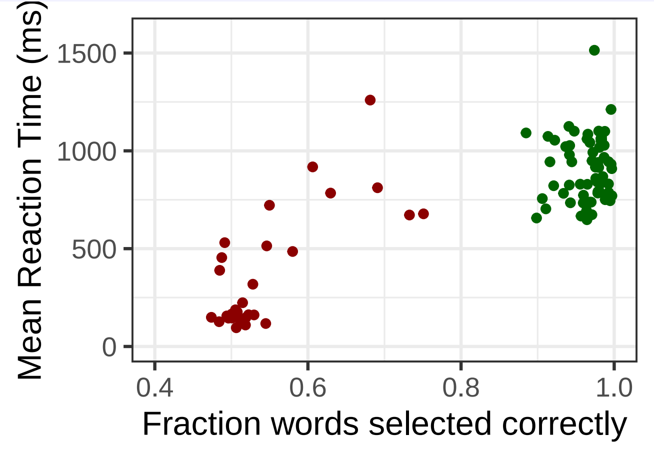
A: 1630's 1730's

Participant accuracy

100 participants each read 1 story

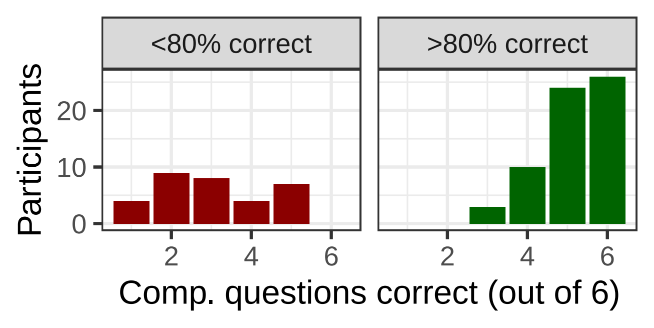
Participant accuracy

100 participants each read 1 story



Comprehension questions

Comprehension questions



Surprisal Effects

Is RT linear in terms of surprisal?

Surprisal Effects

Is RT linear in terms of surprisal?

Estimate surprisal from 3 models:

- smoothed 5-gram
- LSTM-RNN (Gulordava et al. 2018)
- Transformer-XL (Dai et al. 2019)

Surprisal Effects

Is RT linear in terms of surprisal?

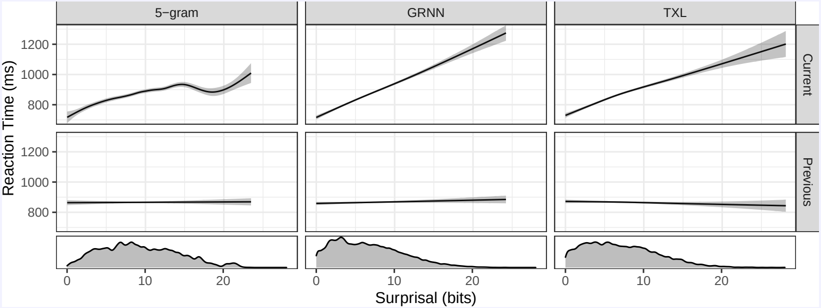
Estimate surprisal from 3 models:

- smoothed 5-gram
- LSTM-RNN (Gulordava et al. 2018)
- Transformer-XL (Dai et al. 2019)

Fit GAMs

- Fit to both current and past word surprisal
- Include frequency, length as predictors

Surprisal Effects



Surprisal Effects

Linear Models

Surprisal Effects

Linear Models

	5-gram	GRNN	TXL
Intercept	**865.3**	**871.1**	**870.8**
Surprisal	**11.7**	**23.7**	**18.5**
Frequency	-2.9	2.9	0.4
Length	**20.5**	**18.5**	**21.4**
Surprisal:Length	**-2.0**	**-1.8**	**-1.4**
Freq:Length	-1.0	-0.1	0.2
Past Surprisal	1.6	**2.7**	0.8
Past Freq	2.6	1.9	1.2
Past Length	**-4.8**	**-6.6**	**-5.2**
Past Surp:Length	-0.2	**-0.9**	-0.6
Past Freq:Length	-1.0	**-1.8**	**-1.5**

Surprisal in bits, Length in characters,
Frequency in log_2 occurrences/billion words

Surprisal Effects

Takeaways:

- Minimal frequency effects (consistent with Shain 2019)
- Large effects of Length, Surprisal
- Very little spillover

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

Consider A-maze!

Easy to use!

Consider A-maze!

Easy to use!

- Runs on command line

Consider A-maze!

Easy to use!

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

Consider A-maze!

Easy to use!

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

Consider 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

Consider 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

Adapt A-maze to your projects:

- 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
- Exp 1 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)

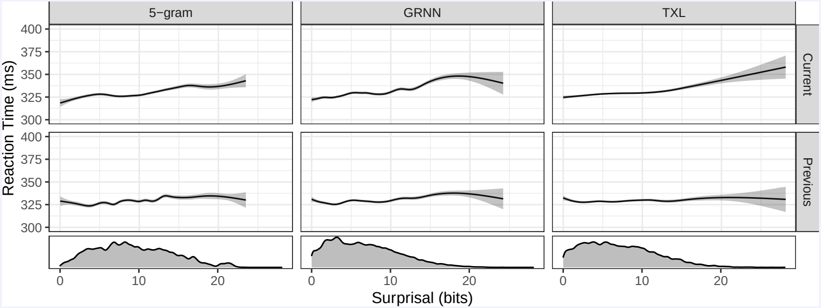
- The cat who the dog scared hid in a box.
  pre-1 pre-2 who art noun verb main-verb post-1 post-2 post-3
- The dog who scared the cat sniffed around the couch.
  pre-1 pre-2 who verb art noun main-verb post-1 post-2 post-3

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]

SPR comparison

Using SPR data collected by Futrell et al. (2017)



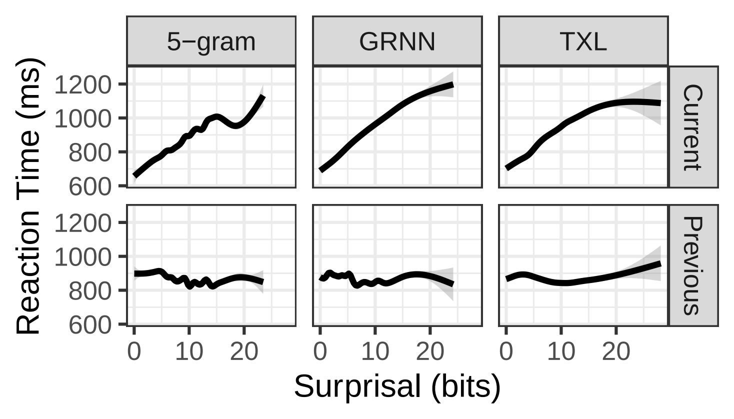
Caveats

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%

What about post-mistake data?

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



Why such large effects?

Bayesian Reader (Norris 2006): Look at words long enough to ID with some threshold of certainty
Possible mechanisms for difference:

- Higher threshold
- Fewer available resources for processing
- Presence of second word