A-maze of Natural Stories:

Texts are comprehensible using the Maze task

Veronica Boyce, Roger Levy

AMLaP 2020

Incremental processing methods

Incremental processing methods

Want a measure of processing difficulty

Incremental processing methods

Want a measure of processing difficulty
Assume that longer RT = more difficulty

Incremental processing methods

Want a measure of processing difficulty
Assume that longer RT = more difficulty

How to measure RT?

Incremental processing methods

Common ways to measure RT

Incremental processing methods

Common ways to measure RT

Eye-tracking

Incremental processing methods

Common ways to measure RT

Eye-tracking

Self-paced reading

Incremental processing methods

Common ways to measure RT

Eye-tracking

Self-paced reading

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

Incremental processing methods

Common ways to measure RT

Eye-tracking

Self-paced reading

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

Incremental processing methods

Common ways to measure RT

Eye-tracking

Self-paced reading

- - - - - - drank - - - -

Incremental processing methods

Common ways to measure RT

Eye-tracking

Self-paced reading

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

Incremental processing methods

Common ways to measure RT

Eye-tracking

Self-paced reading

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

Different methods have different trade-offs

An alternative: Maze

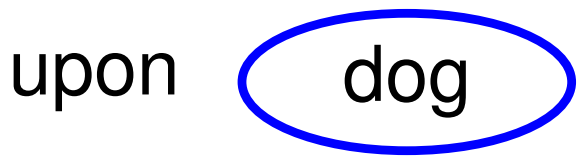
The x-x-x

An alternative: Maze

An alternative: Maze

upon dog

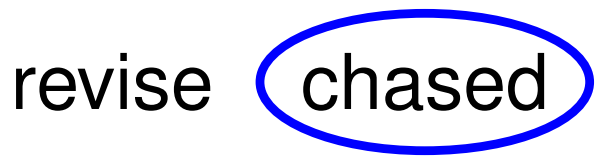
An alternative: Maze



An alternative: Maze

revise chased

An alternative: Maze



An alternative: Maze

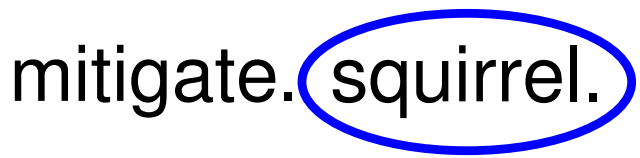
the wish

An alternative: Maze

An alternative: Maze

mitigate. squirrel.

An alternative: Maze

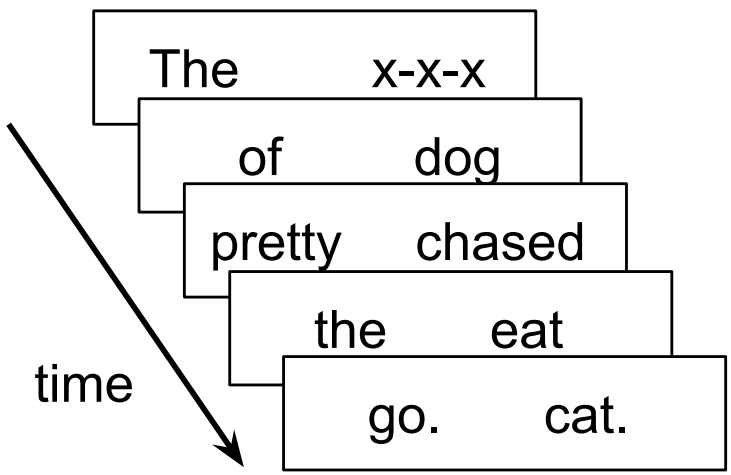


An alternative: Maze

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

G-maze

‘Grammatical’ choices

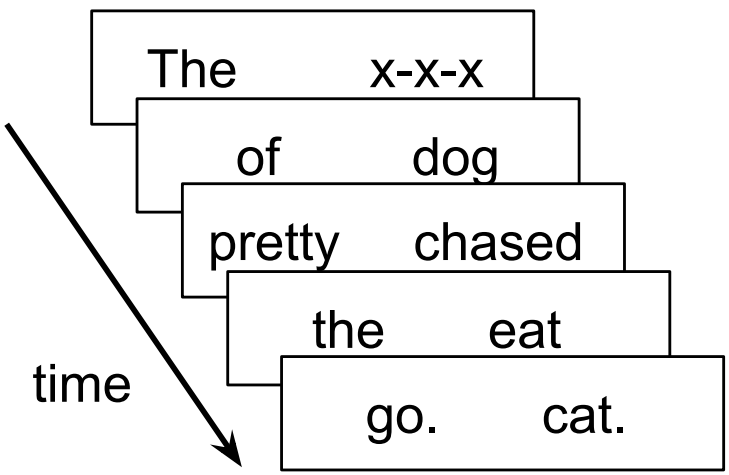


An alternative: Maze

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

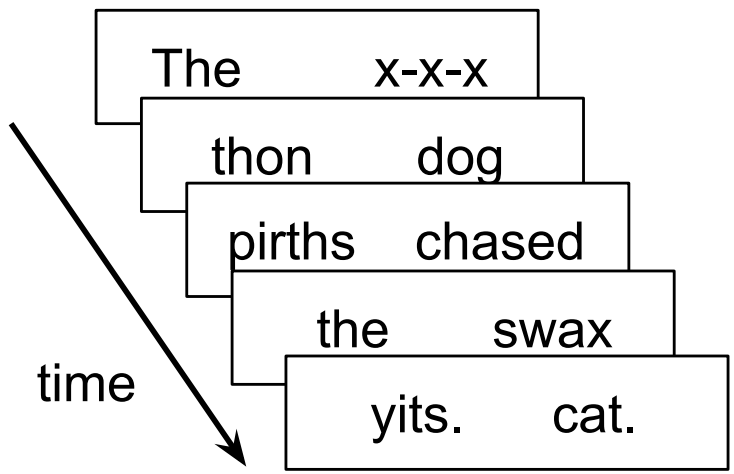
G-maze

'Grammatical' choices



L-maze

'Lexical' choices

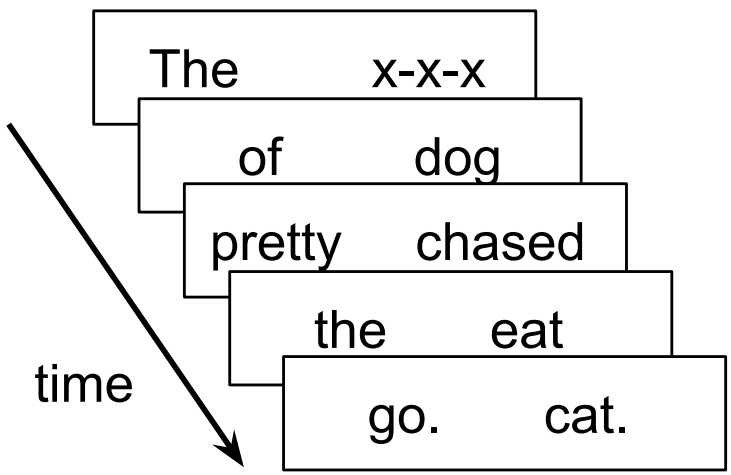


An alternative: Maze

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

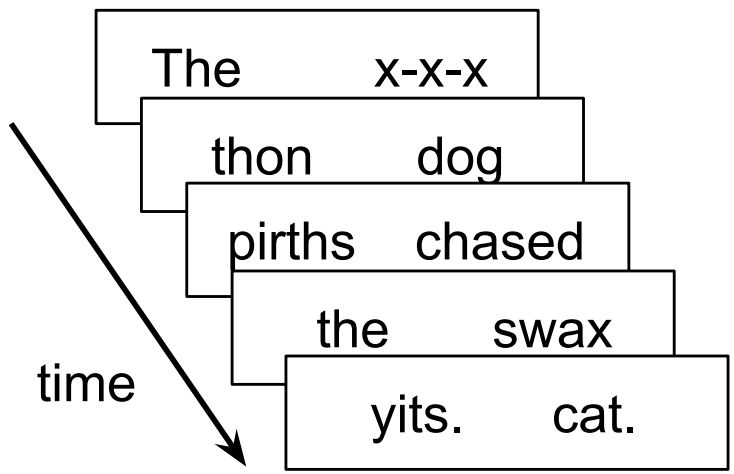
G-maze

'Grammatical' choices



L-maze

'Lexical' choices



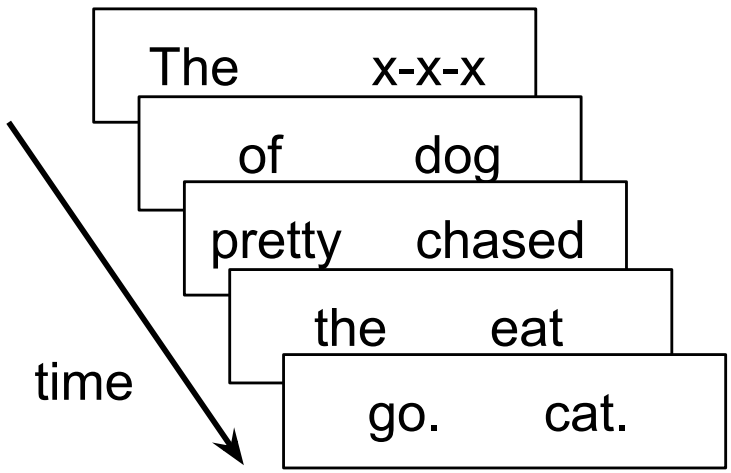
Sentence ends if a mistake is made.

An alternative: Maze

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

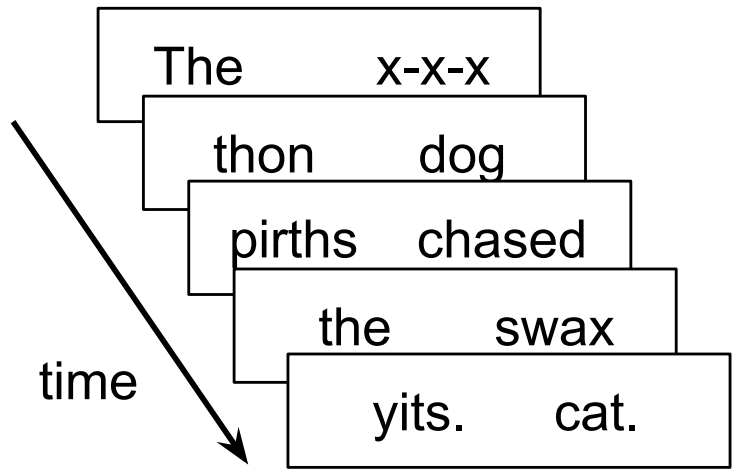
G-maze

'Grammatical' choices



L-maze

'Lexical' choices



Sentence ends if a mistake is made.

Claim: forces incremental processing (no spillover)

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

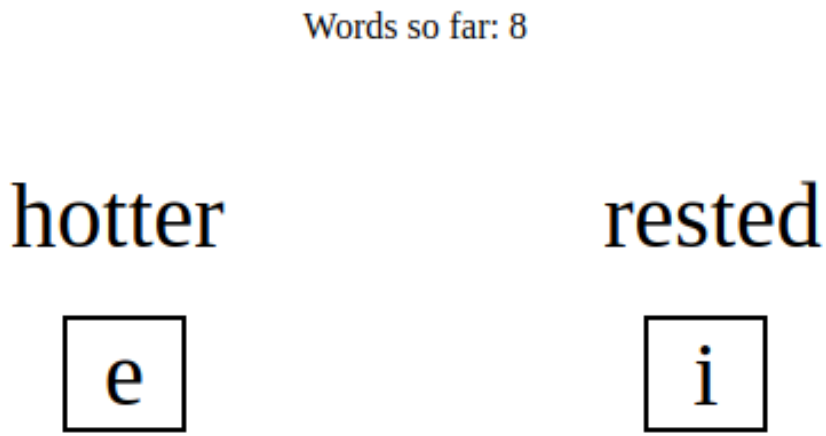
Run on web

Run on web

Wrote an Ibex module

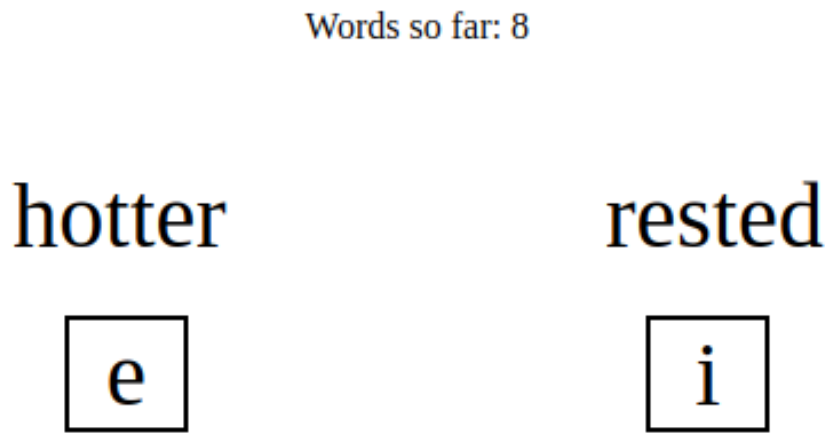
Run on web

Wrote an Ibex module



Run on web

Wrote an Ibex module



Replicated Witzel et al. (2012) results (Boyce et al. 2020)

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?

Can we use LMs?

Can we use LMs?

Language models (LMs)

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

Can we use LMs?

Language models (LMs)

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

Run items through LM, choose high surprisal words as distractors

Can we use LMs?

Language models (LMs)

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

Run items through LM, choose high surprisal words as distractors

Add some restrictions:

- Restrict to a list of possible distractors
- Only consider length, frequency matches

Does it work?

Does it work?

Yes, at least well enough.

Does it work?

Yes, at least well enough.

- Sometimes generates plausible distractors.

Does it work?

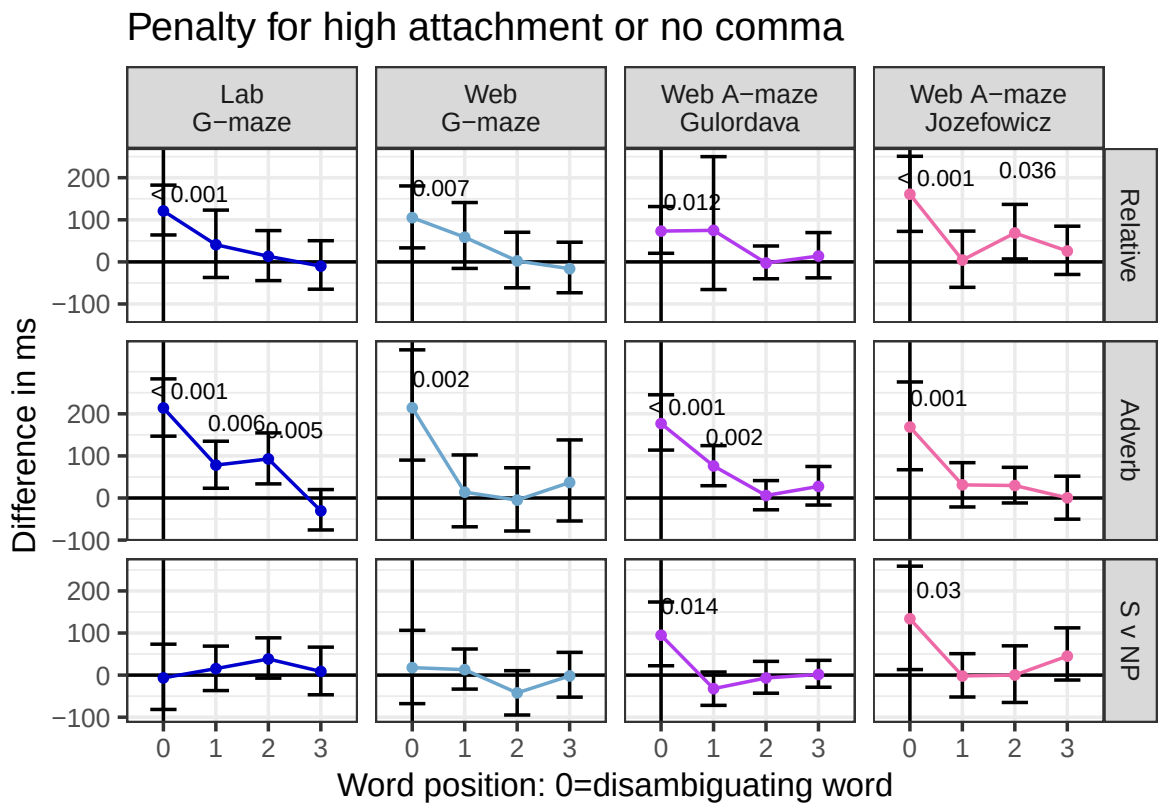
Yes, at least well enough.

- Sometimes generates plausible distractors.
- A-maze results comparable with G-maze (Boyce et al 2020, Sloggett et al 2020)

Does it work?

Yes, at least well enough.

- Sometimes generates plausible distractors.
- A-maze results comparable with G-maze (Boyce et al 2020, Sloggett et al 2020)



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: No one makes it to the end.

Long items

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

- Treat whole story as a unit: No one makes 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: No one makes 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: No one makes 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

Maze with Error Correction

upon dog

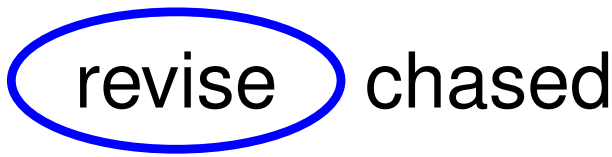
Maze with Error Correction

upon 

Maze with Error Correction

revise chased

Maze with Error Correction



Maze with Error Correction

revise chased

Incorrect. Please try again.

Maze with Error Correction

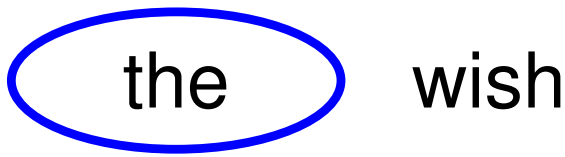
revise chased

Incorrect. Please try again.

Maze with Error Correction

the wish

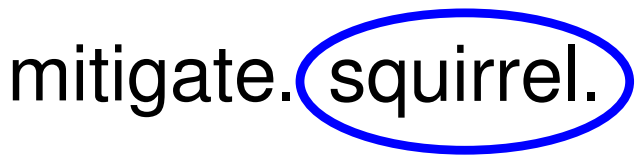
Maze with Error Correction



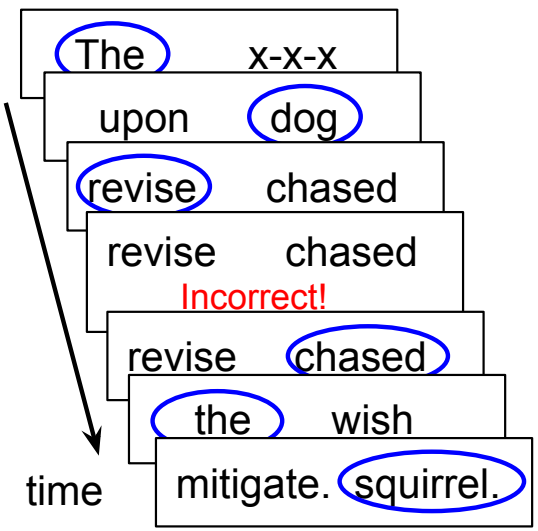
Maze with Error Correction

mitigate. squirrel.

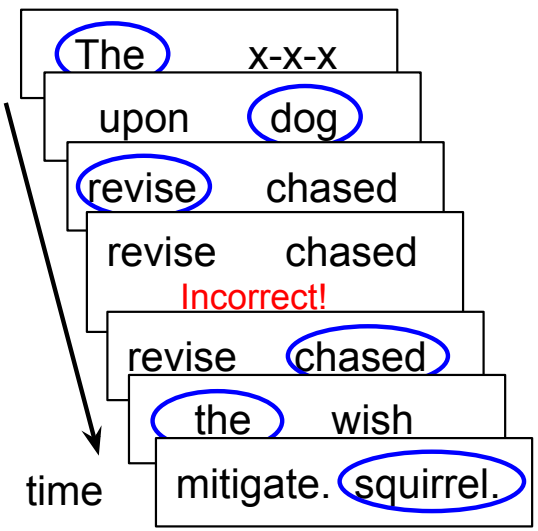
Maze with Error Correction



Maze with Error Correction

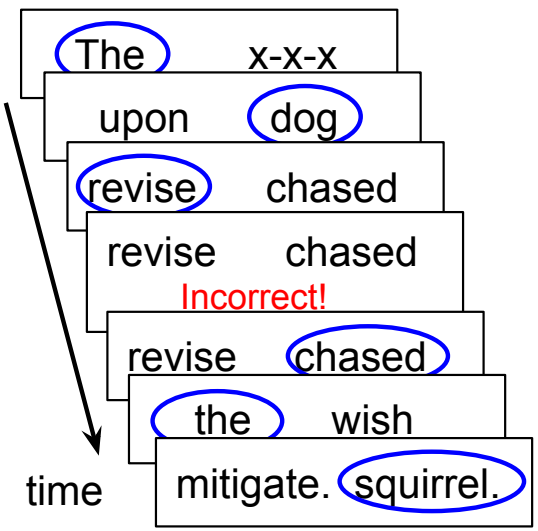


Maze with Error Correction



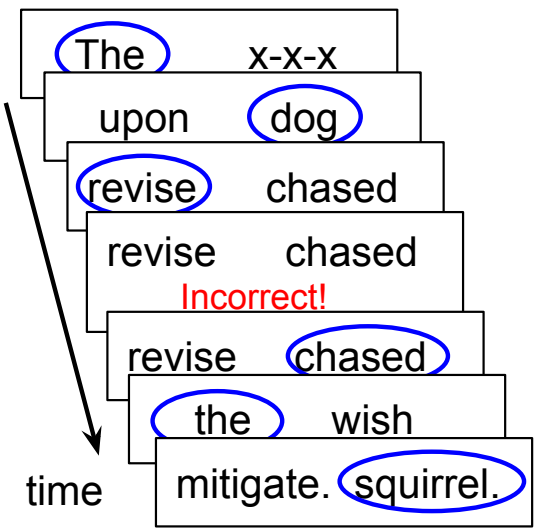
- Can be toggled in Ibex Maze

Maze with Error Correction



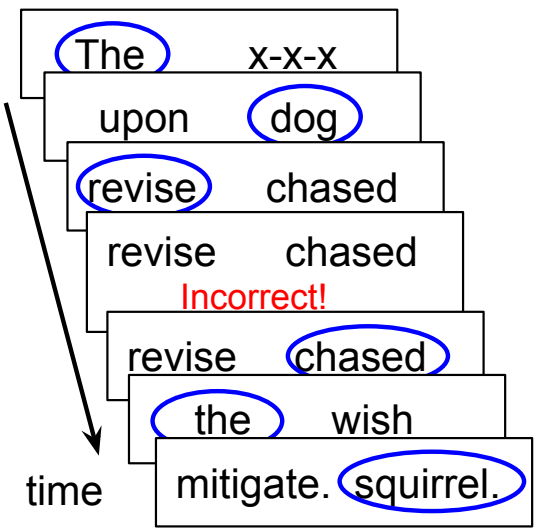
- Can be toggled in Ibex Maze
- Long materials feasible

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 ✓ ?

Current experiment

Various open questions to address

Current experiment

Various open questions to address

- Are people willing to read long texts in this way?

Current experiment

Various open questions to address

- Are people willing to read long texts in this way?
- Do people comprehend what they read?

Current experiment

Various open questions to address

- Are people willing to read long texts in this way?
- Do people comprehend what they read?
- Does the error correction variant work?

Current experiment

Various open questions to address

- Are people willing to read long texts in this way?
- Do people comprehend what they read?
- Does the error correction variant 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. [...]

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. [...]

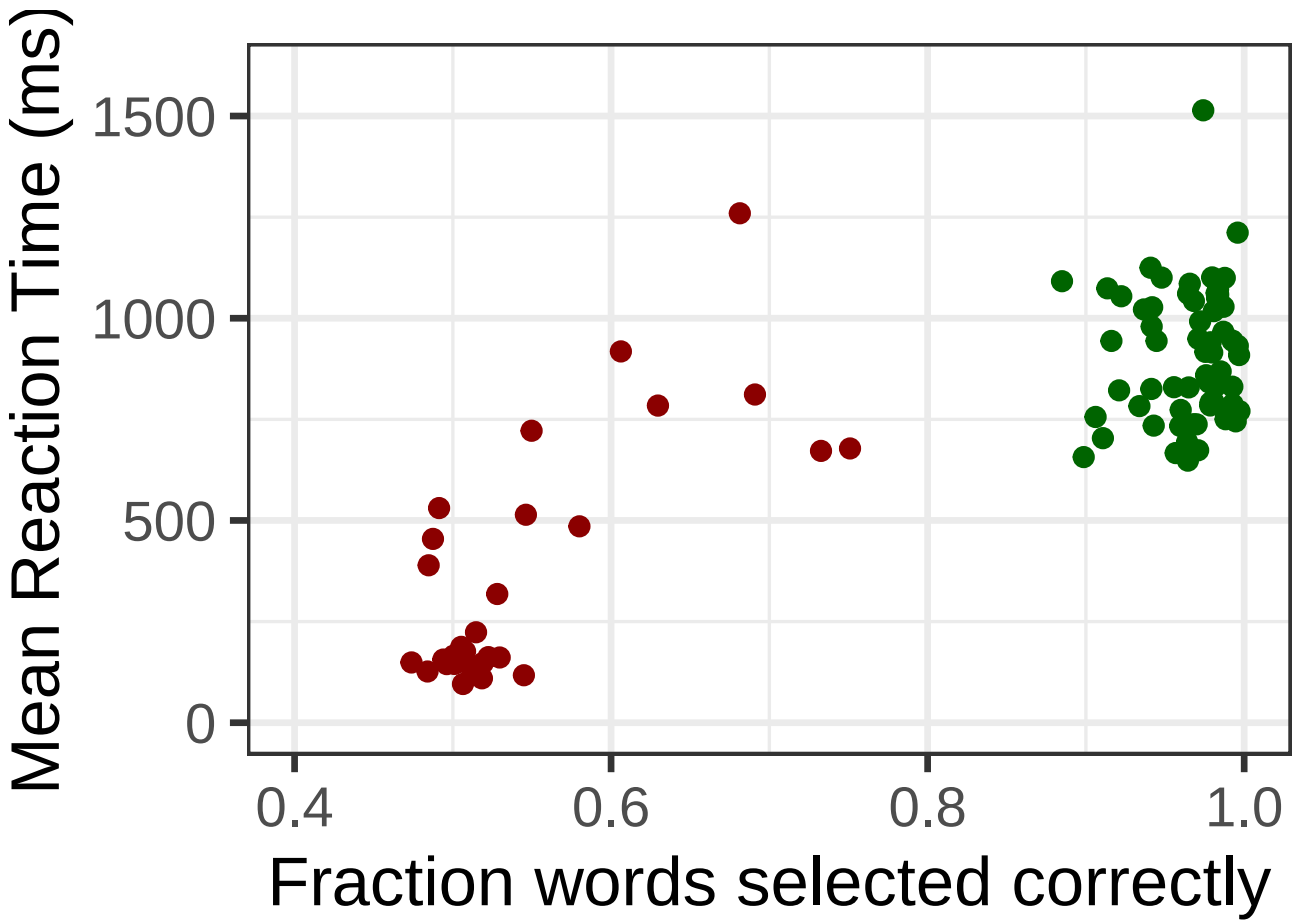
Q: When did tulip mania reach its peak?
A: 1630's 1730's

Participant accuracy

100 participants from MTurk each read 1 story (20 minutes)

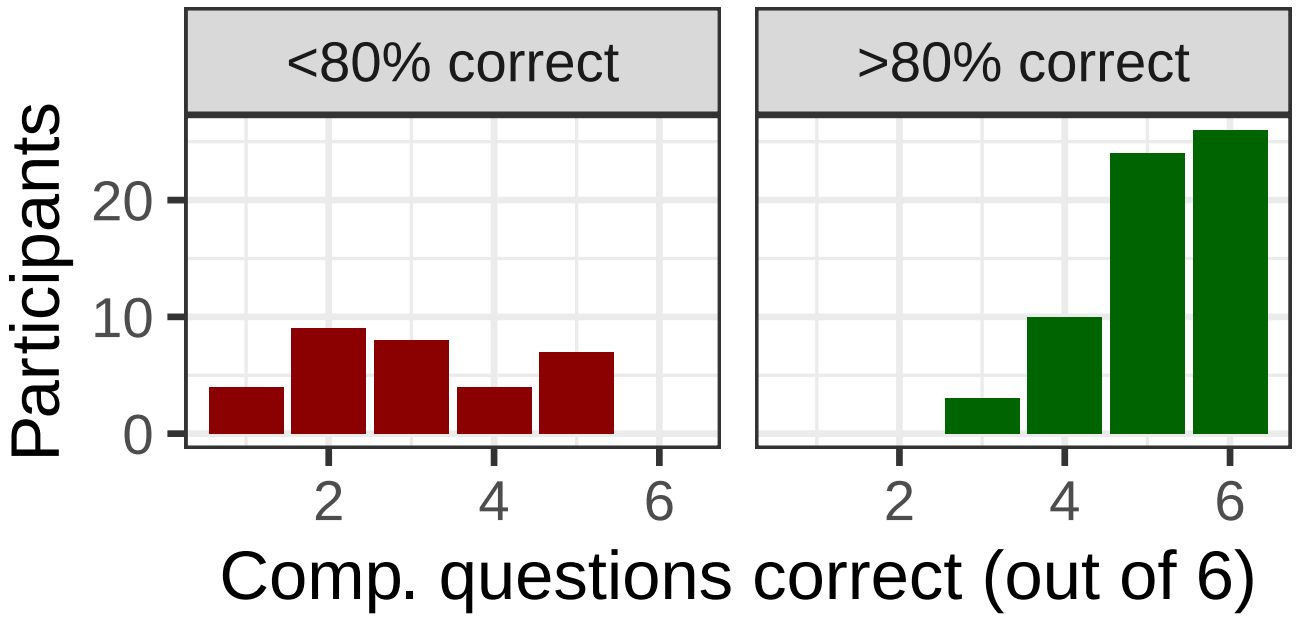
Participant accuracy

100 participants from MTurk each read 1 story (20 minutes)



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?

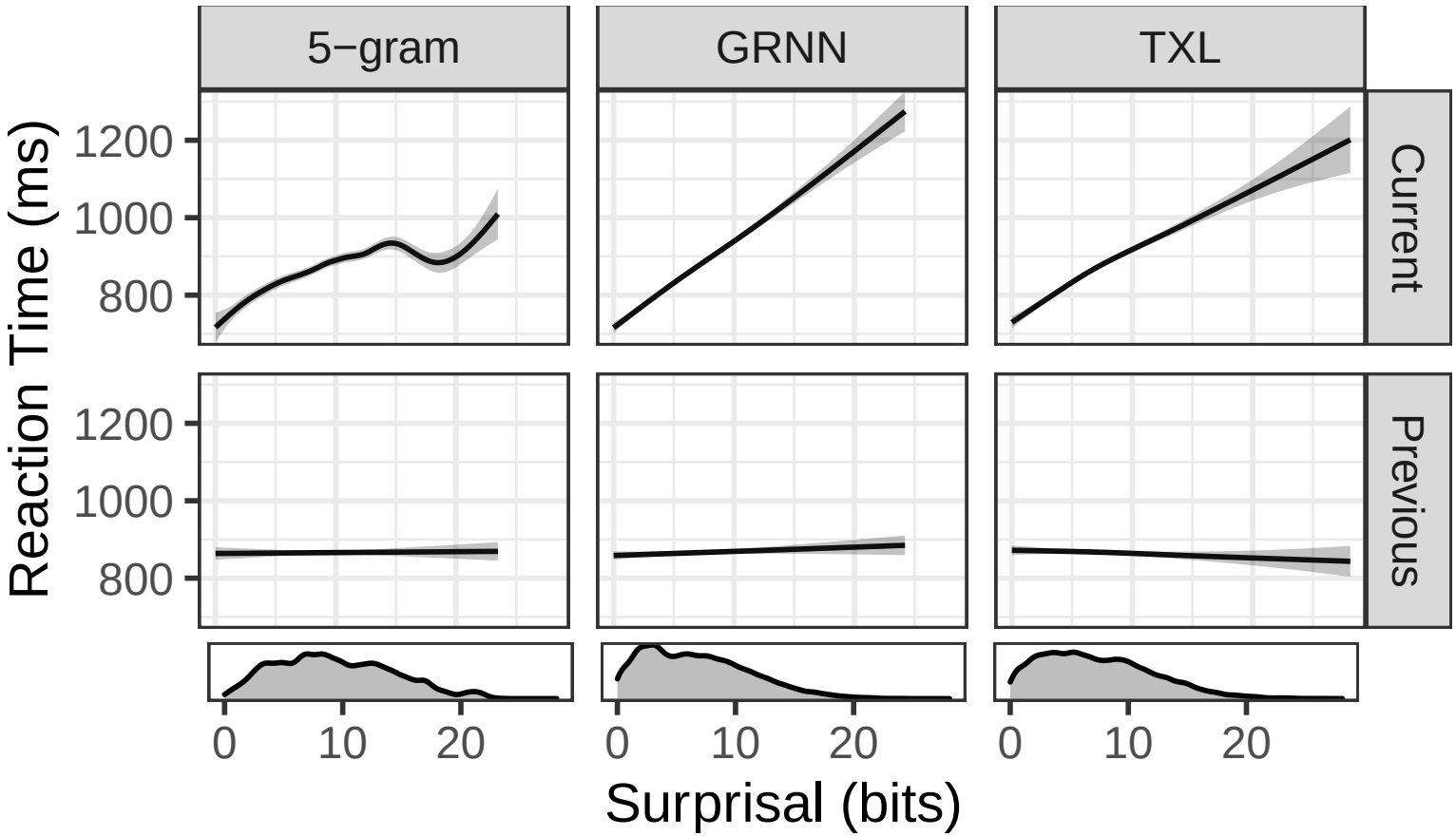
Estimate surprisal from 3 models:

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

Fit GAMs

- Limit to single-token words
- 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:

Surprisal Effects

Takeaways:

- Minimal frequency effects (consistent with Shain 2019)

Surprisal Effects

Takeaways:

- Minimal frequency effects (consistent with Shain 2019)
- Long RTs

Surprisal Effects

Takeaways:

- Minimal frequency effects (consistent with Shain 2019)
- Long RTs
- Large effects of Length, Surprisal

Surprisal Effects

Takeaways:

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

Surprisal Effects

Takeaways:

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

Model comparison: GRNN is best, but TXL complementary

Why such large effects?

Why such large effects?

Framework: Look at words long enough to ID with some threshold of certainty

Why such large effects?

Framework: Look at words long enough to ID with some threshold of certainty
Could hypothesize mechanisms for difference:

Why such large effects?

Framework: Look at words long enough to ID with some threshold of certainty
Could hypothesize mechanisms for difference:

- Task leads to higher threshold

Why such large effects?

Framework: Look at words long enough to ID with some threshold of certainty
Could hypothesize mechanisms for difference:

- Task leads to higher threshold
- Task demands reduce available resources for processing

Why such large effects?

Framework: Look at words long enough to ID with some threshold of certainty
Could hypothesize mechanisms for difference:

- Task leads to higher threshold
- Task demands reduce available resources for processing
- Presence of second word

Conclusion

Conclusion

Consider A-maze!

Conclusion

Consider A-maze!

- Documentation: vboyce.github.io/Maze

Conclusion

Consider A-maze!

- Documentation: vboyce.github.io/Maze
- Versatile

Conclusion

Consider A-maze!

- Documentation: vboyce.github.io/Maze
- Versatile
- Low spillover

Conclusion

Consider A-maze!

- Documentation: vboyce.github.io/Maze
- Versatile
- Low spillover

Natural Stories A-maze:

Conclusion

Consider A-maze!

- Documentation: vboyce.github.io/Maze
- Versatile
- Low spillover

Natural Stories A-maze:

- Participants comprehend what they read

Conclusion

Consider A-maze!

- Documentation: vboyce.github.io/Maze
- Versatile
- Low spillover

Natural Stories A-maze:

- Participants comprehend what they read
- Find linear, large surprisal effects

Surprisal Effects

GAM if we only exclude mistakes (all participants, post-mistake data)