Using Transformer Language Models to Integrate Surprisal, Entropy, and Working Memory Retrieval Accounts of Sentence Processing

Soo Hyun Ryu & Richard L. Lewis

Department of Psychology
Weinberg Institute for Cognitive Science
University of Michigan

March 25, 2022

LSA PSYCHOLOGY
UNIVERSITY OF MICHIGAN

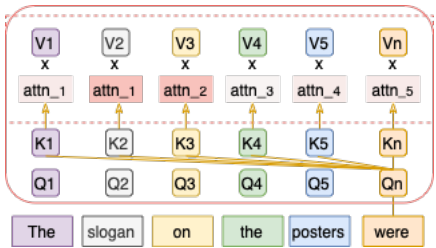
Key ideas and overview

- **Transformer** neural net language models such as GPT are **predictive cue-based retrieval parsers** (Ryu & Lewis, 2021)—and so have the potential to integrate expectation-based and memory-based theories, an important challenge for the field (Futrell et al., 2020).
- We introduce a word-by-word **entropy** metric computed over internal Transformer **attention patterns** and show that it tracks cue-based retrieval theory predictions of **interference**, and is a **predictor of reading times independent of surprisal**.
- We show that Transformer-based **surprisal** and **entropy** metrics together account for a range of phenomena that have posed empirical challenges for conventional surprisal and cue-based accounts.

What is a Transformer?

A Transformer (Vaswani et al., 2017) is a a neural network language model that computes a **probability distribution for the next word** in a sentence given a (large) prefix.

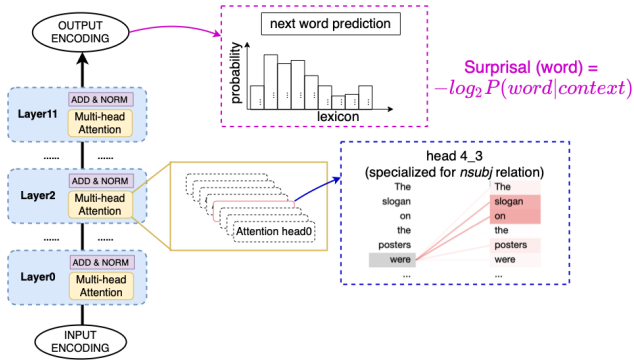
- It solves the problem of long-distance dependencies with an **attention mechanism**.
- For each incoming token it computes a **query** or cue vector that is matched against past **key** vectors that were computed for each previous position.
- Based on the query-key matches, **value** vectors computed for previous positions are retrieved and aggregated, and used to make probabilistic next-token predictions.



What is a Transformer?

A Transformer (Vaswani et al., 2017) is a neural network language model that computes a **probability distribution for the next word** in a sentence given a (large) prefix.

- This *query-match-retrieve-aggregate* process happens in multiple attention heads in multiple layers.
- Because the final output is a probability distribution over the entire lexicon, it is easy to compute **word-by-word surprisal values for arbitrary sentences**.

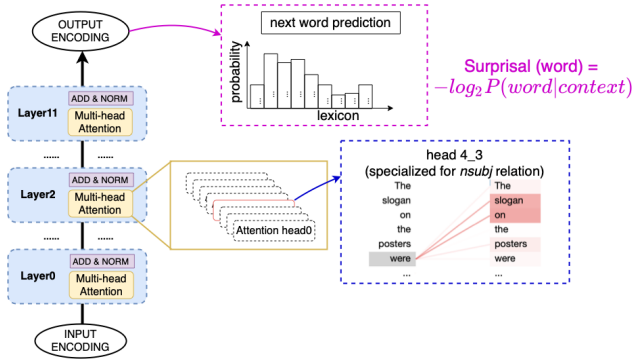


What is a Transformer?

A Transformer (Vaswani et al., 2017) is a neural network language model that computes a **probability distribution for the next word** in a sentence given a (large) prefix.

Transformers are cue-based memory architectures adapted for prediction.

The cues, features, and working memory representations are *learned*.

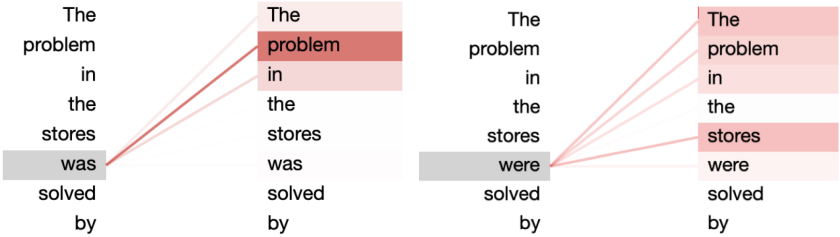


Transformer attention patterns show signs of similarity-based interference

(a) The **problem** in the stores **was** ...
(b) The **problem** in the stores **were** ...

Transformer attention patterns show signs of similarity-based interference

(a) The **problem** in the stores **was** ...
(b) The **problem** in the stores **were** ...



A new word-by-word metric: Attention entropy

We compute the **entropy of the attention** patterns as an index of **similarity-based interference** in the Transformer's internal processes.

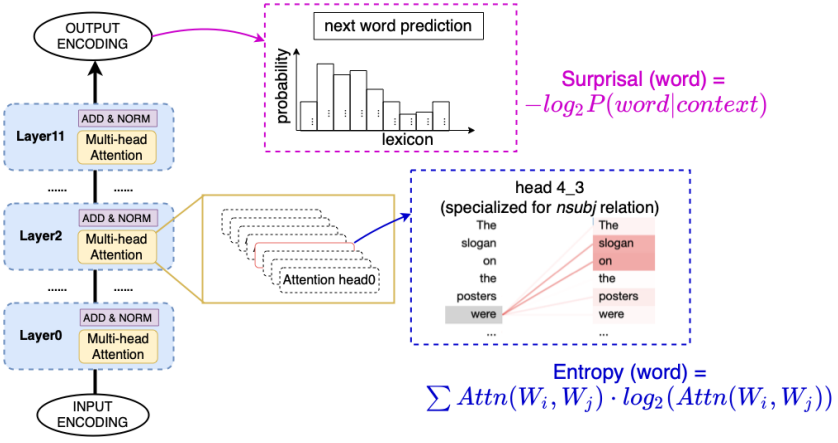
- We can compute this measure for **single attention heads** of interest, and also compute **global measures over multiple or all attention heads**.
- Intuition is: the more diffuse or spread the attention (higher entropy), the greater the similarity-based interference.

$$\text{Entropy}(w_i) = \sum_{j=1}^{i-1} \text{Attn}(w_i, w_j) \times \log_2 \text{Attn}(w_i, w_j) \quad (1)$$

where i refers to the location of the critical word, j are locations of prior words

In other words:

Transformers can be used to estimate processing difficulty from both memory-based and expectation-based theoretical perspectives



Do Transformer-surprisal and attention entropy play distinct roles in predicting reading times?

Modeling self-paced reading times in the Natural Stories Corpus

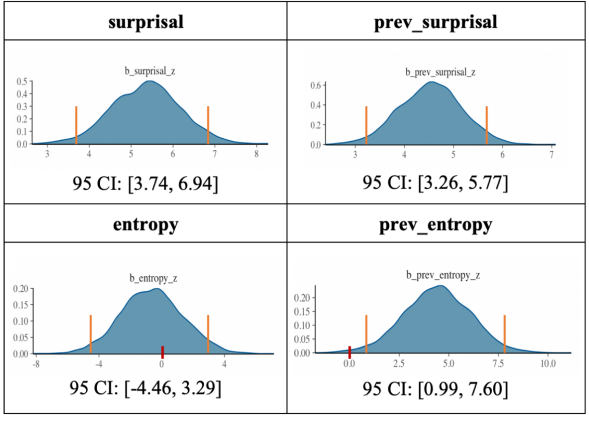
- We use the *Natural Stories Corpus* with reading times for 485 sentences (Futrell, 2017)
- A Bayesian linear mixed model was fit using **Transformer- surprisal** and **global attention entropy** as predictors computed for every word.
  - Surprisal for each word is computed based on full prefix information.
  - Global attention entropy is computed at the point of the word being processed.
  - We include these predictors for the previous word to account for spillover effects

$$
\begin{aligned}
&RT \sim \\
&loc * wlength\_z + \mathbf{surprisal\_z} * wlength\_z + \mathbf{entropy\_z} * wlength\_z + wfreq\_z * wlength\_z + \mathbf{prev\_surprisal\_z} * \\
&prev\_wlength\_z + \ \mathbf{prev\_entropy\_z} * prev\_wlength\_z + prev\_wfreq\_z * prev\_wlength\_z + (1 + surprisal\_z + \\
&entropy\_z + wfreq\_z + prev\_surprisal\_z + prev\_entropy\_z + prev\_wfreq\_z + loc|WorkerId) + (1|word)
\end{aligned}
$$

Modeling self-paced reading times in the Natural Stories Corpus

We find surprisal, surprisal spillover, and **entropy spillover effects**.

The early surprisal effects are consistent with probabilistic expectations influencing early word identification, and the entropy spill-over effects consistent with entropy influencing later dependency formation.



Let's consider Transformer accounts of some phenomena that have challenged conventional grammar-based surprisal and cue-based retrieval models:

- Facilitatory interference in subject-verb agreement
- Strength of garden path effects in NP/S vs. NP/Z ambiguities
- Localized difficulty in object vs. subject relative clauses

---

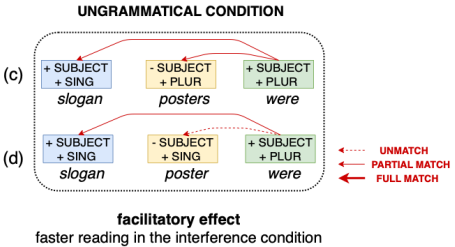
We'll analyse and visualize entropy in a specific attention head specialized for subject-verb relations, and then return to the global entropy measures at the end.

Interference effects in subject-verb agreement

(a) The **slogan** on the poster **was** designed by ...	[interfering, grammatical]
(b) The **slogan** on the posters **was** designed by ...	[non-interfering, grammatical]
(c) The **slogan** on the posters **were** designed by ...	[interfering, ungrammatical]
(d) The **slogan** on the poster **were** designed by ...	[non-interfering, ungrammatical]

Cue-based retrieval theory predicts **facilitatory** interference effects (speedups) in the ungrammatical cases, and **inhibitory** effects (slowdowns) in the gramamtical cases.

Grammar-based surprisal predicts **no effects** because the distractor NP does not participate in the subject relation that affects the prediction of verb number.



Interference effects in subject-verb agreement

(a) The **slogan** on the poster **was** designed by ... [interfering, grammatical]

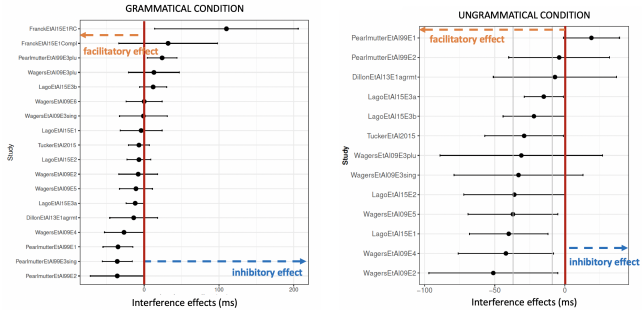
(b) The **slogan** on the posters **was** designed by ... [non-interfering, grammatical]

(c) The **slogan** on the posters **were** designed by ... [interfering, ungrammatical]

(d) The **slogan** on the poster **were** designed by ... [non-interfering, ungrammatical]

But human reading times show **no reliable inhibitory effects** and **reliable facilitatory effects**...

...as shown in this recent meta analysis Vasishth & Englemann (2020)



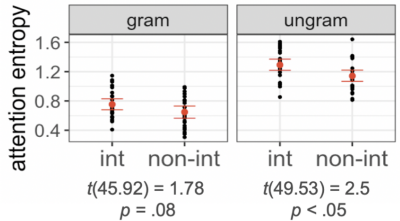
Transformer account of agreement interference

Method: We computed **Transformer(GPT2)-surprisal** and **attention entropy** measures for the critical verbs in all 120 sentences from experiments with available materials from the Vasishth & Englemann (2020) meta-analysis.

	Interference	Grammaticality	Example sentences
Wagers 2009 Exp 2-3	int	gram	The commentator who the **viewer trusts** ...
	unint	gram	The commentators who the **viewer trusts** ...
	int	ungram	*The commentators who the **viewer trust** ...
	unint	ungram	*The commentator who the **viewer trust** ...
Wagers (2009) Exp 4-6	int	gram	The **slogan** on the poster **was** designed ...
	unint	gram	The **slogan** on the posters **was** designed ...
	int	ungram	*The **slogan** on the posters **were** designed ...
	unint	ungram	*The **slogan** on the poster **were** designed ...
Dillon 2013 Exp 1 agrmt	int	gram	The **executive** who oversaw the middle manager apparently **was** dishonest ...
	unint	gram	The **executive** who oversaw the middle managers apparently **was** dishonest ...
	int	ungram	*The **executive** who oversaw the middle managers apparently **were** dishonest ...
	unint	ungram	*The **executive** who oversaw the middle manager apparently **were** dishonest ...

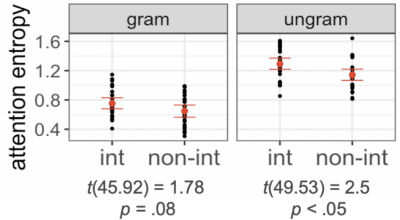
Transformer account of agreement interference

Results: Across three experiments (example at right) found **attention entropy** differences in both grammatical and ungrammatical conditions, consistent with cue-based retrieval predictions.

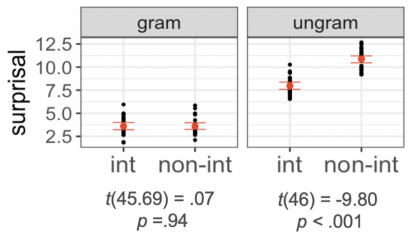


Transformer account of agreement interference

Results: Across three experiments (example at right) found **attention entropy** differences in both grammatical and ungrammatical conditions, consistent with cue-based retrieval predictions.

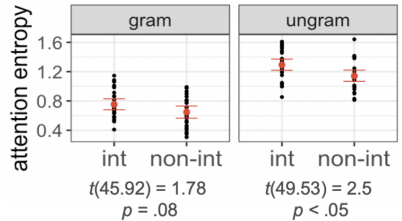


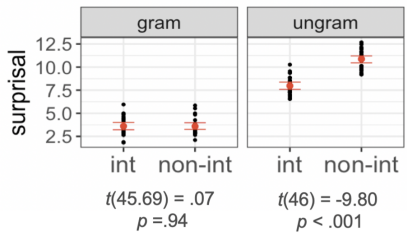
We found **surprisal** differences in the **ungrammatical but not grammatical conditions**, consistent with facilitatory interference in human data (and inconsistent with conventional grammar-based surprisal).



Transformer account of agreement interference

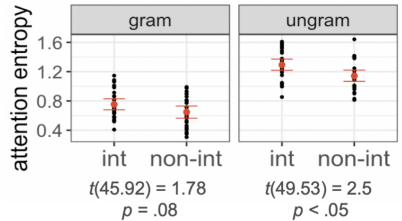
Is the entropy effect in the grammatical condition inconsistent with human reading times? We believe it is not. . . because—as we will see later—the *global entropy* effects here are relatively small.



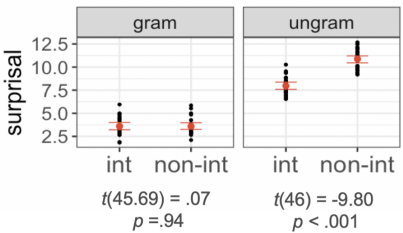


Transformer account of agreement interference

Is the entropy effect in the grammatical condition inconsistent with human reading times? We believe it is not. . . because—as we will see later—the *global entropy* effects here are relatively small.



Is the surprisal effect in ungrammatical conditions due to training on *ungrammatical agreement attraction errors?* We believe it is not, because the actual occurrence of such errors would lead to surprisal effects 3-4x smaller than what we observe.



Transformer account of NP/Z, NP/S garden paths

	Ambiguity	Example sentences
NP/S	amb	The worker maintained the **walls fell** down in a heap before he ...
	unamb	The worker maintained that the **walls fell** down in a heap ...
NP/Z	amb	Even though the girl phoned the **instructor was** upset ...
	unamb	Even though the girl phoned, the **instructor was** upset ...

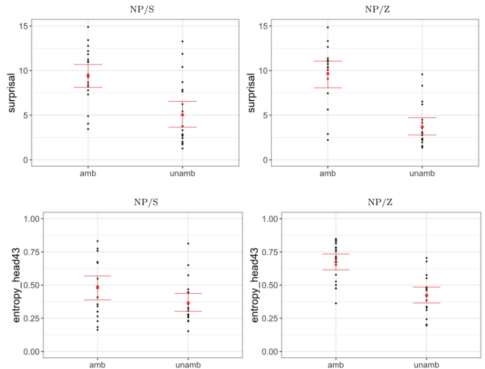
GP effect is stronger in NP/Z than NP/S
(Frazier & Rayner, 1982; Grodner et al, 2003);
previous neural network models do not predict
this contrast (Van Schijndel & Linzen, 2018).

Transformer account of NP/Z, NP/S garden paths

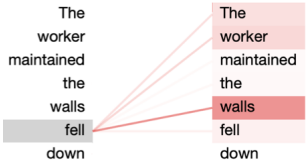
	Ambiguity	Example sentences
NP/S	amb	The worker maintained the **walls fell** down in a heap before he ...
	unamb	The worker maintained that the **walls fell** down in a heap ...
NP/Z	amb	Even though the girl phoned the **instructor was** upset ...
	unamb	Even though the girl phoned, the **instructor was** upset ...

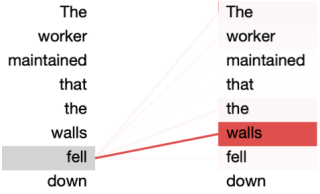
GP effect is stronger in NP/Z than NP/S (Frazier & Rayner, 1982; Grodner et al, 2003); previous neural network models do not predict this contrast (Van Schijndel & Linzen, 2018).

Results: We find both larger Transformer-surprisal effects (≈ 2 bits) and larger subect-verb attention head entropy in NP/Z than NP/S.

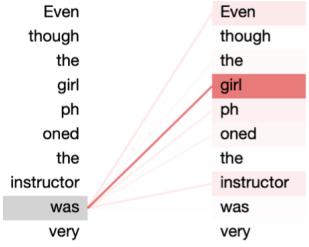


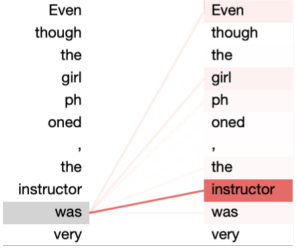
(a) The worker maintained the walls **fell** ... [NP/S ambiguous]
(b) The worker maintained that the walls **fell** ... [NP/S unambiguous]





(c) Even though the girl phoned the instructor **was** ... [NP/Z ambiguous]
(d) Even though the girl phoned, the instructor **was** ... [NP/Z unambiguous]





English object- vs. subject-relative clauses

	Example sentences
SRC	The **reporter**$_i$ that __$_i$ **attacked** the senator admitted the error.
ORC	The **reporter**$_i$ that the senator **attacked** __$_i$ admitted the error.

(e.g, Staub, 2010)

difficulty (ORC) $>$ difficulty (SRC), at embedded verb

Memory-based theories (e.g. Dependency Locality Theory, (Gibson, 2000), cue-based retrieval models) predict difficulty at the embedded verb (*attacked*).

Grammar-based surprisal theory predicts increased difficulty at the RC noun onset (*the*) because that is where the expectation for *that* to be a subject relative pronoun is violated.

English object- vs. subject-relative clauses

	Example sentences
SRC	The **reporter**$_i$ that __$_i$ **attacked** the senator admitted the error.
ORC	The **reporter**$_i$ that the senator **attacked** __$_i$ admitted the error.

(e.g, Staub, 2010)

difficulty (ORC) $>$ difficulty (SRC), at embedded verb

Memory-based theories (e.g. Dependency Locality Theory, (Gibson, 2000), cue-based retrieval models) predict difficulty at the embedded verb (*attacked*).

Grammar-based surprisal theory predicts increased difficulty at the RC noun onset (*the*) because that is where the expectation for *that* to be a subject relative pronoun is violated.

Human data are mostly consistent with what memory-based theories predict; this has been taken as an important explanatory gap for surprisal accounts (Levy & Gibson, 2013).

Transformer account of ORC vs. SRC

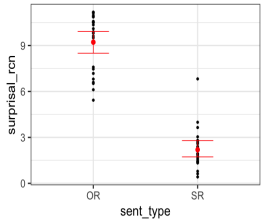
	Example sentences
SRC	The **reporter**$_i$ that __$_i$ **attacked** the senator admitted the error.
ORC	The **reporter**$_i$ that the senator **attacked** __$_i$ admitted the error.

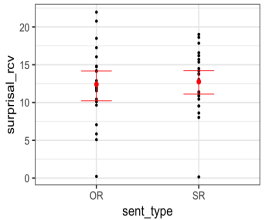
Results: Transformer-surprisals are higher at ORC NP onset (consistent with previous surprisal accounts) and equivalent at the verb, while **Transformer global entropy is higher at the ORC verb** (and lower at ORC NP onset), consistent with human RTs.

Transformer account of ORC vs. SRC

	Example sentences
SRC	The **reporter**$_i$ that __$_i$ **attacked** the senator admitted the error.
ORC	The **reporter**$_i$ that the senator **attacked** __$_i$ admitted the error.

Results: Transformer-surprisals are higher at ORC NP onset (consistent with previous surprisal accounts) and equivalent at the verb, while **Transformer global entropy is higher at the ORC verb** (and lower at ORC NP onset), consistent with human RTs.



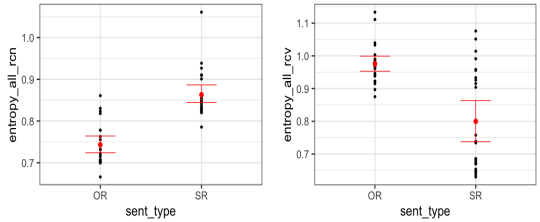


the *attacked*

Transformer account of ORC vs. SRC

	Example sentences
SRC	The **reporter**$_i$ that __$_i$ **attacked** the senator admitted the error.
ORC	The **reporter**$_i$ that the senator **attacked** __$_i$ admitted the error.

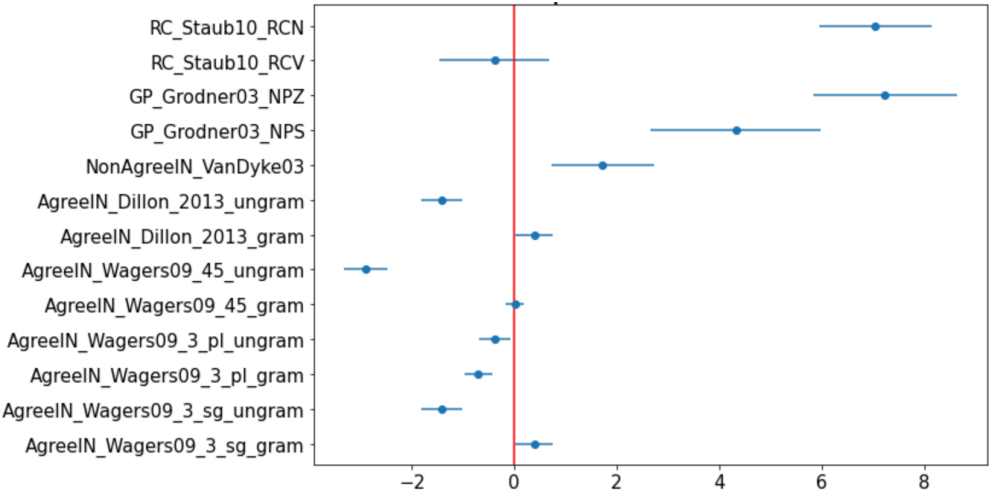
Results: Transformer-surprisals are higher at ORC NP onset (consistent with previous surprisal accounts) and equivalent at the verb, while **Transformer global entropy is higher at the ORC verb** (and lower at ORC NP onset), consistent with human RTs.



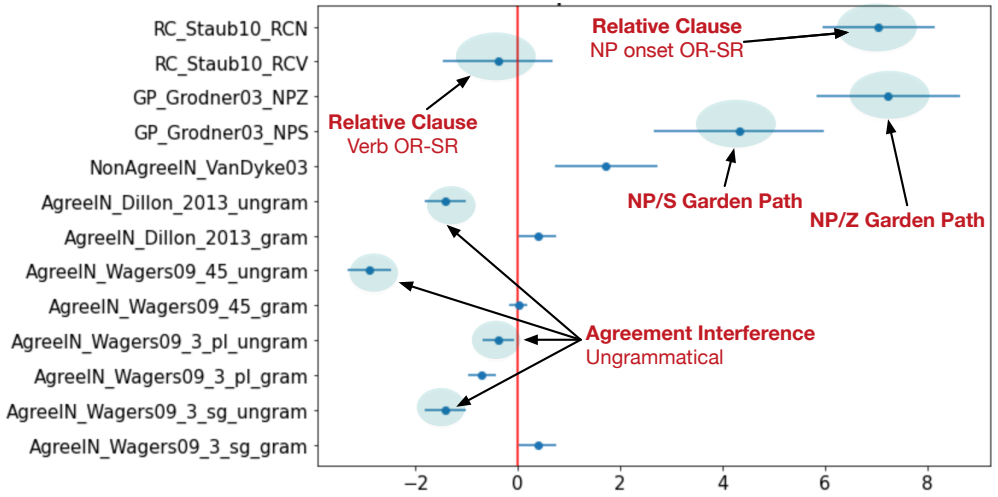
the *attacked*

We conclude now by considering the magnitude of the **surprisal effects across experiments**, and the magnitude of the **global attention entropy effects across experiments**.

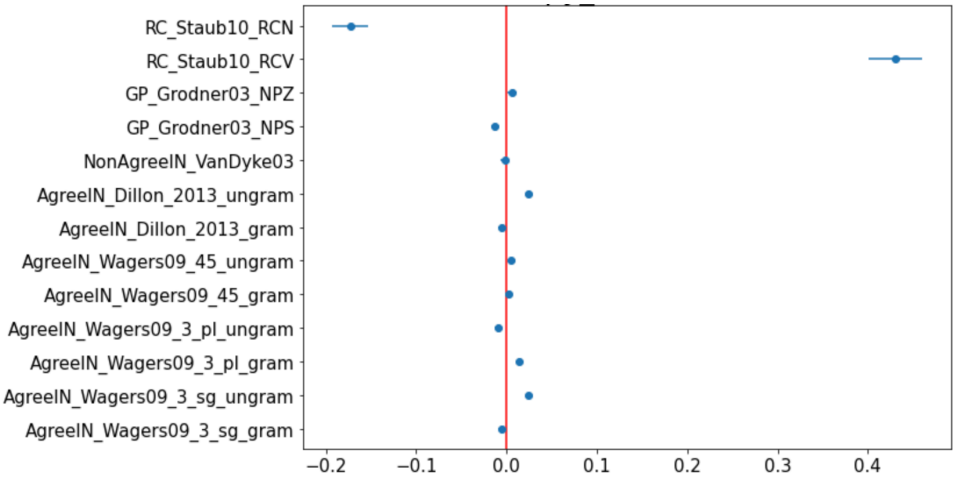
Surprisal effects across the experiments



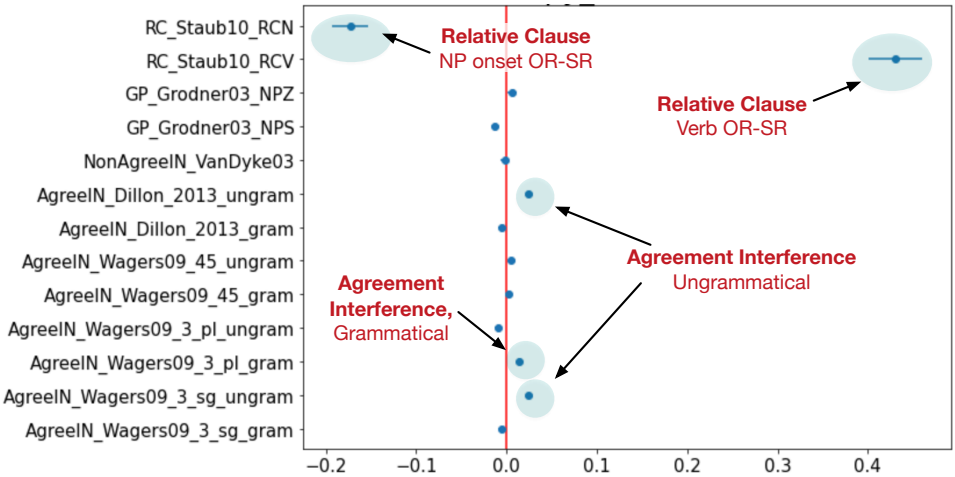
Surprisal effects across the experiments



Global entropy effects across the experiments



Global entropy effects across the experiments



Summary

Transformers are predictive cue-based retrieval parsers.

Summary

Transformers are predictive cue-based retrieval parsers.

Global attention entropy, a processing difficulty metric from the memory-based theoretical perspective, **predicts spillover effects** in naturalistic sentence reading time, independently of surprisal.

Summary

Transformers are predictive cue-based retrieval parsers.

Global attention entropy, a processing difficulty metric from the memory-based theoretical perspective, **predicts spillover effects** in naturalistic sentence reading time, independently of surprisal.

Transformer-surprisal makes different predictions from conventional grammar-based surprisal, in ways that show the **influence of similarity-based interference** as revealed by attention entropy.

Summary

Transformers are predictive cue-based retrieval parsers.

Global attention entropy, a processing difficulty metric from the memory-based theoretical perspective, **predicts spillover effects** in naturalistic sentence reading time, independently of surprisal.

Transformer-surprisal makes different predictions from conventional grammar-based surprisal, in ways that show the **influence of similarity-based interference** as revealed by attention entropy.

Transformer global attention entropy and Transformer-surprisal together provide an integrated account of a range of psycholinguistic phenomena that are not explained by standard expectation-based and memory-based theories alone.

Limitations and future steps

- We should investigate the effects of different aggregations of attention heads; 2/3 of attention heads can be pruned without affecting performance in language tasks. (Voita et al., 2019)
- Parameter estimates from large scale corpora should be used to make quantitative predictions in controlled experiments.
- We should directly manipulate attention patterns to assess their causal role in modulating surprisal.
- Transformers have **no dynamics**; dynamical models should be constructed in which both surprisal and entropy have natural, computationally rational, adaptive effects on eye-movements and button-presses (eliminate "linking assumptions") in a range of task settings.

Thank you!

Speculations

- Similarity-based interference is the **price paid for the capacity for generalization.**
- Similarity-based interference exerts influence on reading times indirectly through **both surprisal** (because probabilistic expectations are predicated on representation similarity) and **attention entropy**.
- Surprisal exerts influence on reading time directly through rational word identification processes.
- Global attention entropy exerts influence on reading time directly through **a global signal used to inhibit/slow eye-movements/button presses** so that dynamic memory processes may have time to reduce interference.

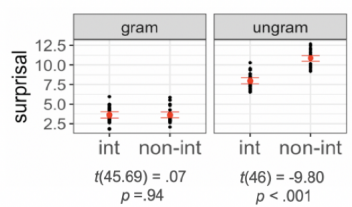
Backup slides

Interfering distractors give rise to agreement attraction errors in production (Bock & Miller, 1991).

Can we explain facilitatory effects with such agreement attraction errors in the training corpus?

- We observed 3 bit difference in surprisals
- Preliminary corpus analysis showed that interfering distractors occur about twice as often as non-interfering distractors, which is 1 bit difference
- **The asymmetry of the numbers of distractor types (interfering or non-interfering) in ungrammatical sentences cannot fully explain the observed facilitatory interference effects**

	singular subj
interfering	80
non-interfering	39



Transformers' accounts on psycholinguistic phenomena as a whole