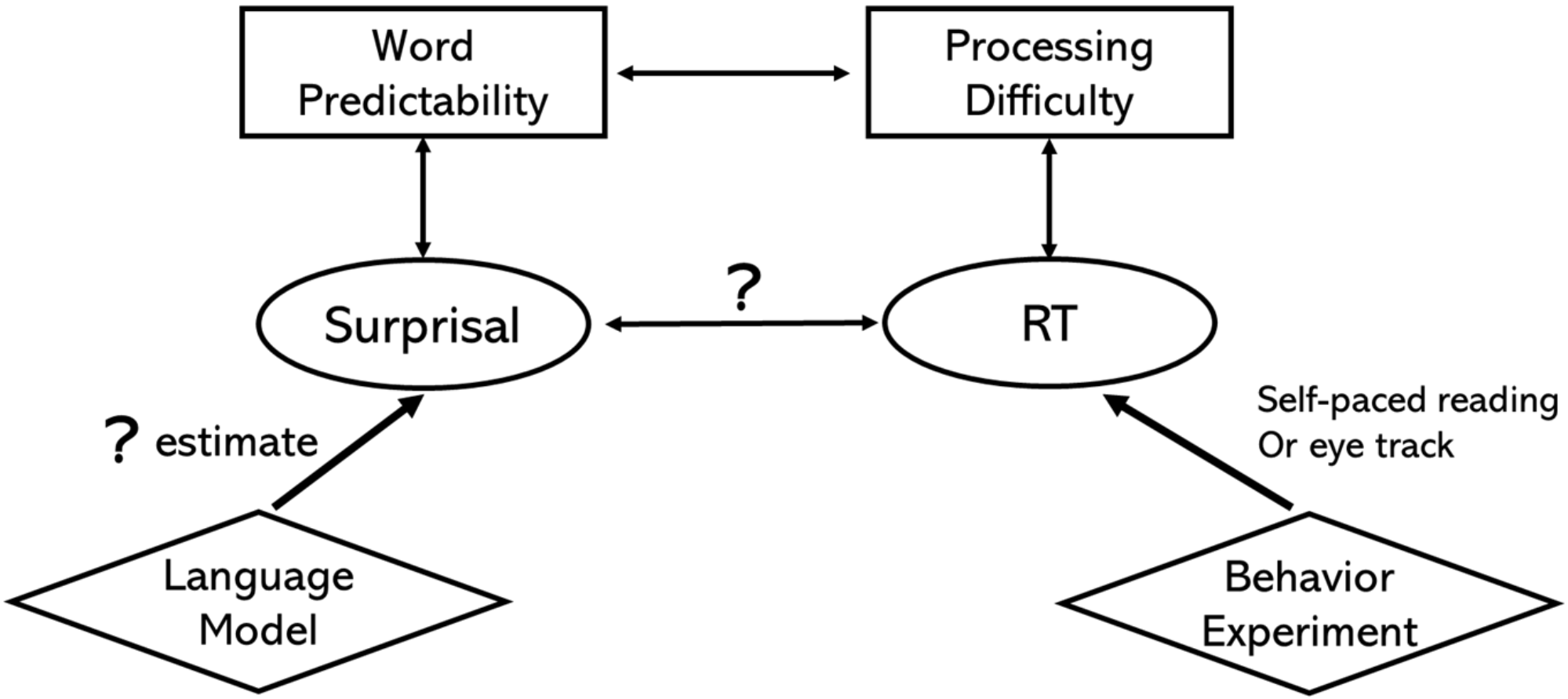
Surprisal Theory & Reading Processing

张子涵
zihan046@outlook.com
2023/12/20

Introduction

- **Surprisal Theory对于Reading Process的基本解释** (Hale, 2001; Levy, 2008) **：**
  - 在人们进行阅读时，每遇到一个新词会根据其加工难度动态地分配给其一定的阅读时间 (Reading Time, RT)，其中对于词的加工难度起关键作用的是词的可预测性 (Predictability)，它可以用惊异度 (Surprisal, $s = log\ p(w|c)$)进行量化表示；
- **既往实证研究的基本结果** (Smith and Levy, 2013; Wilcox et al., 2020; Shain et al., 2022)**：**
  - Surprisal可以预测RT，且二者之间的关系是线性的；
  - 前置词的Surprisal也会影响当前词的RT，即Surprsial对于RT的影响存在溢出效应 (spillover effect);
- **对于Surprisal Theory的扩展** (Pimentel et al., 2023; Cevoli et al., 2022):
  - 加工难度不仅由Surprisal决定，人们的阅读行表现也对Expected Surprisal敏感（即对于一个词的加工难度的预期）。

Procedure


Word Predictability
Processing Difficulty
Surprisal
?
RT
? estimate
Self-paced reading
Or eye track
Language Model
Behavior Experiment

Line of Research

- Surprisal和RT之间的关系以及其深层的认知机制？(Pimentel et al. 2023; Wilcox et al., 2023)
- 语言模型能否可靠地估计surprisal？(Oh et al., 2022, 2023)
- 将阅读加工难度扩展到其他更具体的问题，如消歧的难度等 (Van Schijndel et al., 2021)

Pimentel et al. (2023)

On the Effect of Anticipation on Reading Times

Tiago Pimentel 💭 **Clara Meister** 👓 **Ethan G. Wilcox** 👓 **Roger P. Levy** 📖 **Ryan Cotterell** 👓

💭University of Cambridge 👓ETH Zürich 📖MIT

tp472@cam.ac.uk clara.meister@inf.ethz.ch ethan.wilcox@inf.ethz.ch

rplevy@mit.edu ryan.cotterell@inf.ethz.ch

Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger Levy, and Ryan Cotterell. 2023. On the effect of anticipation on reading times. Transactions of the Association for Computational Linguistics.

Pimentel, T., Meister, C., Wilcox, E. G., Levy, R., & Cotterell, R. (2022). On the Effect of Anticipation on Reading Times. *arXiv preprint arXiv:2211.14301*.

Introduction

- **Responsive reading process**
  - 读者确认一个词，然后根据其需求动态地分配一定的时间用于处理它；
  - 阅读时间(reading time, RT)受到很多因素的影响：如，其长度(length)和惊异度(surprisal)和其呈线性正相关(Hale, 2001; Smith and Levy, 2008; Shain, 2019);
- **Anticipatory reading process**
  - 在读者确认一个词之前，就已经有了对于这个词的预期(expectation)，这种预期（有可能和实际不符）会影响这个词的实际阅读表现（即实际分配的阅读时间）；
  - 支持有预期的阅读的证据：如眼动记录到的word skipping (Ehrlich and Rayner, 1981; Schotter et al., 2012);
  - 研究问题：**对于未确认词的预期如何影响其实际的阅读表现？**

Operationalized Anticipation

- **量化Anticipation时需要考虑的问题**
  - 读者对于未确认词的surprisal进行期望，即对于未确认词需要的RT进行期望；
  - 读者进行预期的策略可能存在差异；
- **Contextual Rényi entropy (Rényi, 1961)**

$$\mathrm{H}_\alpha(W_t \mid \boldsymbol{W}_{<t} = \boldsymbol{w}_{<t}) \stackrel{\text{def}}{=} \lim_{\beta \to \alpha} \frac{1}{1-\beta} \log_2 \sum_{w \in \overline{\mathcal{W}}} \Big(p(w \mid \boldsymbol{w}_{<t}) \Big)^{\beta} \quad (1)$$

*** 三种可以进行直观解释的预期策略**

(2) 读者预期下一个词的surprisal是所有可能的数学期望(Shannon entropy)；

(3) 读者预期下一个词的surprisal时不关注具体的p值，而关注可能出现的词的个数；

(4) 读者预期下一个词的的surprisal是出现概率最高的。

If a = 1

$$\begin{aligned} \mathrm{H}(W_t \mid \boldsymbol{W}_{<t} = \boldsymbol{w}_{<t}) &\stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{w \sim p(\cdot \mid \boldsymbol{w}_{<t})} [h_t(w)] \\ &= -\sum_{w \in \overline{\mathcal{W}}} p(w \mid \boldsymbol{w}_{<t}) \log_2 p(w \mid \boldsymbol{w}_{<t}) \end{aligned} \quad (2)$$

If a = 0

$$\mathrm{H}_0(W_t \mid \boldsymbol{W}_{<t} = \boldsymbol{w}_{<t}) \stackrel{\text{def}}{=} -\log_2 \frac{1}{|\mathrm{supp}(p)|} \quad (3)$$

If a = ∞

$$\begin{aligned} \mathrm{H}_\infty(W_t \mid \boldsymbol{W}_{<t} = \boldsymbol{w}_{<t}) &\stackrel{\text{def}}{=} \min_{w \in \overline{\mathcal{W}}} h_t(w) \\ &= \min_{w \in \overline{\mathcal{W}}} -\log_2 p(w \mid \boldsymbol{w}_{<t}) \end{aligned} \quad (4)$$

$$\mathrm{supp}(p) \stackrel{\text{def}}{=} \{w \in \overline{\mathcal{W}} \mid p(w \mid \boldsymbol{w}_{<t}) > 0\}$$

Hypothesis: Anticipatory Mechanisms

- **Word skipping**

由于读者必须在确认一个词之前决定是否跳过它，因此我们假设当读者对于一个词很确信时（即contextual entropy较低时，下同）会跳过它，因此contextual entropy应该能够很好地预测word skipping；

- **Budgeting**

由于一个词被分配的RT不能无限大，所以可以假设RT的分配在一个词尚未被确认前存在预算，且是可以被contextual entropy预测的；同时，若读者对某个词的RT预算不足（contextual entropy < real surprisal），则在接下来的单词中可能代偿的表现为更大的溢出效应(spillover effect)；

- **Preemptive Processing**

由于对于确定性强的文本，大脑会在抵达下一个词之前对它进行提前处理(e.g., Willems et al., 2015; Goldstein et al., 2022)，因此，我们可以假设surprisal低的词具有较短的RT是因为其RT代偿到了前一个词；

- **Uncertainty Cost**

由于对于一个词的不确定性会导致加工负荷的增长，因此可以假设在头脑中保持众多对于下一个词的不确定性预测是很消耗性的行为，在当前词的surprisal之外造成了新的负荷（即RT的增加）。

Experimental Setup

I 使用LM估计surprisal和contextual entropy

- 本文使用GPT-2 small (Radford et al., 2019)
- 已有实验表明，在LM的surprisal估计中，并非越大的语言模型越好 (Shain et al., 2022; Oh and Schuler, 2022, Oh et al., 2023)
- 由于GPT-2使用字节对编码(Byte Pair Encoding, BPE)，估计单个词的surprisal时需要将subwords的估计值加和，估计contextual surprisal时则直接在sub words水平上进行计算。

II 使用Regressor在不同数据集上评估surprisal等指标对于RT的预测能力

- 本文使用四个数据集，其中两个是自定义步速阅读数据集，两个是眼动数据集
  - Natural Stories Corpus (Futrell et al., 2018)
  - Brown Corpus (Smith and Levy, 2013)
  - Provo Corpus (Luke and Christianson, 2018)
  - Dundee Corpus (Kennedy et al.,2003)
  - 注：眼动数据集可以将skip word去除(×)，也可以保留记作RT=0 (√)

Experimental Setup

II 使用Regressor在不同数据集上评估surprisal等指标对于RT的预测能力

- 基本思路：
  - 以词的RT为因变量（取所有被试的平均值），以影响该词RT的相关特征（如当前词以及前后的词的词长、surprisal，contextual entropy等）为自变量在数据集上构建线性回归模型，以测试集上预测RT对于实际RT的预测效果评估模型，通过更换特征构建不同的模型以评估特征对于RT的预测效果；
  - 主要对比base model和target model：target model比base model多包含一个待检测的特征（如base model只包含当前词的length和surprisal，而target model还包含当前词的contextual entropy，此时两个模型的效果差异即反映了当前词contextual entropy对于其RT的预测能力）
- 模型评估方法：十折交叉验证 (10-fold cross-validation)
- 模型评估指标：**平均对数似然 (average log-likelihood)**

Experimental Setup

连续（左）：y_n服从正态分布

$$\begin{aligned}\mathrm{llh}(f_{\boldsymbol{\phi}}(\mathbf{x})) &= \frac{1}{N}\log\prod_{n=1}^{N}\frac{e^{-\frac{\left(y_n-f_{\boldsymbol{\phi}}(\mathbf{x}_n)\right)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\\ &= -\frac{1}{N}\sum_{n=1}^{N}\left(\log\sqrt{2\pi\sigma^2}+\frac{(y_n-f_{\boldsymbol{\phi}}(\mathbf{x}_n))^2}{2\sigma^2}\right)\\ &= -\log\sqrt{2\pi\sigma^2}-\sum_{n=1}^{N}\frac{(y_n-f_{\boldsymbol{\phi}}(\mathbf{x}_n))^2}{2N\sigma^2}\end{aligned}$$

离散（右）：y_n服从二项分布

$$\mathrm{llh}(f_{\boldsymbol{\phi}}(\mathbf{x})) = \sum_{n=1}^{N}\frac{y_n\log f_{\boldsymbol{\phi}}(\mathbf{x}_n)+(1-y_n)\log(1-f_{\boldsymbol{\phi}}(\mathbf{x}_n))}{N}$$

$$\Delta_{\mathrm{llh}} = \mathrm{llh}(f_{\boldsymbol{\phi}}(\mathbf{x}^{\mathrm{model}})) - \mathrm{llh}(f_{\boldsymbol{\phi}}(\mathbf{x}^{\mathrm{base}}))$$

- 如果Δ_{llh} >0, 则说明target model效果比base model好，待检测的特征能够预测RT；
- 显著性检验：paired permutation test（置换检验）

Experimental Setup

- **例：“张三/爱/吃/苹果。”**
- 通过语言模型得到p(张三|start)、p(爱|start张三)、p(吃|start张三爱)等，并据此计算每个词在相应的语境下的surprisal；另外还可以得到w ~ p(·|start)、w ~ p(·| start张三)、w ~ p(·|start张三爱)等，并据此计算在每个词的contextual entropy；
- 在数据集上，以每个词的RT为因变量，以与这个词有关的特征为自变量（如词长、surprisal，entropy等）构建回归模型（同时构建base model和target model），通过十折交叉验证得到10个Δ_{llh}，最后通过置换检验比较Δ_{llh}和0是否存在显著差异从而得到某个特征对于RT的预测能力；

Experiment 1： Surprisal ~ RT

		Surprisal			
		w_{t-3}	w_{t-2}	w_{t-1}	w_t
Brown		0.33***	0.47***	2.58***	0.50*
Natural Stories		0.20*	0.34*	1.05***	1.54***
Provo	(✓)	0.07	0.18	0.83*	3.22**
Dundee	(✓)	-0.00	0.04**	0.25***	0.89***

$\mathbf{x}_t^{\text{model}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}}$ vs. $\mathbf{x}_t^{\text{base}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}\neq t'}$

Table 1: Δ_{llh} (in 10^{-2} nats) when comparing a model with all surprisal terms against baselines from which a single surprisal term was removed. Green indicates a significantly positive impact of surprisal on the model's predictive power.

- **实验1：不同位置上surprisal对于RT的预测能力评估**
- **Base model**：词长等已被证实对于RT有影响的基本特征 + 去除待检验位置上surprisal后的其他surprisal ；
- **Target model**：base model + 所有位置上的surprisal；
- **结果：**Surprisal在四个数据集上都能很好的预测RT，并且表现出了明显的溢出效应。

Experiment 2：Surprisal vs. Entropy

		w_{t-3}	w_{t-2}	w_{t-1}	w_t
Replace Surprisal with Entropy[1]					
Brown		-0.30*	-0.35**	-1.68***	-0.03
Natural Stories		-0.03	-0.19*	-0.41*	0.37
Provo	(✓)	-0.08	0.18	-0.66*	-2.58*
Dundee	(✓)	-0.00	0.03*	-0.21***	-0.07
Add Entropy[2]					
Brown		-0.03*	-0.01	0.04	0.15*
Natural Stories		0.04	0.01	0.14***	0.89***
Provo	(✓)	-0.04*	0.16	-0.03	-0.06
Dundee	(✓)	-0.00**	0.03**	-0.00	0.25***

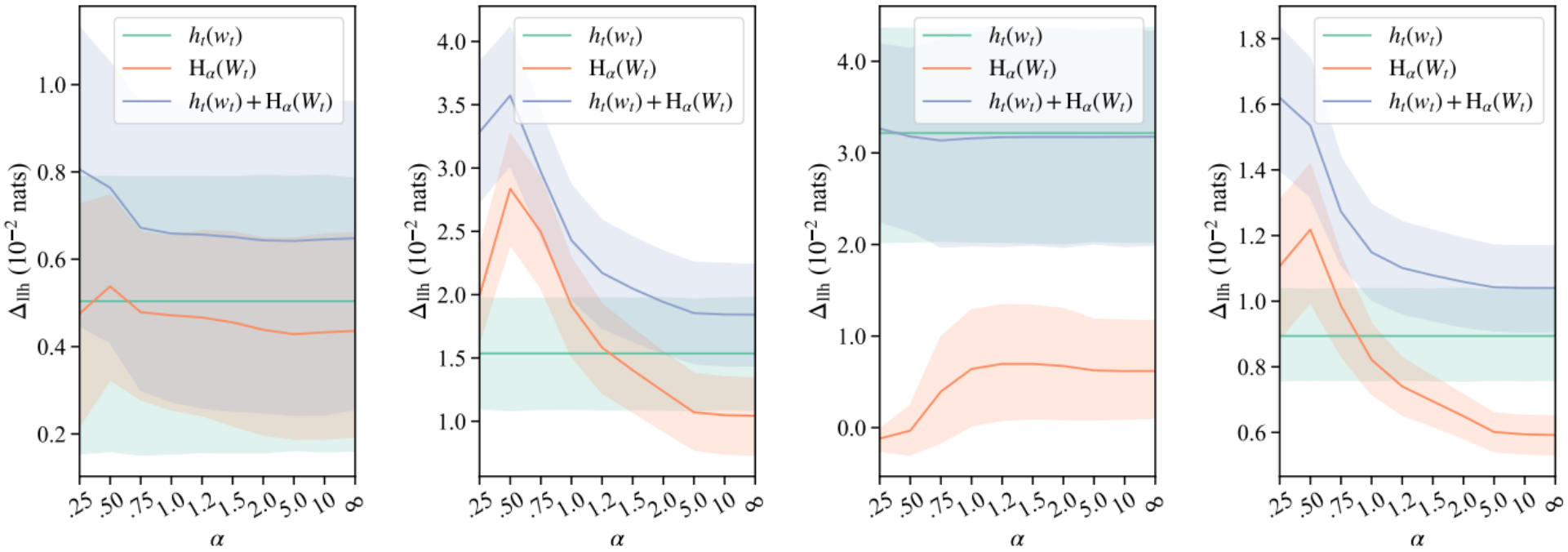
[1] $\mathbf{x}_t^{\text{model}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}\neq t'} \oplus [\mathrm{H}(W_{t'})]^\intercal$
[2] $\mathbf{x}_t^{\text{model}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}} \oplus [\mathrm{H}(W_{t'})]^\intercal$
both $\mathbf{x}_t^{\text{base}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}}$,

- **实验2：不同位置上Contextual Shannon Entropy对于RT的预测能力评估**
- **Base model**：基本特征+所有位置surprisal
- **Target model_1**：base model + 将某一位置的surprisal替换为其Contextual Shannon Entropy
- **Target model_2**：base model + 直接额外增加某一位置的Contextual Shannon Entropy
- **结果：**
- 增加当前词的entropy在3/4的数据集上显著提升了模型的预测能力；替换当前词的surprisal为entropy在1个数据集上使模型的预测能力显著变差，但在其他3个数据集上没有明显差异-> 阅读是responsive & anticipatory；
- 增加前置词的entropy没有显著增加预测能力，但替换后会显著削弱模型的预测能力->entropy的效应是local的，没有surprisal那样的溢出效应。

Experiment 3：Skewed Expectations

- **实验3.1：评估a取不同值时的Contextual entropy对于RT的预测能力**
- **Base model：**基本特征 + 除当前位置外的其他surprisal
- **Target model_1：**base model + 当前位置surprisal
- **Target model_2：**base mode + 当前位置entropy
- **Target model_3：**base model +当前位置surprisal & entropy
- **结果：**
- 除Provo (✓)外，其他三个数据集上趋势基本一致；
- 在a较小时Contextual entropy的预测能力较好，其在约**a=1/2**处取极大值。

Experiment 3: Skewed Expectations



(a) Brown (b) Natural Stories (c) Provo (✓) (d) Dundee (✓)

Figure 1: Δ_{llh} when adding either the current word's surprisal, Rényi entropy, or both on top of a baseline that includes the surprisal of previous words as predictors, i.e., $\mathbf{x}^{\text{base}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}^{\text{surp}\neq t}$. Shaded regions correspond to 95% confidence intervals.

Experiment 3：Skewed Expectations

- **实验3.2：取a=1/2重复实验2（评估entropy对于RT的预测能力）**
- **实验结果：**
- 在除Provo (✓)外的三个数据集上，增加当前词的entropy都能显著提升模型的预测能力（同a=1）；
- 替换当前词的surprisal为entropy能够在两个数据集上显著提升模型的预测能力（与a=1不同）；->为什么entropy比surprisal的效果好？
- entropy(a=1/2)依旧是local的，没有表现出明显的溢出效应（同a=1）。

Experiment 3: Skewed Expectations

		Replace Surprisal with Rényi Entropy[1]				Add Rényi Entropy[2]			
		w_{t-3}	w_{t-2}	w_{t-1}	w_t	w_{t-3}	w_{t-2}	w_{t-1}	w_t
Brown		-0.35***	-0.37**	-1.76***	0.03	-0.01	0.00	0.14	0.26*
Natural Stories		-0.16	-0.27*	-0.19	1.30**	-0.00	-0.00	0.44***	2.04***
Provo	(✓)	-0.05	0.47	-0.89*	-3.25**	-0.01	0.45*	-0.01	-0.04
Dundee	(✓)	0.00	0.07*	-0.25***	0.32*	-0.00	0.08*	0.05	0.64***

[1] $\mathbf{x}_t^{\text{model}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}\neq t'} \oplus [\mathrm{H}_\alpha(W_{t'})]^\intercal$, [2] $\mathbf{x}_t^{\text{model}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}} \oplus [\mathrm{H}_\alpha(W_{t'})]^\intercal$, both $\mathbf{x}_t^{\text{base}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}}$

Table 3: Δ_{llh} (in 10^{-2} nats) achieved after either replacing a surprisal term in the baseline with the contextual Rényi entropy ($\alpha = 1/2$), or adding the Rényi entropy as an extra predictor.

Experiment 4：Word Skipping

- **实验4：word skipping效应在entropy预测RT中的影响**
- **实验4.1：entropy对于word skipping的预测能力评估**
- **数据集：**Self-paced reading无法记录word skip，只使用眼动数据集；
- **基本思路：**使用逻辑回归判断下一个词是否skip（sigmoid的结果其实是一个词被skip的比率，在数据集中通过一个词被skip的人数比例计算），特征与之前相同；
- **结果（in Dudee, not Provo）：**
- Surprisal可以显著的预测word skip的出现；
- Contextual entropy的预测效果比Surprisal更好；
- 在Surprisal的基础上增加Contextual entropy模型效果变好，但反之不会 -> **word skip仅仅是由Contextual entropy驱动的**；

Experiment 4: Word Skipping

		Provo			Dundee		
		$h_t(w_t)$	$\mathrm{H}_\alpha(W_t)$	Both	$h_t(w_t)$	$\mathrm{H}_\alpha(W_t)$	Both
Shannon ($\alpha = 1$)	$\emptyset$	2.60	1.76	2.86	1.30*	2.50***	2.62***
	$h_t(w_t)$	-	-0.84	0.26	-	1.20	1.32***
	$\mathrm{H}_\alpha(W_t)$	-	-	1.10	-	-	0.12
Rényi ($\alpha = 1/2$)	$\emptyset$	2.60	0.84	2.66	1.30*	5.10***	5.14***
	$h_t(w_t)$	-	-1.76	0.06	-	3.79***	3.83***
	$\mathrm{H}_\alpha(W_t)$	-	-	1.82	-	-	0.04

Table 4: Δ_{llh} (in 10^{-4} nats) between a target model (with predictors on columns) vs baseline (with predictors on row) when predicting whether a word was skipped or not. All models also include the surprisal of the previous words as predictors as well as length and unigram frequencies

Experiment 4：Word Skipping

		w_{t-3}	w_{t-2}	w_{t-1}	w_t
Shannon Entropy ($\alpha = 1$)					
Replace[1]	Provo	0.02	-0.03	-0.18	-2.23***
	Dundee	-0.01	-0.02	-0.15***	-0.32**
Add[2]	Provo	-0.02	-0.07	0.03	-0.01
	Dundee	-0.00*	0.01	0.01	0.17**
Renyi Entropy ($\alpha = 1/2$)					
Replace[1]	Provo	0.02	0.10	-0.27	-2.43**
	Dundee	-0.01	0.01	-0.19***	-0.18
Add[2]	Provo	-0.02	0.07	0.01	0.32
	Dundee	-0.00*	0.05*	0.01	0.36***

[1] $\mathbf{x}_t^{\text{model}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}\neq t'} \oplus [\mathrm{H}_\alpha(W_{t'})]^\intercal$,
[2] $\mathbf{x}_t^{\text{model}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}} \oplus [\mathrm{H}_\alpha(W_{t'})]^\intercal$,
both $\mathbf{x}_t^{\text{base}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}}$

Table 5: Δ_{llh} (in 10^{-2} nats) when predicting RTs on eye-tracking datasets where skipped words were removed, i.e., Provo (✗) and Dundee (✗).

- **实验4.2：word skip在entropy对于RT的预测中的效应**
- **基本思路：**之前的实验数据中包含了skip word的数据（即令RT=0ms），由于实验4.1已经证明了entropy可以预测skip，那么之前的实验结果（即entropy可以预测RT）就有可能完全被word skip效应解释。因此在此实验中直接剔除skip word数据，并重复实验2&3；
- **结果：**
- Contextual entropy的预测效果没有surprisal好；
- 在surprisal的基础上增加contextual entropy，依旧能够增强预测效果；
- 因此，contextual entropy对于RT的预测能力确实可以部分地用word skip来解释，然而还有其他机制的影响。

Experiment 5： Budgeting Effects

- **实验5： Budgeting效应在entropy预测RT中的影响**
- Budgeting效应的相关量化特征：

$$h_{t-1}(w_{t-1}) - \mathrm{H}(W_{t-1}) \quad (\Delta\text{-budget}) \quad (14a)$$
$$\mathrm{r}(h_{t-1}(w_{t-1}) - \mathrm{H}(W_{t-1})) \quad (\text{under-budget}) \quad (14b)$$
$$\mathrm{r}(\mathrm{H}(W_{t-1}) - h_{t-1}(w_{t-1})) \quad (\text{over-budget}) \quad (14c)$$
$$|h_{t-1}(w_{t-1}) - \mathrm{H}(W_{t-1})| \quad (|\cdot|\text{-budget}) \quad (14d)$$

r(x) = max(0, x)

以t-1处的suprisal – entropy为例:

如果 > 0，即预期小于实际，就是under-budget

如果 < 0，即预期大于实际，就是over-budget

- **Base model：** 基本特征+所有位置的surprisal+当前位置的entropy；
- **Target model：** base model + 不同位置上的budgeting相关特征；
- **结果：**
- 在两个self-paced corpus和Dundee(√)数据集上观察到了t-1位置的budget effect;
- 在Dundee(×)观察到了t-2位置的效应；
- 只能说明budget效应可能会影响word skip，无法为budget影响RT提供明确的证据；

Experiment 5: Budgeting Effects

	Δ-budget			Over-budget			Under-budget			|·|-budget		
	w_{t-3}	w_{t-2}	w_{t-1}	w_{t-3}	w_{t-2}	w_{t-1}	w_{t-3}	w_{t-2}	w_{t-1}	w_{t-3}	w_{t-2}	w_{t-1}
Shannon Entropy ($\alpha = 1$)												
Brown	-0.03	-0.01	0.02	-0.01	-0.01	0.07	-0.02	-0.01***	-0.02**	0.01	-0.01	0.03
Natural Stories	0.04	0.00	0.05	-0.01	-0.00	0.02	0.04	-0.00	0.02	-0.01	-0.01	-0.02*
Provo (✓)	-0.04***	0.16	-0.04	-0.03**	0.06	-0.03	-0.03	0.07	-0.02	-0.01	-0.04	-0.01
Dundee (✓)	-0.00	0.03**	0.01	-0.00	0.02	0.04	-0.00	0.01	-0.00	-0.00	-0.00	0.03*
Provo (✗)	-0.02***	-0.05	0.07	-0.01	0.05	-0.02	-0.04	-0.06	0.10	-0.03	0.04	0.03
Dundee (✗)	-0.00**	0.01	0.00	-0.00	0.01	-0.00	-0.00	0.00	0.02	0.00	-0.00	0.02
Renyi Entropy ($\alpha = 1/2$)												
Brown	-0.00	-0.00	0.11	0.01	0.00	0.14	-0.01	-0.01*	-0.02	0.01	0.00	0.16
Natural Stories	-0.00	-0.01	0.21***	-0.01	-0.01*	0.23***	0.00	-0.00	-0.01	-0.02	-0.01*	0.21**
Provo (✓)	-0.01	0.48*	-0.03	-0.01	0.42*	-0.02	-0.02	0.05	-0.04**	-0.02	0.26	-0.01
Dundee (✓)	-0.00	0.08*	0.09*	-0.00	0.07*	0.10**	-0.00***	0.01	-0.00***	-0.00	0.06	0.10**
Provo (✗)	-0.01	0.10	0.04	-0.03	0.15	0.01	0.03	-0.03*	0.04	-0.05	0.15	-0.01
Dundee (✗)	-0.00	0.04*	0.00	-0.00	0.04*	-0.00	-0.00	-0.00	0.01	-0.00***	0.04*	-0.00

$\mathbf{x}_t^{\text{base}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}} \oplus [\mathrm{H}_\alpha(W_t)]^\intercal$

Table 6: Δ_{llh} (in 10^{-2} nats) achieved when predicting RTs after adding budgeting effect predictors on top of a baseline with entropy and surprisal as predictors.

Experiment 6： Preemptive Processing

		Entropy[1]		Successor Entropy[2]	
		$\emptyset$ [3]	$[\mathrm{H}_\alpha(W_{t+1})]$[4]	$\emptyset$ [3]	$[\mathrm{H}_\alpha(W_t)]$[5]
Shannon Entropy ($\alpha = 1$)					
Brown		0.15*	0.14*	0.01	-0.01
Natural Stories		0.89***	0.44*	2.27***	1.83***
Provo	(✓)	-0.06	-0.05	-0.06	-0.06*
Dundee	(✓)	0.25***	0.26***	-0.00	-0.00
Provo	(✗)	-0.01	0.01	-0.08*	-0.06
Dundee	(✗)	0.17**	0.16**	0.02	0.00
Renyi Entropy ($\alpha = 1/2$)					
Brown		0.26*	0.27*	-0.01	-0.00
Natural Stories		2.04***	1.52***	1.95***	1.44***
Provo	(✓)	-0.04	-0.01	-0.03	-0.00
Dundee	(✓)	0.64***	0.64***	0.00	-0.00
Provo	(✗)	0.32	0.38	0.06	0.12
Dundee	(✗)	0.36***	0.34***	0.03	0.01

[1] $\mathbf{x}_t^{\text{model}} = \mathbf{x}_t^{\text{base}} \oplus [\mathrm{H}_\alpha(W_t)]^\intercal$, [2] $\mathbf{x}_t^{\text{model}} = \mathbf{x}_t^{\text{base}} \oplus [\mathrm{H}_\alpha(W_{t+1})]^\intercal$,
[3] $\mathbf{x}_t^{\text{base}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}}$, [4] $\mathbf{x}_t^{\text{base}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}} \oplus [\mathrm{H}_\alpha(W_{t+1})]^\intercal$,
[5] $\mathbf{x}_t^{\text{base}} = \mathbf{x}_t^{\text{cmn}} \oplus \mathbf{x}_t^{\text{surp}} \oplus [\mathrm{H}_\alpha(W_t)]^\intercal$,

Table 7: Δ_{llh} (in 10^{-2} nats) after adding the top predictor to a baseline with the predictors in the column. All models include surprisal as a predictor.

- **实验6： Preemptive在entropy预测RT中的影响**
- **基本思路：**若存在超前加工，即如果t+1位置的entropy很小，则会导致t位置的RT代偿性增长；
- **结果：**
- 在含有w_{t+1}的基础上增加w_t位置的contextual entropy会显著提升模型效果；
- 但在含有w_t/不含任何entropy的基础上增加w_{t+1}的entropy只在Natural Stories一个数据集上显著提升模型预测效果；
- Preemptive Processing不是entropy影响RT的主要机制。

Discussion & Conclusion

- **Summary： contextual entropy影响RT的机制是复杂的，还需要进一步探索。**
- 实验1：当前词的surprisal能够预测其RT，同时前置词的surprisal也能预测当前词的RT（即存在溢出效应）；
- 实验2-3：当前词的contextual entropy能够和surprisal一起预测其RT，且在约a=1/2时预测效果最好，说明阅读过程是responsive & anticipatory的，但contextual entropy的作用仅发生在当前词，不存在溢出效应；
- 实验4：当前词的contextual entropy可以用来预测该词是否skip，word skip也可以部分地解释contextual entropy对于RT的预测能力，但不能完全解释（还存在其他机制）；
- 实验5：没有观察到普遍显著的budget效应，不能有效证明人们是通过计算当前词的contextual entropy来对其RT进行预算，即暂时无法用budget解释contextual entropy对于RT的预测能力；但budget效应可能可以预测word skip；
- 实验6：没有观察到普遍的Preemptive Processing，不认为其是contextual entropy影响RT的主要机制。

Discussion & Conclusion

- **本文存在的问题：**
- 不同数据集、不同语言上表现不同 -> 需要测试更多语言的更多数据集；
- 语言模型估计p值的可靠性问题；
- **尚未解决的问题：**
- Contextual entropy和Surprisal的交互是何以实现的？或，一个词的RT是在确认一个词之后决定的，还是在之前决定的？或者是不同的情况下有不同的选择？
- 本文发现contextual entropy可以预测word skip，但不确定是否存在budget effect：由于word skip可以被视为一种特殊的budget，对其的一种猜测是我们在确认下一个词之前计算其contextual entropy，如果其低于某一个阈值，我们决定分配给它RT = 0 ms（即word skip，此时RT时在确认词之前决定的），其他时候，我们还要参考确认这个词后得到的surprisal。

Wilcox et al. (2023)

Testing the Predictions of Surprisal Theory in 11 Languages

Ethan G. Wilcox[1] **Tiago Pimentel**[2] **Clara Meister**[1] **Ryan Cotterell**[1] **Roger P. Levy**[3]
[1]ETH Zürich [2]University of Cambridge [3]MIT
ethan.wilcox@inf.ethz.ch tp472@cam.ac.uk clara.meister@inf.ethz.ch
ryan.cotterell@inf.ethz.ch rplevy@mit.edu

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451-1470.

Introduction

- **既往研究存在的问题：**
- 绝大部分都是英语研究，仅存在少量其他语言的研究，缺乏跨语言普遍性证据 (e.g., Meister et al. 2021 in Dutch and Kuribayashi et al. 2021, 2022 in Japanese) ；
- **本文研究问题：**在11种语言上进行Surprisal Theory相关假设的验证；
- **研究假设:**

(1) Surprisal是否能够预测RT？

(2) Expected Surprisal (Contextual Entropy)是否能够预测RT？

(3) Surprisal和RT之间的关系是否是线性的?

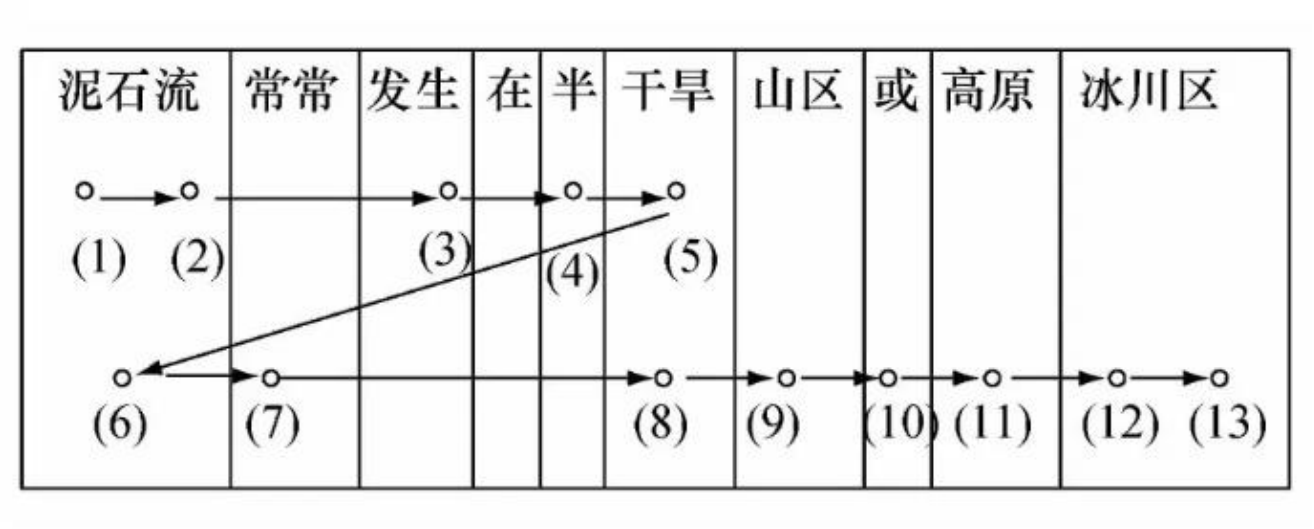
Experimental Setup

- **数据集：** Multilingual Eye Movement Corpus (MECO; Siegelman et al., 2022)
- **眼动指标：** 首次注视时间(first fixation time)、凝视时间(gaze duration)、总注视时间(total fixation time)
- **语言模型（用于估计surprisal等指标）：**
- 单一语言模型 (Monolingual Models)：每种语言使用不同规模的数据集训练两个模型，monoT(all) 使用全部Wiki40B训练(Guo et al., 2020)，monoT(30m) subsample ~ 30 million tokens
- 多语言模型 (Multilingual Model)： 使用mGPT (Shliazhko et al., 2022)，60GB多语言文本在GPT-3构架下训练；

Language	Code	# Training Tokens (mil)
Dutch	du	171
English	en	1,966
Finnish	fi	89
German	ge	883
Greek	gr	57
Hebrew	he	112
Italian	it	376
Korean	ko	75
Russian	ru	488
Spanish	sp	508
Turkish	tr	48

Table 1: Training data information for our monolingual transformer models, noted as *monoT(all)*

Eye-tracking measurements



https://www.zhihu.com/market/pub/120154271/manuscript/1329414252625469440

Skaramagkas, V., Giannakakis, G., Ktistakis, E., Manousos, D., Karatzanis, I., Tachos, N. S., ... & Tsiknakis, M. (2021). Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering*, *16*, 260-277.

- **首次注视时间(first fixation time)：**首次通过某一兴趣区的第一个注视点的注视时间，与兴趣区内的注视点数量无关，如图中的(1)(3)(4)(5)(9)~(12)；
- **凝视时间(gaze duration)：**从第一个注视点开始到注视点首次离开当前兴趣区的持续时间，包括同一兴趣区内的回视，如(1)(2)共同构成了兴趣区“泥石流”的凝视时间；
- **总注视时间(total fixation time)：**落在某一兴趣区的所有注视时间的综合，如(1)(2)(6)共同构成了兴趣区“泥石流”的总注视时间。

Experiment 1 : Surprisal

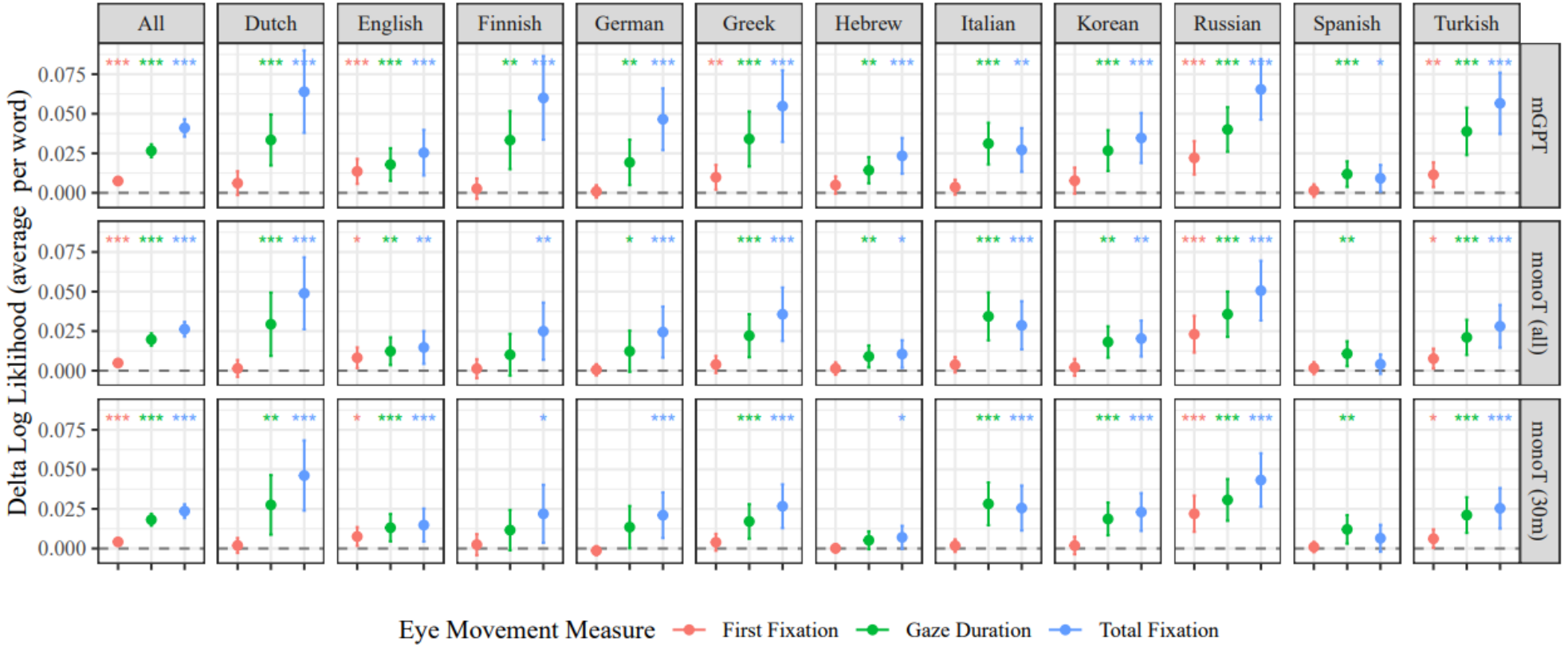


Figure 1: **Psychometric Predictive Power of Surprisal Across Languages:** Positive values mean surprisal

Experiment 1：Surprisal

- **实验1：11种语言的surprisal对于RT的预测能力的评估**
- **结果：**
- gaze duration和total fixation在所有语言上都基本显著，即加入surprisal能够显著跨语言地提升模型预测上述两个指标的能力；
- mGPT估计的surprisal的效应最为稳定，monoT(all)的效果较之monoT(30m)更好->更大的数据规模能够补偿复杂的多模态语言结构预设知识的不足？
- 在预测效果的程度上表现出了语言间的差异(Russian and Dutch上效果明显，Spanish, English, and Hebrew上效果不明显)，但这种差异无法用语言类型学解释；

Experiment 2: Contextual Entropy

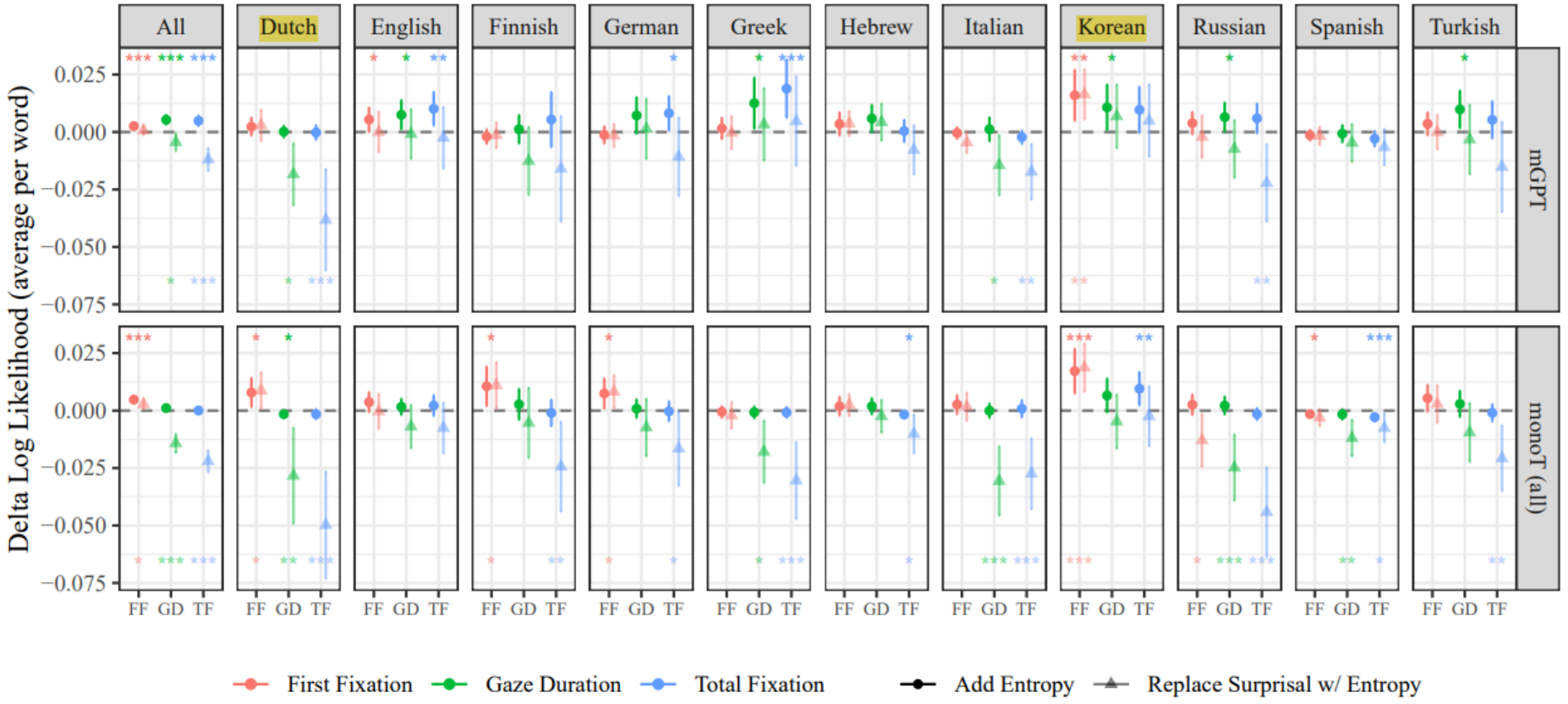


Figure 2: **Psychometric Predictive Power of Contextual Entropy Across Languages:** Positive values mean

Experiment 2：Contextual Entropy

- **实验2：**11种语言的Contextual Entropy对于RT的预测能力的评估
- **结果：**
- 在绝大部分语言中，将surprisal替换为contextual entropy会损害模型对于gaze duration和total fixation的预测能力；
- 同时，额外增加contextual entropy则会带来模型预测能力的提升;
- 这一结果也在语言间表现出了差异：如韩语将surprisal (mGPT)替换为contextual entropy和额外增加entropy都带来了预测效果的提升。

Conclusion & Discussion

- **基本结论：**
- 在所有语言的实验中，surprisal都能提升模型的预测效果；
- 在大部分语言的实验中，替换surprisal为contextual entropy会削弱模型的预测能力，额外加入contextual entropy能够提升模型的预测效果；
- Surprisal和RT之间的底层关系是线性的；
- **尚待解决的问题：**
- 目前的11种语言仍然是Indo-European biased；
- 存在某些跨语言的差异，且这种差异目前无法在语言学上进行解释。

Related Dataset in Chinese

Eye-Movement Measures	Abbreviations	Definition
First fixation duration*	FFD	Duration of the first fixation on the target word
Gaze duration*	GD	Sum of the fixation durations before the target word is exited to the right or left during first-pass reading
First-pass reading fixated proportion*	FPF	Proportion that the target word is fixated during the first-pass reading
Fixation number+	FN	Total number of fixations on the target word
Proportion regression in+	RI	Proportion of regression into the target word
Proportion regression out+	RO	Proportion of regression out from the target word
Saccade length toward the target from the left+	LI_left	Length of saccade into the target word when the word is first fixated from the left side (unit: character)
Saccade length from the target to the right+	LO_right	Length of the saccade from target word to the right after the word first fixated (unit: character)
Total fixation duration+	TT	Sum of the fixation durations on the target word

Table 1. Definitions and Abbreviations of the Nine Eye-Movement Measures. *Note*. *Main measures in the database. +Supplementary measures in the database.

Zhang, G., Yao, P., Ma, G., Wang, J., Zhou, J., Huang, L., ... & Li, X. (2022). The database of eye-movement measures on words in Chinese reading. *Scientific Data*, *9*(1), 411.

Van Schijndel et al. (2021)

Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty

Marten van Schijndel, PhD,[a] Tal Linzen, PhD[b]

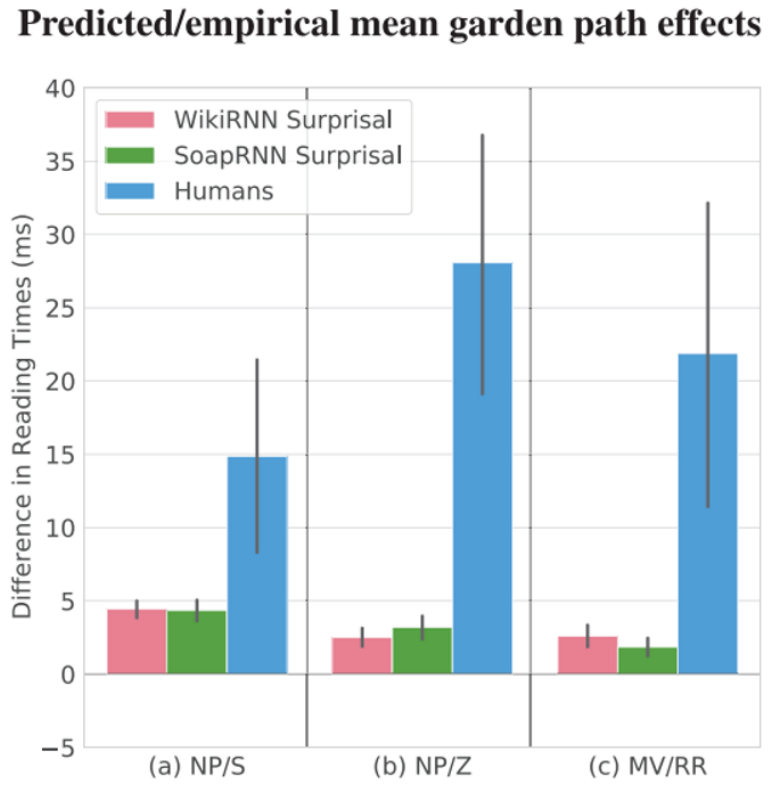
[a]*Department of Linguistics, Cornell University*
[b]*Department of Linguistics and Center for Data Science, New York University*

Received 26 August 2020; received in revised form 21 April 2021; accepted 26 April 2021

Van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive science*, *45*(6), e12988.

Van Schijndel et al. (2021)

- **解释花园路径效应的两种认知理论：**
- **Reanalysis Mechanism：**读者最初只保留所有可能的句法解释的一个子集，随后需要耗时间的重分析来重建一个已经被丢弃的句法解析（two-stage model）；
- **Word Predictability：**花园路径句的的消歧难度被归结于下一个词的可预测性（one-stage model）；
- **结果：**Surprisal成功预测了花园路径效应的存在，但极大地低估了它的程度，并且无法预测花园路径效应在结构间相对的严重性。
- **结论：**对句法消歧难度的充分解释在Word Predictability之外或许仍然需要重分析机制。



Oh et al. (2023)

Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?

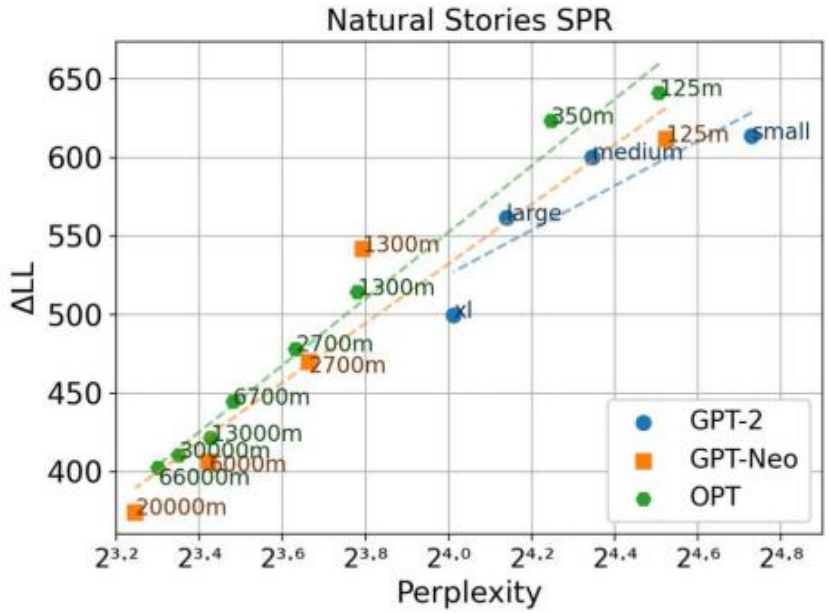
Byung-Doh Oh
Department of Linguistics
The Ohio State University, USA
oh.531@osu.edu

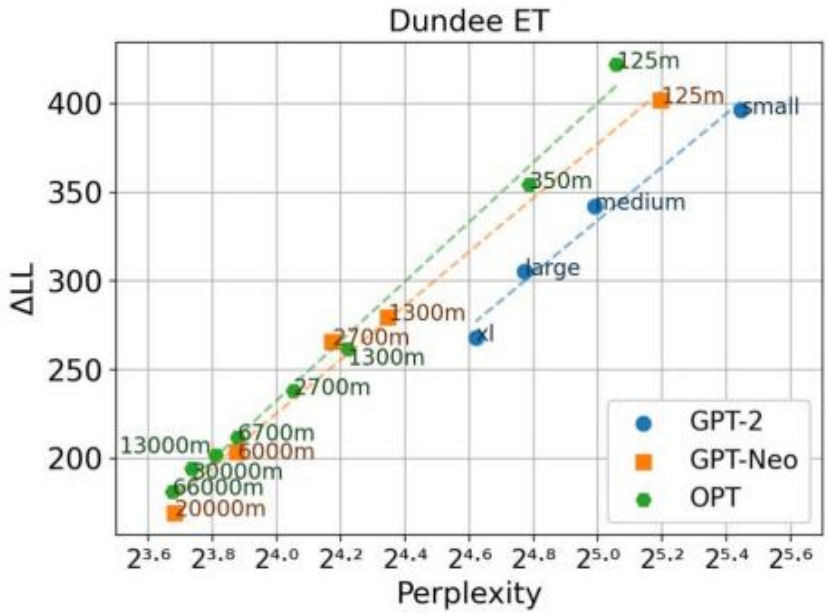
William Schuler
Department of Linguistics
The Ohio State University, USA
schuler.77@osu.edu

Oh, B. D., & Schuler, W. (2023). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?. *Transactions of the Association for Computational Linguistics*, *11*, 336-350.

Model	#L	#H	d_{model}	Parameters
GPT-2 Small	12	12	768	∼124M
GPT-2 Medium	24	16	1024	∼355M
GPT-2 Large	36	20	1280	∼774M
GPT-2 XL	48	25	1600	∼1558M
GPT-Neo 125M	12	12	768	∼125M
GPT-Neo 1300M	24	16	2048	∼1300M
GPT-Neo 2700M	32	20	2560	∼2700M
GPT-J 6B	28	16	4096	∼6000M
GPT-NeoX 20B	44	64	6144	∼20000M
OPT 125M	12	12	768	∼125M
OPT 350M	24	16	1024	∼350M
OPT 1.3B	24	32	2048	∼1300M
OPT 2.7B	32	32	2560	∼2700M
OPT 6.7B	32	32	4096	∼6700M
OPT 13B	40	40	5120	∼13000M
OPT 30B	48	56	7168	∼30000M
OPT 66B	64	72	9216	∼66000M

Table 1: Model capacities of LM families whose surprisal estimates were examined in this work. #L, #H, and d_{model} refers to number of layers, number of attention heads per layer, and embedding size, respectively.





Oh et al. (2023)

- 现象：从更大的LM中得到的surprisal为什么和RT的拟合更差
- 核心问题：Psychological Plausibility
- 即，语言模型是否可用于帮助我们理解语言的心理过程？
- 比较语言模型的训练数据和人类儿童的平均语言经验数量
- 人类儿童每年接受最大约11 million words (Zhang et al., 2021; Hart and Risley, 1995)
- 人类对于词的word predictability的预测，不仅仅是根据上下文条件概率？