processing effort as cost of changing beliefs

or: when unpredictable doesn't mean difficult

Jacob Hoover Vigly

21 February 2025, CLiMB Lab at Stanford

sentence processing

how do we understand what a sentence means?

meaning z = 

utterance $\mathbf{u}$ = Near that bank there is an otter

- sentence unfolds word by word: $\mathbf{u} = u_1, u_2, \ldots$
- with each word, refine guess about the meaning, z

sentence processing

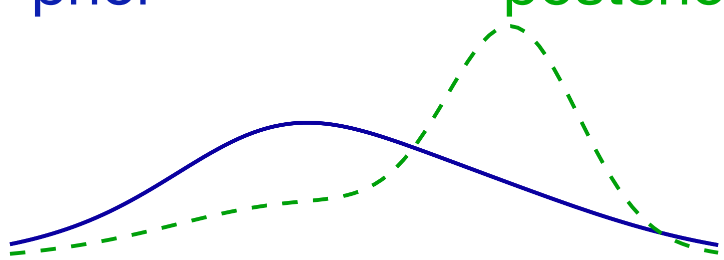
iterative inference problem

$z =$ 

$\mathbf{u} =$ Near that bank htere is an otter …

- observe otterance word by word: $\mathbf{u} = u_1, u_2, \ldots$
- with each word, **update beliefs** about the meaning, z

u_i causes belief update $\underbrace{p(Z \mid u_{1\ldots i-1})}_{\text{prior}} \overset{u_i}{\rightsquigarrow} \underbrace{p(Z \mid u_{1\ldots i-1}, u_i)}_{\text{posterior}}$

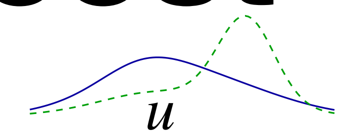


How? …with what processing algorithm?

- important clue: for humans, **unexpected words take more effort**.
- bigger update = more difficult

incremental processing cost

How? …with what processing algorithm?



$p_Z \overset{u}{\mapsto} p_{Z|u}$

- important clue: for humans, **unexpected words take more effort**.

has been formalized as:

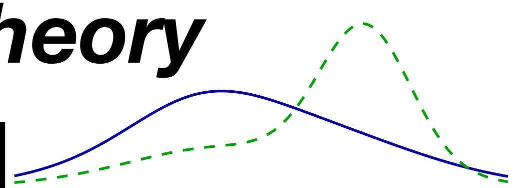
surprisal theory
(Hale '01, Levy '08)

$$\mathrm{cost}(u) \propto \overbrace{\log \frac{1}{p(u)}}^{\mathrm{surprisal}(u)}$$

precise description of phenomenon, … but how? what algorithm?

- refocus idea: difficult = big update (resource allocation cost)

update-size *theory*

$\mathrm{cost}(u) = f(|$  $|)$

size of belief update

hypothesis that cost measured as **<u>bits of information gained</u>** about Z

surprisal theory is special case, by two assumptions:

(a) that $\mathrm{D}(p_{Z|u} \| p_Z) = \mathrm{surprisal}$ (extra term is zero) ⟵ Let's focus on this one

(b) that f is linear

incremental processing cost

How? …with what processing **algorithm**? $\color{blue}{p_Z} \overset{u}{\mapsto} \color{green}{p_{Z|u}}$

- important clue: for humans, **unexpected words take more effort**.
- bigger update = more difficult

most candidate algorithms don’t have this property

- parsing algorithms (Z ranges over trees)
  - non-probabilistic algorithms
  - probabilistic enumerative algorithms
  - neural-parametrized parsing algorithms
- language model inference (e.g. n-gram, RNN, Transformer LLMs …)

… amount of work done during inference **doesn’t depend on probabilistic properties at all**

(so, they don’t directly explain this human behavior)

incremental processing cost

How? …with what processing **algorithm**? $p_Z \overset{u}{\mapsto} p_{Z|u}$

- important clue: for humans, **unexpected words take more effort**.
- bigger update = more difficult

common algorithms don't scale in surprisal / divergence

what kind of algorithm *does*?

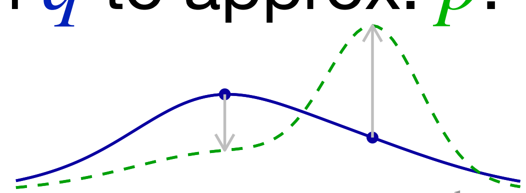
those that somehow prioritize more probable hypotheses:

- **sampling algorithms**

➡ *importance sampling* complexity scales in **divergence**:

sampling from q to approx. p: req #samples $\approx e^{\mathrm{D_{KL}}(p\|q)}$ Chatterjee & Diaconis 2018, …

$\approx \mathrm{D}_{\chi^2}(p\|q)$ Agapiou et al. 2017, Sanz-Alonso 2018, …



importance weight $w(z) \propto \frac{\mathrm{d}p}{\mathrm{d}q}(z)$

$$\mathrm{cost}(u) = f\big(\ \mathrm{D}(p_{Z|u}\|p_Z)\ \big)$$

The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

$$\begin{aligned} \mathrm{cost}(u) &= f\big(\ \mathrm{D}(p_{Z|u} \| p_Z)\ \big) \\ &= f(\text{surprisal}) \\ &\propto\ \text{surprisal} \end{aligned}$$

assumption (a) ⟵ we assumed this held

assumption (b) ⟵ and focused on this one

we show sampling algs may predict

⇒ cost increases **superlinearly**

⇒ with **increasing variance**

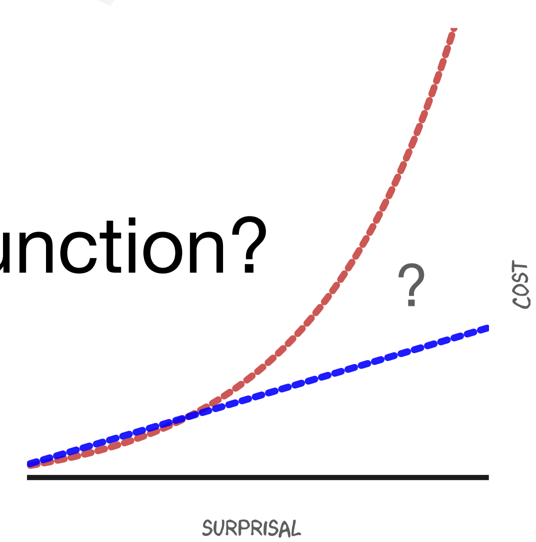
but surprisal theory proposes

⇒ cost increases **linearly**

⇒ (says nothing about variance)

Empirical question:

- what shape is the linking function?



The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing

Jacob Louis Hoover[1,2], Morgan Sonderegger[1], Steven T. Piantadosi[3], and Timothy J. O'Donnell[1,2,4]

linking function: empirical study

is the mean superlinear?

Yes *

- better LM ⇒ more superlinear

linear surprisal theory
(Hale '01, Levy '08)

$$\mathrm{cost}(u) \propto \mathrm{surprisal}(u)$$

***general** surprisal theory*
(Levy '05, Meister '21, Xu '23)

$$\mathrm{cost}(u) = f(\mathrm{surprisal}(u))$$

does variance increase?

Yes

- across LMs

* HOWEVER:

PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Large-scale evidence for logarithmic effects of word predictability on reading time

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy

replicate on NS dataset, not others

=> our empirical results may be idiosyncratic

consistent with sampling algorithms' predictions

$\Longrightarrow$ **motivation:** sampling mechanisms for processing

more precisely what are the empirical predictions?

when surprisal ≠ divergence

now, let’s revisit the other assumption: that surprisal = divergence

surprisal theory

$$\mathrm{cost}(u) = f(\mathrm{surprisal}(u))$$

belief-update theory

$$\boxed{\mathrm{cost}(u) = f(\mathrm{D}_{\mathrm{KL}}(p_{Z|u} \| p_Z))}$$

recall motivation: surprisal as measure of size of belief update

$$\overbrace{\mathrm{D}_{\mathrm{KL}}(p_{Z|u} \| p_Z)}^{\mathbb{E}_{p_{Z|u}}\left[\log \frac{p(z \mid u)}{p(z)}\right]} = \overbrace{\mathrm{surprisal}(u)}^{\log \frac{1}{p(u)}} - \overbrace{R(u)}^{\mathbb{E}_{p_{Z|u}}\left[\log \frac{1}{p(u \mid z)}\right]}$$

$$\mathbb{E}_{p_{Z|u}}\left[\log \frac{p(z \mid u)}{p(z)}\right] = \log \frac{1}{p(u)} - \mathbb{E}_{p_{Z|u}}\left[\log \frac{1}{p(u \mid z)}\right]$$

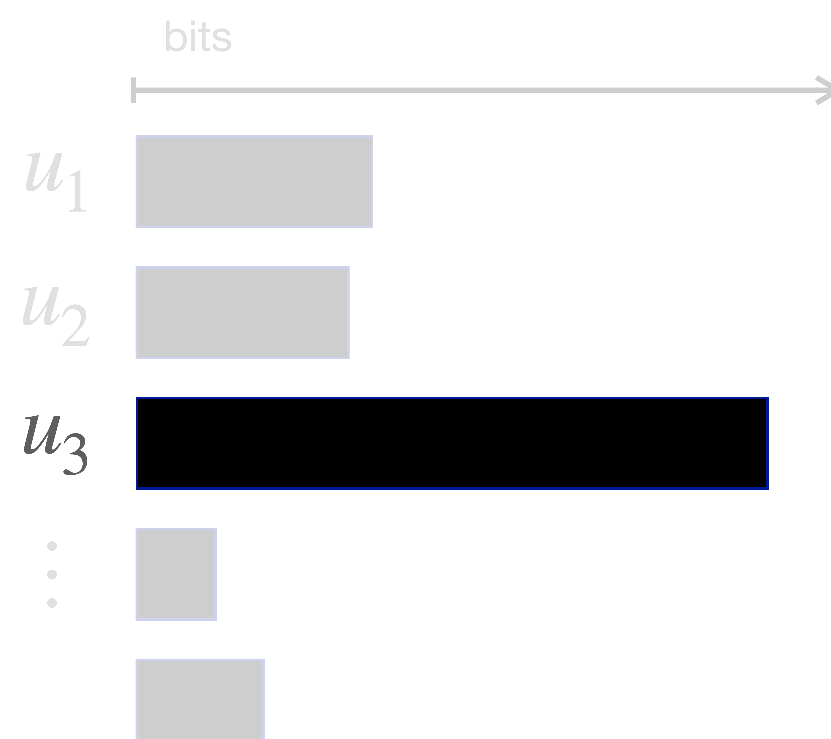
when surprisal > KL divergence

raw amount of info contained in u

size of belief update caused by observing u

reconstruction information: 'extra' bits that don't contribute to update

$$\text{surprisal} = \mathrm{D}_{\mathrm{KL}} + R$$

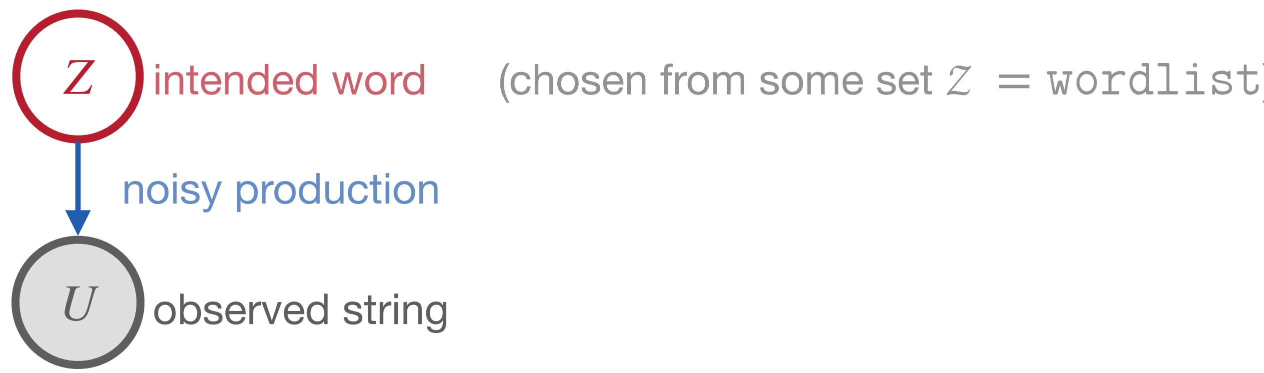


When unpredictable does not mean difficult
(WIP with Peng Qian, Morgan Sonderegger, Tim O'Donnell)

typos as a case study

typos as a case study

surprisal = $\mathrm{D}_{\mathrm{KL}} + R$



For this application,

let latent Z (meaning) range over strings, representing **intended word**

- easy to model prior and likelihood
- narrow application where we might expect **LM surprisal of the observed string is intuitively inadequate** as measure of human processing cost (I'm interested in broader applications to follow!)

typos as a case study

$$\text{surprisal} = D_{KL} + R$$

Example:

- *After tripping on the rug and falling in front of everyone, I felt deeply ______*

condition	target word	surprisal	divergence	
1. expected	*embarrassed*	LOW	LOW	😐
2. unexpected	*innovative*	HIGH	HIGH	🤯
3. expected (typo)	*embarrsased*	HIGH	LOW	😕
4. unexpected (typo)	*innovaitve*	HIGH	HIGH	

(even with correct noise model)

typos as a case study

$$\text{surprisal} = D_{KL} + R$$

Example:

- *After tripping on the rug and falling in front of everyone, I felt deeply* ______

1.	expected	*embarrassed*
2.	unexpected	*innovative*
3.	expected (typo)	*embarrsased*
4.	unexpected (typo)	*innovaitve*

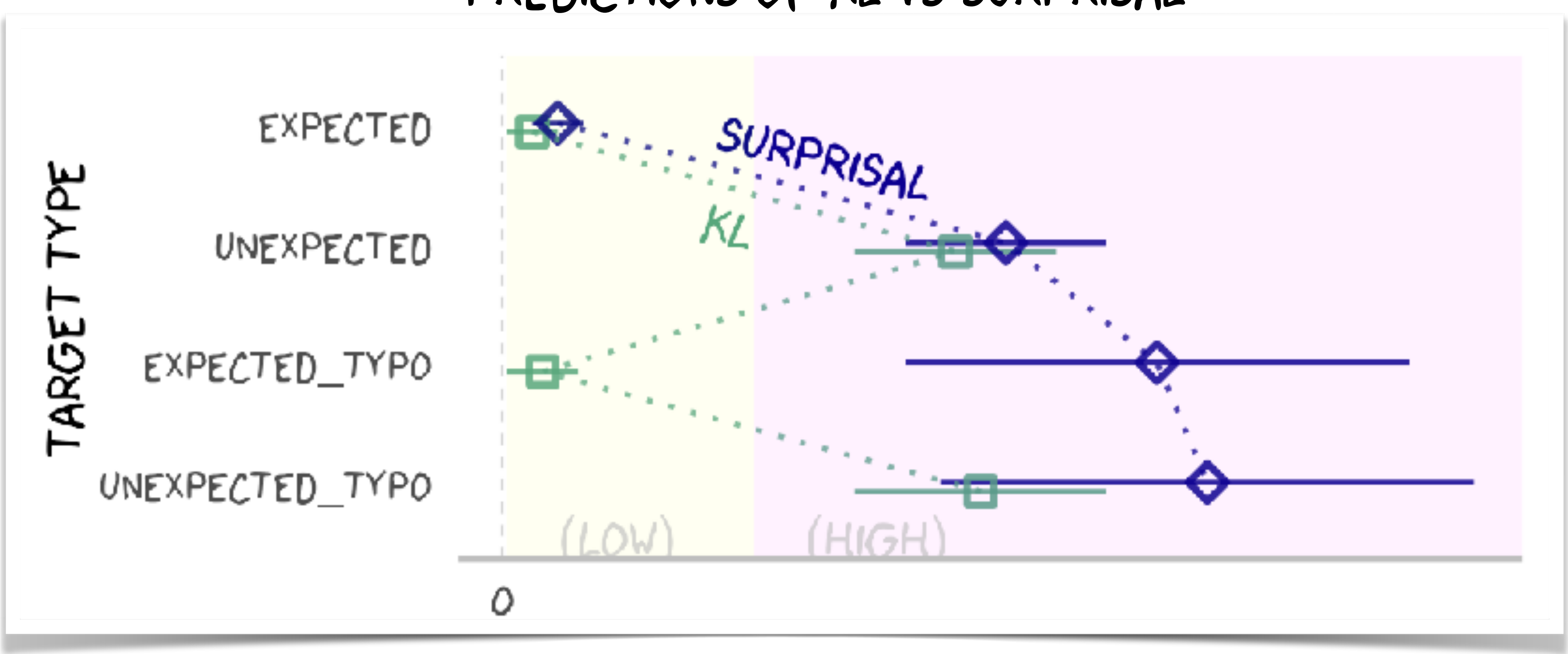
Self-paced reading time study:

- 51 sentences x 4 conditions = 204 unique targets of interest.
- 104 participants on Prolific (post exclusions)

Fit mixed-effect regression models:

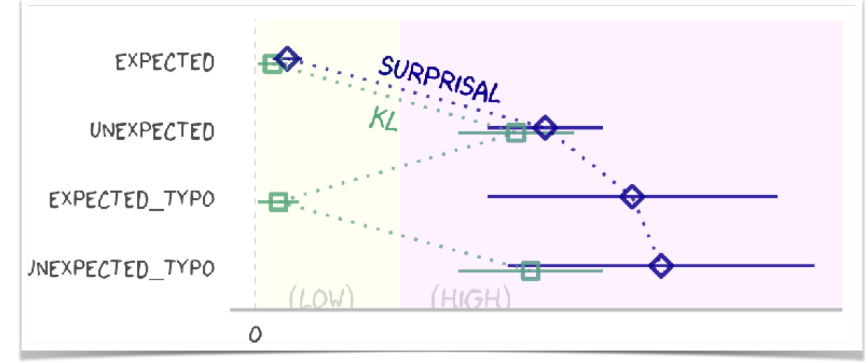
- predict human RT
- predict LLM surprisal (separately)
  - surprisals from collection of LLMs

PREDICTIONS OF KL VS SURPRISAL



When unpredictable does not mean difficult
(WIP with Peng Qian, Morgan Sonderegger, Tim O'Donnell)

typos as a case study



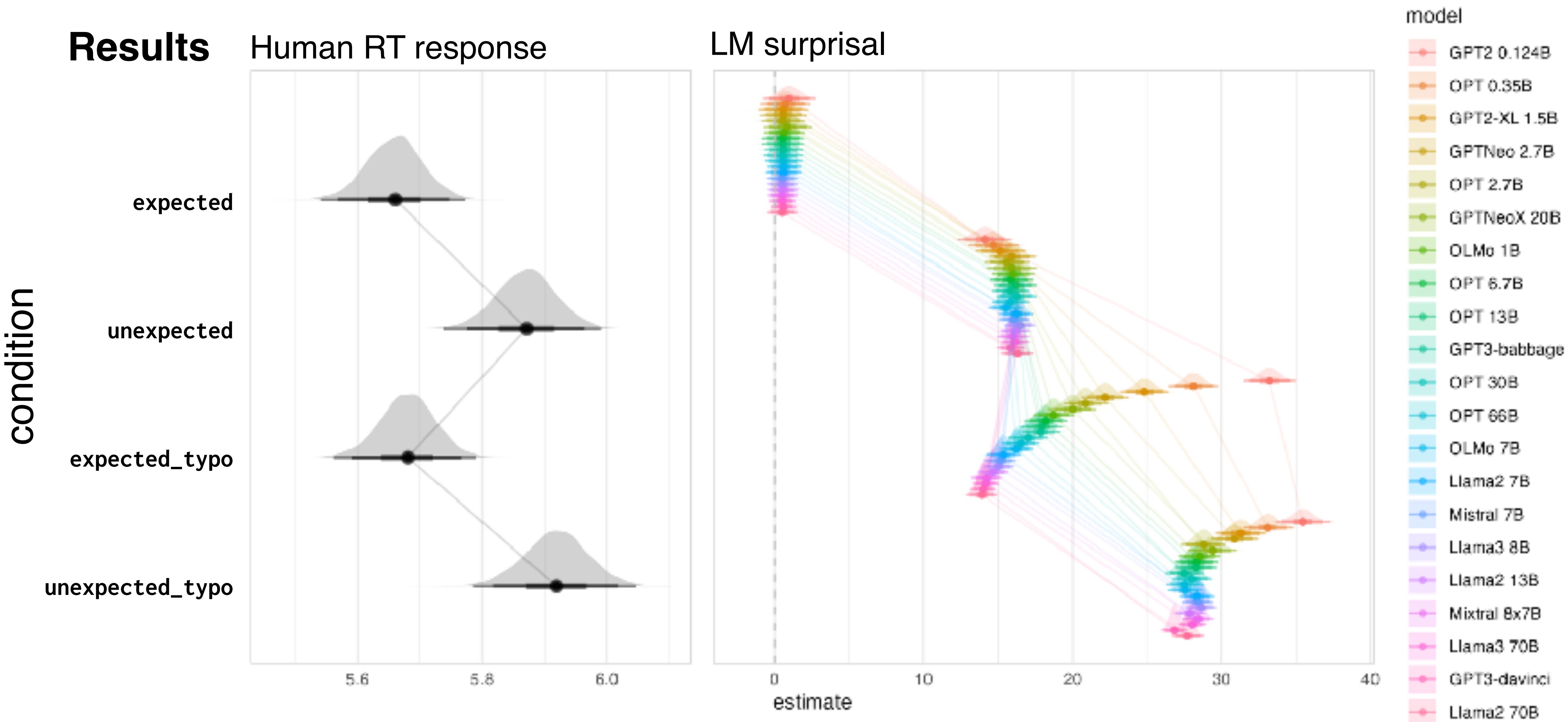
Does surprisal pattern as expected?

Yes. Surprisal is low in `expected` condition, but high in others.

Does human RT pattern like surprisal or divergence?

RTs zig-zag, **as divergence should predict**, contra surprisal.

Results

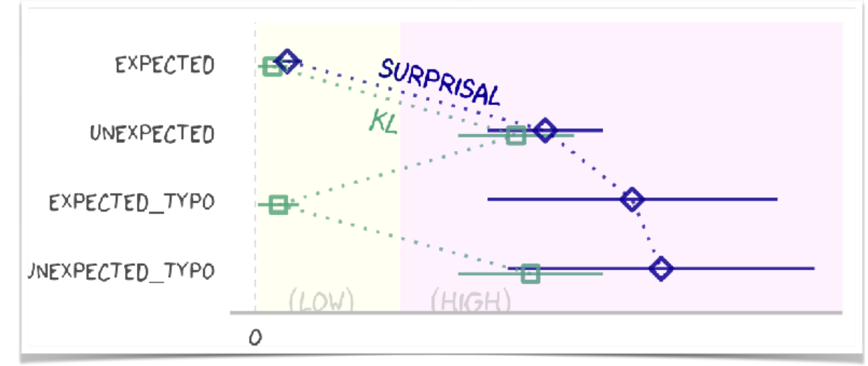


When unpredictable does not mean difficult
(WIP with Peng Qian, Morgan Sonderegger, Tim O'Donnell)

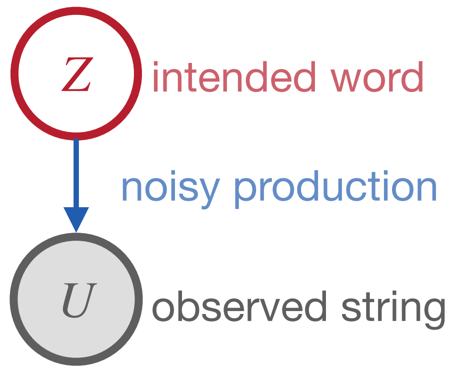
typos as a case study

estimating KL and surprisal

in noisy channel

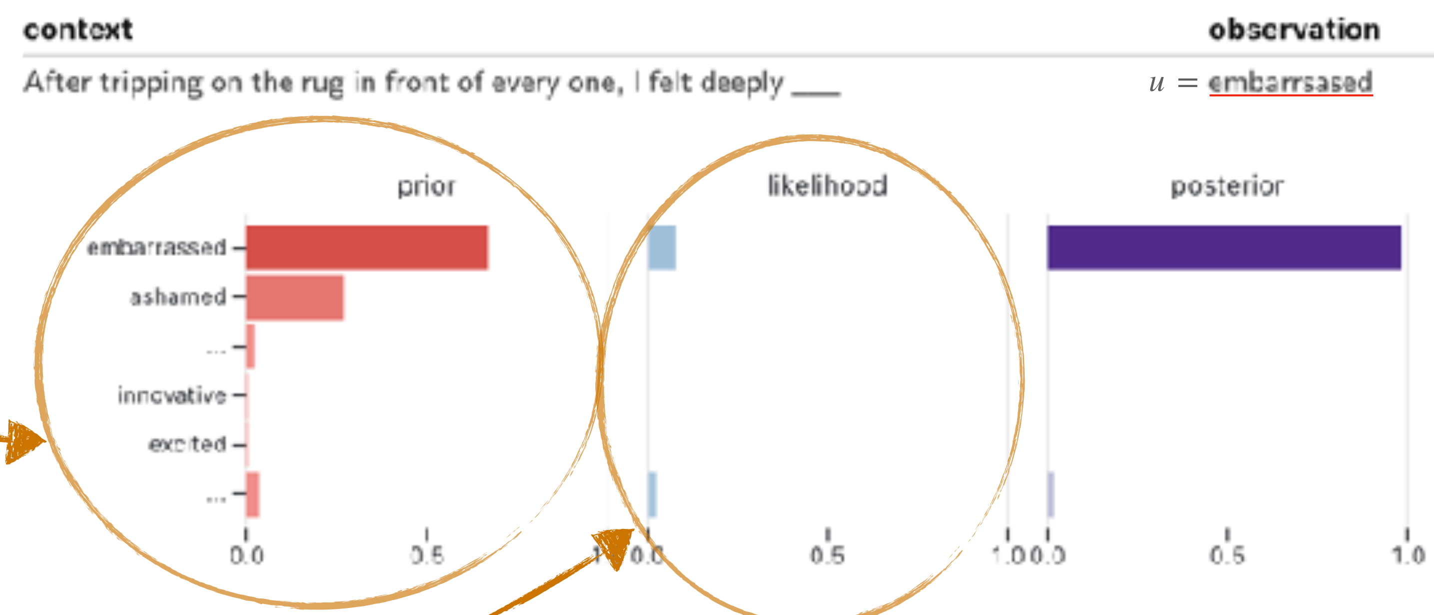


generative model:



(chosen from some set $\mathcal{Z}$ = `wordlist`)

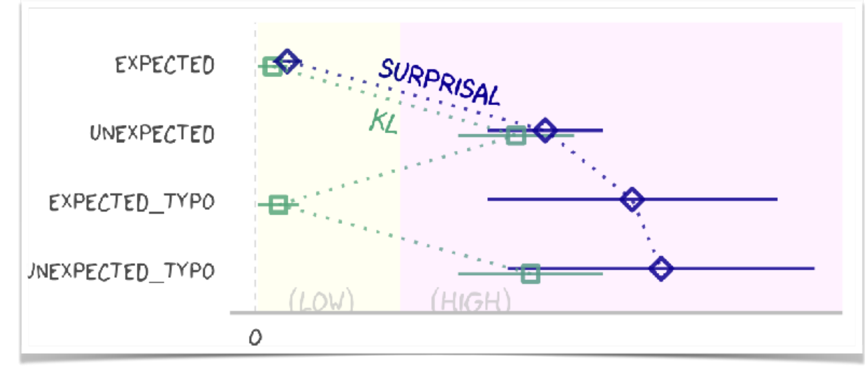
- **prior** over intended words
  $p(z \mid \text{context})$
  = LLM next-seq distribution
  constrained to wordlist
  $\propto p_{\text{LM}}(\text{context}) \odot \mathbf{1}_{\texttt{wordlist}}$



- **likelihood** of observed string:
  $p(u \mid z)$
  = string-edit distance model
  $p(D_{\text{Lev}} \mid z) \cdot p(u \mid D_{\text{Lev}}, z)$

typos as a case study

estimating KL and surprisal



context After tripping over the rug in front of everyone at the party, she quickly got up, but her cheeks turned red and she felt deeply

z	prior
_embarrassed	6.5668e-01
_ashamed	2.6608e-01
_guilty	1.6075e-02
_uncomfortable	1.0753e-02
_shy	7.0945e-03
...	

observation $w =$ embarrassed (expected)

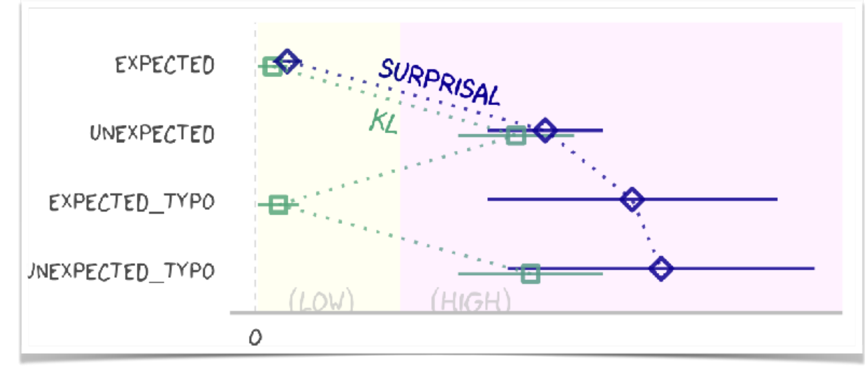
z	prior	likelihood	posterior
_embarrassed	6.5668e-01	8.9583e-01	1.0000e+00
_embraced	3.6091e-06	9.4480e-16	5.7964e-21
_impressed	6.4865e-05	1.6926e-19	1.8663e-23
_arrested	1.8016e-06	1.6229e-19	4.9701e-25
...			

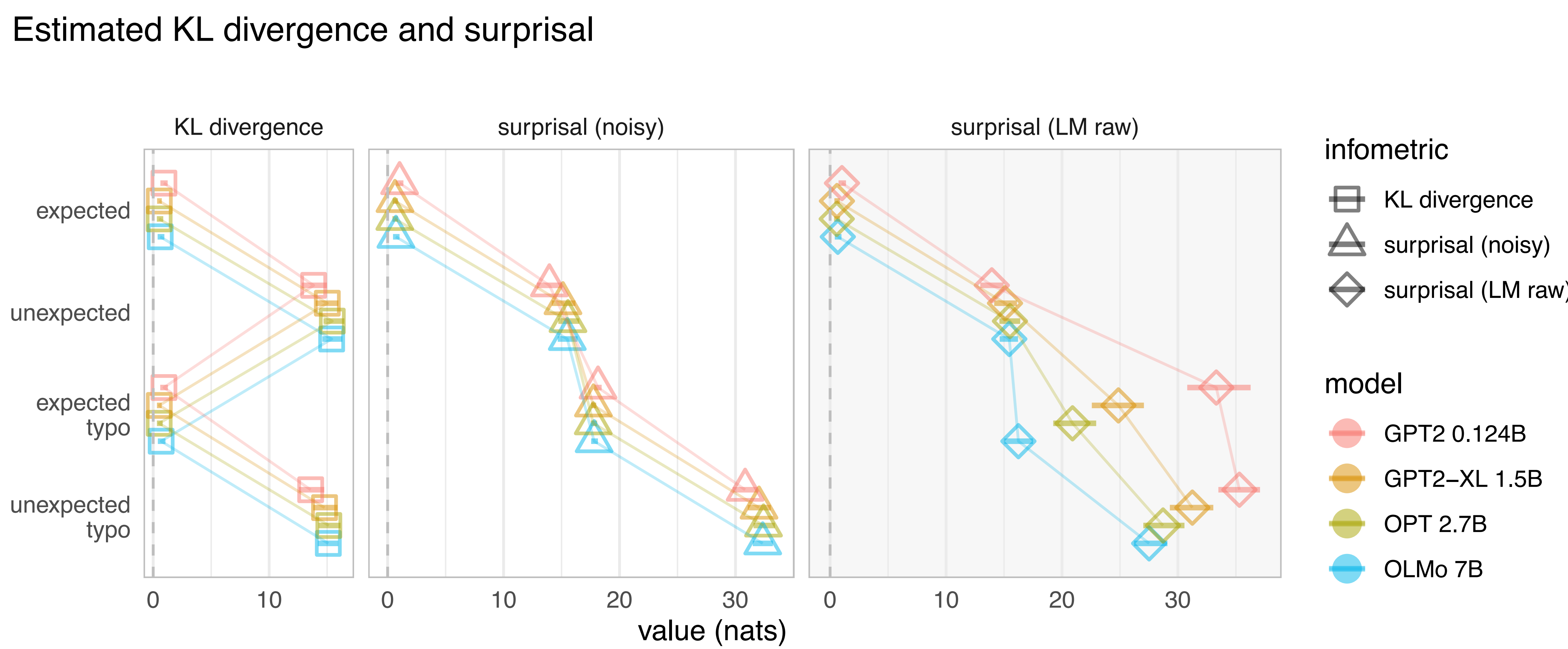
prior over intended words
$p(z)$ = LM

likelihood of observed string:
$p(u \mid z)$ = noisy string model

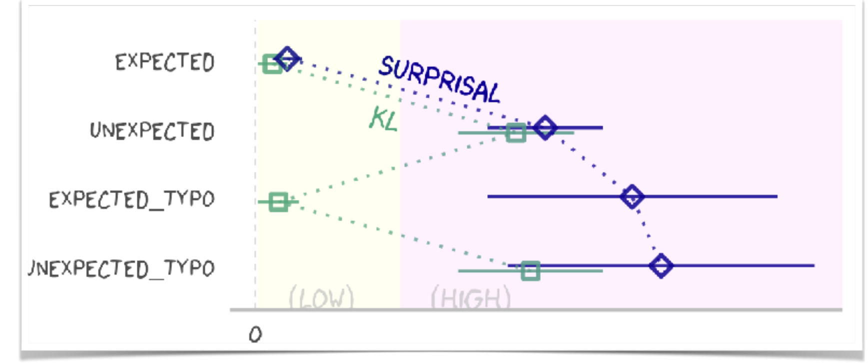
typos as a case study

estimating KL and surprisal





typos as a case study



Does surprisal pattern as expected?

Yes. Surprisal is low in **`expected`** condition, but high in others.

Does human RT pattern like surprisal or divergence?

RTs zig-zag, **as update-size predicts**, contra surprisal.

as estimated in our noisy channel model

surprisal theory
(Levy '08)

$$\mathrm{cost}(u) = f(\mathrm{surprisal}(u))$$

update-size theory

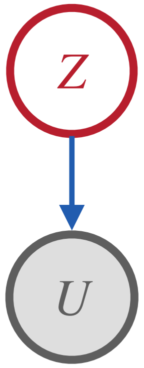
$$\mathrm{cost}(u) = f(\mathrm{D}_{\mathrm{KL}}(p_{Z|u} \| p_Z))$$

divergence (information gain)
connected to sampling complexity

$\Longrightarrow$ **motivates** sampling-based inference algorithms for processing

next steps - better estimates

- for typos
  - more realistic models of typos (using typing statistics)
  - broad-coverage model of KL (not just our materials)
- use **character level LMs** for prior and likelihood models
  - Giulianelli et al. 2024, Vieira et al. 2024
- more broadly: researcher must answer "**what is** Z?"
  - unlike surprisal, requires different models depending on **task**
    - infer intended **words**? **referent**? **sentiment**? etc. (model *task effects*)



next steps - not just typos

other places where we think surprisal $\gg D_{\mathrm{KL}}$ (that is, $R \gg 0$):

any (more interesting) constructions where some target region is processed without difficulty despite being very unpredictable

unexpected ways of communicating expected information

- synonyms: *This living-room furniture set consists of a table, chair, and couch.* (vs *sofa*)
- epithets: *I hate John. From the moment the bastard came in the room ….*

grammatical illusions (as in Yuhan Zhang's talk last week!)

- Moses illusions: *In the biblical story of the Ark, how many animals of each kind did Moses take with him?*
- agreement attraction: *The key to all the cabinets are on the table.*
- NPI illusions: *The bills that no senator voted for will ever become law.*
- depth-charge illusions: *No head injury is too trivial to ignore.*

malapropisms

- *Sure, if I reprehend* (apprehend) *anything in this world it is the use of my oracular* (vernacular) *tongue, and a nice derangement* (arrangement) *of epitaphs* (epithets)*!* *(Sheridan, 1775)*

multilingual codeswitching

- *"Veux-tu rentrer dans ma bubble?"*

next steps - not just typos

other places where we think surprisal $\gg D_{\mathrm{KL}}$ (that is, $R \gg 0$):

any (more interesting) constructions where some target region is processed without difficulty despite being very unpredictable

unexpected ways of communicating expected information

- synonyms: *This living-room furniture set consists of a table, chair, and couch.* (vs sofa)
- epithets: *I hate John. From the moment the bastard came in the room ....*

grammatical illusions (as in Yuhan Zhang's talk last week!)

- Moses illusions: *In the biblical story of the Ark, how many animals of each kind did Moses take with him?*
- agreement attraction: *The key to all the cabinets are on the table.*
- NPI illusions: *The bills that no senator voted for will ever become law.*
- depth-charge illusions: *No head injury is too trivial to ignore.*

malapropisms

- *Sure, if I reprehend* (apprehend) *anything in this world it is the use of my oracular* (vernacular) *tongue, and a nice derangement* (arrangement) *of epitaphs* (epithets)*!* (Sheridan, 1775)

multilingual codeswitching

- *"Veux-tu rentrer dans ma bubble?"*

thanks to

- you!
- my collaborators: Tim O'Donnell, Peng Qian, Morgan Sonderegger, Steve Piantadosi
- National Science Foundation postdoc grant (SMA-2404644)