Event Segmentation In Story Listening Using Deep Language Models

Manoj Kumar[1], Ariel Goldstein[1], Sebastian Michelmann[1], Jeffrey M. Zacks[2], Kenneth A. Norman[1], Uri Hasson[1]

[1]Princeton University, [2]Washington University at St. Louis

Event Segmentation In Naturalistic Settings

We segment our continuous experiences into distinct events (Newtson, 1973; Zacks et al., 2007)

Event Segmentation Theory (EST; Zacks et al., 2007) posits that we mark the boundary of an event at moments when there is a transient increase in prediction error

Prediction error (disfluency) in naturalistic settings is difficult to measure, but is typically operationalized with the probability of expected outcome

In a basketball game (Antony et al., 2021) prediction error was defined as the change in win probability at each moment in the game

In sentence processing, the "Cloze" probability of the sentence ending word is used to index the N400 response (Kutas and Hillyard, 1980) for out-of-context words

Our goal is to test Event Segmentation Theory by using fine-grained measures of disfluency, extracted from deep learning language models, for each word in a narrative

Methods

Behavior: Participants listened to and segmented 3 stories

Monkey In The Middle (~30 minutes, Goldstein et al., 2022)
Pieman (7 and a half minutes; Michelmann et al., 2021)
The Tunnel Under The World (~25 minutes; Lositsky et al., 2016)

GPT-2 Language Model (Radford et al., 2019)

48 hidden layers, ~ 1.5B parameters
Pretrained on 8M webpages
Context window of 1024 tokens
Vocabulary ~ 50K words
Embedding dimension = 1600

Transformers

State-of-the-art Natural Language Processing for PyTorch and TensorFlow 2.0

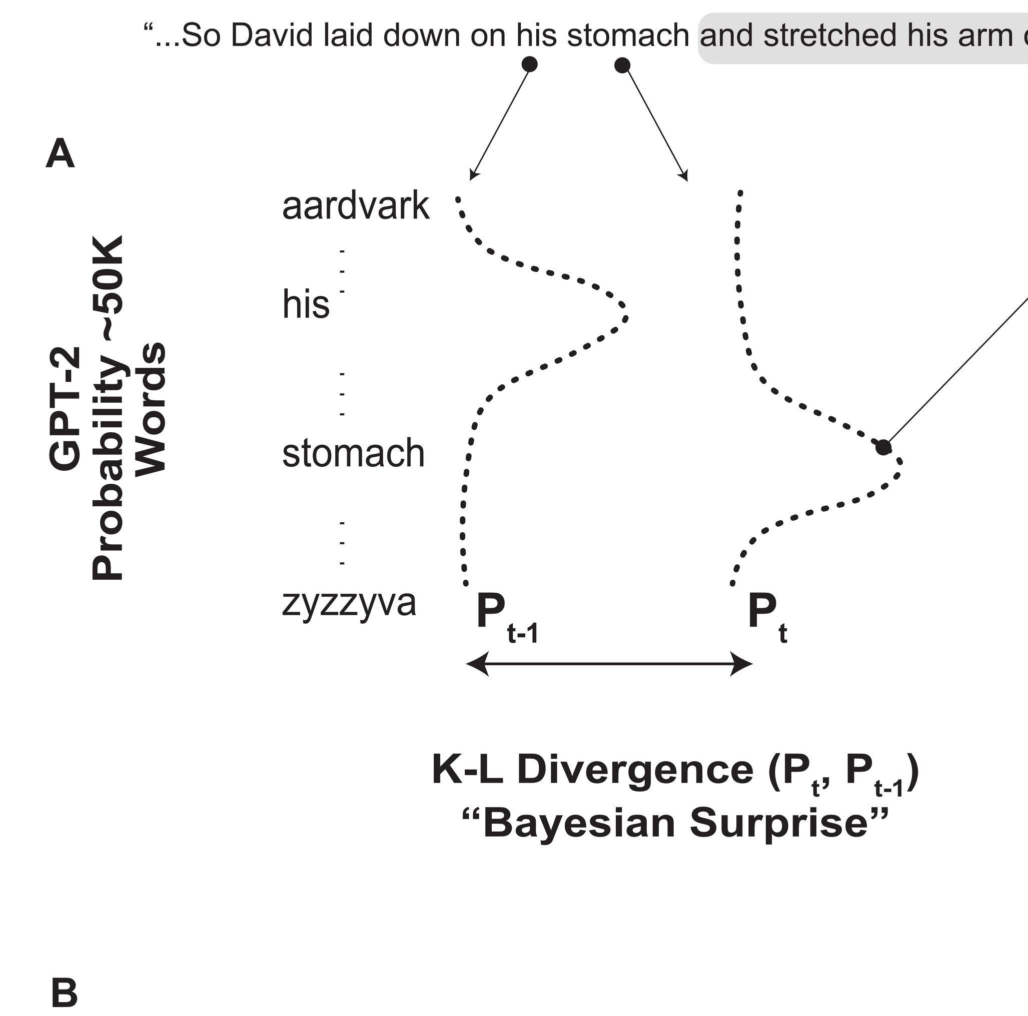
The text from the 3 stories is parsed through GPT-2-XL

Derived Measures: surprise, entropy, Bayesian surprise (K-L Divergence), cosine distance of successive word embeddings from the final layer of GPT-2

Regression (LASSO) using leave-one-story-out cross-validation

Null distribution computed by circularly shifting each word +/- 4000 positions

Disfluency Measures: Surprise, Entropy, and Bayesian Surprise



Surprise: probability of the actual word in the story

$$Surprise_w = -\log(p_w)$$

p_w = probability of the word w

Entropy: how "peaked" is the probability distribution

$$E_t = -\sum_{v=1}^{V} p_v * log(p_v)$$

p_v = probability value for word v in the vocabulary

$V \equiv$ all words in GPT2 vocabulary

Bayesian surprise: how much do the 2 probability distributions differ

$$\text{K-L Divergence}(P_t, P_{t-1}) = -\sum_{v=1}^{V} p_v * log(q_v)$$

$P_{t-1}, P_t \equiv$ probability distribution at word points $t-1, t$

p_v = probability of each word in the vocabulary drawn from P_t

q_v = probability of each word in the vocabulary drawn from P_{t-1}

GPT-2 Embedding: Cosine Distance ("his", "stomach") to measure semantic similarity

Schematic A. The GPT-2 probability distributions at 2 word points and the derived measures of surprise, entropy, and Bayesian surprise (Itti and Baldi, 2009) **B.** The cosine distance between successive word embeddings from the last hidden layer

Transient Increase In Disfluency

(Reynolds et al., 2007)

$$= \frac{\textbf{Current Value (t)}}{\textbf{Average at (t -1)}}$$

$$\text{Average at word 1, } \mu_1 = \frac{1}{W}\sum_{w=1}^{W} S_w$$

W = number of words in the story

$$\mu_2 = \mu_1 + (S_2 - \mu_1) \times drift$$
$$\mu_3 = \mu_2 + (S_3 - \mu_2) \times drift$$
$$\vdots$$
$$\mu_W = \mu_{W-1} + (S_W - \mu_{W-1}) \times drift$$

where $\mu_1, \mu_2, \ldots, \mu_W \equiv$ Average at word $1, 2, \ldots, W$

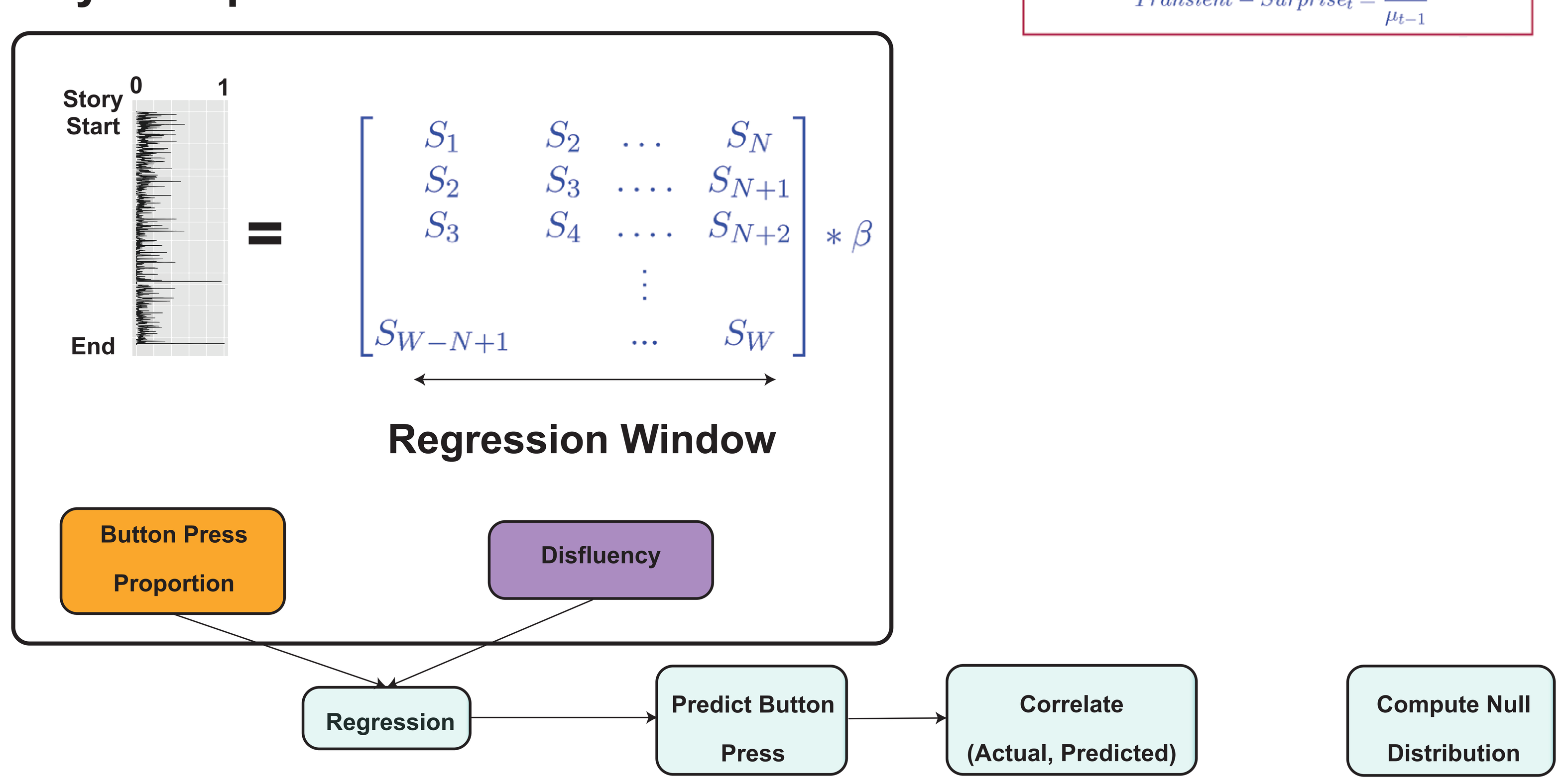
and

$drift$ = smoothing parameter for the running average

The transient Surprise for word at time t,

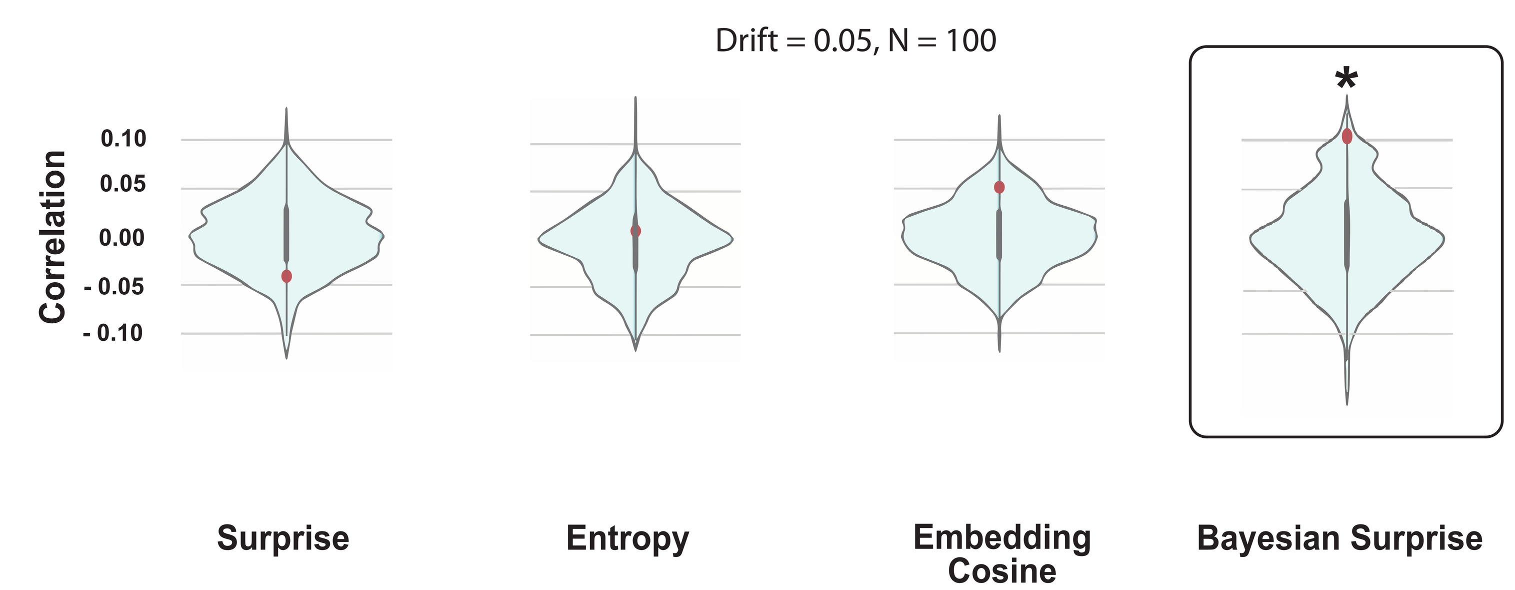
$$Transient - Surprise_t = \frac{S_t}{\mu_{t-1}}$$

Analysis Pipeline



Transient Bayesian Surprise Predicts Event Boundaries

Transient Increase In Disfluency Measures



The transient Bayesian surprise shows the best correlation with human segmentation behavior . The violin plot shows the null distribution with the central bar showing the box plot of the null distribution. The red dot is the unshifted correlation value

Transient Bayesian Surprise Shows Consistent Results Across Regression Windows and Drift

Regression Window	Drift = 0.03		Drift = 0.05		Drift = 0.10	
	Correlation	Percentile	Correlation	Percentile	Correlation	Percentile
20	0.07	97.7	0.06	96.3	0.05	95.4
50	0.07	97.3	0.07	95.0	0.06	96.0
100	0.10	99.6	0.10	98.9	0.09	99.2
150	0.10	99.5	0.11	99.7	0.11	99.8
200	0.10	96.9	0.11	98.2	0.12	99.2

Current (Non-Transient) Bayesian Surprise Shows Less Consistent Results

Regression Window	Correlation	Percentile
20	0.05	86.1
50	0.07	91.0
100	0.09	95.4
150	0.11	97.5
200	0.10	95.7

Current (non-transient) surprise, entropy, and embedding cosine also performed poorly

Conclusion

We extracted fine-grained measures of disfluency for each word in a narrative using GPT-2

Using regression models, we found that the transient **Bayesian surprise** best correlated with human segmentation judgments

References

J. W. Antony, T. H. Hartshorne, K. Pomeroy, T. M. Gureckis, U. Hasson, S. D. McDougle, and K. A. Norman. Neuron, 109(2):377–390.e7, 2021.

A. Goldstein, et al. Nature Neuroscience, 25(3):369–380, 2022.

L. Itti and P. Baldi. Vision Research, 49(10):1295–1306, 2009.

M. Kutas and S. A. Hillyard. Science, 207(4427):203–205. 1980.

O. Lositsky, J. Chen, D. Toker, C. J. Honey, M. Shvartsman, J. L. Poppenk, U. Hasson, and K. A. Norman. eLife, 5:e16070, 2016.

S. Michelmann, A. R. Price, B. Aubrey, C. K. Strauss, W. K. Doyle, D. Friedman, P. C. Dugan, O. Devinsky, S. Devore, A. Flinker, U. Hasson, and K. A. Norman. Nature Communications, 12(1):5394, 2021.

D. Newtson. Journal of Personality and Social Psychology, 28(1):28–38, 1973

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. OpenAI blog. 2019.

J. R. Reynolds, J. M. Zacks, and T. S. Braver. Cognitive Science, 31(4):613–643, 2007.

Acknowledgments

We would like to thank Bobbi Aubrey, Zaid Zada, and Harshvardhan Ghazula for help with the "Monkey In The Middle" embeddings, and Olga Lositsky for sharing data for "The Tunnel Under The World"

This work was supported by ONR MURI grant N00014-17-1-2961 awarded to JZ, KAN, and UH

email: manoj.neuron@gmail.com