Cloze behind: Language model surprisal predicts N400 amplitude better than cloze

James A. Michaelov, Seana Coulson, Benjamin K. Bergen

Department of Cognitive Science, UC San Diego

Introduction

- Language comprehension involves preactivation of expected words (Kutas, DeLong, and Smith 2011; Van Petten and Luka 2012; Kuperberg, Brothers, and Wlotko 2020).
- Expectancy is virtually always operationalized as cloze (Taylor 1953), but:
  - Cloze is an opaque metric: we don't know how it relates to the preactivation involved in comprehension.
  - Language statistics (operationalized by language model surprisal) can also predict N400 amplitude (Frank et al. 2015; Aurnhammer and Frank 2019; Merkx and Frank 2021).
  - Language model surprisal can model the effect of word expectancy on N400 amplitude even when cloze cannot (Michaelov and Bergen 2020).
  - Statistical learning underlies predictive processing in other domains (de Lange, Heilbron, and Kok 2018).
- Can state-of-the-art language models predict the N400 better than cloze?
  - Past research: cloze out-performs language models in predicting processing difficulty (Smith and Levy 2011; Brothers and Kuperberg 2021; Szewczyk and Federmeier 2022).
  - Now: language models continue to advance at a rapid pace, and higher-quality models are better at predicting the N400 (Aurnhammer and Frank 2019)—if language statistics underlie preactivation, a sufficiently high-quality model should capture this.

Methods

- Stimuli from Nieuwland et al. (2018) were truncated until the target noun, which was either more or less contextually predictable.
- These stimuli were run through 8 neural network language models:
  - Two LSTM **recurrent neural network** language models (Gulordava et al. 2018; Jozefowicz et al. 2016).
  - Three **autoregressive transformer** language models (Dai et al. 2019; Radford et al. 2019; Brown et al. 2020)
  - Three **masked language model transformers** (Devlin et al. 2019; Liu et al. 2019; Lan et al. 2020)
- Contextual probability $P(w_i|w_{1\ldots i-1})$ calculated by each model for each target word in the vocabulary was recorded and transformed into surprisal:

$$\text{Surprisal} = -\log(P(w_i|w_{1\ldots i-1}))$$

- Analyses were run comparing how well **cloze** (raw probability and surprisal) and language model surprisal fit the single-trial N400 amplitudes (mean over 200-500ms range) from (Nieuwland et al. 2018).
- We further investigated how much variance in N400 amplitude is explained by cloze vs. language models.

Results

How good is each metric at predicting single-trial N400 amplitude?

- Assessed by comparing the AICs of regressions including each predictor as a main effect:

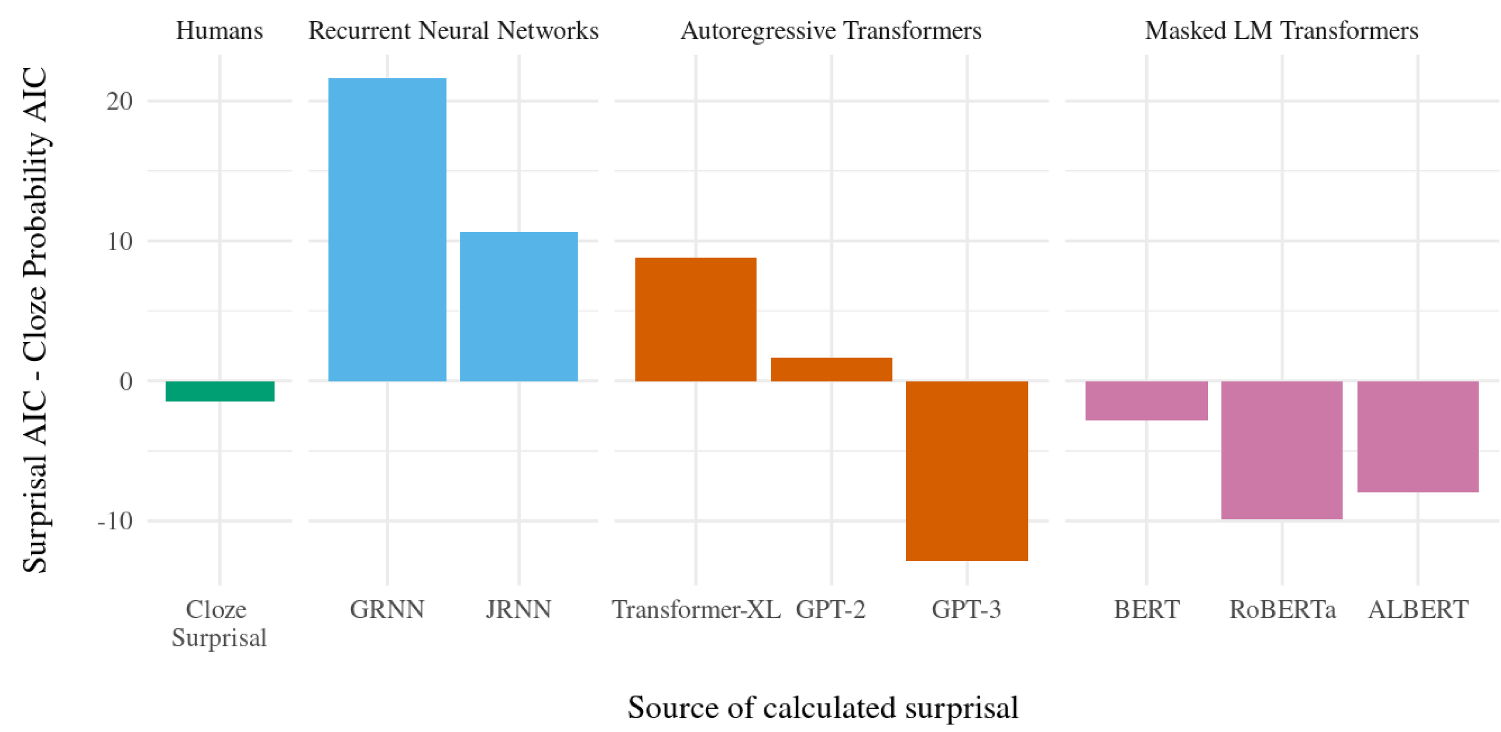

Figure 1: Relative AIC of regressions including each predictor. The cloze probability regression was used as the baseline, so regressions with AIC values below zero have an improved fit relative to cloze probability.

What is the variance in N400 amplitude explained by language model surprisal and cloze?

- Assessed by investigating whether adding cloze surprisal to a regression already including language model surprisal significantly improves model fit (Table 1), and vice-versa (Table 2):

Table 1:
Predictor + Cloze Surprisal

Predictor	χ^2	df	p
GRNN	**23.103**	**1**	**<0.001**
JRNN	**16.056**	**1**	**0.001**
Tranformer-XL	**13.277**	**1**	**0.002**
GPT-2	**8.178**	**1**	**0.025**
GPT-3	0.754	1	1
BERT	**8.282**	**1**	**0.025**
RoBERTa	3.276	1	0.351
ALBERT	1.935	1	0.757

Table 2:
Cloze Surprisal + Predictor

Predictor	χ^2	df	p
GRNN	0.056	1	1
JRNN	3.982	1	0.24
Tranformer-XL	3.031	1	0.392
GPT-2	5.088	1	0.131
GPT-3	**12.168**	**1**	**0.004**
BERT	**9.639**	**1**	**0.013**
RoBERTa	**11.72**	**1**	**0.005**
ALBERT	**8.45**	**1**	**0.024**

Summary of Results

- The surprisals calculated from the predictions of four language models—GPT-3, BERT, RoBERTa, and ALBERT—fit single-trial N400 amplitude better than cloze probability or surprisal (Figure 1).
- GPT-3, BERT, RoBERTa, and ALBERT surprisal explains variance in N400 amplitude not explained by cloze surprisal (Table 2).
- Cloze surprisal does not explain variance in the N400 above and beyond that explained by GPT-3, RoBERTa, and ALBERT surprisal (Table 1).

Conclusions

- The surprisals calculated from three of the highest-quality language models predict N400 amplitude better than cloze on all fronts, making them the best predictors of N400 amplitude to date (based on this study).
- Provides evidence for the idea that the statistics of language drives the preactivation underlying the N400 response.
- Suggests that researchers should use high-quality language models for norming when designing stimuli.

References

Aurnhammer, Christoph, and Stefan L. Frank. 2019. "Evaluating Information-Theoretic Measures of Word Prediction in Naturalistic Sentence Reading." *Neuropsychologia* 134 (November): 107198. https://doi.org/10.1016/j.neuropsychologia.2019.107198.
Brothers, Trevor, and Gina R. Kuperberg. 2021. "Word Predictability Effects Are Linear, Not Logarithmic: Implications for Probabilistic Models of Sentence Comprehension." *Journal of Memory and Language* 116 (February): 104174. https://doi.org/10.1016/j.jml.2020.104174.
Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." In *Advances in Neural Information Processing Systems*, 33:1877–1901. Curran Associates, Inc.
Dai, Zihang, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context." *arXiv:1901.02860 [Cs, Stat]*, June. https://arxiv.org/abs/1901.02860.
de Lange, Floris P., Micha Heilbron, and Peter Kok. 2018. "How Do Expectations Shape Perception?" *Trends in Cognitive Sciences* 22 (9): 764–79. https://doi.org/10.1016/j.tics.2018.06.002.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.
Frank, Stefan L., Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. "The ERP Response to the Amount of Information Conveyed by Words in Sentences." *Brain and Language* 140 (January): 1–11. https://doi.org/10.1016/j.bandl.2014.10.006.
Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. "Colorless Green Recurrent Networks Dream Hierarchically." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1195–205. New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1108.
Jozefowicz, Rafal, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. "Exploring the Limits of Language Modeling." *arXiv:1602.02410 [Cs]*, February. https://arxiv.org/abs/1602.02410.
Kuperberg, Gina R., Trevor Brothers, and Edward W. Wlotko. 2020. "A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation." *Journal of Cognitive Neuroscience* 32 (1): 12–35. https://doi.org/10.1162/jocn_a_01465.
Kutas, Marta, Katherine A. DeLong, and Nathaniel J. Smith. 2011. "A Look Around at What Lies Ahead: Prediction and Predictability in Language Processing." In *Predictions in the Brain: Using Our Past to Generate a Future*, edited by Moshe Bar, 190–207. New York, NY, US: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195395518.003.0065.
Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." In *International Conference on Learning Representations*.
Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv:1907.11692 [Cs]*, July. https://arxiv.org/abs/1907.11692.
Merkx, Danny, and Stefan L. Frank. 2021. "Human Sentence Processing: Recurrence or Attention?" In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 12–22. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.cmcl-1.2.
Michaelov, James A., and Benjamin K. Bergen. 2020. "How Well Does Surprisal Explain N400 Amplitude Under Different Experimental Conditions?" In *Proceedings of the 24th Conference on Computational Natural Language Learning*, 652–63. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.conll-1.53.
Nieuwland, Mante S, Stephen Politzer-Ahles, Evelien Heyselaar, Katrien Segaert, Emily Darley, Nina Kazanina, Sarah Von Grebmer Zu Wolfsthurn, et al. 2018. "Large-Scale Replication Study Reveals a Limit on Probabilistic Prediction in Language Comprehension." Edited by Barbara G Shinn-Cunningham. *eLife* 7 (April): e33468. https://doi.org/10.7554/eLife.33468.
Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models Are Unsupervised Multitask Learners."
Smith, Nathaniel J, and Roger Levy. 2011. "Cloze but No Cigar: The Complex Relationship Between Cloze, Corpus, and Subjective Probabilities in Language Processing." In *Proceedings of the Annual Meeting of the Cognitive Science Society, 33*, 7.
Szewczyk, Jakub M., and Kara D. Federmeier. 2022. "Context-Based Facilitation of Semantic Access Follows Both Logarithmic and Linear Functions of Stimulus Probability." *Journal of Memory and Language* 123 (April): 104311. https://doi.org/10.1016/j.jml.2021.104311.
Taylor, Wilson L. 1953. "'Cloze Procedure': A New Tool for Measuring Readability." *Journalism Quarterly* 30 (4): 415–33. https://doi.org/10.1177/107769905303000401.
Van Petten, Cyma, and Barbara J. Luka. 2012. "Prediction During Language Comprehension: Benefits, Costs, and ERP Components." *International Journal of Psychophysiology*, Predictive information processing in the brain: Principles, neural mechanisms and models, 83 (2): 176–90. https://doi.org/10.1016/j.ijpsycho.2011.09.015.