The predictive power of Cloze probability vs. LLM surprisal for N400 amplitude is task-dependent

Miriam Schulz (mschulz@lst.uni-saarland.de), Masato Nakamura, Francesca Delogu, Matthew W. Crocker

Department of Language Science and Technology, Saarland University

The N400 is widely taken to reflect processing effort related to a word's contextual predictability. Predictability can be estimated using human cloze probabilities or large language model (LLM) surprisal, with current debates centering on which better predicts neurophysiological processing [1, 2]. We test the hypothesis that the relative predictive power of cloze vs. surprisal depends on the extent to which participants generate strong, lexically specific predictions. We analyze German ERP data showing greater attenuation of N400 amplitudes for predictable words through a task manipulation enhancing engagement of the production system during comprehension [4]. If surprisal has a general advantage, as advocated by some [2, 3], it should outperform cloze regardless of task-modulated prediction strength. Alternatively, we suggest that this advantage may diminish when the task induces stronger, more specific lexical predictions. Indeed, because it is based on single-word production responses, cloze may be better suited to capture more categorical distinctions between predictable and unpredictable continuations in highly constraining contexts when the task encourages stronger commitment to the most predictable lexical candidates.

The ERP dataset from [4] used a 2x3 design crossing Task and Expectancy: Reading-for-Comprehension (RfC; comprehension questions) vs. Reading-for-Production (RfP; cloze-style sentence completion after each trial), and three expectancy levels (high cloze, HC; low cloze but plausible, LC; implausible, IM). Crucially, RfP showed a stronger N400 attenuation for HC vs. LC compared to RfC (Fig. 2a), indicating that the best lexical candidate receives a disproportionately large processing advantage. We compared cloze and surprisal as predictors of N400 amplitude by fitting linear mixed-effects models and comparing predictor models to baseline models via AIC. Predictors included surprisal from four German transformer LLMs (GerPT2 small, GerPT2 large, LeoLM, LLäMmlein) and two cloze measures (raw cloze, log-cloze). We first fit models to the full dataset containing both tasks, where log-cloze showed the largest improvement in model fit (Fig. 1a). Crucially, task-specific analyses revealed a dissociation: the largest LLM trained on German-only data (LLäMmlein) best predicted N400 amplitude in RfC (Fig. 1b), whereas both cloze measures outperformed all surprisal predictors in RfP (Fig.1c). We further conducted rERP analyses [5] using the best-performing surprisal and cloze predictors to re-estimate ERPs, showing that log-cloze closely captured the full N400 pattern (Fig. 2b), whereas surprisal underestimated the increased HC-LC gap in RfP (Fig. 2c). This pattern further held in GAMMs and when trials from the implausible condition were excluded, indicating that surprisal systematically fails to capture the task-induced increase in the gap between highly predictable and less predictable continuations. Indeed, it appears that the task manipulation made categorical HC-LC differences more relevant than within-condition variability in RfP; however, surprisal from all LLMs tested here conflates HC and LC conditions, therefore underestimating the processing advantage of HC words in a production-like task regime (Fig. 3).

Our results suggest that the predictive power of surprisal and cloze dissociates depending on the strength and mechanisms of prediction: surprisal best captures graded probabilistic expectations during standard reading, whereas cloze best captures stronger lexical predictions when comprehension becomes more production-like. Taken together, these findings highlight that cloze and surprisal provide complementary accounts of N400 amplitude in controlled psycholinguistic experiments by capturing distinct aspects of prediction, underscoring that the two measures should not be treated as interchangeable, and that claims to replace cloze with LLM-based surprisal should be treated with caution.

Figure 1: Improvements in AIC relative to the baseline model ($\Delta AIC = AIC_{baseline} - AIC_{predictor}$) across all predictors. All models included standard lexical control predictors and random intercepts for Item, Subject, and Electrode.

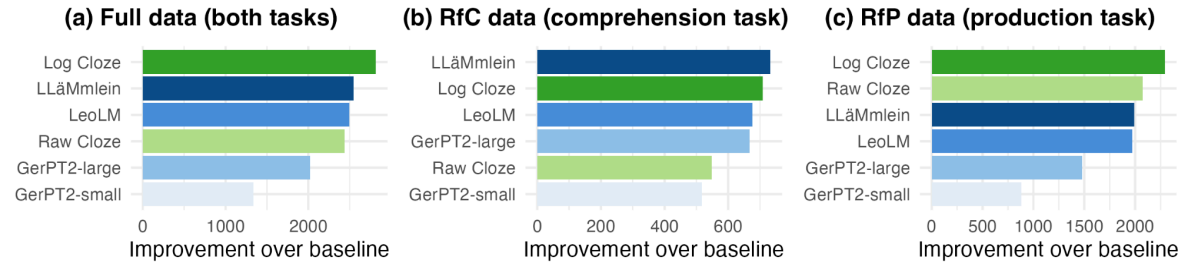


Figure 2: *Left:* Observed N400 amplitudes (in μV) in the experimental data from [4]. *Middle & right:* rERP estimates obtained by re-estimating the observed data with the best-performing cloze (log-cloze; middle) and LLM surprisal (LLäMmlein; right) metric, and corresponding residuals for the rERP estimates (below).

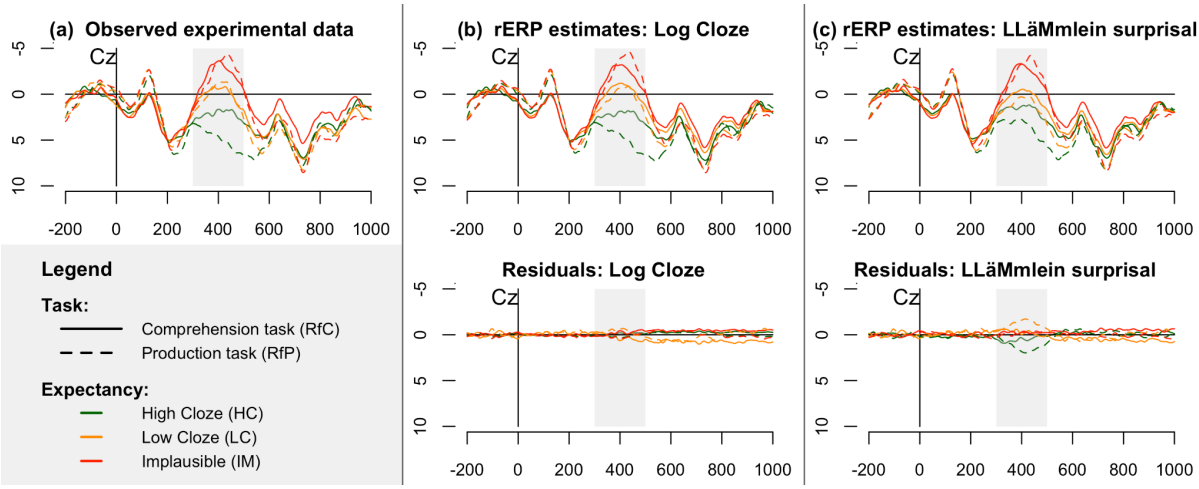
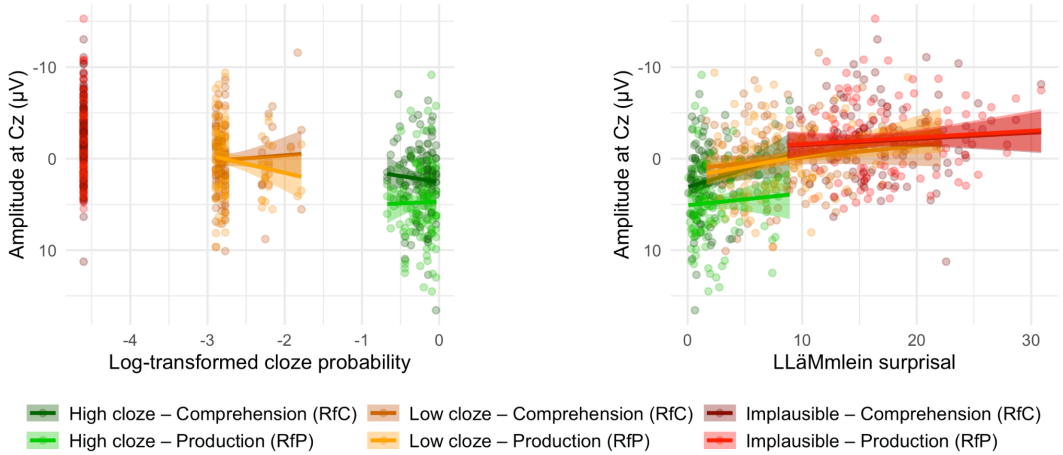


Figure 3: Correlations of mean N400 amplitude (300 to 500 ms) per item, task and condition vs. log cloze (smoothed with α=0.01; left) or vs. LLäMmlein surprisal (right).



References

[1] Arkhipova, Y., Lopopolo, A., Vasishth, S., & Rabovsky, M. (2025). *BioRxiv* preprint.

[2] Michaelov, J. A., Coulson, S., & Bergen, B. K. (2022). *IEEE TCDS.*

[3] Nair, S., & Oh, B. D. (2026). *arXiv* preprint.

[4] Schulz, M., Nakamura, M., Delogu, F., & Crocker, M. W. (2026). *Proc. CogSci.*

[5] Smith, N. J., & Kutas, M. (2015). *Psychophysiology*.