Revisiting Joke Comprehension with Surprisal and Contextual Similarity: Implication from N400 and P600 Components

Haoyin Xu (hyx002@ucsd.edu) and Masaki Nakanishi (masaki@sccn.ucsd.edu)
Swartz Center for Computational Neuroscience, University of California San Diego
9500 Gilman Drive, La Jolla, CA 92093, USA

Seana Coulson (scoulson@ucsd.edu)
Department of Cognitive Science, University of California San Diego
9500 Gilman Drive, La Jolla, CA 92093, USA

Abstract

Recent studies link *surprisal*—a measure of conditional probability of words in context—to the N400 component size in event-related potentials (ERP), supporting a role for predictive coding in language comprehension. An alternative account argues that N400 variations are better explained by a retrieval mechanism sensitive to the semantic similarity between a word and its preceding context. Because jokes often rely on the presence of unexpected words that relate to the prior context multiple ways, they afford observation of the relative importance of contextual predictability and contextual similarity. We employed state-of-the-art machine learning to assess the surprisal and contextual semantic similarity of critical words in jokes and control stimuli. Using regression models to predict ERP, we found contextual similarity best explains N400 and P600 responses, supporting the semantic similarity account. Additionally, jokes elicit enhanced N400 and P600 responses that go beyond that attributable to their surprisal and contextual semantic similarity.

Keywords: EEG, Joke Comprehension, Contextual Similarity, Surprisal

Introduction

One amazing ability of human cognition is to comprehend language by transforming streams of linguistic input into high-level representations of the world. Imagine hearing someone telling you "You need help." Depending on the speaker and circumstances, it can be a genuine concern or an aggressive judgment. While it is possible to assign abstract meanings to words and sentences, the meanings they assume in particular utterances can be quite different. Mainstream research in the language comprehension domain has been primarily concerned with how context influences the processing of individual words, yet how words contribute to the message level representation was often overlooked. In fact, many sentences we encounter in daily life are not straightforward. Consider another joke example: *"I let my accountant do my taxes because it saves time: last spring it saved me ten years."* To fully understand this sentence, one would need to reorganize the existing elements and perform a semantic and pragmatic reanalysis to construct meaning at the message-level, which is referred to as the *frame-shifting* operation (Coulson, 2001). This operation is commonly seen in the comprehension of figurative language where speakers exploit cognitive operations such as metaphorical mapping and conceptual blending to construct enriched meanings in context.

To probe the neurocognitive mechanism for joke comprehension, one popular method in language research is the use of event-related potential (ERP) components. Two ERP components of particular interest for language researchers include the N400, a negative-going component peaking approximately 400ms after the onset of a visually presented word; and the P600, a positive response evident approximately 600 to 1000ms after word onset. According to the retrieval-integration model, the N400 component indexes the context-sensitive retrieval of word meaning from long-term memory, and the P600 components indexes the integration of this meaning into the unfolding utterance interpretation (Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Brouwer, Fitz, & Hoeks, 2012; Van Berkum, Sauerland, & Yatsushiro, 2009). This model can potentially accommodate extant ERP findings on joke processing that suggest jokes elicit larger N400 and P600 responses than control stimuli, presumably due to demands on retrieval and integration of the word that triggers frame-shifting.

However, the functional significance of the N400 is still under debate. Some have argued that the N400 component is driven by the conditional probability of words in their linguistic context, an account that posits an underlying predictive coding mechanism (Brothers, Wlotko, Warnke, & Kuperberg, 2020). Others have suggested N400 amplitude is driven by a context-sensitive retrieval mechanism and as such indexes the semantic similarity of incoming words to the semantic features of prior words in the context (DeLong & Kutas, 2020). Both views are supported by a substantial body of empirical evidence (Kuperberg & Jaeger, 2016; Aurnhammer & Frank, 2019; Merkx & Frank, 2021; Kutas & Federmeier, 2011; Chwilla & Kolk, 2005; Van Petten, 2014; Chwilla, Kolk, & Vissers, 2007), yet the conclusion remains unclear due to the fact that contextual predictability is inevitably highly correlated with contextual similarity.

Fortunately, recent advances in generative language models can allow language scientists to measure the contextual predictability and contextual similarity of words in a large variety of linguistic contexts and thus provide a great opportunity to revisit these questions. For example, Michaelov et al. (2024) used state-of-the-art computational tools – neural language models and word embeddings – to determine whether contextual predictability or similarity provides a better account of N400 variance. They operationalized contextual pre-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

dictability by extracting surprisal values from large language models, and operationalized contextual semantic similarity as the cosine distance between co-occurrence-derived vector-based meaning representations for words. Through statistical model comparison, they found that the GPT-3 surprisal provided the best account of N400 amplitude and multiple N400 effects can be reduced to variation in the predictability of words (Michaelov et al., 2024). Based on these results, Michaelov and colleagues concluded that the link between contextual probability and N400 supports the use of predictive coding in the human language network.

Current Study

As jokes have been argued to provide an excellent test-case for any comprehensive account of language comprehension (Coulson, 2001), here we ask whether the enhanced response to jokes in the ERPs can be better explained by the contextual predictability or the contextual similarity of the critical words. To answer this question, we conducted a study that mirrors Michaelov's experimental design, including jokes in the stimuli. We aim to examine whether the predictive or the contextual similarity account provides a better explanation for the N400 and P600 amplitudes. If the predictive account provides a better explanation of N400 variance, then surprisal should show significant effects in regression models of its amplitude. If contextual similarity provides a better explanation, the contextual similarity measure should show significant effects. Further, the same analysis was applied to P600 amplitudes to investigate the neurocognitive mechanism behind the integration stage.

Materials and Methods

The original experimental materials and EEG data were provided by the authors (Coulson & Lovett, 2004). We analyzed the stimuli to obtain both surprisal and contextual similarity measures for final words, and ran a series of linear mixed effect models to examine which were more closely related to the EEG (as in Michaelov et al., 2024). Details of the original experiment and analysis can be found below.

Participation

19 UCSD undergraduate students (8 female) were recruited to complete the study. All participants were right-handed, monolingual native English speakers with normal or corrected to normal vision and no history of reading difficulties or neurological/psychiatric disorders. All participants received academic credit or cash as compensation.

Stimuli

The stimuli included 400 sentences in total, comprised of 160 pairs of joke, straight sentences, and 80 expected sentences. The straight sentences shared the same sentence set-up with the jokes, but ended with a cloze-matched non-funny ending. (e.g. Joke: "It is hard to raise a family nowadays, especially in the *morning*." vs. Straight: "It is hard to raise a family nowadays, especially in the *country*."). For each pair of experimental sentences, one version was randomly assigned to one of two lists so that each participant saw only a single version. However, across participants, both versions were presented a similar number of times over the course of the study. The expected sentences were filler sentences with predictable endings (e.g. "Most cats see very well at *night*.") These filler sentences were identical in both lists. Each individual participant thus saw 240 sentences (80 jokes, 80 straight, and 80 expected sentences). Following each sentence, participants answered a true-false question that was intended to test comprehension of the materials.

Experimental Procedure

The EEG experiment consisted of a single session, with words presented centrally using RSVP presentation. Before each sentence, a fixation cross appeared to guide visual attention the the center of the screen where sentences were presented one word at a time. Each word was presented for a fixed duration of 200ms with an inter-stimulus interval (ISI) that varied as a function of word-length (i.e. 200ms $\pm$ 32ms / character). The final (critical) word of each sentence was followed by a blank screen for 2500ms before the presentation of the comprehension probe. Comprehension probes were presented for 4s, allowing participants to answer by pressing a button, and followed by a blank screen for 2s until the next trial began.

EEG Data Acquisition

Participants' EEG was recorded from 29 tin electrodes mounted in an Electro-Cap organized in the International 10-20 configuration. Additional electrodes were placed below the eye and near the external canthi to detect eye movements and blinks. Scalp electrodes were referenced online to an electrode on the left mastoid, and later re-referenced to an average of the left and right mastoid electrodes. The EEG was amplified using an SA Instruments bio-electric amplifier, and digitized online at 250 Hz.

EEG Pre-processing

EEG was time-locked to the onset of each sentence's final word. Mean voltage during the 200ms interval preceding each word's appearance was used to baseline epochs spanning 200ms before until 1000ms after word onset. Trials containing artifacts due to blinks, eye movements, or amplifier saturation were removed prior to analysis.

Computation Metrics

Two computational metrics were derived from NLP tools – GPT-3 surprisal and GloVe cosine dis-similarity. Both models (GPT-3, and GloVe) are trained on the Common Crawl corpus (https://commoncrawl.org/), albeit using subsets of different sizes. GPT-3 is trained on 300 billion tokens and GloVe is trained on 840 billion tokens.

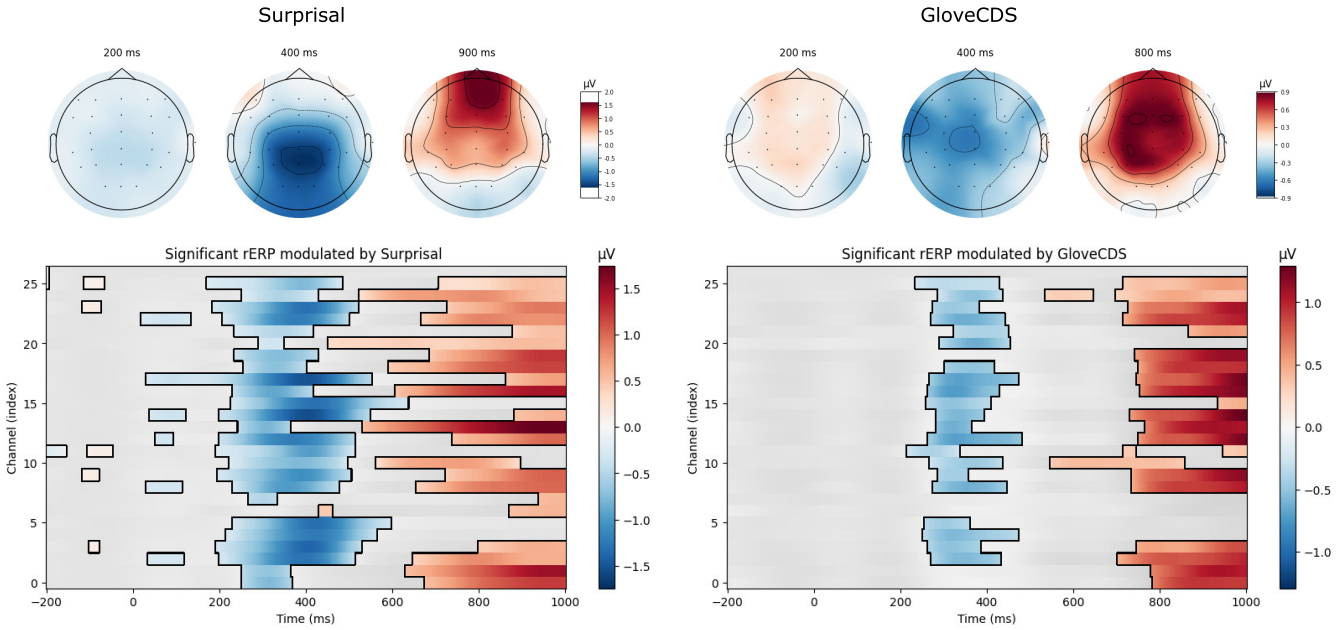

Figure 1: Topography and significant modulation of single-trial EEGs from surprisal and GloVe contextual cosine dis-similarity (GloveCDS). All p values are corrected for multiple comparisons based on false discovery rate. Left top: topography of surprisal effect at 200ms, 400ms, and 900ms after the onset of final words. Left bottom: Significant positive or negative modulation from surprisal throughout the epoch. Blue indicates significant negative effects and red indicates significant positive effects. Right top: topography of GloveCDS effect at 200ms, 400ms, and 900ms after the onset of final words. Right bottom: Significant positive and negative modulations from GloveCDS throughout the epoch.

GPT-3 Surprisal Surprisal of a word denotes the log-transformed conditional probability of a word based on the preceding context. It has previously been found to index human lexical prediction processing as it correlates with increased N400 amplitude (Frank, Otten, Galli, & Vigliocco, 2015). As in Michaelov et al., (2024), here we operationalize the predictive coding account by computing surprisals for materials using the *davinci-text-002* model from the OpenAI API (Brown et al., 2020). To access the conditional probability of sentence final words, each sentence stimulus was input into the API, and the davinci model was used to access the log probability of the final word. The log probabilities were then multiplied by -1 to yield their surprisals.

GloVe contextual cosine dis-similarity (GloveCDS) GloVe is an unsupervised learning algorithm trained on global, aggregated word-word co-occurrence statistics that yields vector representations for words. In this study, it is used to operationalize the semantic similarity account. The GloVe contextual cosine similarity was computed by using the GloVe vectors (Pennington, Socher, & Manning, 2014) available on their official website. Specifically, we used the version with a 2.2 million word vocabulary and 300-dimensional vectors trained on 840 billion tokens from the Common Crawl corpus. The contextual vector of all the words preceding the final word in each sentence was computed by averaging the vectors for each of the individual words in the sentence frame. Then the cosine similarity between this vector and the vector representing the final word was computed by using Scipy package on python (Virtanen et al., 2020). To better compare with surprisal effect, we transformed the Glove cosine contextual similarity (GloveCCS) to GloVe cosine contextual dis-similarity (GloveCDS) by multiplying the GloveCCS by negative one. Higher GloveCDS indicates a greater difference in the meaning of the sentence final word and the preceding words in the sentence frame.

The average surprisals for joke, straight, and expected conditions are 9.42, 7.29 and 0.91, respectively. The paired t-test indicates that the surprisals are significantly different between the two unexpected conditions (joke vs. straight: $t(159) = 5.08$, p-value < 0.001) and Welch's t-test indicated the straight condition differed from the expected: $t(229) = 16.01$, p-value < 0.001). The average GloveCDS for joke, straight, and expected conditions are -0.34, -0.37, and -0.42, respectively. Similarly, the paired t-test indicates a significant difference in the GloveCDS across the conditions (joke vs. straight:$t(159) = -2.04$, p-value = 0.02) and Welch's t-test revealed differences in the comparison of straight vs. expected: $t(220) = -2.64$, p-value = 0.0089). The overall Pearson's correlation between GPT-3 surprisal and GloveCDS is 0.32. ($r = 0.32$, p-value < 0.01).

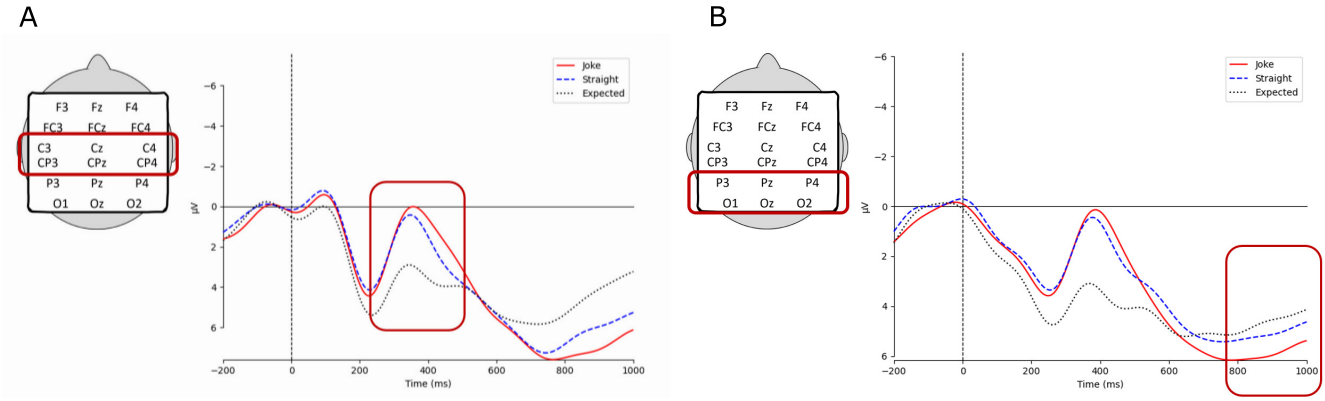

Figure 2: Event Related Potential (ERPs) waveforms with different groups of channels. (A) ERP waveforms in each condition (Joke, Straight, and Expected) at centro-parietal channels selected to best visualize the N400 component. In the N400 window (300-500ms), the joke condition elicited the largest N400 amplitude, followed by the straight condition, and the expected condition elicited the least negative N400 amplitude. (B) ERP waveforms in each condition (Joke, Straight, and Expected) at parietal channels selected to highlight the P600 component. In the P600 window (800-1000ms), the jokes elicited the largest P600 amplitude, followed by the straight condition, and the expected condition elicited the least positive P600.

rERP Analysis

Our initial analysis was intended visualize the topographic profile associated with EEG effects of predictability and contextual dissimilarity when each factor is treated as the sole driver of the brain response. Accordingly, we derived the rERP to observe each effect on a fine-grained spatial-temporal level (Smith & Kutas, 2015).

Method

The rERP analysis is a method conducted by constructing linear regression models to predict the single-trial EEG amplitude value at each time point and each channel. In this study, the independent variables were the selected predictors (i.e. surprisal and contextual dis-similarity measures). Then, the regression coefficient (β) of the model was plotted like ERPs to visualize the temporal and spatial distribution of the effects of predictors. All reported p values are corrected for multiple comparisons via false discovery rate (Benjamini & Yekutieli, 2001). In this study, data from all conditions were used, and single predictor regression model (e.g. amplitude $\sim$ intercept + surprisal and amplitude $\sim$ intercept + GloveCDS) was used to examine the modulation effect of predictors over time.

Results

Figure 1 shows the topography and the significant effect of surprisal and GloveCDS obtained from the rERP analysis. Overall, both surprisal and contextual dis-similarity exerted similar effects on the single trial EEGs. Both metrics were associated with negative modulation to the brain response in the N400 window and positive modulation during the late positivities window (600-1000ms). In other words, both unexpectedness and drastic semantic changes at the end of the sentence elicit large negative and positive modulation in N400 and late positivities window, respectively.

One major difference is the scalp distribution of the effects indicated by the topography. The N400 effect from surprisal is mostly observed in the posterior region, while the late positive effect is focused in the frontal region. Compared to surprisal, the GloveCDS effects were more broadly distributed during both the N400 and the P600 windows. Moreover, the surprisal effects seem to last longer than those of GloveCDS. Figure 1 suggests the negative modulation from surprisal begins at roughly 200ms and ends around 500ms, whereas the negative modulation from GloveCDS ends at around 400ms. Similarly, the late positive modulation from surprisal begins as early as 600ms after the onset of the critical word, while gloveCDS does not elicit positive effects until after approximately 750ms.

In summary, the rERP results suggest surprisal and contextual dis-similarity potentially both modulate the single trial EEG in a similar time window but with somewhat different scalp topographies (and potentially somewhat different underlying neural generators). Surprisal and contextual dissimilarity may independently modulate single trial EEG to words, but investigation of this possibility requires the additional power of mixed effects models as described below.

ERP Analysis

Given that the rERP analysis revealed activity from both factors within both components' time windows, we can now proceed to construct additive models that incorporate both surprisal and GloveCDS.

N400

ERP Waveform The N400 amplitude was operationalized as the mean voltage 300-500 ms post-onset recorded from six centroparietal electrodes: C3, Cz, C4, CP3, CPz, and CP4.

Figure 2 (left) shows the grand average ERP waveform for

Table 1: Linear Mixed Effect Regression Models and AICs

LME Models	Normalized AIC	
	N400	P600
M1: RV + Int + GloveCDS	-14	-8
M2: RV + Int + Surprisal	-30	0
M3: RV + Int + Condition	-32	-5
M4: RV + Int + GloveCDS + Condition	-37	**-10**
M5: RV + Int + Surprisal + Condition	-35	-3
M6: RV + Int + GloveCDS + Surprisal	-33	-6
M7: RV + Int + GloveCDS + Surprisal + Condition	**-39**	**-10**

RV: Random variables, Int: Random Intercept

Table 2: Linear Mixed Effect Regression Model Result

N400				P600			
BM: Amp $\sim$ RV + GloveCDS + Surprisal + Condition				BM: Amp $\sim$ RV + GloveCDS + Condition			
Predictors	Estimates	CI	p-value	Predictors	Estimates	CI	p-value
(Intercept)	11.52	[-3.49, 26.52]	0.133	*(Intercept)*	51.32	[30.85, 71.79]	$<$**0.001**
GloveCDS	-25.95	[-47.76, -4.15]	**0.020**	GloveCDS	30.02	[7.94, 52.10]	**0.008**
Condition (J)	-16.08	[-26.82, -5.33]	**0.003**	Condition (J)	9.95	[1.59, 18.32]	**0.020**
Condition (S)	-8.62	[-18.17, 0.93]	0.077	Condition (S)	5.91	[-2.32, 14.14]	0.159
Surprisal	-3.71	[-7.99, 0.56]	0.088				

BM: Best model, RV: Random variables, CI: Confidence Interval, J: Joke, S: Straight

final words in each condition (joke, straight, and expected) as typically measured in the ERP language literature (see (Kutas & Federmeier, 2011) for a review). In the N400 window, as predicted, the joke condition elicited the greatest (most negative) N400 amplitude, and the expected condition elicited the least negative (most positive) N400 amplitude. The straight condition fell in between, though its amplitude was much closer to the joke condition. This reflects the fact that both joke and straight sentences ended with low-cloze (high surprisal) words, that have traditionally been shown to elicit larger amplitude N400 than expected endings (Coulson & Lovett, 2004; Coulson & Kutas, 2001; Kutas & Federmeier, 2011).

Linear Mixed Effect (LME) Regression Models A series of linear mixed effect (LME) regression models were constructed to examine the relationship between surprisal and contextual dis-similarity with ERP component amplitudes. All models shared the same random effect structure, including random intercepts for the final words, subjects, and EEG channels. The regression model with the lowest Akaike information criterion (AIC) was selected as the best model (Akaike, 1973). The AIC of each regression, normalized by the AIC of the null regression (which includes the same random effects structure as the other regressions, and only has an intercept term as the fixed variable) is shown in Table 1. All fixed variables were normalized by z-scoring before fitting.

The best fitting model for N400 amplitude appears to be M7 (i.e. N400 amplitude $\sim$ Random Variables + GloveCDS + Surprisal + Condition). Model output is shown in Table 2. The model suggests N400 amplitude was significantly modulated both by GloveCDS (Estimate: -25.95, Confidence Interval (CI): [-47.76, -4.15], $p = 0.020$) and the Joke condition (Estimate: -8.62, CI: [-26.82, -5.33], $p = 0.003$); neither the straight condition (Estimate: -8.62, Confidence Interval (CI): [-18.17, 0.93], $p = 0.077$) nor the GPT-3 surprisal effect (Estimate: -3.71, CI: [-7.99, 0.56], $p = 0.088$) reached significance. This result suggests that both contextual dissimilarity and joke condition predict larger (more negative) N400 components when the variances of other predictors are controlled. More importantly, it shows that the joke effect persists when both GloveCDS and surprisal are controlled, suggesting that it is not attributable to either factor. The non-significant effect from the straight condition suggests that the larger N400 amplitude in response to straight endings compared to expected endings may be attributable to differences in GloveCDS and GPT3-surprisal for words in each of the conditions. It is worth pointing out that in the second-best model of N400 amplitude (M4, normalized AIC = -37), the straight condition effect was significant (Estimate: -13.29, CI: [-21.22, -5.37], $p = 0.001$). Thus, it is possible that the inclusion of surprisal in M7 explained away the main effect from the straight condition effect evident in M4.

P600

ERP Waveform P600 amplitude was operationalized as the mean voltage 800-1000 ms post-onset recorded from six occipital-parietal electrodes: P3, Pz, P4, O1, Oz, and O2.

Figure 2 (right) shows the grand average for final words in each condition (joke, straight, and expected). In the P600

windows, the joke condition elicited the greatest (most positive) P600 amplitude while the expected condition elicited the least positive P600 amplitude (Figure 2). The straight condition fell in the middle of the two conditions. The large P600 amplitudes elicited by joke and straight conditions are also consistent with previous studies of jokes and other unexpected, yet, plausible sentence completions (Coulson & Williams, 2005; Kuperberg, Brothers, & Wlotko, 2020; DeLong & Kutas, 2020).

LME Regression Models A series of LME models were constructed and compared in a similar fashion as in N400 analysis to predict P600 amplitudes. As shown in Table 1, both M4 and M7 shared the same AIC score, suggesting the addition of surprisal did not significantly improve the model quality. In fact, M7 suggests that after controlling for contextual similarity, surprisal was not significantly related to P600 amplitudes (Estimate: -3.33, CI: [-7.79, 1.13], p = 0.14). Thus, we selected M4 (i.e. P600 amplitude $\sim$ Random Variables + GloveCDS + Condition) as the best model for interpretation. The model indicated that there are main effects from intercept (Estimate: 51.32, CI: [30.85, 71.79], $p < 0.001$), GloveCDS (Estimate: 30.02, CI: [7.94, 52.10], $p < 0.008$), and joke conditional effect (Estimate: 9.95, CI: [1.59, 18.32], $p < 0.020$). The straight conditional effect (Estimate: 5.91, CI: [-2.32, 14.14], $p = 0.159$) was not significant. Both the contextual dis-similarity and joke condition effects reflect larger amplitude P600. Similar to the N400 results, these results also suggest the joke effect is not attributable to the contextual dis-similarity of joke endings to the preceding sentence frames. By contrast, larger P600 to straight than expected endings can be explained by differences in the contextual dis-similarity of the sentence final words in each.

Discussion

Here we aim to investigate whether the strong account of predictive coding found in Michaelov's study can be leveraged in sentences that require reorganization of the existing elements to achieve full comprehension (i.e., frame-shifting). To address this question, we applied a series of linear mixed-effect regression models to EEG signals of adults reading sentences with endings that vary in their contextual predictability (i.e., expected vs. straight sentences), their contextual similarity, and whether they require frame-shifting to fully understand their meaning (i.e., the jokes). In contrast to the findings reported by Michaelov and colleagues (2024), our results suggest contextual (dis)similarity elicits a stronger effect on both N400 and P600 amplitude.

One seemingly obvious reason for the discrepancy is the difference in the stimuli in the two studies – namely, the inclusion of one-line jokes in the present study. The straight sentences in the present study are similar to the Related condition in Michaelov's study in that they were chosen to be consistent with the frame or schema evoked by the sentence context. In keeping with their study, our results do suggest ERP amplitude differences between straight and expected sentences can be explained by their differing surprisals. Presumably, when language comprehension does not require the additional demands posed by frame-shifting, N400 amplitude is mainly driven by surprisal of the critical word. In the case of the jokes, when re-constructing the meaning of the preceding words, the contextual (dis)similarity of the eliciting word has additional relevance. Consequently, the contextual (dis)similarity of critical words was more closely tethered to N400 amplitude. This raises the possibility that rather than relying exclusively on either predictive coding or semantic activation, language comprehension may recruit both mechanisms to provide flexible processing of incoming language.

Szewczyk and Federmeier's work (2022) on context-based facilitation and predictability argue for both linear and logarithmic relationship between N400 and word predictability provided by word predictability and context-based facilitation, thereby establishing the probability that feature- and word-related information is both activated online (Szewczyk & Federmeier, 2022). More recently, Federmeier has suggested that during "active comprehension", the brain may dynamically switch between both systems to select the best way to bind semantic information and afford immediate action (e.g. constructing predictions) (Federmeier, 2022). Findings of the present study are in keeping with their dynamic combined account of the N400.

The present study also has implications for joke comprehension and the neurocognitive basis of frame-shifting. Our analyses showed that the large N400 elicited by words in the joke condition was not fully accounted for by either their slightly higher surprisals than words in the straight condition or by their greater contextual dissimilarity. By contrast, differences between N400s to straight and expected sentences were better captured by surprisal and contextual dissimilarity. We suggest that while these metrics provide a fairly good measure of the demands of semantic retrieval, they do not fully capture the way that higher level knowledge structures impact real time retrieval during language processing. Similarly, the LMER analysis of the P600 also indicated that the joke effects were larger than could be accounted for by their contextual (dis)similarity, and may thus reflect processes related to frame-shifting. That is, neural indices of the integration of the joke endings into readers' situation models suggested the associated cognitive demands were not captured by contextual semantic distance as operationalized here, and thus in keeping with the suggestion that joke comprehension requires the reorganization of the extant situation model.

In sum, we found that relative to surprisal, contextual (dis)similarity provides a better account of the size of the N400 and P600. These findings suggest that predictive activation and contextual similarity both influence human language comprehension. However, more research is needed with language materials that vary in their contextual predictability, contextual similarity, and the extent that they prompt reanalysis of prior elements of the context.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Proceedings of the 2nd international symposium on information theory* (pp. 267–281). Akadémiai Kiadó. doi: 10.1007/978-1-4612-1694-0$_1$5

Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107198. doi: 10.1016/j.neuropsychologia.2019.107198

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188. Retrieved 2024-01-26, from http://www.jstor.org/stable/2674075

Brothers, T., Wlotko, E. W., Warnke, L., & Kuperberg, G. R. (2020). Going the extra mile: Effects of discourse context on two late positivities during language comprehension. *Neurobiology of Language*, *1*(1), 135–160. doi: 10.1162/nol$_{a0}$0006

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A neurocomputational model of the n400 and the p600 in language processing. *Cognitive Science*, *41*, 1318–1352. doi: 10.1111/cogs.12461

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the p600 in language comprehension. *Brain Research*, *1446*, 127–143. doi: 10.1016/j.brainres.2012.01.055

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Chwilla, D. J., & Kolk, H. H. (2005). Accessing world knowledge: evidence from n400 and reaction time priming. *Brain Research. Cognitive Brain Research*, *25*(3), 589–606. doi: 10.1016/j.cogbrainres.2005.08.011

Chwilla, D. J., Kolk, H. H., & Vissers, C. T. (2007). Immediate integration of novel meanings: N400 support for an embodied view of language comprehension. *Brain Research*, *1183*, 109-123. doi: https://doi.org/10.1016/j.brainres.2007.09.014

Coulson, S. (2001). *Semantic leaps: Frame-shifting and conceptual blending in meaning construction*. Cambridge University Press. doi: 10.1017/CBO9780511551352

Coulson, S., & Kutas, M. (2001). Getting it: human event-related brain response to jokes in good and poor comprehenders. *Neuroscience Letters*, *316*(2), 71-74. doi: https://doi.org/10.1016/S0304-3940(01)02387-4

Coulson, S., & Lovett, C. (2004). Handedness, hemispheric asymmetries, and joke comprehension. *Cognitive Brain Research*, *19*(3), 275-288. doi: https://doi.org/10.1016/j.cogbrainres.2003.11.015

Coulson, S., & Williams, R. F. (2005). Hemispheric asymmetries and joke comprehension. *Neuropsychologia*, *43*(1), 128–141. doi: 10.1016/j.neuropsychologia.2004.03.015

DeLong, K. A., & Kutas, M. (2020). Comprehending surprising sentences: sensitivity of post-n400 positivities to contextual congruity and semantic relatedness. *Language, cognition and neuroscience*, *35*(8), 1044–1063. doi: 10.1080/23273798.2019.1708960

Federmeier, K. D. (2022). Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology*, *59*(1), e13940. doi: https://doi.org/10.1111/psyp.13940

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1-11. doi: https://doi.org/10.1016/j.bandl.2014.10.006

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (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. doi: 10.1162/jocn$_{a0}$1465

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, *31*(1), 32–59. doi: 10.1080/23273798.2015.1102299

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, *62*, 621–647. doi: 10.1146/annurev.psych.093008.131123

Merkx, D., & Frank, S. L. (2021, June). Human sentence processing: Recurrence or attention? In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 12–22). Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.2

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2024). Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of Language*. doi: 10.1162/nol$_{a0}$0105

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543). doi: 10.3115/v1/D14-1162

Smith, N. J., & Kutas, M. (2015). Regression-based estimation of erp waveforms: I. the rerp framework. *Psychophysiology*, *52*(2), 157–168. doi: 10.1111/psyp.12317

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, *123*, 104311. doi: https://doi.org/10.1016/j.jml.2021.104311

Van Berkum, J. J. A., Sauerland, U., & Yatsushiro, K. (2009). *Semantics and pragmatics: From experiment to theory*. Palgrave Macmillan.

Van Petten, C. (2014). Examining the n400 semantic context effect item-by-item: Relationship to corpus-

based measures of word co-occurrence. *International Journal of Psychophysiology*, *94*(3), 407–419. doi: 10.1016/j.ijpsycho.2014.10.012

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... others (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2