THE N400 IS NOT A SEMANTIC ANOMALY RESPONSE: MORE EVIDENCE FROM ADJECTIVE-NOUN COMBINATION

Ellen F. Lau
Anna Namyst
Allison Fogel\textsuperscript{1,2}
Tania Delgado

\textsuperscript{1} – University of Maryland, Department of Linguistics, College Park, MD 20742
\textsuperscript{2} – Tufts University, Department of Psychology, Medford, MA 02155

Corresponding author:
Ellen Lau (ellenlau@umd.edu)
1 301 405 4934
University of Maryland
Department of Linguistics
College Park, MD 20742
Abstract

Previous work has shown that the N400 ERP component is elicited by all words, whether presented in isolation or in structured contexts, and that its amplitude is modulated by semantic association and contextual predictability. Although these data suggest that the N400 reflects core semantic computations involved in processing any meaningful stimulus, the broader literature continues to maintain the alternative idea that the N400 response primarily indexes the recognition of semantic incongruity, in part because many designs have confounded predictability with congruity. In the current study, we evaluate the hypothesis that the N400 response indexes semantic anomaly with an adjective-noun paradigm that allows us to precisely control predictability through corpus counts. In three experiments, we find small and unreliable N400 effects of semantic congruity (yellow bag vs. innocent bag), and yet we find massive and reliable N400 effects of predictability (runny nose vs. dainty nose) under the same conditions. While these data alone cannot determine the functional interpretation of the N400 effect, they provide very clear evidence against the common characterization that ‘semantic anomaly’ is one of its primary determinants.

Introduction

One of the core components of language comprehension is accessing the stored meanings of the individual words in a sentence or discourse and combining them to construct a global interpretation of the speaker’s message. Many critical questions about this process—questions about speed, automaticity, interactivity, and neural implementation—have been explored in the last 30 years through work with the ERP component known as the N400, which is modulated by a vast array of lexical and contextual factors thought to influence the processing of meaning (see Kutas & Federmeier, 2011, for review). This work has been so important and influential that the N400 response is familiar not only to specialists in the electrophysiology of language, but to many in the fields of cognitive psychology and cognitive neuroscience at large, appearing in broader reviews and textbook chapters. In this larger realm, the N400 is most often characterized as a response to ‘semantic anomaly’ (Ward, 2010), ‘semantic violation’ (Purves et al., 2013), ‘semantic incongruity’ (Harley, 2008) or ‘semantic mismatch’ (Gazzaniga et al., 2009) when the current word doesn’t ‘make sense’ in the context. In keeping with this, the example most often cited in textbooks comes from Kutas and Hillyard’s seminal 1980 work first reporting contextual modulation of the N400, in which they compared the response to congruous and incongruous sentence endings such as It was his first day at work and He spread the warm bread with socks.

The goal of the current work, following many previous investigators beginning with Kutas and Hillyard themselves (1984), is modest: to provide a simple and concise illustration of why this common characterization of the N400 as a ‘semantic anomaly’ response is misleading and fundamentally incorrect. In a review of past work and in data
from three new ERP experiments, we highlight the fact that N400 amplitude (a) is modulated by manipulations that do not involve what is colloquially understood to be ‘semantic anomaly’ and (b) is not modulated by at least some manipulations that do involve what is colloquially understood to be ‘semantic anomaly’. The important consequence of these facts is that N400 amplitude cannot be used by psycholinguists as a reliable and unambiguous indicator that comprehenders have computed a full message-level interpretation of the sentence and/or detected that such an interpretation violates semantic or world knowledge.

We must note at the outset that this conclusion is certainly not a new one, and that in fact to most of those investigators who specialize in research involving the N400 component, the idea that the N400 effect is primarily a semantic anomaly response has been clearly repudiated by a large body of work beginning with Kutas and Hillyard’s own 1984 follow-up. However, we have found the N400-as-semantic-anomaly-response hypothesis to be so pervasive among the larger body of researchers who work on the psychology and cognitive neuroscience of language, and so costly in terms of mistakes in experimental design and interpretation when investigators rely on this idea, that we believe that making the argument against it even more explicit is important and necessary. And, while this endeavor may appear to pertain only to the narrower segment of language processing researchers that use ERPs, we believe that a better understanding of the functional interpretation of an ERP response like the N400 effect among psycholinguists more broadly would make such responses even more useful in the larger effort of understanding language comprehension.

We also note that among N400 specialists, debate about the functional interpretation of the response over the last decade has centered on two competing accounts, an ‘activation’ view (e.g. Kutas & Federmeier, 2000; Lau, Phillips, & Poeppel, 2008) and an ‘integration’ view (e.g. Osterhout & Holcomb, 1992; Brown & Hagoort, 1993; Hagoort, 2008). Although we think the current data is relevant to this debate, it is important to be clear that the experiments that follow were not designed to decide between these accounts. Rather, the narrow goal of this study is to directly evaluate the common conception outside of this core of specialists that N400 effects primarily reflect semantic anomaly. It seems to us that a commonsense prediction of any such characterization is that robust N400 effects should be observed when semantically anomalous phrases are presented, and that such effects should be at least as large as the N400 effects observed for contextual manipulations of non-anomalous phrases conducted under the same conditions. Although many alternative accounts can explain why this prediction is not fulfilled, we do not think that any of them could reasonably maintain that the N400 effect primarily reflects semantic anomaly.

*The N400 response*

The N400 is a negative deflection in the ERP that peaks at around 400 ms after stimulus presentation and is largest over centro-parietal sites, with a slightly rightward focus when visual presentation is used (Kutas & Van Petten, 1994; Van Petten & Luka, 2006). This
component first came to the attention of the field with Kutas and Hillyard’s (1980) study. As noted above, their classic example sentences were:

(congruous) It was his first day at work.
(incongruous) He spread the warm bread with socks.

Kutas and Hillyard observed a large negative deflection in response to words that were incongruous in a sentence context, whereas the words that were congruous in the sentence context showed no negative deflection at all. Because of the strikingly anomalous interpretation associated with the incongruous sentences and the complete absence of a negative deflection for the congruous completions, Kutas and Hillyard’s first hypothesis was that ‘N400 is not a general response to all linguistic or meaningful stimuli…Rather, the N400 seems to reflect the interruption of ongoing sentence processing by a semantically inappropriate word’.

However, in another landmark study four years later, Kutas and Hillyard (1984) demonstrated evidence that directly undermined this earlier hypothesis. They showed that large N400 responses could be observed for words that were completely congruous in their context (‘Don’t touch the wet dog). The variation in N400 amplitude associated with different contexts appeared to rather be due to the degree to which the context predicted the target, as in this comparison:

(predicted) Don’t touch the wet paint
(not predicted) Don’t touch the wet dog

Kutas and Hillyard (1984) observed that N400 amplitude appeared to vary parametrically with the degree of predictability, such that a negative deflection was observed in all but the most predicted condition, but where the size of this negative deflection was inversely related to predictability. They noted that predictability could also have explained their earlier (1980) results, as the congruous words were probably more predictable than the incongruous words. Therefore, Kutas and Hillyard (1984) concluded that ‘These results are in agreement with the hypothesis that the N400 component reflects the extent to which a word is semantically primed, rather than its being a specific response to contextual violations’ (p. 163).

Many subsequent studies replicated and extended these results. In one example, Federmeier and Kutas (1999) found that sentence completions that were equally semantically incongruous showed differences in N400 amplitude that corresponded to how closely related the incongruous completion was to the contextually predictable ending (e.g. He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of monopoly/baseball). Relatedly, Hoeks et al. (2004) showed that sentence completions that were equally semantically incongruous showed N400 differences as a function of how related the incongruous word was to other words and events in the context (e.g. [Dutch] The bread has the baker baked/summarized). In another example, DeLong, Urbach, and Kutas (2005) showed differences in N400 amplitude corresponding to the predictability of the semantically empty a/an alternation
conditioned on the predictability of the subsequent noun (e.g. *The day was breezy so the boy went outside to fly a / an …*). More generally, a large body of ERP language studies have shown that it is not only anomalous words that elicit a negative deflection peaking at 400ms, but in fact almost all words show this response, whether they are presented in isolation or in context (Kutas, Van Petten, & Kluender, 2006; Barber & Kutas, 2007). Finally, work on brain localization of the N400 effect suggest that the effect is primarily generated in areas of anterior and posterior temporal cortex that have been linked to storage and access of lexical and conceptual representations (Patterson, Nestor & Rogers, 2007; Martin, 2007), as would be predicted if the N400 component reflected typical processes involved in activating meanings (e.g. McCarthy, Nobre, Bentin & Spencer, 1995; Halgren et al., 2002; Van Petten & Luka, 2006; Lau, Phillips, & Poeppel, 2008; Lau, Weber, Gramfort, Hämäläinen, & Kuperberg, 2014; cf Hagoort, Hald, Bastiaansen, & Petersson, 2004); although we note that other work implicating anterior temporal cortex in combinatorial processes could be consistent with a semantic anomaly account (e.g. Vandenberghe, Nobre, & Price, 2002; Bemis & Pylkkänen, 2011; Jeffries, 2013; Molinaro, Paz-Alonso, Duñabeitia, & Carreiras, 2015).

All these results would be difficult to explain if the N400 component indexed a response to semantic violations, but follow directly from the hypothesis that the N400 reflects processes involved in comprehending all words that are facilitated when words are predictable in context. Despite this massive body of work, however, the initial idea that the N400 reflects the response to semantic incongruity has had a strong and long-lasting impact on the field. This view has important theoretical consequences. In an early example, Fischler and colleagues (1983) argued that sentential negation takes considerable time to process because *A sparrow is not a bird* demonstrated a smaller N400 than *A sparrow is not a vehicle*. However, if the N400 indexes normal access of lexical meaning rather than a response to semantic violations, this result can easily be explained by semantic priming or contextual prediction (Nieuwland & Kuperberg, 2008). More recently, the lack of an N400 difference between role-reversal sentences such as ‘The meal was devoured’ and ‘The meal was devouring’ has been argued by many to constitute evidence that readers do not immediately recognize the semantic violation in the second sentence (Kolk, Chwilla, van Herten & Oor, 2003; Kim & Osterhout, 2005). Again, however, if the N400 indexes normal access of lexical meaning rather than a response to semantic violations, this result could also be easily accounted for by the high semantic association between the argument(s) and the verb independent of argument structure (Hoeks et al., 2004; Stroud, 2008; Brouwer, Fitz, & Hoeks, 2012; Chow & Phillips, 2013). While these cases illustrate the incorrect conclusions that may be drawn if the N400 is erroneously taken to reflect a response to semantic incongruity, another consequence is that, for many years, less effort was devoted to identifying other neural indices of sentence- and discourse-level interpretation (van Berkum, 2009; Pylkkänen, Oliveri, & Smart, 2009; Brouwer et al., 2012; Hoeks & Brouwer, 2014).

If N400 amplitude does not index the presence of semantic incongruity, two further questions that arise are a) is there an ERP response that *does* index semantic incongruity, and b) what kinds of computations *should* be modulated by semantic incongruity? In the last decade, many researchers have demonstrated that semantic incongruity often, but not
always, elicits a late posterior positivity or ‘P600’. Although the functional interpretation of this response has been debated (Kolk et al., 2003; Kim & Osterhout, 2005; Kuperberg, 2007; Brouwer et al., 2012), Van Petten and Luka (2012) review a number of good arguments in favor of the hypothesis that semantic incongruity invokes memory retrieval processes associated with attempted reanalysis, which elicits a late posterior positivity. First, late posterior positivities are famously observed to syntactic violations and disambiguation in garden path sentences (Osterhout & Holcomb, 1992; Hagoort, Brown, & Groothusen, 1993), both of which could also plausibly invoke reanalysis. Second, in at least one study, semantic incongruity that was not detected (as assessed by behavioral report) did not evoke a late positivity (Sanford et al., 2011). Third, late positivities are also reported in studies of episodic memory as an old/new effect: items that are correctly recognized as previously studied (and are thus successfully retrieved from memory) elicit a larger late positivity than new items or unrecognized old items (Van Petten & Senkfor, 1996). Van Petten and Luka’s meta-analysis further supports this hypothesis, showing that while incongruity manipulations do not always elicit a late positivity, those that do elicit positivities with a parietal distribution like the syntactic late positivity, in contrast to predictability manipulations in which positivities are usually frontal. However, since not all cases of incongruity will elicit reanalysis in all subjects (some cases may be hard to notice, some cases may be so obviously bad that no reanalysis is attempted, some subjects may not be motivated to attempt reanalysis), even the late positivity cannot yet be taken as a straightforward index of whether semantic incongruity was detected.

The current study

We believe that the previous work reviewed above constitutes sufficient evidence to support the claim that the N400 effect primarily reflects predictability or lexical/semantic priming rather than a response to semantic violations. However, this argument mainly relies on the negative fact that most of the literature that interpreted the N400 as a semantic violation response systematically confounded semantic incongruity with contextual predictability. Even given the evidence that N400 amplitude tracks predictability and semantic association, one might still maintain that N400 amplitude also tracks semantic incongruity according to the hypothesis that a common process is modulated by both incongruity and predictability. For example, it could be that, rather than or in addition to invoking reanalysis as discussed above, incongruity invokes a prolonged search through semantic long-term memory; this increased activation of the semantic network would then increase N400 amplitude in the same way that decreased activation of the semantic network through priming or predictability reduces N400 amplitude. Alternatively, a multi-process theory could hold that N400 effects comprise multiple functionally disparate computations, such that some are due to contextual facilitation and some are due to error or repair processes invoked by semantic anomaly. In fact, there is surprisingly little existing evidence on the question of whether incongruity modulates N400 amplitude when contextual facilitation is held constant, and we speculate that this may be partially responsible for the failure of the ‘conventional wisdom’ on the N400 to change. Therefore, the goal of our study was to determine whether the N400 semantic anomaly effect remains when the critical congruous and incongruous words are controlled to be equally unpredictable.
Constructing materials that unconfound predictability and congruity in full sentences is challenging because in order to minimize the effects of contextual predictability, incongruous completions must be compared with extremely low probability congruous completions. However, it is difficult to accurately estimate probabilities on the lower end of the probability range for sentences with standard materials norming using the Cloze sentence completion task (Taylor, 1953), because this task encourages participants to respond with their most preferred completion. If individuals in fact maintain a probability distribution over possible endings, less probable endings may be underrepresented by the Cloze task (e.g. Smith & Levy, 2013). Longer sentence contexts could also contain semantic associates to the critical word. A recent sentence study by DeLong et al. (DeLong, Quante, & Kutas (2014) illustrates some of these challenges. While this study was targeted at dissociating frontal and posterior late positivities, DeLong et al. also showed that the N400 response to implausible sentence continuations was slightly but significantly larger than the response to unpredictable sentence continuations. However, they note that a small but significant difference in cloze for the two sentence continuations could also have been responsible for this difference.

In the current study we directly investigate whether semantic incongruity results in an N400 effect by using an adjective-noun paradigm that allows us to more precisely estimate these probabilities with corpus counts (Smith & Levy, 2011). In order to minimize effects of predictability on our congruity comparison, we compare incongruous adjective-noun combinations to congruous adjective-noun combinations in which the probability of the noun given the adjective is very low (p < .005) (note that it is the particular noun that is not predictable; in English, adjectives strongly predict that the word class of the following word is a noun). To create balanced congruous and incongruous sets, we crossed animate nouns and inanimate nouns with adjectives that must modify animate nouns and with adjectives that usually modify inanimate nouns, as in (1). We note that it is likely impossible to create semantically congruous items that are not slightly more predictable or more semantically associated than completely incongruous items. However, to the extent that this is the case, it could only act to work against our key hypothesis that N400 differences will be minimal when comparing congruous and incongruous items.

<table>
<thead>
<tr>
<th>congruous</th>
<th>incongruous</th>
</tr>
</thead>
<tbody>
<tr>
<td>(inanimate) yellow bag</td>
<td>(inanimate) innocent bag</td>
</tr>
<tr>
<td>(animate) healthy cat</td>
<td>(animate) empty cat</td>
</tr>
</tbody>
</table>

Previous authors (e.g., Hagoort et al., 2004; Pylkkänen et al., 2009) have pointed out that ‘semantic anomaly’ or ‘semantic incongruity’ are vague terms that can be taken to refer to various properties, from violations of world knowledge (e.g., in our world, people don’t usually put socks on bread) to mismatches between formal semantic features (e.g. inanimate entities clash with predicates that require animacy). We agree, and are not committed to a particular ontology of semantic well-formedness. Our incongruous items almost certainly vary in the properties that lead them to feel anomalous (e.g. one might argue that a bag cannot be innocent, but that a cat could be empty, much like socks can be
put on bread). Given the flexibility of language—e.g. the common use of metaphors, irony, and jokes; the ability to discuss fantastical or impossible worlds—it is difficult to construct grammatical examples that cannot be assigned some kind of coerced or accommodated interpretation. For example, in an appropriate ‘cartoon’ context, an animate predicate like falling in love is non-anomalous and perhaps even predictable given an inanimate subject like peanut (Nieuwland & Van Berkum, 2007). In the current study, our goal is to test the existing hypothesis that the N400 is impacted by semantic anomaly, and since this term has not received a precise definition in prior literature, we aimed to test items that would fit the conventional understanding of this term and which participants confirm to ‘not make sense’ in the absence of other context.

Several recent EEG and MEG studies have used adjective-noun manipulations to examine prediction and semantic combination mechanisms, as these designs make it possible to maintain tight control over the relevant (single-word) context being manipulated. In a series of MEG studies, Bemis and Pykkänen (e.g. 2011; 2013) have contrasted adjective-noun sequences such as red boat with noun-only sequences such as xkq boat in order to isolate the brain areas involved in semantic and syntactic combination. In another recent MEG experiment, Fruchter and colleagues use adjective-noun sequences varying in predictability to demonstrate neural activity associated with predicting the noun in the time-window prior to noun presentation (Fruchter, Linzen, Westerlund, & Marantz, 2015). These data confirm that single adjectives are enough to engender prediction of the subsequent noun, even outside of sentence contexts.

Most relevant for the current study, Molinaro and colleagues used EEG to examine the response to neutral, atypical, and anomalous noun-adjective sequences (monstruo solitario, monstruo hermoso, and monstruo geográfico, trans.: lonely monster, lovely monster, and geographical monster) embedded in sentences in Spanish (Molinaro, Carreiras, & Duñabeitia, 2012). Importantly, in all three conditions they used adjectives that elicited almost zero cloze in an offline sentence completion task, and yet still demonstrated a larger N400 response for the anomalous condition relative to the other three conditions. Therefore, these results appear to constitute evidence that N400 amplitude in fact does index semantic anomaly when predictability is controlled for. However, one potential caveat about this conclusion has to do with whether participants in the offline completion task made different predictions than the participants in the ERP study. Because adjectives are optional modifiers that come after the noun in Spanish, only 5% of offline completions contained any adjective at all; therefore the zero cloze for the critical adjectives reflected not that different adjectives were predicted but that no adjective was predicted. But in the ERP study, all of the experimental items and many of the fillers contained nouns modified by adjectives. While Molinaro et al. (2012) report that participants did not report being conscious of any particular construction, participants may have nonetheless implicitly recognized this regularity (Fine, Jaeger, Farmer, & Qian, 2013) such that in the experiment they would expect that a noun would be followed by an adjective. If semantic features of the neutral and atypical adjectives were predicted slightly more often than features of the anomalous adjectives given the noun during the ERP experiment, then this could explain the difference in N400 amplitude observed here. More generally, as in the DeLong et al. (2014) study discussed above, small residual
differences in cloze probability could have driven the N400 differences, and because the critical adjectives were embedded in multi-word sentence contexts it is more difficult to use corpora to estimate probabilities precisely. In the current study, the word category order of English ensured that participants were very likely to predict some noun after the adjective, and presenting the adjective-noun sequences in isolation allowed us to estimate predictability purely on the basis of the adjective (at the cost of less naturalistic presentation).

Here we report the results of three ERP experiments. Experiment 1 was designed to confirm that N400 effects of predictability are observed in isolated adjective-noun sequences, where predictability is computed through corpus counts rather than offline Cloze tasks. Experiment 2 investigated whether semantic incongruity alone could affect N400 amplitude when items were uniformly unpredictable, as in (1). Experiment 3 directly contrasted the predictability and semantic incongruity effects of Experiments 1 and 2 in a within-subjects design. If N400 amplitude is sensitive to semantic incongruity independent of contextual predictability, we should expect to see robust differences between congruous and incongruous items across centro-parietal electrodes in the 300-500ms time-window. If instead N400 amplitude is primarily sensitive to contextual predictability and not semantic incongruity, we should expect to see minimal differences across centro-parietal electrodes, as the conditions are only minimally different in noun probability.

Experiment 1

Materials

Adjective-noun pairs were selected from the Corpus of Contemporary American English (COCA; Davies, 2009). As a first step, we identified 120 highly constraining adjectives using the following procedure. We extracted all adjective-noun bigrams from the full list of bigrams in COCA, and next selected the subset of these bigrams for which \( p(noun \mid adjective) > .50 \) and which appeared at least 10 times in the corpus. Items that were judged unlikely to be familiar to our participant population (e.g. *peroneal nerve, gordian knot*), items that were too constraining (for which only one noun seemed felicitous, e.g. *iodized salt*), and items that contained repetitions of words used in other items were excluded from the set, resulting in 120 highly constraining adjectives.

120 high probability items were created simply by combining the strongly constraining adjectives (e.g. *runny*) with the noun which they most frequently occurred with (e.g. *nose*), such that \( p(noun \mid adjective) > .50 \). 120 low probability items were created by pairing these same nouns (e.g. *nose*) with weakly constraining adjectives (e.g. *dainty*) for which \( p(noun \mid adjective) < .02 \), and for which the maximum \( p(noun \mid adjective) \) across all nouns was less than .15. Candidate adjectives which satisfied this criteria were automatically identified for each item and in case more than one such adjective existed, one was selected by hand. The stimuli properties for this comparison are presented in Table 1. <<Table 1 about here>>
We note that in the current design, the bigram frequency of the high probability condition was much higher than in the control condition. This was not a primary concern here because the probability comparison in the current experiment was mainly aimed at replicating the effects of probability that have been observed in sentence paradigms, where an analogous confound between lexical probability and probability of the overall event being described also holds; however, it does mean that this design cannot distinguish between activation and integration accounts of the N400, as discussed further below in the General Discussion.

The frequency of the adjectives in the preceding context also differed across conditions, such that adjectives in the high probability condition were more frequent than the adjective in the low probability condition. Past authors have reported effects of lexical frequency on the N400 (e.g. Rugg, 1990; Van Petten & Kutas, 1990). Although the stimulus-onset asynchrony between adjective and noun was long enough (600ms) in the current study that frequency differences at the adjective appear unlikely to impact time-windows of interest at the noun (starting at 300ms, 900ms after adjective onset), we conducted pairwise comparisons in the 500:600ms time-window following the adjective to confirm that this frequency difference on the adjective led to no baseline differences prior to the onset of the noun.

30 items from each of the two conditions were distributed across lists in a Latin Square design, and each list was presented in 4 different random orders for a total of 8 presentation lists. 260 additional low adjective-noun bigrams for which \( p(\text{noun} | \text{adjective}) < .50 \) were drawn from the subset of COCA bigrams and added to each set of 60 experimental items for a total of 320 items per list. Since these additional items are not relevant to the current question of interest, they are not reported here. All items were designed to be semantically congruous, and no participant saw any word more than once in an experimental session.

Participants

Participants were University of Maryland students who participated in the study for monetary compensation. Prior written consent was obtained from all participants according to the established guidelines of the Institutional Review Board of the University of Maryland. All participants were right-handed as assessed by the Edinburgh Handedness Inventory (Oldfield, 1971). In total, 38 participants took part in the study, but two datasets were excluded due to excessive artifact, two datasets were excluded due to low accuracy (less than 60%) on the concurrent behavioral measure, and six datasets were excluded because after data collection it was discovered that the participants had significant exposure to a language other than English prior to the age of 5. Of the 28 participants whose datasets were included in the study, 19 were females and 9 were males, with a mean age of 21.3 years.

Procedure

Each experimental session was divided into 4 blocks, with 15 target items and 65 other items presented in each block. Participants were asked to complete a memory recognition test administered on paper after every block. This quiz consisted of 20 bigrams, of which
10 had appeared in the preceding block and 10 were mismatched adjective-noun pairs from the stimulus set. Participants were asked to circle the phrases that they remembered seeing during the previous experimental block.

During the experiment, participants were seated in a chair in a dimly lit room. Stimuli were visually presented on a computer monitor in white 24-point case Arial font on a black background. Each trial began with a fixation cross presented at the center of the screen for 700 ms, followed by a 200 ms blank screen. An adjective was then presented for 500 ms, followed by a 100 ms blank screen, then a noun was presented for 900 ms, followed by another 100 ms blank screen. Each participant began the experiment with a short practice session and was offered the opportunity to take a break between each testing block. In total, the stimulus presentation portion of the experiment lasted 20-25 minutes.

Electrophysiological Recording

Twenty-nine tin electrodes were held in place on the scalp by an elastic cap (Electro-Cap International, Inc., Eaton, OH) in a 10-20 configuration (O1, Oz, O2, P7, P3, Pz, P4, P8, TP7, Cp3, CPz, CP4, TP8, T7, C3, Cz, C4, T8, FT7, FC3, FCz, FC4, FT8, F7, F3, Fz, F4, F8, FP1) Bipolar electrodes were placed above and below the left eye and at the outer canthus of the right and left eyes to monitor vertical and horizontal eye movements. Additional electrodes were placed over the left and right mastoids. Scalp electrodes were referenced online to the left mastoid and re-referenced offline to the average of left and right mastoids. Impedances were maintained at less than 5 kΩ for all scalp electrode sites, less than 2 kΩ for mastoid sites, and less than 10 kΩ for ocular electrodes. The EEG signal was amplified by a NeuroScan SynAmps® Model 5083 (NeuroScan, Inc., Charlotte, NC) with a bandpass of 0.05-100 Hz and was continuously sampled at 500 Hz by an analog-to-digital converter.

Analysis

Averaged ERPs time-locked to adjectives and nouns were formed off-line from trials free of ocular and muscular artifact using preprocessing routines from the EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) toolboxes. Two participants with fewer than 60% surviving trials were excluded from further analysis and are not included in the 28-participant data set presented here. Across the 28 participants included in the analysis, approximately 12% of the trials were rejected because of artifact. A 100-ms prestimulus baseline was subtracted from all waveforms before statistical analysis, and a 40-Hz low-pass filter was applied to the ERPs offline.

Analyses were conducted on mean ERP amplitudes for the critical nouns in the 300-500ms time-window in which the N400 effect is usually observed. In order to quantify the topography of the effects observed, we focused on a subset of 16 electrodes (left anterior: F7, F3, FT7, FC3; right anterior: F4, F8, FC4, FT8; left posterior: TP7, CP3, P7, P3; right posterior: CP4, TP8, P4, P8) and used R (R Development Core Team, 2010) to conduct a quadrant analysis consisting of a $2 \times 2 \times 2$ (probability $\times$ anteriority $\times$ hemisphere) Type III SS repeated-measures ANOVA. Because N400 effects often peak at midline electrodes, we also conducted a $2 \times 2$ ANOVA (probability $\times$ anteriority) on
the 6 midline electrodes (Fz, FCz, Cz, CPz, Pz, Oz). As adjectives differed across conditions, analyses were also conducted on mean ERP amplitudes for the preceding adjectives in the 500-600ms time-window (baselined to the 100ms prior to the adjective) in order to rule out baseline differences in the responses to the adjectives prior to presentation of the critical word.

Results and Discussion

Total mean accuracy on the memory tests was 71.2% (mean d’ = 1.96). ERP waveforms are presented in Figure 1, and the scalp map in Figure 3a illustrates the topographical distribution of the probability effect in the N400 time-window.

No significant effects of condition were observed in the 100ms time-window prior to noun onset (ps > .1), suggesting that responses were well-matched prior to the presentation of the critical noun. In the 300-500ms time-window the high probability items showed a reduced N400 response relative to the low probability items, as revealed by a main effect of probability ($F(1,27) = 10.3, \text{MSE} = 7.4, p < .05$; mean high probability = .77µV, mean low probability = -.39µV). This effect was largest over posterior electrodes, resulting in a significant interaction between probability and anterior-posterior distribution ($F(1,27) = 12.6, \text{MSE} = 1.2, p < .05$). We followed up with probability × hemisphere ANOVAs in anterior electrodes and posterior electrodes separately. We observed a significant main effect of probability in posterior electrodes ($F(1,27) = 22.4, \text{MSE} = 3.6, p < .05$; mean high probability = .98µV, mean low probability = -.71µV) but no significant effects involving probability in anterior electrodes (ps > .1). Similarly, in midline electrodes we observed a main effect of probability ($F(1,27) = 15.2, \text{MSE} = 6.7, p < .05$) and a significant interaction between probability and anteriority ($F(1,27) = 7.3, \text{MSE} = 1.0, p < .05$).

These results confirm that, just as in sentence paradigms in which contexts are more or less predictable of an upcoming word (as assessed by offline completions), N400 amplitude is strongly modulated by probability (as assessed by corpus counts) in a paradigm in which noun phrases are presented in isolation. In Experiment 2, we investigated whether N400 amplitude in this paradigm would similarly be modulated by semantic incongruity, when probability (as assessed by corpus counts) was held relatively constant.

Experiment 2

Materials

The materials for Experiment 2 were also drawn from the Corpus of Contemporary American English (COCA; Davies, 2008); however, all congruous nouns had a low probability (p < .005) in their adjective context, and all adjectives were relatively unconstraining. Each noun was paired with a “well-fitted” adjective to create 80 semantically/pragmatically congruent phrases and a “poorly-fitted” adjective to create 80 semantically incongruent phrases. Specifically, this was accomplished by selecting 40 animate nouns, 40 inanimate nouns, 40 adjectives that seemed to us to describe a property
that conventionally requires animacy (e.g. *innocent*), and 40 adjectives that seemed to us to most describe a property that conventionally applies to inanimate objects (e.g. *striped*), and pairing each noun with one congruous adjective and one incongruous adjective. Items were distributed across two lists so that each participant saw each word exactly once. The stimulus properties are presented in Table 2. Note that a few of our incongruous items did actually occur in the corpus (e.g. *yellow boy*), such that the mean bigram frequency for the incongruous condition was not zero as might have been otherwise expected.

In order to confirm that our items did indeed differ in semantic congruity, we conducted an offline rating study with participants recruited through Amazon Mechanical Turk. 32 participants were asked to rate noun phrases on a scale from 1 to 7 according to what degree they ‘made sense’. Each participant saw items from only one of the two lists, so that each participant saw each word exactly once. The results of the ratings showed that, as expected, our congruous items were rated much higher (mean = 6.59, s.d. = .35, range of item means = 4.3 – 7.0) than our incongruous items (mean = 1.75, s.d. = .5, range of item means = 1.1 - 3.7), such that the ratings differed significantly ($t(1,31) = 43.7$, $MSE = .2$, $p < .05$).

**Participants**

As in Experiment 1, participants were University of Maryland students who participated in the study for monetary compensation. Prior written consent was obtained from all participants. Participants adhered to the eligibility guidelines applied in Experiment 1, and had not participated in any prior studies using the same materials. In total, 38 participants took part in the study, but 10 datasets were excluded. Of the 10 datasets excluded, 4 were excluded due to excessive artifact, and 6 were excluded for accuracy below 60%. Of the 28 participants whose datasets were included in the study, 16 were females and 12 were males, with a mean age of 21.9 years.

**Procedure**

Because this experiment contained only 80 items, the materials were presented in one block. Participants were asked to complete a memory recognition test directly following the conclusion of the block. This quiz consisted of 20 bigrams, of which 10 had appeared in the preceding block and 10 were mismatched adjective-noun pairs from the stimulus set. For this experiment, we implemented the memory quiz electronically rather than on pencil and paper for easier post-processing. Participants were asked to respond with button presses on a keyboard to indicate whether or not they saw the phrases during the previous experimental block. The current experiment was presented as the second part of an EEG session in which the first part consisted of a sentence experiment described in Chow, Smith, Lau and Phillips (submitted). The sentence experiment contained congruous and incongruous sentences and required participants to make congruity judgments.

The same presentation parameters were used as in Experiment 1. In total, the stimulus presentation portion of the experiment lasted 10-15 minutes.
Electrophysiological Recording and Analysis

Experiment 2 used the same recording procedure as described in Experiment 1, with the exception that one dataset was erroneously sampled at 1000 Hz and resampled offline to 500 Hz with the EEGLAB toolbox.

Data was preprocessed following the same procedure described for Experiment 1. Four participants with fewer than 60% trials surviving the artifact rejection procedure were excluded from further analysis and are not included in the 28-participant data set presented here. Across the 28 participants included in the analysis, approximately 7.3% of the trials were rejected because of artifact. A 100-ms prestimulus baseline was subtracted from all waveforms and a 40-Hz low-pass filter was applied to the data before statistical analysis.

As in Experiment 1, quadrant and midline analyses were conducted on mean ERP amplitudes for the critical nouns in the 300-500ms time-window in which the N400 effect is usually observed. We also conducted analyses in the 600-800ms time-window in which a late positivity is often observed to incongruous sentence continuations (Van Petten & Luka, 2012).

Results and Discussion

Total mean accuracy on the memory tests was 72.6% (mean d’ = 1.22). ERP waveforms are presented in Figure 2, and the scalp map in Figure 3b illustrates the topographical distribution of the congruity effect in the N400 time-window.

In the 300-500ms time-window the incongruous items were more negative than the congruous items, resulting in a marginally significant effect of congruity ($F(1,27) = 4.0$, $MSE = 7.4$, $p = .05$; mean congruous = -.64μV, mean incongruous = -1.36μV). However, the topographical distribution of this effect was more widespread than the probability effect of Experiment 1, such that there was no interaction between congruity and anterior/posterior distribution ($p > .5$). No other contrasts including congruity yielded significant effects. For comparison with Experiment 1 we followed up with congruity × animacy × hemisphere ANOVAs in anterior electrodes and posterior electrodes separately. These showed a significant main effect of congruity in anterior electrodes ($F(1,27) = 4.6$, $MSE = 2.9$, $p < .05$; mean congruous = -.79 μV, mean incongruous = -1.47 μV) and a marginal main effect of congruity in posterior electrodes ($F(1,27) = 3.0$, $MSE = 5.5$, $p = .09$; mean congruous = -.49 μV, mean incongruous = -1.26 μV), again suggesting that the effect of congruity was broadly distributed and if anything more robust in anterior electrodes. Similarly, in midline electrodes we observed a marginal main effect of congruity ($F(1,27) = 3.4$, $MSE = 17.0$, $p = .08$) and no interaction between congruity and anteriority ($p > .1$). In the 600-800ms time-window there were no significant effects of congruity ($ps > .1$).

In summary, in Experiment 2 we examined the effect of semantic congruity on the amplitude of the N400 when probability was held relatively constant, such that the incongruous noun phrases that mostly did not occur in the corpus were compared with congruous noun phrases in which the probability of the noun given the adjective in the
corpus was extremely low (< .005). We observed a marginally significant effect of congruity in the N400 time-window. However, this effect had a different topographical distribution than what has classically been reported for the N400, as the difference between conditions was equally prominent in anterior and posterior electrodes. Furthermore, in the posterior electrodes where the N400 effect is typically strongest, the congruity effect size in Experiment 2 (~75µV) was a little less than half as large as the probability effect size in Experiment 1 (~1.7µV).

These results indicate that manipulating probability within semantically congruous items has a stronger effect on N400 amplitude than manipulating semantic congruity within items of uniform low probability, which would be unexpected if N400 amplitude primarily indexed semantic anomaly. However, these results are ambiguous as to whether semantic incongruity alone can modulate the N400 component when probability is reasonably well controlled. Assuming that the widespread increased negativity for incongruous relative to congruous items is reliable, one interpretation is that this reflects the modulation of a different component driven by different brain areas than the N400, but the other interpretation is that this reflects the simultaneous modulation of two components, the N400 in posterior electrodes and a separate anterior negativity in anterior electrodes.

A related concern about interpreting these results with respect to the results of Experiment 1 is that the two experiments used different sets of participants, who may have had inherent differences in the amplitude or topography of the N400 effect. In Experiment 3, we included both the probability manipulation and the semantic congruity manipulation in the same participants, in order to more directly compare the two effects and to attempt to replicate the marginally significant negativity of Experiment 2.

**Experiment 3**

*Materials*

All 80 item sets from Experiment 2 were included in Experiment 3. Because Experiment 1 included 120 item sets, a subset of 80 item sets was selected for Experiment 3. The stimulus properties for this subset of items is presented in Table 3. Materials from Experiment 3 are available at [http://ling.umd.edu/~ellenlau/supplementary_materials/Lau_LCN_SuppMaterials.pdf](http://ling.umd.edu/~ellenlau/supplementary_materials/Lau_LCN_SuppMaterials.pdf). Items were randomized (such that Experiment 1 and Experiment 2 items were completely mixed) and distributed across four lists in a Latin Square design so that each participant saw each word exactly once.

*Participants*

As in Experiments 1 and 2, participants were University of Maryland students who participated in the study for monetary compensation. Prior written consent was obtained from all participants. Participants adhered to the eligibility guidelines applied in Experiments 1 and 2, and had not participated in any prior studies using the same
materials. In total, 38 participants took part in the study, but 10 datasets were excluded. Of the 10 datasets excluded, 7 were excluded due to excessive artifact, and 3 were excluded for accuracy below 60%. Of the 28 participants whose datasets were included in the study, 16 were females and 12 were males, with a mean age of 22.0 years.

Procedure

The materials were presented in two blocks of 80 items each. Participants were asked to complete a memory recognition test directly following the conclusion of each block. This quiz consisted of 20 bigrams, of which 10 had appeared in the preceding block and 10 were mismatched adjective-noun pairs from the stimulus set. Participants were asked to respond with button presses on a keyboard to indicate whether or not they saw the phrases during the previous experimental block. The current experiment was presented as the second part of an EEG session in which the first part consisted of a sentence experiment. The sentence experiment contained congruous and incongruous sentences and required participants to answer comprehension questions.

The same presentation parameters were used as in Experiments 1 and 2. In total, the stimulus presentation portion of the experiment lasted 10-15 minutes.

Electrophysiological Recording and Analysis

Experiment 3 used the same recording procedure as described in Experiments 1 and 2, except that data was recorded using a slightly different electrode configuration in which electrode Oz was replaced by electrode FP2. Data was preprocessed following the same procedure described for Experiments 1 and 2. Seven participants with fewer than 60% trials surviving the artifact rejection procedure were excluded from further analysis and are not included in the 28-participant data set presented here. Across the 28 participants included in the analysis, approximately 20% of the trials were rejected because of artifact. A 100-ms prestimulus baseline was subtracted from all waveforms and a 40-Hz low-pass filter was applied to the data before statistical analysis.

As in Experiment 2, quadrant and midline analyses were conducted on mean ERP amplitudes for the critical nouns in the 300-500ms time-window in which the N400 effect is usually observed and in the 600-800ms time-window. Since the probability comparison and the congruity comparison consisted of different factors and different materials, we conducted separate analyses for each, which were identical to the analyses conducted in Experiments 1 and 2, respectively. Since this electrode configuration had only 5 midline electrodes, we conducted the midline analysis on 4 of the 5 in order to maintain equal numbers of anterior/posterior observations (Fz, FCz, CPz, Pz).

Results and Discussion

Total mean accuracy on the memory tests was 74.8% (mean d’ = 1.37). ERP waveform plots depicting the two pairwise-comparisons are presented in Figures 4 and 5; Figure 6 illustrates the topographical distribution of the probability effect and the congruity effect in the N400 time-window. Planned quadrant ANOVAs in the time-window that served as
the baseline for the critical noun ERP, 500-600ms following the presentation of the adjective, showed no significant effects involving congruity or probability (ps > .1).

No significant effects of probability or congruity were observed in the 100ms time-window prior to noun onset, suggesting that responses were well-matched prior to the presentation of the critical noun. In the probability comparison, we again observed that N400 amplitude was reduced for highly probable items compared to less probable items, and that this effect had the central-posterior distribution characteristic of the N400, resulting in a significant main effect of probability in the 300-500ms time-window ($F(1,27) = 19.4, MSE = 8.4, p < .05$) and a significant interaction between probability and anteriority ($F(1,27) = 6.0, MSE = 1.4, p < .05$). In the congruity comparison, the incongruous items were slightly more negative than congruous items in right hemisphere electrodes and slightly more positive than congruous items in left hemisphere electrodes, resulting in a significant interaction between congruity and hemisphere ($F(1,27) = 5.60, MSE = .7, p < .05$). Although ERPs to incongruous items were numerically larger than congruous items in a number of posterior electrodes (e.g. O1, O2, Pz), no other contrasts including congruity yielded significant effects. Similarly, in midline electrodes we observed a significant effect of probability ($F(1,27) = 29.8, MSE = 6.7, p < .05$) and a marginally significant interaction between probability and anteriority ($F(1,27) = 4.0, MSE = 1.1, p < .05$), but no significant effects involving congruity (ps > .1).

For comparison with Experiments 1 and 2 we followed up both comparisons with ANOVAs in anterior electrodes and posterior electrodes separately. In the probability manipulation there was a significant main effect of probability in both anterior ($F(1,27) = 8.70, MSE = 5.6, p < .05$) and posterior ($F(1,27) = 28.9, MSE = 4.25, p < .05$) electrodes. In the congruity manipulation there was a significant interaction between congruity and hemisphere in anterior electrodes ($F(1,27) = 6.59, MSE = .5, p < .05$), but no significant effects involving congruity in posterior electrodes ($p > .1$). In the 600-800ms time-window, there were no significant effects involving probability, and the only effect of congruity was a continuation of the interaction between congruity and hemisphere ($F(1,27) = 6.08, MSE = 1.5, p < .05$).

Although we observed no significant effects of congruity on the amplitude of the N400 or the late posterior positivity, we note that visual inspection indicates that ERPs were numerically more positive for incongruous nouns than congruous nouns in the 150-300ms time-window, particularly visible in frontal electrodes (Figure 5). This difference is most likely due to random noise, because it has not been reported in prior literature and

---

1 An alternative perspective is that Experiment 3 constituted a 2 × 2 design (something akin to manipulation: predictability/congruity × context: good/bad) such that if the N400 is more sensitive to predictability than to congruity, an interaction is predicted. Although we were concerned about whether grouping the conditions in this way was fully justified, we conducted a supplementary 2 × 2 × 2 × 2 (manipulation × context × hemisphere × anteriority) analysis of variance on all four conditions in the 300-500ms time-window to confirm that the N400 effect of predictability was larger than the N400 effect of congruity. Indeed, we observed a significant interaction between manipulation and context ($F(1,27) = 12.2, MSE = 7.6, p < .01$).
also was not observed in Experiment 2, which used the same stimuli (in that case, the reverse pattern was observed, such that congruous nouns were numerically more positive than incongruous nouns). A reviewer suggests that it is possible that whatever random noise caused this early deviation of the incongruous nouns in the positive direction in Experiment 3 also acted to diminish a true increase in the N400 to the incongruous condition. However, as the positive shift observed is small, at most this could only have masked a small N400 effect of incongruity. Therefore, this scenario remains consistent with the qualitative pattern of results reported in Experiment 2: a large N400 effect of probability and at most a very small N400 effect of incongruity.

To summarize, in Experiment 3 we again observed a robust N400 effect of probability with a classic central posterior focus, but did not observe the same for the semantic congruity manipulation. Although numerically there was a slightly larger N400 peak amplitude for incongruous relative to congruous items across posterior electrodes, this effect was not reliable. We did not replicate the marginal, broadly distributed congruity effect observed in Experiment 2, and instead we observed a significant interaction between congruity and hemisphere, with a larger negativity for incongruous items in right hemisphere electrodes and a larger negativity for congruous items in left hemisphere electrodes. We speculate that the differences in the congruity effect from Experiment 2 to Experiment 3 may be due to the fact that Experiment 2 was presented in the same session as a sentence paradigm with a plausibility judgment task, while Experiment 3 was presented in the same session as a sentence paradigm with a comprehension question task. Importantly, when both the probability manipulation and the congruity manipulation were presented in the same participants as part of the same session, a large and robust N400 probability effect was observed while no significant N400 congruity effect was observed.

**General Discussion**

Despite much prior evidence to the contrary, the idea that the N400 is primarily a semantic anomaly response persists in much of cognitive psychology and cognitive neuroscience. In the current study, we used an adjective-noun paradigm that allowed us to contrast the effects of contextual predictability and semantic congruity on the N400 while precisely controlling predictability through the use of corpus counts. Across three ERP experiments, we found that while predictability had a massive and reliable effect on the amplitude of the N400 component, semantic congruity had at best minor and inconsistent effects. While these results thus do not indicate that semantic congruity does not or cannot impact N400 amplitude, they are clearly inconsistent with the hypothesis that semantic congruity is the primary contributor to N400 amplitude. In contrast, these results support the hypothesis that the N400 effect primarily reflects lexical or conceptual predictability (Kutas & Hillyard, 1984; Federmeier, 2007).

Although many studies have observed larger N400 responses for semantically incongruous words relative to semantically congruous words (e.g. Kutas & Hillyard, 1980; Holcomb & Neville, 1991; van Berkum, Hagoort, & Brown, 1999), these studies have tended to systematically confound congruity and predictability, such that the
congruous endings were also highly predicted by the context (see Kutas & Hillyard, 1984; Kutas & Federmeier, 2000; van Berkum, 2009, for relevant discussion). Avoiding this confound is challenging in sentence contexts on the assumption that predictions can be distributed across several options, because even in large corpora, tokens of full sentence contexts are too sparse to yield robust corpus estimates of predictability unless numerous assumptions are made about the language model. At the same time, offline completion norming tasks are likely to return the most predictable word for a given context but may not provide accurate information about somewhat lower probability words that are still predictable enough to impact N400 amplitudes. Using a single adjective context for the critical noun in the congruity comparison here allowed us to ensure that the nouns in the semantically congruous items were in fact quite unpredictable, although they were necessarily slightly more predictable than the nouns in the semantically incongruous items. When predictability was controlled in this way, we observed weak to non-existent N400 effects of congruity. Therefore, we argue that most apparent N400 responses to semantic incongruity in previous work are likely due instead to predictive facilitation in the semantically congruous conditions. These results are similarly consistent with recent work by DeLong et al. (2014) demonstrating only a small N400 effect of semantic congruity in sentence materials for which the incongruous continuations were slightly but significantly lower in cloze probability. We suggest that the N400 effect of semantic congruity observed by Molinaro et al. (2012) may also have been due to small differences in predictability, which were perhaps magnified in the experimental context by the relatively frequent occurrence of nouns followed by an adjective modifier (which is otherwise optional in Spanish). Future work could evaluate this prediction of our account by examining the response to similar Spanish noun-adjective sequences outside of sentence contexts and assessing predictability through corpus counts of these isolated sequences.

The conclusion that semantic anomaly does not elicit robust N400 effects has important consequences, because N400 amplitude has been used by many investigators exactly in order to assess whether comprehenders have recognized semantic anomalies in particular contexts. One important example is work arguing that comprehenders are ‘attracted’ to a sentence-level meaning inconsistent with the syntax in ‘role-reversal’ sentences such as The hearty meal was devouring or The cop that the thief arrested... because no N400 effect is observed when comparing these sentences with their congruous controls The hearty meal was devoured or The thief that the cop arrested... (Kolk, Chwilla, van Herten & Oor, 2003; Kim & Osterhout, 2005). The current work suggests that N400 amplitude is determined primarily by predictability rather than whether a semantically incongruous meaning has been computed or not. Consistent with this, other investigators have noted that (a) in many of these paradigms the critical word is likely to be primed by lexical items or event schemas, creating the potential for a floor effect on N400 amplitude (Kuperberg, 2007; Stroud, 2008; Brouwer et al., 2012; Chow & Phillips, 2013) and (b) N400 effects re-emerge when more time is available for predictions to be computed (Chow, Lau, Wang, & Phillips, submitted). Although these results do not indicate that syntax-independent interpretations are not constructed, it suggests that the presence or absence of N400 effects cannot be taken to support this hypothesis.
A related debate over the interpretation of the N400 effect has concerned whether the effect reflects the impact of context on ‘activation’ or ‘integration’ processes. According to the ‘activation’ view, the N400 component reflects neural activity associated with activating long-term memory representations of lexical or conceptual features, and its amplitude is reduced in predictive contexts because accessing or selecting these stored representations is made easier (e.g. Kutas & Federmeier, 2000; Lau, Phillips, & Poeppel, 2008). For example, when a word is encountered in isolation, many orthographic or phonological neighbors may also initially be activated as competitors before the correct word is selected, and multiple senses of the word and associated conceptual features may also be activated, such that a relatively broad portion of the lexical/conceptual memory network is initially active. However, when a word is encountered in a predictive context, pre-activation may result in more selective activation of the correct lexical representation and the corresponding conceptual features that are appropriate to that context, such that a relatively small portion of the network is activated. In contrast, according to the ‘integration’ view the N400 component reflects neural activity associated with building an integrated, ‘working memory’ representation of the intended situation or message on-the-fly from the individual words and associated concepts (Osterhout & Holcomb, 1992; Brown & Hagoort, 1993; Hagoort, 2008). On this view, N400 amplitude is affected by predictability because it is easier to build an integrated representation for messages that are predicted by background knowledge than messages that are not. One kind of argument in favor of the integration view is the fact that sentences that require semantic coercion of a noun entity into an event have been shown to elicit larger N400 responses (Baggio, Choma, van Lambalgen, & Hagoort, 2010; Kuperberg, Choi, Cohn, Paczynski, & Jackendoff, 2010), which might be interpreted as integration cost (although see Delogu & Crocker, 2012 for arguments that these effects might be attributable to differential predictability). Another recent finding in favor of the integration view is that infelicitous repeated names elicit larger N400 responses than felicitous ones (John went to the store so that John… vs. John and Mary went to the store so that John…) even though the offline cloze probability of the repeated name is quite low in both cases (Swaab, Camblin, & Gordon, 2004; Ledoux, Gordon, Camblin & Swaab, 2007, although see Hoeks & Brouwer, 2014). Finally, Hagoort and colleagues have suggested a unified account in which N400 effects reflect both activation and integration (Baggio & Hagoort, 2011; Hagoort, 2013).

The N400 predictability effect observed here is consistent with either the activation or the integration view, as predictable phrases that drove the N400 predictability reduction were more frequent in the corpus than the unpredictable phrases, and therefore may describe objects that are better predicted by background knowledge and are thus easier to integrate. However, the absence of a reliable N400 effect of congruity is more difficult to explain on the integration view, as it would seem that it should be more difficult to form a message-level interpretation for the incongruous phrases. One possible explanation for the current results consistent with the integration account is that the semantically incongruous phrases in the current study may not have allowed higher-level integration at all, if a mismatch in core semantic features such as animacy precludes further semantic processing (Kuperberg, 2007; Paczynski & Kuperberg, 2011). At the same time, if the amplitude of the N400 component indicates integration effort, it would then seem
surprising that a congruous item that allows phrase-level integration would not elicit a larger N400 than an incongruous item that does not.

A second possible explanation consistent with the integration account is that when any noun phrase is presented in isolation, the meaning of the adjective and noun are not integrated as they are in connected text or speech. On such an account, the N400 effects observed by Molinaro et al. (2012) for Spanish noun-adjective sequences do reflect integration difficulty due to implausibility rather than small differences in predictability. This hypothesis would predict that congruity effects on the N400 would also emerge in the current materials if these phrases were embedded in sentence contexts. This explanation would seem unable to explain our full pattern of results: if N400 effects reflect integration processes and participants are essentially not integrating any of the noun phrases in this paradigm, we should not have seen huge N400 reductions in the predictable (but on this view, un-integrated) noun phrases. However, one possible move would be to adopt a hybrid activation-and-integration account of the N400 (Baggio & Hagoort, 2011), which could posit that the N400 effects in the current paradigm are due to activation processes and that N400 effects in sentence contexts are due to integration processes, and/or that the earlier portion of the N400 effect reflects activation processes and the later portion integration processes that were minimized in the current phrasal paradigm (Molinaro et al., 2012; Brothers et al., 2015).

A related concern is that the incongruous adjective-noun sequences used here are not actually semantically incongruous, because a congruous meaning could be coerced or accommodated, e.g. by imagining them in the context of a fairy tale or a metaphor. Although, as noted in the introduction, we believe that an interpretation can be coerced or accommodated for almost any string, we think this also holds for the kinds of incongruous sentences classically used in N400 paradigms (She spread the warm bread with socks) as well as the adjective-noun sequences used here. In an offline task, participants reliably judged the incongruous adjective-noun sequences used here as ‘not making sense’, suggesting that these items were not literally congruous in the absence of a broader context and would require such an accommodation process to be viewed as congruous. While it could be the case that participants automatically and rapidly computed this kind of accommodation during the ERP experiment, this would be the kind of ‘beyond pure predictability’ response to semantic anomaly that the current experiment was designed to investigate. Our results show that whatever extra processes are elicited by semantic anomaly, they do not impact N400 amplitude.

While we take these results to indicate that semantic congruity is not the primary factor contributing to N400 amplitude, we definitely do not think these results rule out the possibility that congruity has some impact. In the current study, we observed a widespread negativity in Experiment 2 that could have been partially driven by an increased N400 to incongruous items, and in Experiment 3 we observed a small, nonsignificant increased negativity to incongruous items with a posterior N400-like distribution (it is also possible these small differences were driven by predictability; see below). However, even if these are interpreted as N400 congruity effects, the fact that
they are small and unreliable relative to predictability effects raises interesting questions about the mechanisms by which congruity impacts N400 amplitude.

One promising explanation is the idea that N400 amplitude indexes only the activation of stored lexical and conceptual representations, but the separate process of semantic integration sometimes modulates the activation of these stored representations, for example when activating the rarely used ‘floatable’ feature of basketball is necessary to verify the sentence *The child survived in the water by clinging to a basketball* (Barsalou, 1982; Van Berkum, 2011; Wilson et al., 2014). If integration processes only impact N400 amplitude through this indirect route, one might expect the impact of congruity manipulations on the N400 to be quite subtle and dependent on the extent to which the incongruous materials encourage attempts to activate less commonly used conceptual features of the word. Thus, further exploration of the effects of congruity on the N400 is an important avenue for future work that may shed light on how long-term memory representations are manipulated by combinatorial processes. However, the current results reiterate the point that all investigations hoping to target N400 effects of congruity must strive more than ever to control predictability and semantic association, and highlight the significance of this challenge. Even in creating pared-down single-word context items, we found it impossible to ensure that our congruous items were uniformly congruous without making them very slightly predictable (p < .005), and thus the possibility remains that the small and unreliable ‘congruity’ effects we observed were actually predictability effects given recent arguments that the relationship between predictability and processing difficulty is logarithmic (Smith & Levy, 2013). Future work may begin to address this challenge both by systematically examining the shape of the N400-predictability function at the lower end of the predictability scale and by refining computational models of predictability in larger contexts (e.g., Frank et al., 2013).

Finally, we note that some authors have argued that the relatively long-lasting difference in the ERP response to different conditions that is traditionally referred to as the N400 effect (often ranging from ~200-550ms) actually reflects the composite effect of multiple computations pushing the ERPs in different directions. For example, it has been observed that the early part of the N400 amplitude difference appears to show an increased positive deflection (relative to the pre-stimulus baseline) in the predictable condition, rather than just less negativity than the unpredictable condition (Roehm, Bornkessel-Schlesewsky, Roseler, & Schlesewsky, 2007; Molinaro & Carreiras, 2010; Vespignani, Canal, Molinaro, Fonda, & Cacciari, 2010). However, because it is not possible to draw strong conclusions from the absolute position of peaks in the ERP response (Luck, 2014), the contribution of multiple components to differences between two ERP responses can only be assessed by methods such as principal components analysis or source localization. We believe that this is an important question for future work on the underlying generators of the N400 effect, but as the primary goal of the current work is simply to assess the impact of semantic anomaly on the N400 effect as traditionally described, we do not attempt such analyses here.

In conclusion, the current study failed to show reliable effects of semantic incongruity on the N400 when predictability was held relatively constant, but showed very large N400
effects in the same paradigm when predictability was varied and congruity was held constant. These results are consistent with much prior work in suggesting that references to the N400 as a ‘semantic anomaly’ response are bad practice, as contextual manipulation of non-anomalous phrases can have a much larger impact on the N400 than semantic incongruity itself (when measured in the same participants under the same conditions) and the evidence that semantic anomaly modulates the response when confounding factors are controlled is so far minimal. We hope that this study will help to discourage this common misconception, and that general reviews in the future will instead emphasize the key empirical generalizations that the N400 reflects normal semantic processing, that N400 effects typically arise through a reduction in amplitude for facilitated items, and that contextual predictability is the most well-established factor governing N400 effects.

Acknowledgments

We thank Josh Falk, Wing Yee Chow, Glynis MacMillan, Zoe Schlueter, Julia Buffinton, and Stephen Devilbiss for assistance with materials, data collection, and analysis. We thank Bill Idsardi, Ming Xiang, and Colin Phillips for useful discussion.

References


<table>
<thead>
<tr>
<th></th>
<th>High Probability</th>
<th>Low Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(noun</td>
<td>adjective)</td>
<td>0.650 (.14)</td>
</tr>
<tr>
<td>p(adjective</td>
<td>noun)</td>
<td>0.073 (.09)</td>
</tr>
<tr>
<td>bigram frequency</td>
<td>649 (1107)</td>
<td>3 (5)</td>
</tr>
<tr>
<td>adjective constraint (max p(noun</td>
<td>adjective) across nouns)</td>
<td>0.650 (.12)</td>
</tr>
<tr>
<td>adjective frequency</td>
<td>961 (1534)</td>
<td>618 (983)</td>
</tr>
<tr>
<td>noun frequency</td>
<td>15013 (18329)</td>
<td></td>
</tr>
<tr>
<td>adjective length</td>
<td>8.3 (2.1)</td>
<td>8.2 (1.8)</td>
</tr>
<tr>
<td>noun length</td>
<td>5.5 (1.9)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Stimulus properties for the 120 pairs of items used to instantiate the probability comparison in Experiment 1, all derived from the Corpus of Contemporary American English. Standard deviations are presented in parentheses. Simple frequencies represent the total number of occurrences of the lemma across all adjective-noun bigrams that occurred in COCA.
<table>
<thead>
<tr>
<th></th>
<th>Congruous</th>
<th>Incongruous</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(noun</td>
<td>adjective)</td>
<td>.002 (.001)</td>
</tr>
<tr>
<td>p(adjective</td>
<td>noun)</td>
<td>.007 (.023)</td>
</tr>
<tr>
<td>bigram frequency</td>
<td>9 (6)</td>
<td>.3 (1.3)</td>
</tr>
<tr>
<td>adjective constraint (max p(noun</td>
<td>adjective) across all nouns)</td>
<td>.08 (.08)</td>
</tr>
<tr>
<td>adjective frequency</td>
<td>6475 (7069)</td>
<td></td>
</tr>
<tr>
<td>noun frequency</td>
<td>9458 (17824)</td>
<td></td>
</tr>
<tr>
<td>adjective length</td>
<td>6.9 (1.8)</td>
<td></td>
</tr>
<tr>
<td>noun length</td>
<td>5.8 (1.9)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Mean stimulus properties for the 80 pairs of items used to instantiate the congruity comparison in Experiment 2, all derived from the Corpus of Contemporary American English. Standard deviations are presented in parentheses. Simple frequencies represent the total number of occurrences of the lemma across all adjective-noun bigrams that occurred in COCA.
<table>
<thead>
<tr>
<th></th>
<th>High Probability</th>
<th>Low Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(noun</td>
<td>adjective)</td>
<td>.664 (.12)</td>
</tr>
<tr>
<td>p(adjective</td>
<td>noun)</td>
<td>.091 (.10)</td>
</tr>
<tr>
<td>bigram frequency</td>
<td>812 (1299)</td>
<td>4 (6)</td>
</tr>
<tr>
<td>adjective constraint (max p(noun</td>
<td>adjective) across nouns)</td>
<td>.664 (.12)</td>
</tr>
<tr>
<td>adjective frequency</td>
<td>1188 (1786)</td>
<td>734 (1152)</td>
</tr>
<tr>
<td>noun frequency</td>
<td>14253 (19828)</td>
<td></td>
</tr>
<tr>
<td>adjective length</td>
<td>8.1 (1.9)</td>
<td>8.1 (1.8)</td>
</tr>
<tr>
<td>noun length</td>
<td>5.6 (1.8)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Mean stimulus properties for the 80 pairs of items used to instantiate the probability comparison in Experiment 3, all derived from the Corpus of Contemporary American English. Standard deviations are presented in parentheses. Simple frequencies represent the total number of occurrences of the lemma across all adjective-noun bigrams that occurred in COCA. Note that the stimulus properties for the congruity comparison in Experiment 3 were identical to those in Experiment 2.
Figure 1. ERP waveforms for 28 scalp electrodes for the predictability contrast in Experiment 1. ERPs are time-locked to the onset of the critical noun.

Figure 2. ERP waveforms for 28 scalp electrodes for the semantic congruity contrast in Experiment 2. ERPs are time-locked to the onset of the critical noun.

Figure 3. Scalp maps and selected electrode waveforms demonstrating measured N400 effects of predictability and congruity during the 300-500ms time-window in Experiments 1 and 2. Scalp maps illustrate the mean difference between low probability and high probability conditions in Experiment 1 (A) and between semantically incongruous and congruous conditions in Experiment 2 (B).

Figure 4. ERP waveforms for 29 scalp electrodes for the predictability contrast in Experiment 3. ERPs are time-locked to the onset of the critical noun.

Figure 5. ERP waveforms for 29 scalp electrodes for the semantic congruity contrast in Experiment 3. ERPs are time-locked to the onset of the critical noun.

Figure 6. Scalp maps and selected electrode waveforms demonstrating measured N400 effects of predictability and congruity during the 300-500ms time-window in Experiment 3. Scalp maps illustrate the mean difference between low probability and high probability conditions (A) and between semantically incongruous and congruous conditions in (B).
Figure 1. ERP waveforms for 28 scalp electrodes for the predictability contrast in Experiment 1. ERPs are time-locked to the onset of the critical noun.
Figure 2. ERP waveforms for 28 scalp electrodes for the semantic congruity contrast in Experiment 2. ERPs are time-locked to the onset of the critical noun.
Figure 3. Scalp maps and selected electrode waveforms demonstrating measured N400 effects of predictability and congruity during the 300-500ms time-window in Experiments 1 and 2. Scalp maps illustrate the mean difference between low probability and high probability conditions in Experiment 1 (A) and between semantically incongruous and congruous conditions in Experiment 2 (B).
Figure 4. ERP waveforms for 28 scalp electrodes for the predictability contrast in Experiment 3. ERPs are time-locked to the onset of the critical noun.
Figure 5. ERP waveforms for 28 scalp electrodes for the semantic congruity contrast in Experiment 3. ERPs are time-locked to the onset of the critical noun.
Figure 6. Scalp maps and selected electrode waveforms demonstrating measured N400 effects of predictability and congruity during the 300-500ms time-window in Experiment 3. Scalp maps illustrate the mean difference between low probability and high probability conditions (A) and between semantically incongruous and congruous conditions in (B).