Published in final edited form as:
Lang Cogn Neurosci. 2015 July 1; 30(6): 648–672. doi:10.1080/23273798.2014.995679.

Reversing expectations during discourse comprehension

Ming Xiang[a] and **Gina Kuperberg**[b,c]

[a]Language and Processing Lab, Linguistics Department, University of Chicago, IL, 60615

[b]NeuroCognition Laboratory, Department of Psychology, Tufts University, Medford, MA, 02155

[c]Department of Psychiatry, Massachusetts General Hospital, Bldg 149, 13th Street, Charlestown, MA, 02129

Abstract

In two ERP experiments, we asked whether comprehenders used the concessive connective, *even so*, to predict upcoming events. Participants read coherent and incoherent scenarios, with and without *even so*, e.g. “Elizabeth had a history exam on Monday. She took the test and aced/failed it. (Even so), she went home and celebrated wildly.”, as they rated coherence (Experiment 1) or simply answered intermittent comprehension questions (Experiment 2). The semantic function of *even so* was used to reverse real-world knowledge predictions, leading to an attenuated N400 to coherent versus incoherent target words (“celebrated”). Moreover, its pragmatic communicative function enhanced predictive processing, leading to more N400 attenuation to coherent targets in scenarios with than without *even so*. This benefit however, did not come for free: the detection of failed event predictions triggered a later posterior positivity and/or an anterior negativity effect, and costs of maintaining alternative likelihood relations manifest as a sustained negativity effect on sentence-final words.

Keywords

Discourse processing; concessive connectives; prediction; event structures; N400; P600; late negativity; ERP

General Introduction

Successful language comprehension draws heavily upon our experience in the real world. This real-world knowledge, stored within long-term memory, is recruited by the comprehension system to aid the construction of a discourse model. It tells us whether what we hear or read is plausible, implausible, true or false. Moreover, as language unfolds online, we continually draw upon this stored knowledge to facilitate our comprehension of sentences describing familiar events or states (Marslen-Wilson, Brown and Tyler, 1988; Warren and McConnell 2007; McRae et al. 1997), as well as discourse describing familiar relationships between events and states (Singer et al. 1996; Keenan et al., 1984; van Dijk and Kintsch, 1983).

Corresponding author: Ming Xiang, Language and Processing Lab, Linguistics Department, University of Chicago, IL, Chicago, 60615, mxiang@uchicago.edu, phone: 773-702-8023, fax: 773-834-0924.

At the same time, however, language offers us the remarkable ability to construct discourse models that do not necessarily conform to our real-world knowledge. Moreover, there is emerging evidence that such models can sometimes (although not always) facilitate the processing of incoming words during comprehension (Nieuwland and Van Berkum, 2006; Nieuwland and Martin, 2012; Nieuwland, 2013). For simplicity, we will refer to any such discourse model as an 'alternative world model'— a separate set of events and relations that are established in some mental space that is different from the default real-world knowledge, stored within long-term memory.[1] For example, when reading Harry Potter, we may well expect to encounter events of magic and wizardry that are quite different from our everyday reality. And, if asked to entertain the possibility, "if humans were living on the moon…", we would not be surprised to hear about facts and events that are quite different from what happens on planet Earth.

Concessive Connectives and "Even so"

Importantly, we do not only construct alternative world models when reading fiction or carrying out counterfactual reasoning. We use such models all the time during everyday communication through our use of small words or phrases, like *but, however, even so*, and *although.* These so-called *concessive connectives* set up an alternative world model by introducing a presupposition (Lakoff, 1971; Lagerwerf, 1998) or conventional implicature (Grice, 1975) that an upcoming proposition will contrast with, or contradict, a previously held assumption or expectation based on world knowledge (Blakemore, 2002; Lakoff, 1971). The rich inherent meaning of these lexical items also provides the comprehender with explicit information about *how* the upcoming proposition should be linked to its preceding discourse context. In this sense, they function to *pragmatically constrain* the incremental process of discourse comprehension, helping us to infer the relevance of any upcoming information (Blakemore, 2002; Wilson and Sperber,1993). This pragmatic constraining function sets concessive connectives apart from the fictional scenarios or counterfactuals mentioned above. These also set up alternative world models. However, they do not necessarily provide sufficient linguistic information about exactly how upcoming propositions will link to the discourse context.

Although concessive connectives are often used in everyday communication, there have been very few psycholinguistic studies examining their influences on online comprehension (but see Murray, 1994 and 1997). In the present study, we examine the effects of one particular concessive connective on word-by-word discourse comprehension—*even so. Even so* is a concessive connective that is commonly used to link propositions across sentence boundaries. It is derived from the scalar term, *even*, which introduces a presupposition that the event under discussion is very low in its likelihood, but asserts that the utterance is nevertheless true (Karttunen and Peters 1979). At the discourse level, *even so* inherits the scale-reversing property of *even*: it functions to establish an alternative world model in which the possible causal relationships between events are reversed from what would follow from our real-world knowledge. In addition, as discussed above, *even so*, like other concessive connectives, acts to pragmatically constrain the discourse context by 'narrowing

[1] Our use of 'alternative world model' in this sense should not be confused with the concept of "possible worlds" in formal semantics.

down' the number of potential relationships to those that causally opposite-to-expected (see Noveck and Spotorno, 2013, for a more general discussion of this type of 'narrowing down' effect in communication, and discussed further below).

To illustrate the scale-reversing function of *even so*, consider the four types of three-sentence scenarios shown in Table 1. In the scenarios where the final sentence does not begin with "Even so" (the 'plain scenarios'), coherence arises purely from the real-world relationship between the particular event described in the final sentence and the events and states described in the preceding context. For example, in the plain coherent scenario, "… Elizabeth took the test and aced it. She went home and celebrated…", our real-world knowledge tells us that these events are causally related, whereas in the plain incoherent scenario, "…Elizabeth took the test and failed it. She went home and celebrated …", it tells us that these events are unlikely to follow on from one another.

In the scenarios where the final sentence begins with "Even so" (the 'even-so scenarios'), these relationships are reversed: coherence is evaluated in relation to the alternative world model set up under *even so*, rather than real-world knowledge. For example, the even-so coherent scenario, "…Elizabeth took the test and failed it. Even so, she went home and celebrated…", is coherent, despite the fact that the relationships between the events/states described do not match our long-term real-world knowledge. And the even-so incoherent scenario, "…Elizabeth took the test and aced it. Even so, she went home and celebrated…", is incoherent, despite the fact that the events described are consistent with our real-world knowledge.

In the present study, we asked participants to read discourse scenarios similar to those described above. Based on a large psycholinguistic literature showing that we are able to integrate multiple linguistic cues quickly and incrementally during online word-by-word language comprehension (e.g. Altmann and Steedman, 1988; MacDonald, Pearlmutter, and Seidenberg, 1994; Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy, 1995; van Berkum, 2009, Traxler, Bybee and Pickering, 1997), we expected that comprehenders would integrate *even so* relatively quickly to establish an alternative world model. Our main questions concerned whether, when and how comprehenders would use this alternative world model to predict and process incoming information as it unfolded, word by word.

Prediction, generative models and event-related potentials

The term *prediction* has been used by different researchers in different ways. While early models assumed that it necessarily entailed committing to specific lexical items (Forster, 1981) in a strategic, all-or-nothing fashion (we either predict or we don't) (Becker, 1980 Becker, 1985), here we make no such assumptions. We use the term *prediction* and *expectation* interchangeably and conceive of prediction as influencing a Bayesian prior — an assessment of the probability of accessing information at a particular representational level ahead of encountering all the linguistic information required to activate, retrieve or compute this representation. We assume that predictions are generated at multiple levels of representation and that rather than being deterministic, they are probabilistic in nature: that is, multiple predictions at a particular representational level are held with differing probabilities that add up to 100% in total. Thus, a very strong prediction corresponds to a near-certain (e.g. 99 %) probability of belief in a particular upcoming representation, and a weak prediction corresponds to many low-probability beliefs in multiple possible upcoming representations.

We view such probabilistic predictions as a consequence of a *dynamic hierarchical generative* process by which our brains draw upon high-level stored representations and contextual information to construct a generative model that best explains the sensory input we encounter. This type of framework is proving powerful not only for understanding language processing (e.g. Farmer, Brown, & Tanenhaus, 2013; Feldman, Griffiths, & Morgan, 2009; Fine, Jaeger, Farmer, & Qian, 2013; Hale, 2001; Kleinschmidt & Jaeger, In press; Levy, 2008; Norris, 2006; Norris & McQueen, 2008, Kuperberg, 2014), but also many other aspects of perception and cognition (Clark, 2013; Jacobs and Kruschke, 2011; Griffiths et al., 2008; Rao & Ballard, 1999; Friston, 2005). According to this framework, probabilistic predictions are propagated from higher to lower level representational layers, and any residual error between these predictions and the input to each layer (implicit prediction error) serves as the feed-forward signal from lower to higher-level representational layers. This implicit prediction error (or Bayesian surprise) is, in turn, used to update our predictions in an ongoing attempt to refine the generative model and ‘explain away’ the bottom-up input (see Kuperberg, 2014).

In this study, we are primarily concerned with activity at three layers of representation (see Kuperberg, 2013 for discussion): (1) *event sequences*, which describe our knowledge about the necessary and likely temporal, spatial, causal and other relationships that link multiple events and states together to form sequences of events, sometimes known as scripts, frames or narrative schemas (Fillmore, 2006; Schank & Abelson, 1977; Sitnikova, Holcomb, & Kuperberg, 2008; Zwaan & Radvansky, 1998); (2) *event structures*, which describe our fine-grained knowledge about specific events (e.g. in an “arresting” event, it is more likely that a policeman arrests a burglar than the other way around, McRae, Ferretti, and Amyote, 1997), our knowledge about similar events (e.g. the similarities between a ‘teaching event’ and an “instructing” or ‘mentoring’ event), as well as our coarser-grained knowledge about the prototypical semantic-thematic roles (Agent, Patient, Experiencer, Stimulus etc) played by participants in actions and states (Jackendoff, 2002); and (3) *semantic (or conceptual) features*, which describes our knowledge of the perceptual features and functional properties of conceptual entities and categories, e.g. our knowledge that a “boy” has the properties of being <human>, <male>, <young> etc.

These representational layers interface with one another through statistical dependencies that describe the regularities between them, and, during language processing, predictions generated at higher layers influence the priors at lower layers through these dependencies. Thus, at any given time, the situation-level representation of context will interact with our stored knowledge of event sequences, influencing probabilistic predictions about upcoming event structures, which will, in turn influence probabilistic predictions about upcoming semantic features (indeed, this propagation of implicit predictions may sometimes continue down to lower representational layers, including word-form, pre-lexical and perceptual representations).

We conceive of *even so* as exerting its influence at the *event sequence* layer of representation by explicitly signaling to the comprehender to expect an opposite-to-expected causal relationship. Under the framework we just described, this prediction will propagate down to constrain the prior probability distribution at the event structure layer, narrowing it down from including many different kinds of event structures (with causal, spatial, temporal and other relationships with the context) that are each held with relatively low probabilities, to a specific type of event structure that is held with higher probability. This, in turn, will lead to strong (high probability) predictions about upcoming semantic features at the representation layer below.

One way of examining how these types of probabilistic predictions interact with incoming information as it unfolds in real time is through event-related potentials (ERPs) — an online neural measure of cognitive processing. It has been proposed that ERPs associated with auditory speech processing are a direct reflection of implicit prediction error within a hierarchical predictive coding system (Friston, 2005; Wacongne, Changeux, & Dehaene, 2012; Wacongne et al., 2011), and recent evidence suggests that this may also be true of semantic processing (Rabovsky & McRae, 2014), with different ERP components reflecting both the representational level of a prediction error (Kuperberg, 2014), as well as the certainty of our original predictions (Kuperberg, 2013, 2014). In this study, we focus on three sets of ERP components: the N400, the posterior late positivity or P600, and the late negativities.

The N400 is a negative-going waveform with a centro-parietal scalp distribution, seen between 300-500ms after word onset. It reflects changes in activity within semantic memory that are induced by incoming stimuli (Kutas and Federmeier, 2011), and it can be formalized as reflecting implicit prediction error at the level of semantic features (Rabovsky & McRae, 2014). Of particular relevance to the present study is evidence that the amplitude of the N400 can be influenced by implicit predictions generated at higher representational levels (e.g. event sequences or event structures). If an incoming word's coarse-grained (Bornkessel-Schlesewsky & Schlesewsky, 2009; Paczynski & Kuperberg, 2011, 2012) or finer-grained (Ferretti, Kutas, & McRae, 2007; Metusalem et al., 2012; Paczynski & Kuperberg, 2012) semantic features match these implicit predictions, the N400 evoked by that word is attenuated. Moreover, pragmatic communicative cues (for example, commas, e.g. Nieuwland, Ditman, and Kuperberg, 2010, Experiment 2, or speech dysfluencies, e.g. Corley, MacGregor, and Donaldson, 2007) can play an important role in determining whether or not comprehenders are able to fully use a discourse context and stored event knowledge to generate implicit predictions leading to facilitated semantic processing during online comprehension.

Although the N400 is influenced by predictions generated at the event structure layers, and it reflects implicit prediction error at the level of semantic features, it is actually not directly sensitive to prediction errors at the event structure layer itself (see Kuperberg, 2013 for discussion and Federmeier et al, 2007; Lau et al, 2013; Van Petten and Luka, 2012 for consistent evidence). Rather, the costs of disconfirmed predictions at the level of event structure manifest as waveforms that peak after the N400 time window.

The first of these is a posteriorly-distributed late positive-going ERP component — the so-called P600, which peaks between 500-800ms after stimulus onset (Kuperberg, 2007). Although the P600 was originally characterized as the ERP component produced by words that violated the syntactic constraints of predicted events structures (Hagoort, Brown, & Groothusen, 1993; Osterhout & Holcomb, 1992; 1993), it is now clear that it can also be evoked by violations of strong semantic constraints on event structure (see Kuperberg, 2007 for a review). Specifically, it is triggered when the sentential or discourse context encourages a strong *near-certain* prediction for a particular event structure (a prediction with near-100% probability), and this *conflicts* with the event structure that is computed by initial attempts to integrate the bottom-up input. It may reflect prolonged attempts to (re)-analyze the context and input to come up with a new discourse model (Kuperberg, 2007, Paczynski & Kuperberg, 2012). In Bayesian terms, it may reflect ‘unexpected surprise’ that triggers a switch to a new generative cause at the level of event sequences that better explains (and allows us to rapidly adapt to) the input (see Courville, Daw, & Touretzky, 2006; Qian, Jaeger, & Aslin, 2012; Yu, 2007; Yu & Dayan, 2005; discussed by Kuperberg, 2013, 2014).

The second set of waveforms that can be seen when event structure predictions are disconfirmed by an input is a group of late negativities, which also peak past the N400 time window and often have a more widespread or frontal distribution than the N400. Unlike the P600, these late negativities are not evoked by violations of a single *near-certain* event structure prediction. Rather, they are seen when the context constrains for one event structure with medium-high probability and another event structure with lower probability, and the bottom up input leads the less probable event structure to be selected (e.g. Lee and Federmeier, 2006, 2009, Baggio, Lambalgen, and Hagoort, 2008; Bott, 2010; Wlotko and Federmeier, 2012; Paczynski, Jackendoff and Kuperberg, 2014; Wittenberg, Paczynski, Wiese, Jackendoff, and Kuperberg, 2014). In this sense, they correspond to implicit prediction error at the level of event structures (see Kuperberg, 2014).

In the present study, we examined these ERPs to ask four questions about how and when predictions established by *even so* are used during online discourse processing.

1. Reversed and enhanced semantic expectations under *even so?*

Our first question was whether, under *even so*, our implicit predictions about upcoming semantic features, if any, would be based on an alternative world model or based on the stored long-term real-world knowledge. Previous studies using other constructions that set up alternative world models have provided different answers to this question. Sometimes, stored long-term real-world knowledge appears to dominate the initial stages of semantically processing incoming words, as indexed by the N400. For example, in a study of counterfactuals, Fergurson et al. (2008, Experiment 2) asked participants to read sentences like, “If cats were not carnivores, families could feed their cats a bowl of fish/carrots…”. They showed that the real-world consistent (but alternative world inconsistent) word, *fish*, elicited a smaller N400 than the real-world inconsistent (but alternative world consistent) word, *carrots* (see also Ferguson et al. 2008, Experiment 1, and Ferguson and Sanford, 2008, for related eye tracking results). At other times, however, an alternative world model can override long-term real-world knowledge to facilitate semantic processing of incoming words (e.g. Nieuwland and Van Berkum, 2006; Nieuwland and Martin, 2012; Nieuwland, 2013; also see Hald, Steenbeek-Planting, and Hagoort 2007, and Ferguson and Breheny, 2011). For example, in another study of counterfactuals, Nieuwland and Martin (2012; also see Nieuwland, 2013) asked participants to read sentences like, "If NASA had not developed its Apollo Project, the first country to land on the moon would have been Russia/America surely", and showed that the N400 was smaller to *Russia* than to *America*.

One factor that seems to be crucial in determining whether we draw upon long-term real-world knowledge or an alternative world model to facilitate subsequent semantic processing is whether the discourse context is pragmatically constraining—that is, whether it provides explicit information about *how* the upcoming proposition will be linked to it. This is nicely illustrated by the counterfactual experiments described above. In the study by Ferguson et al. (2008, Experiment 2), the preceding discourse context does not constrain for a particular event or state: people will have quite different opinions (or little to say) about the most likely thing that people will feed a non-carnivorous cat. In the studies by Nieuwland and Martin (2012) and Nieuwland et al. (2013), however, the discourse is more constraining: given how well-known the America-Russia space race is, most people will expect the upcoming event to describe the opposite of what actually happened.

Returning to the present study, we hypothesized that the pragmatic constraining function of *even so* would lead comprehenders to draw upon the alternative world model established and generate strong (high probability) predictions about an upcoming real-world inconsistent event and semantic features consistent with this event, leading to more semantic facilitation of congruous incoming words. Indeed, given this pragmatic constraining function, these predictions might be stronger (higher probability) than those generated in the plain scenarios.

To test these hypotheses, we examined the modulation of the N400 across two contrasts. First, we compared the coherent and incoherent even-so scenarios. If comprehenders reverse their expectations under the scale-reversing function of *even so*, then we should see a smaller N400 (more semantic facilitation) to critical words like *celebrated* in the even-so coherent (but real-world inconsistent) scenarios (e.g. "…Elizabeth took the test and failed it. Even so, she went home and celebrated…"). Second, we contrasted the even-so and the plain coherent scenarios. If comprehenders enhance their semantic expectations under the pragmatic constraining function of *even so*, then we should also see a smaller N400 to *celebrated* in the even-so coherent scenarios than in the coherent plain scenarios (e.g. "… Elizabeth took the test and aced it. She went home and celebrated…").

2. Costs of disconfirmed event structure predictions under *even so?*

Our second question was whether we would see any evidence in the ERP waveform of encountering input that disconfirmed any high probability predictions of upcoming event structures established under *even so.* As noted above, the costs of violating event structure predictions do not manifest directly on the N400 itself (Federmeier et al, 2007; Lau et al, 2013; Van Petten and Luka, 2012; Kuperberg, 2013), but appear later, on components that peak past the N400 time window — a stage at which the event structure has been fully computed from the bottom-up input.

We considered two possibilities. The first was that *even so* would lead comprehenders to generate a single strong *high certainty* prediction for one specific type of event that has a real-world inconsistent causal relationship with the preceding event. On this account, if integration of the bottom-up input yields a real-world consistent causal relationship, the resulting *conflict* would trigger prolonged attempts to (re)-analyze the context and input to come up with a new discourse model (Kuperberg, 2007 & 2013). This might manifest as a larger posteriorly distributed late positive-going P600 to targets like *celebrated* in incoherent even-so scenarios than in incoherent plain scenarios.

The second possibility was that, rather than commit with near-100% certainty to a single real-world inconsistent event structure under *even so*, comprehenders would consider the possibility of encountering a real-world consistent event structure with some lower probability. On this account, integration of the bottom-up input would ultimately select the less probable but real-world consistent structure over the more probable but real-world inconsistent structure, and this selection cost might manifest as a larger late anteriorly-distributed sustained negativity to targets in the incoherent even-so scenarios than in the incoherent plain scenarios (e.g. Baggio, Lambalgen, and Hagoort, 2008; Bott, 2010; Paczynski, Jackendoff & Kuperberg, In press; Wittenberg, Paczynski, Wiese, Jackendoff, & Kuperberg, 2014). Note that the two possibilities outlined above are not mutually exclusive because the certainty of event structure prediction might vary between participants and/or between trials.

3. Wrap-up costs of assessing overall discourse coherence against the alternative world established under *even so*

Our third question was whether *even so* would lead to independent costs associated with assessing a discourse model that is coherent under a set of likelihood relations that differ from our default real-world knowledge. These costs might not necessarily be apparent at the point of the critical word itself. However, they might manifest later as a sustained negativity on the sentence-final word—the point at which overall discourse coherence is ‘wrapped up’ and evaluated.

A larger sustained negativity is often seen on the final words of sentences that are implausible (versus plausible) in relation to real-world knowledge, even when the implausibility or anomaly occurs mid-sentence (e.g. Hagoort and Brown, 2000; Hagoort, 2003; Ditman, Holcomb and Kuperberg, 2007; De Grauwe, Swain, Holcomb, Ditman and Kuperberg, 2010). This is presumably because it is harder to assess overall coherence if the overall discourse model mismatches long-term real-world knowledge than if it matches real-world knowledge. If similar wrap-up costs are incurred when the comprehension system evaluates overall coherence against a set of reversed likelihood relationships established under *even so* (the alternative world model), this would predict a larger sustained negativity effect on sentence-final words in the even-so scenarios than in the plain scenarios, even when both are coherent. Moreover, it should be even harder to come up with a final representation of meaning when overall coherence is evaluated against a set of reversed likelihood relationships, *and* the scenario turns out to be incoherent, predicting the largest sustained negativity on the final words of the even-so incoherent scenarios.

4. Effects of task

Our final set of questions concerned the effects of task on comprehending both the plain and the even-so scenarios. Task can influence processing in several different ways. It can influence the degree to which comprehenders attend to different aspects of discourse, including the semantic relationships between propositions. This can, in turn, influence the strength/certainty of our predictions, which, as discussed above, can influence the neural mechanisms engaged at multiple stages of processing. To examine the effects of task in this study, we carried out two experiments. In Experiment 1, participants were asked to explicitly rate the coherence of each discourse scenario. In Experiment 2, participants were asked to read the scenarios and to answer intermittent comprehension questions about their content.

Experiment 1

In Experiment 1, participants read the four types of three-sentence scenarios described in Table 1. The factors of Coherence and Even-so were fully crossed, and we measured ERPs on both critical words (e.g. *celebrated*) as well as on the final words of the third sentence. After each scenario, participants explicitly judged the coherence of the final sentence in relation to the previous context. This encouraged them to pay close attention to the internal semantic relationships between the propositions.

Importantly, we matched general schema-based semantic relationships between the critical word and the ‘bag of words’ in the context across all four conditions using Latent Semantic Analysis (LSA, see Methods). This allowed us to determine whether readers based their expectations on specific types of relationships between events, e.g. the fact that students are likely to celebrate after doing well on exams (Yang et al. 2007; St. George et al. 1997; Kuperberg et al. 2011), or on more general, unstructured word associations based around a particular schema (Schank and Abelson, 1977), e.g. the general association between successful/failed exams and parties afterwards (see Otten and Van Berkum, 2007; Paczynski and Kuperberg, 2012 for discussion).[2]

Our starting point was the plain scenarios. We asked whether, with these task instructions, readers would generate predictions about likely upcoming events/states, leading to facilitated semantic processing of incoming words whose semantic features were associated with these events. Based on a previous study in which we contrasted causally coherent and incoherent plain discourse scenarios while participants carried out a similar explicit

[2] In fact, most studies of sentence and discourse processing have not matched general schema-based lexical relationships in this way and have therefore not been able to distinguish between these two possibilities (see Otten & Van Berkum, 2007; Kuperberg et al. 2011, and Paczynski & Kuperberg, 2012 for discussion). For example, even in the classic example of an N400 effect, “She liked to take her coffee with cream and sugar/dog” (Kutas and Hillyard, 1980), the attenuation of the N400 to *sugar* (versus *dog*) could, in theory, be driven by its closer semantic relationship with the general schema of *coffee drinking*, rather than by a more specific expectation of the most likely thing, after cream, that someone would put inside her coffee. Those studies that have used such schema-matched stimuli reveal a mixed picture, as discussed further in Experiment 2.

coherence judgment task. (Kuperberg et al., 2011), we hypothesized that we would see a smaller N400 on critical words in the plain coherent than in the plain incoherent scenarios.

Having established how the plain scenarios were processed, we then turned to the effects of *even so*. We make the following hypotheses based on the discussion in the General Introduction: (1) If comprehenders are able to draw upon the alternative world model previously generated under *even so*, and use this to *reverse* their predictions about the real-world likelihood of upcoming events, the N400 should be smaller to critical words in the coherent than the incoherent even-so scenarios, just as in the plain scenarios; indeed, if predictions established under *even so* are stronger (higher probability) than in the plain scenarios, then the N400 to critical words in the coherent even-so scenarios should be even smaller than than in the coherent plain scenarios. (2) If *even so* leads to strong reverse predictions about upcoming events, then disconfirmation of these event predictions by the bottom-up input should lead to prolonged neural costs (past the N400 time window) to critical words in the even-so incoherent versus the plain incoherent scenarios. Finally, (3) if, during sentence-final wrap-up, evaluating coherence against the temporary alternative world model established under *even so* incurs more costs than evaluating coherence against long-term real-world knowledge, then sentence-final words in the even-so coherent scenarios should produce a larger negativity than sentence-final words in the plain coherent scenarios; moreover, this negativity should be still larger on sentence-final words in the incoherent even-so scenarios.

Methods

Construction and characterization of stimuli—One-hundred-and-eighty sets of two-sentence scenarios were constructed, each with four conditions (45 scenarios per condition), see Table 1. In all scenarios, a critical word appeared in the final sentence but before the sentence-final word. The number of words between the critical word and the sentence-final word varied between 0 and 3 across trials, but was matched between conditions within a scenario. The number of words between the critical word and the sentence initial word (excluding “Even so”) mostly varied between 1 to 4 words (with only a few scenarios with more than 4 words), but once again this was matched between conditions within a scenario.

The four conditions were constructed by crossing two factors: Even-so (the presence or absence of the phrase, “Even so” at the beginning of the final sentence: plain or even-so scenarios) and Coherence (the coherence of the critical word in relation to its preceding context: coherent or incoherent scenarios). In the plain coherent and the plain incoherent conditions, the final sentence was identical: differences in coherence arose because of differences in the first two sentences. In the even-so scenarios, coherence was reversed: the original plain coherent scenarios became the even-so incoherent scenarios, and the original plain incoherent scenarios became the even-so coherent scenarios. The critical word was thus identical in all four conditions.

Lexical predictability of critical words and lexical constraint of discourse contexts: Offline cloze norming—The 180 sets of scenarios were counterbalanced across four lists using a Latin Square design. For each scenario, the critical word and all words following it were removed and replaced by an ellipsis, e.g. "Elizabeth had a history exam on Mon. She took the test and aced it. She went home and ….". Cloze ratings of these stems were conducted as an online survey using SurveyMonkey.com, with participants recruited from Tufts University and other neighboring areas. Participants were asked to read the scenario stems and to complete the unfinished last sentence by writing down the most likely ending. Initially, 40 native English speakers (30 female, 10 male, average age: 23.3) participated (ten per list). Cloze probabilities for each of the four scenario types were calculated based on the percentage of respondents who produced a word that matched the critical word exactly. Based on the initial results, we modified 13 of the scenarios that didn't show any difference between coherent and incoherent scenarios, and carried out a second cloze study on these 13 new items with another set of 40 native English speakers (29 female, 11 male, average age =26.2). We used these 13 new items to replace the old 13 items, and then recalculated the cloze probability for each item across the entire stimulus set.

Cloze probabilities for each scenario type are given in Table 1. A 2 × 2 repeated measures ANOVA with Coherence and Even-so as within-items factors revealed a main effect of Coherence (F(1,179)=329, p < .001) and a main effect of Even-so (F(1,179)=22, p<.001). There was also an interaction between the two factors (F(1, 179)=21, p<.001). Follow-up paired t-tests examining effects of Coherence at each level of Even-so showed that, as expected, the even-so and the plain coherent scenarios had significantly higher cloze probabilities than their corresponding incoherent conditions (plain coherent vs. plain incoherent: t(179)=16, p<0.001; even-so coherent vs. even-so incoherent: t(179)=13, p<0.001), but that the *difference* in cloze probability between the coherent and incoherent conditions was larger in the plain than in the even-so scenarios. Follow-up t-tests examining effects of Even-so at each level of Coherence showed that the cloze probability of critical words in the plain coherent scenarios was significantly higher than in the even-so coherent scenarios (t(179)=5, p<0.001), but that there was no significant difference in cloze probability between the plain and even-so incoherent scenarios (t(179)=0.16, p>0.8).

In addition to calculating cloze probabilities, we also calculated the contextual lexical constraint for each type of scenario context by finding the most common completion across participants who saw that context, regardless of whether or not it matched the critical word, and tallying the number of subjects who provided this completion, see Table 1. For example if, for a given scenario, our designed critical word was "disappointed" and 3 out of 10 people provided "disappointed" as their answer, then the cloze probability would be 0.3 for this scenario, but if 5 out of the 10 people provided "confused" as their answer, then the lexical constraint probability for this context would be 0.5. A 2 × 2 repeated measures ANOVA with Coherence and Even-so as within-items factors again revealed a main effect of Coherence (F(1, 179)=22, p < .001), a main effect of Even-so (F(1, 179)=5.7, p<.05), and an interaction between the two (F(1, 179)=7.4, p<.01). Follow-up paired t-tests showed that the lexical constraint of the plain coherent contexts was greater than the plain incoherent contexts (t(179)=5, p<0.001), while the difference between the even-so coherent and incoherent contexts was only marginally significant (t(179)=1.9, p<0.06). In addition, the lexical constraint of the plain coherent contexts was greater than the even-so coherent scenario contexts (t(179)=3.6, p<0.001), but there was no difference in lexical constraint between the plain and even-so incoherent scenario stems (t(179)=0.76, p>0.9).

Latent Semantic Analysis—Latent Semantic Analysis (LSA, a measure of semantic relatedness, Landauer and Dumais 1997; Landauer et al. 1998) was carried out on the final stimulus set, on a term-to-term basis, to examine the Semantic Similarity Values (SSVs) between the critical word and previous content words across the context of each scenario. A paired t-test revealed no significant differences between the plain coherent and plain incoherent scenarios (t(179)=1.55, p > 0.10). Note that, because of how the stimuli were constructed, SSVs were the same for the plain coherent and even-so incoherent conditions, and for the plain incoherent and even-so coherent conditions. See Table 1 for SSVs in all four scenario types.

Set-up of lists for the ERP experiment—The final set of experimental scenarios was divided into four lists, counterbalanced using a Latin Square design. During the ERP experiment, each participant viewed only one list and therefore only one condition of each scenario, but across all participants, each scenario and critical word was seen in all four conditions. Each list had 180 scenarios, 45 from each condition. The order of items was randomized within each list separately.

Participants in the ERP experiment—Twenty-nine native English speakers initially participated in the ERP study (two participants were subsequently excluded for extensive ocular and muscular artifacts, see below). All participants were undergraduate students recruited from local universities. They were all right-handed, as assessed using the Edinburgh handedness inventory (Oldfield 1971), with normal or corrected-to-normal vision, and no history of neurological disorders. Participants were paid for their participation and gave full consent according to the guidelines of the Tufts University Human Subjects Committee. The 27 subjects included in the data analysis had an average age of 20 years (SD: 1.7) and 14 were males.

Stimulus presentation—Participants sat in a quiet and dimly-lit room, separated from the experimenter and control computers. Their task was to rate each scenario on a 1 to 5 scale, based on how naturally the third sentence followed on from the previous two sentences. For half of the participants, a score of 1 meant “it does not follow at all” and 5 meant “it follows very naturally”; and for the other half, the scoring was reversed for counterbalancing purpose. Before starting the experiment, each participant read twelve practice scenarios to ensure that they understood the task.

Stimuli were presented on a computer monitor, in white font, centered on a black background Participants were randomly assigned to one of the four lists. Each trial began with the word “READY” on the screen, which cued the participant to press a button to begin reading the three-sentence scenario. The first two context sentences were presented one after another as whole sentences. Participants read these two sentences at their own pace, pressing a button to move on to the second sentence. They then saw a fixation cross (“+”) in the middle of the screen for 500ms, followed by a blank screen for 100ms, and then the last sentence was presented word by word. Each word was centered in the middle of the screen, and was presented for 350ms, followed by an interstimulus interval (ISI) of 150ms. In the even-so scenarios, the phrase “Even so” was presented as a whole. Because it consists of two words, it appeared on the screen for 400ms followed by an ISI of 150ms. The last word of the final sentence appeared with a period and was presented for 800ms. A 400ms ISI followed this final word, and then a “?” appeared on the screen which cued participants to make their rating responses. Participants indicated their responses by pressing one of the five buttons on the response pad.

ERP recording—The EEG response was recorded from 29 electrodes (Electro-Cap International, Inc., Eaton, OH; see Figure 1 for montage). Additional electrodes were placed below the left eye and at the outer canthus of the right eye to monitor vertical and horizontal eye movements. There were also two mastoid electrodes (A1, A2) and the EEG signal was referenced to the left mastoid online. The EEG signal was amplified by an Isolated Biometric Amplifier (SA Instrumentation Co., San Diego, California) with a band pass of 0.01-40 Hz. It was continuously sampled at 200Hz and the impedance was kept below 5kOhm.

ERP analysis—Trials contaminated with eye artifact (with max-min amplitudes exceeding 70μv, as well as visual inspection) or amplifier blockage were excluded from analyses. After artifact rejection, averaged ERPs, time-locked to critical words, were obtained by calculating the mean amplitude (relative to a 100ms pre-stimulus baseline). At the critical word, we carried out analyses across two time windows. To capture the peak of the N400 in all four conditions, we used a 350-450ms time window. To capture the late positivity/P600 in all four conditions, and to avoid component overlap with the earlier N400 effect, we used a 600-800ms time window. Examination of the waveforms also revealed a late, sustained anteriorly-distributed negativity effect, which we captured between 800-1000ms. At the sentence-final word, visual inspection of the waveform revealed a prolonged negativity, which was captured with a 300-1000ms time window.

We initially carried out two omnibus repeated-measures ANOVAs in which the scalp was subdivided into several 3-electrode regions along its anterior–posterior distribution, at both mid and peripheral sites (each region contained three electrode sites, see Figure 1). In the mid-regions omnibus ANOVA, the within-subject variables were Coherence (2 levels: coherent, incoherent), Even-so (2 levels: plain, even-so), and Region (5 levels: prefrontal, frontal, central, parietal, occipital). In the peripheral regions omnibus ANOVA, the within-subjects variables were Coherence (2 levels: coherent, incoherent), Even-so (2 levels: plain, even-so), Region (2 levels: frontal, parietal) and Hemisphere (2 levels: left, right). For further follow-ups, we focused on the subgroup of regions that showed most modulation across conditions (smallest p values; largest F values in omnibus ANOVAs analyses carried out in each region).

In each subgroup of regions, we carried out 2 (Coherence) × 2 (Even-so) × Region ANOVAs; any interactions between Coherence and Even-so were followed up by (a) by examining the effects of Coherence at each level of Even-so, and (b) by examining the effects of Even-so at each level of Coherence. In all analyses, the Greenhouse and Geisser

(1959) correction was applied to repeated measures with more than one degree of freedom, and a significance level of alpha = .05 was used for all comparisons.

Results

Behavioral results—The coherence ratings for each of the four scenario types are given in Table 1. A 2 × 2 ANOVA confirmed a significant main effect of Coherence (F(1, 26)=131, p<.001). It also revealed a main effect of Even-so, reflecting higher overall coherence ratings in the plain than the even-so scenarios (F(1, 26)=33, p<.001). In addition, there was a significant interaction between these two variables (F(1, 26)=33, p<.001). Planned follow-up comparisons examining the effects of Coherence on the plain and even-so scenarios separately indicated that, as expected, the coherent scenarios were always rated as significantly more coherent than the incoherent scenarios (plain: coherent vs. incoherent: t(26) = 32, p < .001; even-so: coherent vs. incoherent: t(26) = 2.45, p < .05), although the difference was larger in the plain than in the even-so scenarios. Follow-ups examining the effects of Even-so in the coherent and incoherent scenarios separately showed that the coherent plain scenarios were rated as *more coherent* than the coherent even-so scenarios (t(26)=7.6, p<.001), and the incoherent plain scenarios were rated as *more incoherent* than the incoherent even-so scenarios (t(26)=-2.9, p<.01). In other words, the plain coherent and plain incoherent scenarios were rated as more coherent and incoherent respectively than their corresponding even-so scenarios, which received ratings that were in between these two extremes.

Event related potentials

Critical Word: At the critical word, 20% of trials were rejected for artifact (plain coherent: 19%; plain incoherent: 19%; even-so coherent 20%; even-so incoherent: 20%). A 2 × 2 within-subjects ANOVA showed that the rejection rate did not differ between the coherent and incoherent scenarios (no main effect of Coherence F(1,26)<.1, p>.9), or between the even-so and plain scenarios (no effect of Even-so F(1,26)=.9, p>.3). There was also no interaction between these two factors (F(1,26)<.1, p>.9).

N400: 350-450ms: There was a main effect of Coherence, reflecting a widespread N400 effect across the even-so and plain scenarios (mid-regions: F(1,26)=40.8, p<.001; peripheral regions: F(1,26)=36.7, p<.001), which was largest in frontal, central and parietal mid-regions (interaction between Coherence and Region in the mid-regions analysis, F(4, 104)=4.3, p<.01, with follow-ups showing the effect of Coherence in each of these three regions, Fs>23, p<.001). To determine how effects of Coherence were modulated by Even-so, we collapsed across these three frontal, central and parietal mid-regions (9 electrode sites in total).

In this 9-electrode region, Coherence was modulated by Even-so (Coherence × Even-so interaction: F(1,26)=5.4, p<.05). There were significant N400 effects of Coherence in both the plain scenarios (F(1,26)=7, p<.05) and the even-so scenarios (F(1, 26)=40, p<.001). However, the magnitude of the N400 effect in the even-so scenarios was larger than in the plain scenarios, see Figure 2B. This larger N400 Coherence effect was driven by a *smaller* N400 to *coherent* critical words in the even-so than the plain coherent scenarios (F(1, 26)=5.3, p<.05),[3] see Figure 3A (note that voltage map in Figure 3A shows a positivity between 350-450ms because the plain condition was subtracted from the even-so condition). In contrast, there was no difference in the N400 evoked by *incoherent* critical words in the plain and even-so incoherent scenarios (F(1, 26)<1, p>.4), see Figure 3B.

Late Posterior Positivity/P600: 600-800ms: Collapsed across the even-so and plain scenarios, there was a P600 effect over parietal (left, right and mid) and mid-occipital regions (interactions between Coherence and Region in the mid-regions analysis, F(4, 104)=9.5, p<.001, and in the peripheral regions analysis, F(3,78)=14.2, p<.001, with follow-ups showing the effect of Coherence in each of these four regions, all Fs > 6.8, ps < 0.05). To determine how the effect of Coherence was modulated by Even-so, we collapsed across these four regions: left, mid and right parietal and mid-occipital regions (12 electrode sites in total).

In this 12-electrode parietal-occipital region, there was a three-way interaction between Coherence, Even-so and Region (F(2, 52)=4.9, p<.05). Follow-ups showed a P600 effect of Coherence in both the plain scenarios (at left and right parietal regions, F(1,26)s>4.7, ps<.05, Figure 2A), and in the even-so scenarios (at the right parietal region and the mid-occipital region, F(1,26)s>6.4, ps<.05, Figure 2B). Once again, the effect was larger in the even-so scenarios than in the plain scenarios. This time, however, the larger effect in the even-so scenarios was driven by a larger late positivity to *incoherent* even-so than incoherent plain critical words (Figure 3B, in the occipital region, F(1,26)=8.2, p<.01, and in the right parietal region, F(1,26)=4.3, p<.05); there was no significant difference in the late positivity evoked by coherent critical words in the even-so and plain coherent scenarios in any of these regions (all ps>.05, Figure 3A).

Late Anterior Negativity: 800-1000ms: Analysis within this time window revealed an Even-so × Coherence × Region interaction in the mid-regions analysis (F(4, 104)=16.6, p<.001). Some of this effect was driven by the late positivity effect continuing into the late time window at the posterior-parietal site. But in the prefrontal region (electrode sites FP1, FP2, FPz), there was a larger negativity to critical words in the even-so incoherent scenarios than in the other three scenarios (all ts(26)>2.6, ps<.05, see Figure 2 and 3).

Sentence-final word: At the sentence-final word, 28% of trials were rejected for artifact (plain coherent 24%; plain incoherent: 27%; even-so coherent: 29%; and even-so incoherent: 32%). There was a near-significant effect of Coherence (F(1, 26)=4.0, p=.056) because there were slightly more rejected incoherent than coherent trials, and an effect of Even-so (F(1,26)=9.8, p<.01) because there were slightly more rejected even-so than plain trials, but there was no interaction between the two factors.

Sentence-final negativity: 300-1000ms: Activity on the sentence-final words was captured over a prolonged 300-1000ms time window, shown in Figure 4. As expected, there was a

[3] To exclude the possibility that the smaller N400 to critical words in the even-so coherent versus the plain coherent scenarios was driven purely by their later position in the sentence (see Van Petten and Kutas, 1990, for effects of word position on the N400), we looked at the N400 on the word following the critical word. We found no difference between these two conditions between 350-450ms (F(1, 26)=0.66, p>.4), suggesting word position alone did not produce this N400 difference.

larger sustained negativity on sentence-final words in the incoherent than the coherent scenarios (main effects of Coherence: mid-regions, F(1,26)=24, p<.001; peripheral regions, F(1,26)=15, p<.01 analyses). In addition, there was a main effect of Even-so, with a larger negativity on sentence-final words in the even-so than the plain scenarios (mid-regions: F(1,26)=24, p<.001; peripheral regions: F(1,26)=15, p<.01). Finally, in the mid-regions analysis, there was an interaction between Even-so, Coherence and Region (F(4, 104)=3.4, p<.05). To follow-up this three-way interaction, we carried out pair-wise comparisons at two 6-electrode regions: posterior (combining the parietal and occipital 3-electrode regions) and frontal (combining the frontal and central 3-electrode regions).

In the posterior 6-electrode region, the final words of the even-so incoherent scenarios evoked the largest negativity (differing significantly from the three other conditions, Fs > 13, ps < .01), while the final words of the plain coherent scenarios evoked the smallest negativity. The final words of the plain incoherent and even-so coherent scenarios each evoked medium-sized sustained negativities, which were smaller than in the even-so incoherent scenarios (Fs > 13, ps < .01), but larger than in the plain coherent scenarios (Fs > 4, ps < .05). In the frontal 6-electrode region, however, there was no difference in the amplitude of the negativities evoked by sentence-final words of the even-so coherent, even-so incoherent and plain incoherent scenarios (Fs<3, ps>.05), which all evoked a more negative waveform than the sentence-final words of the coherent plain scenarios (Fs > 6, ps < .05).

Discussion

In this experiment, participants made explicit judgments about the coherence of each scenario. In both the plain and the even-so scenarios, we saw both N400 and P600 effects of coherence on the critical words. In the even-so scenarios, however, both these effects were larger than in the plain scenarios. In addition, in the even-so scenarios, we also saw an effect of coherence on a late anterior negativity between 800-1000ms. At the sentence-final word, we saw effects of both Coherence and Even-so. Finally, the pattern of these ERP findings across the four conditions dissociated from participants' offline behavioral coherence judgments of the same sentences. We discuss each of these findings in turn.

The plain scenarios—The significant N400 effect of Coherence in the plain scenarios is consistent with our previous findings (Kuperberg et al., 2011), in which we contrasted causally coherent and incoherent plain discourse scenarios while participants carried out a similar explicit coherence judgment task. Just as in this previous study, the N400 in this study was attenuated on the coherent critical words, even though its general schema-based semantic relatedness with the preceding context (as operationalized by LSA) was matched with the incoherent scenarios (see Table 1). This tells us that, when participants actively attend to discourse relationships, they are able to construct a full discourse model from the context, use this model to access stored information about event relationships, and generate predictions about likely upcoming events and their associated semantic features, leading to facilitated processing of incoming words whose semantic features match these predictions.

In the present study, the N400 effect was followed by a small P600 effect, which we did not see in our previous study (Kuperberg et al., 2011). The reason for this is unclear, but one possibility is that the presence of the even-so scenarios themselves encouraged comprehenders to engage in additional analysis in attempts to make sense of all incoherent sentences (see footnote 6).

The even-so scenarios—The N400 attenuation to critical words in the coherent versus incoherent even-so scenarios indicates that readers were able to draw upon the alternative world model set up under *even so* rather than stored long-term real-world knowledge, to facilitate semantic processing of upcoming words. Moreover, our finding that the N400 Coherence effect was actually *larger* in the even-so than the plain scenarios, and that this was due to an attenuation of the N400 to the even-so coherent (versus plain coherent) critical words, indicates that the event structure and semantic predictions, set up under *even so*, were *stronger* (higher probability) than those generated in the plain scenarios on the basis of default world knowledge. We argue that this is because *even so* narrowed down the number of potential upcoming event structures to those that were causally real-world inconsistent, unlike in the plain scenarios where they may have considered multiple possible upcoming event structures (based on causal, spatial and temporal relationships with the preceding context), each with lower probability.

The *larger* P600 on critical words in the incoherent even-so than in the incoherent plain scenarios suggests that, when these high probability predictions were violated, prolonged neural costs were incurred. More specifically, we argue that, under *even so*, at least on some trials, comprehenders predicted a real-world inconsistent event with high certainty (near 100% probability), and that the P600 was triggered by conflict between this event structure and the structure produced by initial attempts to integrate incoming critical words. We suggest that the P600 itself reflected prolonged attempts to integrate the context and critical words to construct a new discourse model (see Kuperberg, 2007; Paczynski & Kuperberg, 2012; Kuperberg, 2013).[4]

[6] A potential concern is that the absence of a N400 effect in Experiment 2 was due to a lack of statistical power due to the smaller number of participants run in Experiment 2 than in Experiment 1 (20 vs. 27). It is also possible that, in both Experiments 1 and 2, the P600 seen to the plain scenarios was due to the presence of so many even-so incoherent scenarios in the experimental environment, which may have encouraged comprehenders to engage in additional analysis to *all* types of incoherent critical words. Finally, the presence of this P600 in the plain scenarios may have masked any N400 on the scalp surface due to component overlap. To address these three concerns, we carried out an additional Experiment 3 in which a new group of 16 individuals read the same plain scenarios, but with no even-so scenarios in the experimental set. They carried out the same comprehension task as used in Experiment 2. Results are reported in Supplementary material at: http://www.nmr.mgh.harvard.edu/kuperberglab/publications.htm. To summarize our findings: (1) we saw no N400 or P600 coherence effect (although, as in Experiments 1 and 2, we did see a prolonged negativity effect on the sentence-final word of the incoherent versus coherent plain scenarios, indicating that participants were engaged in comprehending the sentences). (2) When we pooled the ERPs evoked by critical words in the plain scenarios of Experiments 2 and 3 to give a total of 36 participants (exceeding the number of participants in Experiment 1), we still did not see any sign of a N400 effect. Based on these results, it seems unlikely that statistical power alone explains the absence of N400 in Experiment 2, or that the absence of the N400 effect in Experiment 2 was simply an artifact of component overlap. Finally, the absence of a P600 effect in Experiment 3 provides some preliminary support for the idea that the large number of even-so scenarios in the wider experimental environment, leading to the P600 effect to the plain scenarios in both Experiments 1 and 2, although this is speculative and requires systematic follow-up.

[4] Note that the enhanced P600 was driven entirely by a larger P600 to the *incoherent* critical words in the even-so (versus plain) scenarios, and that there was no P600 difference between the coherent even-so and plain scenarios. This indicates that the enhanced N400 reduction in the coherent even-so (versus the plain) scenarios cannot be simply explained by temporal and spatial overlap between the N400 and the P600 components at the scalp surface: if the N400 to coherent critical words in the even-so (versus plain) scenarios was being artificially ‘pulled down’ by an overlapping P600, we would have seen a larger (more positive-going) P600 to these coherent even-so (versus coherent plain) critical words.

In addition to producing the late posteriorly distributed positivity/P600 effect, we also saw evidence of a late sustained anteriorly distributed negativity effect on critical words in the incoherent even-so (versus the incoherent plain) scenarios. We suggest that, at least on some even-so trials, comprehenders kept alive the possibilities of *both* a *real-world inconsistent* and a r*eal-world consistent* upcoming event structure. That is, the late sustained anterior negativity effect reflected the maintenance of these alternative event structures and the process of selecting the (less probable) real-world consistent event structure as the bottom-up information from the critical word was integrated (see Baggio, van Lambalgen, and Hagoort, 2008; Paczynski, Jackendoff & Kuperberg, 2014; Wittenberg, Paczynski, Wiese, Jackendoff, & Kuperberg, 2014; and Lee and Federmeier, 2009; for similar interpretations of late sustained negativities in other situations).

The sentence-final word—Our findings at the sentence-final word suggest that the costs of generating an alternative world model under *even so* did not come for free: in addition to the prolonged sentence-final negativity effect in the incoherent (versus coherent) plain scenarios, we also saw a larger sentence-final negativity effect in the even-so coherent scenarios than in the plain coherent scenarios. Moreover, at posterior sites, the effects of Coherence and Even-so appeared to be additive: the sentence-final negativity effect was maximal in the even-so incoherent scenarios, indicating that still more wrap-up costs were incurred when the integration of the sentence-final word mismatched the alternative world model that had been originally anticipated under *even so* (at frontal sites, the negativities evoked by the sentence-final words in the even-so coherent, even-so incoherent and plain incoherent scenarios were all of the same magnitude).

Dissociation between online ERP findings and offline behavioral data—The pattern of ERP findings at the critical word dissociated from the pattern of offline discourse coherence ratings across the four conditions: despite clear semantic facilitation (a smaller N400) on critical words in the even-so versus the plain coherent scenarios, participants rated the same even-so coherent scenarios as *less* coherent than the plain coherent scenarios (3.3 vs. 4.8, Table 1). Moreover, despite prolonged neural costs associated with processing the even-so (versus plain) incoherent critical words, participants rated the same even-so incoherent scenarios as *less* incoherent than the plain incoherent scenarios (2.4 vs. 1.7, Table 1).

We think that these dissociations arose because participants' offline retrospective coherence judgments about the coherence of the even-so scenarios were influenced by their long-term real-world knowledge. For example, the even-so incoherent scenario, “Elizabeth took the test and aced it. Even so, she went home and celebrated” sounds quite odd, but the relationship between the events themselves matches real-world knowledge, and this may have led participants to rate it as less incoherent than the incoherent plain scenarios. Similarly, although the even-so coherent scenario, “Elizabeth took the test and failed it. Even so, she went home and celebrated” sounds quite natural, the real-world relationship between the two main events mismatches our real-world knowledge; given time to think about its coherence, this mismatch may have led participants to rate it as less coherent than the plain coherent scenarios.

We also suggest that similar factors influenced the cloze ratings of the even-so scenarios, which also dissociated from the online ERP data. For example, when asked to produce a specific word after the context, “Alice was walking home from work at night. A stranger was following her. Even so, she felt…”, although 5 out of the 10 participants that were given this particular item produced the word “safe”, one participant actually produced “uncomfortable”, suggesting that, during these offline cloze judgments, some participants might have ignored “even so” altogether and continued the sentence on the basis of the default real-world event relations.

Another factor to consider is that, on a substantial subset of the items, although participants clearly had committed to a continuation that was semantically inconsistent with the default real-world knowledge, they did not necessarily converge on exactly the same lexical item. For example, for the scenario, “John tried out for the comedy troupe. His lines were all mumbled and unintelligible. Even so, the director asked him to…”, the target critical word was “join”, but some participants produced other coherent options, such as “stay”.

Together, these findings underline two important points. First, online neural activity does not always mirror participants’ offline coherence judgments and cloze ratings: our ERP results suggest that, with limited time before the next word appeared, *online* semantic facilitation of the mid-sentence critical words, as reflected by the amplitude of the N400, was driven by the alternative world model established under *even so*, whereas our offline ratings, and, as discussed above, neural activity at the sentence-final word, were partially driven by long-term real-world knowledge. Second, they show that enhanced prediction for a particular type of event doesn’t necessarily imply prediction or commitment to one specific lexical item. We will return to the relationships between cloze ratings, lexical prediction, event structure prediction, semantic facilitation and N400 modulation in the General Discussion.

Experiment 2

In this second experiment, we presented a new set of participants with the same experimental stimuli. This time, however, participants were asked simply to read the scenarios and answer intermittent, randomly-dispersed comprehension questions about these scenarios.

Once again, our starting point was the plain scenarios. In Experiment 1, and in our previous study examining the use of real-world causal knowledge across sentence boundaries (Kuperberg et al. 2011), we showed that, when explicitly asked to judge coherence, readers are able to retrieve and use quite specific knowledge about likely relationships between real-world events to predict upcoming events and facilitate semantic processing of incoming words consistent with these events, as reflected by an attenuation of the N400. This, however, does not imply that we *always* use event knowledge to facilitate semantic processing of incoming words during online comprehension; there are times in which semantic processing can be driven primarily by more general stored knowledge about unstructured schema-based semantic relationships between words or concepts (see Otten & Van Berkum, 2007; Kuperberg et al. 2011; Paczynski & Kuperberg, 2012; and Lau, Holcomb & Kuperberg, 2013 for discussion). Whether we use context to go beyond this more general unstructured schema-based knowledge depends on whether we are able to establish a deep discourse representation of the context and use this to retrieve and predict stored upcoming event representations before the semantic features of the incoming word becomes available. This, in turn, depends on many factors (see General Discussion). One of these factors is task demands.

Thus, in Experiment 2, our first question was whether, in the plain scenarios, without an explicit task requirement to judge coherence, readers would still be able to construct a deep discourse model to predict upcoming events and facilitate semantic processing of incoming critical words. Alternatively, semantic processing of critical words might simply be driven by an interaction between whatever contextual representation had been constructed at the time the bottom-up semantic input became available and more general, unstructured schema-based stored semantic relationships. As LSA was matched between the coherent and incoherent plain scenarios, this would predict no difference between the N400 produced by critical words in the two conditions.

Having established how the plain scenarios were processed with a comprehension task, we then asked about the effect of *even so*. Our question here was whether readers would still be able to reverse and enhance their semantic expectations, as they did in Experiment 1. If so, then we should see similar semantic facilitatory effects on the N400 to those seen in Experiment 1: a smaller N400 on critical words in even-so coherent versus even-so incoherent scenarios, and a smaller N400 on critical words in even-so coherent versus plain coherent scenarios, despite these being matched on schema-based lexical relationships. We might also see evidence of costs of disconfirmed event structure predictions, manifested as prolonged ERP effects past the N400 time window. Finally, we asked whether, in the absence of an explicit coherence judgment, we would still see a sentence-final wrap cost associated with *even so*.

Methods

Participants and procedure—Twenty-five undergraduates initially participated in Experiment 2 (inclusion criteria were the same as in Experiment 1). Three subjects were subsequently excluded from data analysis because of extensive ocular movements. Two additional subjects were excluded because we later found out that they had participated in our earlier stimuli norming studies. We report data from the remaining 20 subjects (9 males and 11 females; mean age: 20).

The procedure of this experiment was largely the same as for Experiment 1. The only difference was that, after the 400ms ISI that followed the final word of each trial, no prompt appeared to cue participants to give an explicit coherence rating. Rather, on 25% of trials, a comprehension question appeared, requiring a yes/no response, which probed participants' understanding of the scenarios. For instance, in the example scenario in Table 1, participants received the question, "Did Elizabeth care how she did on the test?" The planned correct answer for this particular question was "yes" for the plain coherent and the even-so incoherent versions, and "no" for the plain incoherent and the even-so coherent versions. Throughout the stimuli set, the "yes" and "no" answers were counterbalanced so that participants didn't anticipate a "yes" or "no" answer for a particular condition. Participants were told in advance that only some of the scenarios would be followed by a comprehension question, and that they simply needed to read and understand each discourse scenario and answer the questions when they came up.

Results

Behavioral results—Overall accuracy in answering the comprehension questions was 87% (SD: 4.41%). A 2 × 2 repeated measure ANOVA revealed no main effects of either Coherence or Even-so (all Fs < 1, ps >.6), but there was a significant interaction between the two factors (F(1,19)=22, p<.001). Paired t-tests showed that participants were significantly more accurate on questions following the plain coherent scenarios (91.72%, SD: 8.07) than the plain incoherent scenarios (80.88%; SD: 14.56), t(19)=3.2, p<0.01, but they were less accurate on questions following the even-so coherent (82.51%; SD: 8.97) than the even-so incoherent scenarios (92.08%; SD: 5.43), t(19)=3.7, p<0.01. Tests examining the effect of Even-so on the coherent and incoherent scenarios separately confirmed this pattern: questions following the plain coherent scenarios were answered more accurately than those following the even-so coherent scenarios, t(19)=3.2, p<0.01, but questions following the plain incoherent scenarios were answered less accurately than those following the even-so incoherent scenarios, t(19)=3.2, p<0.01.

Event related potentials

Critical Word: At the critical word, 17% of trials were rejected for artifact. A 2 × 2 repeated measures ANOVA showed no effect of Coherence or Even-so on the rejection rate, and no interaction between the two factors (all Fs<1, p>.4).

N400: 350-450ms: Within the 350-450ms time window, there was a 3-way interaction between Coherence, Even-so and Region (approaching significance in the mid-regions ANOVA, F(4,76)=2.5, p=.09, and significant in the peripheral regions ANOVA, F(1,19)=5.5, p<.05). Follow-ups showed a two-way interaction between Coherence and Even-so in the mid-parietal and the two peripheral parietal regions (all Fs > 4.8, ps <.05).

Further follow-ups, collapsing across the mid-parietal and the two peripheral parietal regions (9 electrodes), revealed no N400 effect of Coherence in the plain scenarios (F(1,19)<.1, p>.6; Figure 5A),[5] but a clear N400 effect of Coherence in the even-so scenarios (F(1, 19)=9.3, p<.01, Figure 5B). As in Experiment 1, this larger N400 coherence effect in the even-so scenarios was largely driven by an *attenuation* of the N400 to critical words in the coherent even-so (vs. coherent plain) scenarios (F(1,19)=4.2, p<.05; see Figure 6A; again the voltage

[5] A potential concern is that the lack of a N400 may be due to overlap of an earlier positivity. Visual inspection of more electrodes did indeed show that, at some of the central-posterior electrode sites, there was an early positivity between approximately 200-350ms that appeared larger to critical words in the plain incoherent scenarios than the plain coherent scenarios. We carried out statistics within this time window to contrast these two conditions, collapsing across the 5 electrode sites where this effect seemed to be largest (CP1, CP2, Pz, P3, P4), but found no significant effect (F(1, 19)=2.2, p>.1). In addition, we also redid the N400 analysis within the 350-450ms time window, with a new baseline between 200-300ms, i.e. we re-baselined right before the N400 time window. If the early difference masked the N400 in some way, we might expect to see a N400 difference emerge between the plain coherent and plain incoherent scenarios. This is not what we found: the effect of Coherence in the plain scenarios was non-significant (F(1,19)<.1, p>.5), but remained significant in the even-so scenarios (F(1, 19)=6.1, p<.05). Based on these two analyses, we think it is unlikely an early divergence in the waveforms masked the N400 effect in the plain scenarios.

map in Figure 6A shows a positivity between 350-450ms because the waveforms evoked in the plain scenarios was subtracted from those evoked in the even-so scenarios); there was no significant difference in the N400 evoked by the critical words in the incoherent even-so and the incoherent plain scenarios (F(1,19)=3.4, p>.05, Figure 6B).

P600: 600-800ms: Collapsed across the plain and even-so scenarios, there was a Coherence × Region interaction in the 600-800ms window (mid-regions: F(4,76)=4.3, p<.05; peripheral regions: F(1,19)=11.3, p<.01), with follow-ups showing P600 effects in mid-parietal, mid-occipital and right occipital regions (F(1,19)s>7, p<.05). There were no further interactions involving Coherence and Even-so over these regions (p>.4), indicating that the magnitude of the P600 Coherence effect did not differ between the even-so and the plain scenarios.

Late Anterior Negativity: 800-1000ms: Analysis within this time window revealed an Even-so × Coherence × Region interaction in the mid-regions analysis (F(4, 76)=5.1, p<.05). Follow-ups at the prefrontal region showed marginally larger negativities to critical words in the even-so incoherent scenarios, relative to the even-so coherent scenarios (t(19)=1.9, p<.08) and the plain incoherent scenarios (t(19)=2, p<.06), but not to the plain coherent scenarios (t(19)<1, p>.3), see Figure 5 and 6.

<u>**Sentence-final word: 300-600ms:**</u> At the sentence-final word, 26% of trials were rejected for artifact. A 2 × 2 within-subjects ANOVA showed no effect of Coherence, Even-so, and no interaction between the two factors (all Fs<2, p>.2). Visual inspection of the ERP waveform on the sentence-final word suggests that there were effects of both Coherence and Even-so, but these effects were not as prolonged as in Experiment 1. We therefore chose the 300-600ms time window for data analysis.

The waveforms for the sentence-final word are presented in Figure 7. The mid-regions omnibus ANOVAs revealed a main effect of Coherence (mid-regions: F(1,19)=4.8, p<.05) while the peripheral regions ANOVA showed a main effect of Coherence (F(1,19)=6.0, p<.05) and a marginal interaction between Coherence and Region (F(1,19)=4.1, p<.06). There were also main effects of Even-so (mid-regions: F(1,19)=4.7, p<.05; peripheral: F(1,19)=5.0, p<.05), as well as interactions between Even-so and Region (mid-regions: F(4,76)=3.7, p<.05; peripheral: F(1,19)=4.6, p<.05).

To determine how this negativity was modulated across the four sentence types, we carried out pair-wise comparisons in the same 6-electrode posterior (parietal-occipital) and 6-electrode anterior (central-frontal) regions as in Experiment 1. In the posterior region, the pattern of effects was similar to that seen in Experiment 1, but modulation was generally weaker. Sentence-final words of the even-so incoherent scenarios again appeared to evoke the largest negativity. This was significantly larger than the negativity produced by sentence-final words in the plain coherent scenarios (F(1,19)>12, p<.01), and marginally larger than the sentence-final negativities of the plain incoherent and even-so coherent scenarios (Fs<4, ps<.08). These latter conditions each produced negativities that were marginally larger than in the plain coherent scenarios (Fs3, ps.1), but that did not differ from one another (F(1,19)<1, p>.9). In the anterior 6-electrode region, the pattern was similar to that seen in Experiment 1: the final words of even-so incoherent, even-so coherent and plain incoherent scenarios each produced significantly larger sustained negativities than the final words of the plain coherent scenarios (all Fs>6, ps<.05), and the amplitude of these negativities did not differ from one another (all Fs<2, ps>.2).

Discussion

In this experiment, a different set of participants read the same stimuli as in Experiment 1, this time answering intermittent comprehension questions about the scenarios, rather than explicitly rating their causal coherence. On these intermittent trials, comprehension questions following the plain coherent scenarios were answered more accurately than those following the even-so coherent scenarios, but questions following the plain incoherent scenarios were answered less accurately than those following the even-so incoherent scenarios. This mirrors the pattern of offline behavioral judgments seen in Experiment 1 (see Discussion of Experiment 1 for a possible interpretation). We also saw some similarities, as well as some differences, between the two experiments in the pattern of ERP results.

The plain scenarios—In the plain scenarios, we saw no N400 effect of coherence at all. This contrasts with the findings of an N400 coherence effect in two previous experiments that examined processing of LSA-matched plain scenarios using a coherence judgment task: Experiment 1 and Kuperberg et al. (2011). We suggest that, in the absence of any explicit requirement to judge discourse coherence, comprehenders did not construct a deep representation of context in time to access stored information about likely upcoming event relationships and predict upcoming events and semantic features before the semantic features of the incoming words became available (see Kuperberg et al., 2011 and Paczynski & Kuperberg, 2012 for discussion). This is not to say that context wasn't used at all; however, it served mainly to activate more general schema-based semantic relationships rather than specific knowledge about events. As LSA—a measure of these schema-based semantic relatedness between the critical word and its context—was matched between the coherent and incoherent plain scenarios, this meant that there was no difference in N400 amplitude evoked by critical words in the plain coherent and incoherent scenarios (see Kuperberg et al. 2003, 2006, 2007; Hoeks et al. 2004; Kolk et al. 2003; Nieuwland & Van Berkum, 2005; Paczynski & Kuperberg, 2012, for other examples of the N400 not patterning with overall coherence or plausibility). Of note, however, as in Experiment 1, we did see a P600 effect on the critical words in the incoherent versus coherent plain scenarios.[6]

The even-so scenarios—Unlike in the plain scenarios, we did see a reduction of the N400 to coherent versus incoherent critical words in the even-so scenarios. In other words, despite the coherent and incoherent scenarios also being matched on lexical schema-based relationships, and there being no explicit requirement to judge coherence, comprehenders *were* able draw upon the alternative world model established under *even so* and use this to predict upcoming real-world unexpected events and facilitate the semantic processing of incoming words.

Unlike in Experiment 1, we did not see an enhanced P600 effect to the incoherent even-so versus incoherent plain critical words. We suggest that this was because, with a comprehension task, comprehenders were less likely to commit, with near certainty, to one specific type of event continuation under *even so*, leading to less conflict when the even-so incoherent critical words came to be integrated. There was, however, a near-significant late sustained anterior negativity effect for this contrast.

Sentence-final words: Finally, on sentence-final words, we observed effects of both Even-so and Coherence. At posterior sites, these effects again seemed to be additive, with the largest negativity produced by sentence-final words in the even-so incoherent scenarios. This suggests that additional wrap-up costs were incurred as participants drew upon a different likelihood scale (the alternative world established under *even so*), rather than real-world knowledge, to evaluate final discourse coherence, even when participants did not explicitly judge discourse coherence.

General Discussion

In two experiments, we asked when and how comprehenders use the concessive connective, *even so* during online discourse comprehension. Our findings on the N400 were clear: in both experiments, the N400 was smaller to critical words in the even-so coherent than the even-so incoherent scenarios. Moreover, the N400 was also smaller to critical words in the even-so coherent than the plain coherent scenarios. To our knowledge, this is the first study to show that a concessive connective can lead both to a reversal and an enhancement of online semantic expectations during discourse comprehension. Here we return to the four sets of questions outlined at the end of the General Introduction. We will then consider the more general implications of our findings for understanding the neurocognitive mechanisms engaged in online discourse comprehension, as well as for understanding the functional significance of the N400, and subsequent ERP components.

1. *Even so* reverses and enhances semantic predictions during online discourse comprehension

Our first question was whether the reversed set of likelihood relations set up under the scalar reversal function of *even so* would lead to *reversed* semantic expectations, thereby facilitating the initial stages of accessing or retrieving the semantic features of congruous incoming words. Our findings clearly indicate that they could: in both experiments, the N400 in the even-so scenarios was smaller to critical words that were coherent than those that were incoherent in relation to their context, even though the message conveyed in the coherent even-so scenarios mismatched long-term real-world knowledge.

These findings are consistent with previous studies of fictional scenarios and counterfactuals, showing that comprehenders can use alternative world models set up by a discourse context to modulate (Hald, Steenbeek-Planting, and Hagoort 2007) and even fully reverse expectations based on real-world knowledge, so long as these contexts are pragmatically constraining (cf. Fergurson et al., 2008, Experiment 2, and Nieuwland et al., 2012, 2013). What distinguishes the present findings from this previous work is that both the setup of the alternative world model and the pragmatic discourse constraint were determined by a simple, but yet semantically rich, concessive connective.

Indeed, *even so* seemed to have set up *stronger* expectations that those generated in the plain scenarios: in both experiments, the N400 was smaller to critical words in the coherent even-so than in the coherent plain scenarios (see Murray, 1994, for consistent behavioral results). In the plain scenarios, where there was no pragmatic cue to indicate what type of upcoming event they might encounter, comprehenders are likely to have entertained multiple possibilities, each with a relatively low probability. We argue that, under *even so*, they are more likely to have predicted a specific type of event structure (one that was causally inconsistent with real-world knowledge) with fairly high probability. This constraining function of *even so*, in turn, led to a higher probability of predicting the semantic features associated with incoming words in the even-so coherent scenarios and more semantic facilitation when these features were accessed (see Lau, Holcomb and Kuperberg, 2013, for evidence that top-down semantic prediction can lead to a reduced N400 to expected words, even in single word contexts). This, of course, raises the question of whether this type of enhanced expectation effect is unique to concessive connectives, or whether it is also induced by other discourse connectives like “and so” or “because”, which do not reverse expectations but also specify a specific relationship between events and states. We are currently carrying follow-up studies to address this question directly.

2. Costs of disconfirmed event predictions under *even so*

Further evidence that *even so* led comprehenders to predict a particular event type with higher probability than in the plain scenarios comes from the prolonged processing costs (activity past the N400 time window) observed when such predictions were disconfirmed by the input. Of note, we saw two such effects — a late posteriorly-distributed P600 effect (seen just in Experiment 1) and a late anteriorly-distributed negativity effect (significant in Experiment 1 and near-significant in Experiment 2).

We suggest that the P600 was produced on trials in which comprehenders committed with high certainty (near 100% probability), to a *real-world inconsistent* event. When this highly certain prediction was violated, the resulting conflict between the predicted real-world inconsistent and bottom-up real-world consistent event structure triggered prolonged attempts to integrate the incoming word to construct a new situation model. The interpretation of the late anterior negativity effect is less clear. However, one possibility is that it was produced on trials in which comprehenders considered both the possibility that they would encounter a *real-world inconsistent* upcoming event structure (with relatively high, but not near-certain probability) and a real-world consistent event structure (with lower probability), and that it reflected the cost of selecting a (less probable) real-world consistent event structure and suppressing the more probable real-world inconsistent event structure, as the critical word was integrated. Similar late negativity effects have been associated with selecting relatively low probability specific events (Wlotko & Federmeier, 2012) as well as selecting ‘alternative’ or non-canonical event structures, such as in aspectual coercion (Bott, 2010; Paczynski, Jackendoff & Kuperberg, 2014), aspectual shift (Baggio, van Lambalgen, and Hagoort, 2008), light verb constructions (Wittenberg, Paczynski, Wiese, Jackendoff, & Kuperberg, 2014), ambiguous noun-verb homographs (Lee and Federmeier, 2006, 2009), and non-literal language or jokes (e.g. Coulson and Van Petten, 2007; Coulson and Kutas, 2001).

3. The assessment of overall coherence against the ‘alternative world’ established under *even so* leads to costs at sentence-final wrap-up

Our third question was whether the establishment of an alternative world model under *even so*, would lead to delayed processing costs. We found that it did. These costs manifest as a prolonged negativity effect of *even so* on sentence-final words. This sentence-final negativity effect was qualitatively similar to the wrap-up effect of coherence seen in both the even-so and the plain scenarios, and, at posterior sites, the two effects were additive. This suggests that some wrap-up costs were incurred when the final discourse model was evaluated against a set of non-default likelihood event relations established under *even so*, and that even more wrap-up costs were incurred when this overall discourse model was found to be incoherent.

In both experiments, the additive effects of Even-so and Coherence were evident primarily at posterior sites. At more frontal sites, the final words of the coherent even-so, incoherent plain, and incoherent even-so scenarios all evoked a negativity effect of the same magnitude, which was larger than that seen in the coherent plain scenarios. We suggest that this more frontal component of the sentence-final negativity reflected a general engagement of working memory resources that maintained multiple representations, rather than actually evaluating them against each other for coherence. If this interpretation is correct, then it would suggest that there would be similar working memory costs associated with maintaining both the real-world and alternative world model under *even so* much earlier in the sentence (see King & Kutas, 1995; Kluender and Kutas, M.1993; Nieuwland & van Berkum, 2008; van Berkum et al., 2003; Münte, Schiltz, and Kutas, 1998 for evidence of early onset and sustained negativity effects associated with maintaining multiple representations within working memory over multiple words). In the present study, we were unable to address this question because corresponding words in the even-so scenarios and the plain scenarios were confounded by baseline differences (the even-so scenarios started with a connective, whereas the plain scenarios did not). It will therefore be important to address this in future studies. It will also be important to determine whether there are costs associated with *even so* itself—the point at which the presupposition of a reversed likelihood scale was calculated—and to determine how such costs might manifest in the ERP waveform.

4. Task can impact multiple stages of online discourse comprehension

Finally, our findings show that task made an important difference to how both the plain and and even-so scenarios were processed, at multiple stages of comprehension.

In the plain scenarios, task influenced the type of stored real-world knowledge that was used to influence semantically processing of incoming words. With a requirement to explicitly judge discourse coherence (Experiment 1), comprehenders were able to construct a deep situation-level representation of context and use it to access their stored knowledge of real-world event relationships to predict upcoming events and semantic features, thereby facilitating semantic processing of incoming coherent words. With no such requirement, however, comprehenders simply drew upon their more general stored knowledge about unstructured relationships between words/concepts within a particular schema. Because these schema-based relationships were matched (through LSA) between conditions, we saw no N400 amplitude difference between critical words of the plain coherent and incoherent scenarios (Experiment 2; also see Experiment 3 in footnote 6, in which we also saw no hint of a coherence N400 effect on the plain scenarios when we pooled together results on the two plain scenarios from a larger number of participants (n=36)).

In the even-so scenarios, task made less of a qualitative difference to N400 modulation. This is because the pragmatic communicative constraint of *even so* led comprehenders to narrow down their expectations and anticipate a *real-world inconsistent* event structure, regardless of whether there was an explicit task requirement to focus on discourse relationships. On the other hand, task did make some difference to the neural mechanisms that comprehenders engaged when these event structure predictions were violated: the enhancement of the P600 on the even-so (versus plain) incoherent critical words was only seen when participants made active judgments (Experiment 1). We suggest that this is because the requirement to judge coherence, together with the discourse pragmatic function of *even so*, led comprehenders to commit with near certainty to a specific type of event structure, ahead of integrating the critical word, at least on some trials. As discussed above, this, in turn, led to more conflict as the real-world consistent critical word was integrated, which triggered additional bottom-up attempts to integrate the incoming word, leading to increased P600. This is consistent with frameworks holding that task is one of several factors that can lower the threshold for triggering a P600 effect to semantically incoherent words that conflict with alternative analyses (e.g. Kuperberg, 2007). It is also consistent with previous work suggesting that the P600 is closely linked to comprehenders' detection of incoherence (e.g. Sanford et al. 2011).

General implications and functional relevance of the N400 and P600 in discourse comprehension

Beyond speaking to the specific role of concessive connectives in discourse comprehension, our findings highlight some general points about the functional significance of the N400 and later ERP components.

First, our findings underline the fact that, while N400 modulation often patterns with offline ratings of semantic discourse coherence, this is by no means always the case (see Paczynski & Kuperberg, 2011, 2012 and references therein for many other examples). In the present study, we saw no N400 effect of coherence at all in the plain scenarios in Experiment 2. Furthermore, in both experiments, the N400 coherence effect was larger in the even-so than the plain scenarios, despite the difference in coherence ratings being larger in the plain scenarios.

The reason why the N400 does not always pattern with offline coherence or plausibility ratings is because it is *not* a direct reflection of a process of evaluating a representation of discourse meaning against stored knowledge in order to asses 'plausibility' or 'coherence' during online comprehension (see also Kuperberg, Choi, Cohn, Paczynski and Jackendoff, 2010; Paczynski and Kuperberg, 2011, 2012 for further discussion). Rather, is best characterized as reflecting changes in the activity within semantic memory induced by incoming words (Kutas & Federmeier, 2011), or the implicit semantic prediction error in shifting from a prior to a posterior distribution of semantic features on the basis of new input (Rabovsky & McRae, 2014). Whether or not the N400 will pattern with coherence or plausibility will therefore depend on the prior distribution of semantic features, which as we have argued, will depend on the probability of predicting a particular event structure just before a target's semantic features become available. The probability/strength of such prediction at the event structure layer will depend on many factors (see Kuperberg et al. 2011, and references therein for discussion). In this study, we have highlighted the pragmatic constraining function of certain words (here, *even so)* as well as task demands.

This study also underlines the fact that the N400 is sensitive to implicit predictions of *semantic* features, but that this does not always or necessarily equate to *lexical* pre-activation or prediction (the co-activation of semantic features together with phonological or orthographic form and sometimes syntactic features). Of course, semantic and lexical expectations often go hand-in-hand: we have known for some time that lexical predictability/constraint, as operationalized by cloze probability, is an important determinant of N400 amplitude (e.g. Kutas and Hillyard 1984; Federmeier et al. 2007). However, the two can be dissociated. For example, a context can constrain for a particular group of semantic features without necessarily constraining for a specific lexical item (e.g. Federmeier & Kutas, 1999; Paczynski & Kuperberg, 2011; Paczynski & Kuperberg, 2012), and it is possible for comprehenders to predict a particular event structure, without necessarily predicting a specific lexical item (see Kuperberg, 2013, for discussion). Moreover, as discussed in Experiment 1, offline cloze ratings may not necessarily reflect fast word-by-word lexical prediction.

Finally this study also adds to the growing evidence that activity within the N400 time window does not always reflect the final stages of semantically integrating an incoming word into its context. We have argued that the late posterior positivity/P600 ERP component is triggered by the *conflict* between a strong (near-certain) prediction and bottom-up input that violates this prediction, and that it reflects a switch to a new (generative) model (representation relationships between events) as the incoming word is integrated into the context[7]. We have also suggested that the late negativity effect is evoked when we predict more than one event or event structure, and the less probable of these event structures is selected as the incoming word is integrated. If these interpretations are correct, then this would imply that different individuals, at different times, might use exactly the same context to predict event structures with different strengths/certainties, leading them to mount either a P600 or a late anterior negativity response if these predictions are violated. It would also imply that these responses are likely to vary *within* individuals, with a given comprehender engaging quite different mechanisms depending on the degree to which their wider linguistic (and non-linguistic) environment encourages strong or weak prediction (see Kuperberg, 2013, 2014 for discussion; and Kleinschmidt and Jaeger (In press) for a formalization in the

[7] In some of our previous discussions of the P600 component, we have described these type of near certainty event predictions as arising from a 'semantic memory based analysis', i.e. stemming from the interaction between context and real-world knowledge stored in long-term semantic memory. By definition, incoming words that violate strong semantic memory-based predictions will, when fully integrated, yield an event representation that is highly implausible/impossible (see Paczynski & Kuperberg, 2012 for recent discussion). What the present study shows is that a P600 can also be produced when event predictions are based on an 'alternative world model' and full integration of a target word outputs a discourse representation that is highly *incoherent*, but not necessarily 'implausible' with respect to real-world knowledge.

domain of speech perception). It will be important for future studies to examine this type of inter- and intra-individual variation more closely.

Conclusion

To conclude, we have shown that the concessive connective, *even so*, leads to a reversal of expectations such that comprehenders can use discourse-internal information to override the effects of stored long-term world knowledge. Moreover, comprehenders are able to make maximal use of the pragmatic discourse constraint of *even so* to enhance their expectations about the upcoming events, leading to facilitated semantic processing of incoming words, even in contexts that are not highly lexically constraining, as indexed by cloze probability. This benefit of *even so*, however, did not come for free: we also observed global costs of constructing and maintaining an alternative world under *even so*, manifesting on the final word of the scenarios. Together, our results show that, although stored knowledge provides an important background for language comprehension, we are nonetheless able to use concessive connectives to construct, constrained and integrate new information into an abstract mental model amazingly quickly, even when this model mismatches our real-world experience.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Abigail Swain, Eric Fields, Kristina Fanucci for their assistance in data collection, and Ellen Lau and Einat Shetreet for their helpful comments. This work was funded by the National Institute of Mental Health (R01MH071635 to G. R. K.).

References

Altmann GTM, Steedman M. Interaction with context during human sentence processing. Cognition. 1988; 30:191–238. [PubMed: 3215002]

Baggio G, Van Lambalgen M, Hagoort P. Computing and recomputing discourse models: An ERP study. Journal of Memory and Language. 2008; 59:36–53.

Becker CA. Semantic context effects in visual word recognition: An analysis of semantic strategies. Memory and Cognition. 1980; 8:493–512. [PubMed: 7219170]

Becker, CA. What do we really know about semantic context effects during reading?. In: Besner, D.; Waller, TG.; McKinnon, EM., editors. Reading research: Advances in theory and practice. Vol. 5. Academic Press; Toronto: 1985. p. 125-169.

Blakemore, D. Relevance and Linguistic Meaning. Cambridge University Press; Cambridge: 2002.

Bott, O. The processing of events. John Benjamins Publishing; 2010.

Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences. 2013; 36(3):181–204. [PubMed: 23663408]

Corley M, MacGregor LJ, Donaldson DI. It's the way that you, er, say it: Hesitations in speech affect language comprehension. Cognition. 2007; 105:658–668. [PubMed: 17173887]

Coulson S, Van Petten C. A special role for the right hemisphere in metaphor comprehension: An ERP Study. Brain Research. 2007; 1146:128–145. [PubMed: 17433892]

Coulson S, Kutas M. Getting it: Human event-related brain response to jokes in good and poor comprehenders. Neuroscience Letters. 2001; 316:71–74. [PubMed: 11742718]

De Grauwe S, Swain A, Holcomb PJ, Ditman T, Kuperberg GR. Electrophysiological insights into the processing of nominal metaphors. Neuropsychologia. 2010; 48:1965–1984. [PubMed: 20307557]

DeLong KA, Urbach TP, Groppe DM, Kutas M. Overlapping dual ERP responses to low cloze probability sentence continuations. Psychophysiology. 2011; 48(9):1203–1207. [PubMed: 21457275]

Ditman T, Holcomb P, Kuperberg G. An investigation of concurrent ERP and self paced reading methodologies. Psychophysiology. 2007; 44:927–935. [PubMed: 17850242]

Farmer TA, Brown M, Tanenhaus MK. Prediction, explanation, and the role of generative models in language processing [Commentary article]. Behavioral and Brain Sciences. 2013; 36:211–212. [PubMed: 23663410]

Federmeier KD, Kutas M. A roase by any other name: long-term memory structure and sentence processing. Journal of Memory and Language. 1999; 41:469–495. 1999.

Federmeier KD, Wlotko EW, De Ochoa-Dewald E, Kutas M. Multiple effects of sentential constraint on word processing. Brain Research. 2007; 1146:75–84. [PubMed: 16901469]

Feldman NH, Griffiths TL, Morgan JL. The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. Psychological Review. 2009; 116(4):752–782. [PubMed: 19839683]

Ferguson HJ, Sanford AJ, Leuthold H. Eye-movements and ERPs reveal the timecourse of processing negation and remitting counterfactual worlds. Brain Research. 2008; 1236:113–125. [PubMed: 18722356]

Ferguson HJ, Sanford AJ. Anomalies in real and counterfactual worlds: An eyemovement investigation. Journal of Memory and Language. 2008; 58:609–626.

Ferguson HJ, Breheny R. Eye movements reveal the time-course of anticipating behaviour based on complex, conflicting desires. Cognition. 2011; 119(2):179–196. [PubMed: 21353214]

Fillmore, C. Frame semantics. In: Geeraerts, D., editor. Cognitive Linguistics. Basic readings. Mouton de Gruyter; Berlin & New York: 2006. p. 373-400.Linguistic society of Korea (ed.). , editor. Originally published in *Linguistics in the morning calm*. Hanshin Publishing Company; Seoul: 1982. p. 111-137.

Fine AB, Jaeger TF, Farmer TA, Qian T. Rapid expectation adaptation during syntactic comprehension. PLoS ONE. 2013 DOI: 10.1371/journal.pone.0077661.

Forster KI. Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. Quarterly Journal of Experimental Psychology. 1981; 33:465–495.

Friston K. A theory of cortical responses. Philosophical Transactions of the Royal Society B: Biological Sciences. 2005; 360(1456):815–836.

Grice, HP. Logic and conversation. In: Cole, P.; Morgan, J., editors. Syntax and Semantics. Vol. 3. Academic Press; New York: 1975. p. 41-58.

Griffiths, TL.; Kemp, C.; Tenenbaum, JB. Bayesian models of cognition. In: Sun, Ron, editor. The Cambridge handbook of computational cognitive modeling. Cambridge University; 2008.

Hagoort P, Brown C. ERP effects of listening to speech: Semantic ERP effects. Neuropsychologia. 2000; 38:1518–1530. [PubMed: 10906377]

Hagoort P. Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. Journal of Cognitive Neuroscience. 2003; 15:883–899. [PubMed: 14511541]

Hagoort, P.; Brown, C.; Groothusen, J. The syntactic positive shift (SPS) as an ERP measure of syntactic processing. In: Garnsey, SM., editor. Language and Cognitive Processes. Special Issue: Event-Related Brain Potentials in the Study of Language. Vol. 8. Lawrence Erlbaum Associates; Hove: 1993. p. 439-483.

Hald L, Steenbeek-Planting E, Hagoort P. The interaction of discourse context and world knowledge in online sentence comprehension. Evidence from the N400. Brain Research. 2007; 1146:210–218. [PubMed: 17433893]

Hale, J. A probabilistic Earley parser as a psycholinguistic model; Paper presented at the Proceedings of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL ’01); 2001;

Hoeks JC, Stowe LA, Doedens G. Seeing words in context: The interaction of lexical and sentence level information during reading. *Brain Research*: *Cognitive Brain Research*. 2004; 19:59–73. [PubMed: 14972359]

Jacobs RA, Kruschke JK. Bayesian learning theory applied to human cognition. Wiley Interdisciplinary Reviews: Cognitive Science. 2011; 2:8–21.

Karttunen, L.; Peters, S. Conventional implicature. In: Oh, CK.; Dinneen, DA., editors. Syntax and Semantics 11: Presuppositions. Academic Press; New York: 1979. p. 1-55.

Keenan JM, Baillet SD, Brown P. The effect of causal cohesion on comprehension and memory. Journal of Verbal Learning and Verbal Behavior. 1984; 23:115–126.

King JW, Kutas M. Who did what and when? Using word- and clause-related ERPs to monitor working memory usage in reading. Journal of Cognitive Neuroscience. 1995; 7:378–397.

Kleinschmidt D, Jaeger F. Robust speech perception: Recognizing the familiar, generalizing to the similar, and adapting to the novel. Psychological Review. In press.

Kluender R, Kutas M. Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. Journal of Cognitive Neuroscience. 1993; 5:196–214. [PubMed: 23972154]

Kolk HH, Chwilla DJ, van Herten M, Oor PJ. Structure and limited capacity in verbal working memory: a study with event-related potentials. Brain and Language. 2003; 85:1–36. [PubMed: 12681346]

Kuperberg GR, Holcomb PJ, Sitnikova T, Greve D, Dale AM, Caplan D. Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. Journal of Cognitive Neuroscience. 2003; 15:272–293. [PubMed: 12676064]

Kuperberg G, Caplan D, Sitnikova T, Eddy M, Holcomb P. Neural correlates of processing syntactic, semantic and thematic relationships in sentences. Lanuage and Cognitive Processes. 2006; 21:489–530.

Kuperberg GR. Neural mechanisms of language comprehension: Challenges to syntax. Brain Research. 2007; 1146(Special Issue):23–49. [PubMed: 17400197]

Kuperberg GR, Paczynski M, Ditman T. Establishing causal coherence across sentences: an ERP study. Journal of Cognitive Neuroscience. 2011; 23(5):1230–1246. [PubMed: 20175676]

Kuperberg GR, Choi A, Cohn N, Paczynski M, Jackendoff R. Electrophysiological correlates of Complement Coercion. Journal of Cognitive Neuroscience. 2010; 22(12):2685–2701. [PubMed: 19702471]

Kuperberg, GR. The pro-active comprehender: What event-related potentials tell us about the dynamics of reading comprehension. In: Miller, B.; Cutting, L.; McCardle, P., editors. Unraveling the Behavioral, Neurobiological, and Genetic Components of Reading Comprehension. Paul Brookes Publishing; Baltimore: 2013.

Kuperberg G. What event-related potentials tell us about predictive coding in language comprehension: A commentary on Rabovsky & McRae. Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. 2014 manuscript under review.

Kutas M, Hillyard SA. Brain potentials during reading reflect word expectancy and semantic association. Nature. 1984; 307:161–163. [PubMed: 6690995]

Kutas M, Hillyard SA. Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity. Science. 1980; 207:203–205. [PubMed: 7350657]

Kutas M, Federmeier KD. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu. Rev. Psychol. 2011; 62:621–647. [PubMed: 20809790]

Lagerwerf, L. Causal Connectives Have Presuppositions. Catholic Univ. of Brabant, Holland Academic Graphics; The Hague, The Netherlands: 1998. Ph.D. thesis

Lakoff, G. On generative semantics. In: Steinberg, D.; Jakobovits, L., editors. Semantics: An Interdisciplinary Reader. Cambridge University Press; 1971.

Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review. 1997; 104:211–240.

Landauer TK, Foltz PW, Dumais ST. Introduction to Latent Semantic Analysis. Discourse Processes. 1998; 25:259–284.

Lau EF, Phillips C, Poeppel D. A cortical network for semantics: (de)constructing the N400. Nature Reviews Neuroscience. 2008; 9(12):920–933.

Lau E, Holcomb PJ, Kuperberg GR. Dissociating N400 effects of prediction from association in single word contexts. Journal of Cognitive Neuroscience. 2013; 25(3):484–502. [PubMed: 23163410]

Lee C, Federmeier KD. Wave-ering: An ERP study of syntactic and semantic context effects on ambiguity resolution for noun/verb homographs. Journal of Memory and Language. 2009; 61:538–555. [PubMed: 20161361]

Levy R. Expectation-based syntactic comprehension. Cognition. 2008; 106(3):1126–1177. [PubMed: 17662975]

MacDonald MC, Pearlmutter NJ, Seidenberg MS. The lexical nature of syntactic ambiguity resolution. Psychological Review. 1994; 101:676–703. [PubMed: 7984711]

Marslen-Wilson W, Brown C, Tyler L. Lexical representations in spoken language comprehension. Language and Cognitive Processes. 1988; 3:1–16.

McRae K, Ferretti TR, Amyote L. Thematic Roles and verb-specific concepts, Language and Cognitive Processes. 1997; 12(2-3):137–176.

Münte TF, Schiltz K, Kutas M. When temporal terms belie conceptual order: an electrophysiological analysis. Nature. 1998; 3(395):71–73. [PubMed: 9738499]

Murray, JD. Logical connectives and local coherence. In: Lorch, RF.; O'Brien, E1., editors. Sources of cohesion in text comprehension. Erlbaum; Hillsdale, NJ: 1994. p. 107-125.

Murray JD. Connectives and narrative text: the role of continuity. Memory and cognition. 1997; 25(2): 227–236. [PubMed: 9099073]

Nieuwland MS. "If a lion could speak …": Online sensitivity to propositional truth value of unrealistic counterfactual sentences. Journal of Memory and Language. 2013; 68(1):54–67.

Nieuwland MS, van Berkum JJA. When peanuts fall in love: N400 evidence for the power of discourse. Journal of Cognitive Neuroscience. 2006; 18(7):1098–1111. [PubMed: 16839284]

Nieuwland MS, Van Berkum JJA. Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary change deafness in discourse comprehension. Cognitive Brain Research. 2005; 24(3):691–701. [PubMed: 15894468]

Nieuwland MS, van Berkum. The neurocognition of referential ambiguity in language comprehension. Language and Linguistics Compass. 2008:603–630.

Nieuwland MS, Ditman T, Kuperberg GR. On the incrementality of pragmatic processing: An ERP investigation of underinformative scalar sentences. Journal of Memory and Language. 2010; 63:324–346. [PubMed: 20936088]

Nieuwland MS, Martin AE. If the real world were irrelevant, so to speak: The role of propositional truth-value in counterfactual sentence comprehension. Cognition. 2012; 122:102–109. [PubMed: 21962826]

Norris D. The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. Psychological Review. 2006; 113(2):327–357. [PubMed: 16637764]

Norris D, McQueen JM. A Bayesian model of continuous speech recognition. Psychological Review. 2008; 115(2):357–395. [PubMed: 18426294]

Noveck, I.; Spotorno, N. Narrowing. In: Goldstein, Laurence, editor. Brevity. Oxford University Press; Oxford: 2013.

Oldfield RC. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia. 1971; 9:97–113. [PubMed: 5146491]

Otten M, van Berkum JJA. What makes a discourse constraining? A comparison between the effects of discourse message and priming on the N400. Brain Research. 2007; 1153:166–177. [PubMed: 17466281]

Osterhout L, Holcomb PJ. Event-related potentials elicited by syntactic anomaly. Journal of Memory and Language. 1992; 31:785–806.

Osterhout L, Holcomb PJ. Event-related potentials and syntactic anomaly: Evidence of anomaly detection during the perception of continuous speech. Language and Cognitive Processes. 1993; 8:413–437.

Paczynski M, Kuperberg GR. Electrophysiological evidence for use of the animacy hierarchy, but not thematic role assignment, during verb argument processing. Language and Cognitive Processes. 2011; 26(9):1402–1456. [PubMed: 22199415]

Paczynski M, Kuperberg GR. Multiple influences of semantic memory on sentence processing: distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. Journal of Memory and Language. 2012; 67(4):426–448. [PubMed: 23284226]

Paczynski M, Jackendoff R, Kuperberg GR. When events change their nature. Journal of Cognitive Neuroscience. in press.

Rabovsky M, McRae K. Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. Cognition. 2014; 132(1):68–89. [PubMed: 24762924]

Rao RPN, Ballard D. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience. 1999; 2(1):79–87.

Sanford AJ, Leuthold H, Bohan J, Sanford AJS. Anomalies at the borderline of awareness: an ERP study. Journal of Cognitive Neuroscience. 2011; 23:514–523. [PubMed: 19925201]

Schank, RC.; Abelson, RP. Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Erlbaum; Hillsdale, NJ: 1977.

Singer M, Halldorson M. Constructing and validating motive bridging inferences. Cognitive Psychology. 1996; 30:1–38. [PubMed: 8660781]

Sitnikova, T.; Holcomb, P.; Kuperberg, GR. Neurocognitive mechanisms of human comprehension. In: Shipley, TF.; Zacks, JM., editors. Understanding Events: How Humans See, Represent, and Act on Events. Oxford University Press; 2008. p. 639-683.

Stewart AJ, Pickering MJ, Sanford AJ. The time course of the influence of implicit causality information: focusing versus integration accounts. Journal of Memory and Language. 2000; 42(3):423–443.

St. George M, Mannes S, Hoffman JE. Individual differences in inference generation: An ERP analysis. Journal of Cognitive Neuroscience. 1997; 9:776–787. [PubMed: 23964599]

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information during spoken language comprehension. Science. 1995; 268:1632–1634. [PubMed: 7777863]

Traxler MJ, Bybee MD, Pickering MJ. Influence of connectives on language comprehension: Eye-tracking evidence for incremental interpretation. Quarterly Journal of Experimental Psychology: Human Experimental Psychology. 1997; 50(3):481–497.

van Dijk, TA.; Kintsch, W. Strategies of discourse comprehension. Academic Press; New York: 1983.

van Berkum JJA, Brown CM, Hagoort P, Zwitserlood P. Event-related brain potentials reflect discourse-referential ambiguity in spoken language comprehension. Psychophysiology. 2003; 40(2):235–248. [PubMed: 12820864]

van Berkum, JJA. The neuropragmatics of “simple” utterance comprehension: An ERP review. In: Sauerland, U.; Yatsushiro, K., editors. Semantics and pragmatics: From experiment to theory. Palgrave Macmillan; Basingstoke: 2009. p. 276-316.

Van Petten C, Luka BJ. Prediction during language comprehension: Benefits, costs, and ERP components. International Journal of Psychophysiology. 2012; 83:176–190. [PubMed: 22019481]

Van Petten C, Kutas M. Interactions between sentence context and word frequency in event-related brain potentials. Memory and Cognition. 1990; 18(4):380–393. [PubMed: 2381317]

Wacongne C, Labyt E, van Wassenhove V, Bekinschtein T, Naccache L, Dehaene S. Evidence for a hierarchy of predictions and prediction errors in human cortex. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108(51):20754–20759. [PubMed: 22147913]

Wacongne C, Changeux JP, Dehaene S. A neuronal model of predictive coding accounting for the mismatch negativity. Journal of Neuroscience. 2012; 32(11):3665–3678. [PubMed: 22423089]

Warren T, McConnell K. Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. Psychonomic Bulletin & Review. 2007; 14:770–775. [PubMed: 17972747]

Wilson D, Sperber D. Linguistic form and relevance. Lingua. 1993; 90:1–25.

Wittenberg E, Paczynski M, Wiese H, Jackendoff R, Kuperberg G. The difference between “giving a rose” and “giving a kiss”: a sustained anterior negativity to the light verb construction. Journal of Memory and Language. 2014; 73:31–42. [PubMed: 24910498]

Yang CL, Perfetti CA, Schmalhofer F. Event-related potential indicators of text integration across sentence boundaries. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2007; 33:55–89.

Zwaan RA, Radvansky GA. Situation models in language comprehension and memory. Psychological Bulletin. 1998; 123(2):162–185. [PubMed: 9522683]

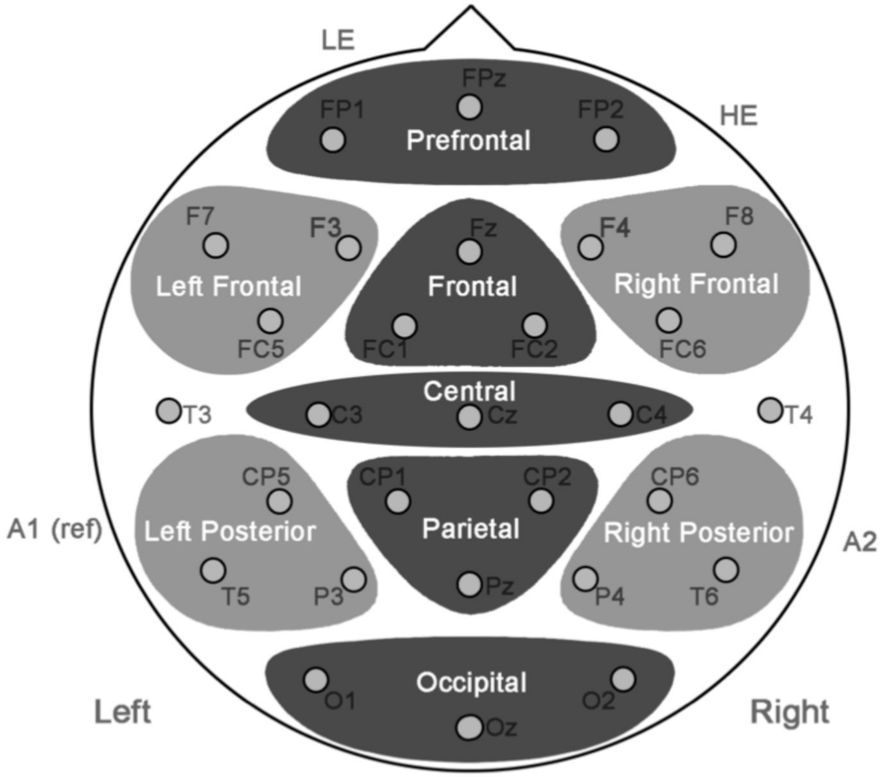

Figure 1.

Electrode montage, showing each 3-electrode region used for analysis. Regions in dark grey were part of the mid-regions omnibus ANOVA and regions in light grey were part of the peripheral regions omnibus ANOVA.

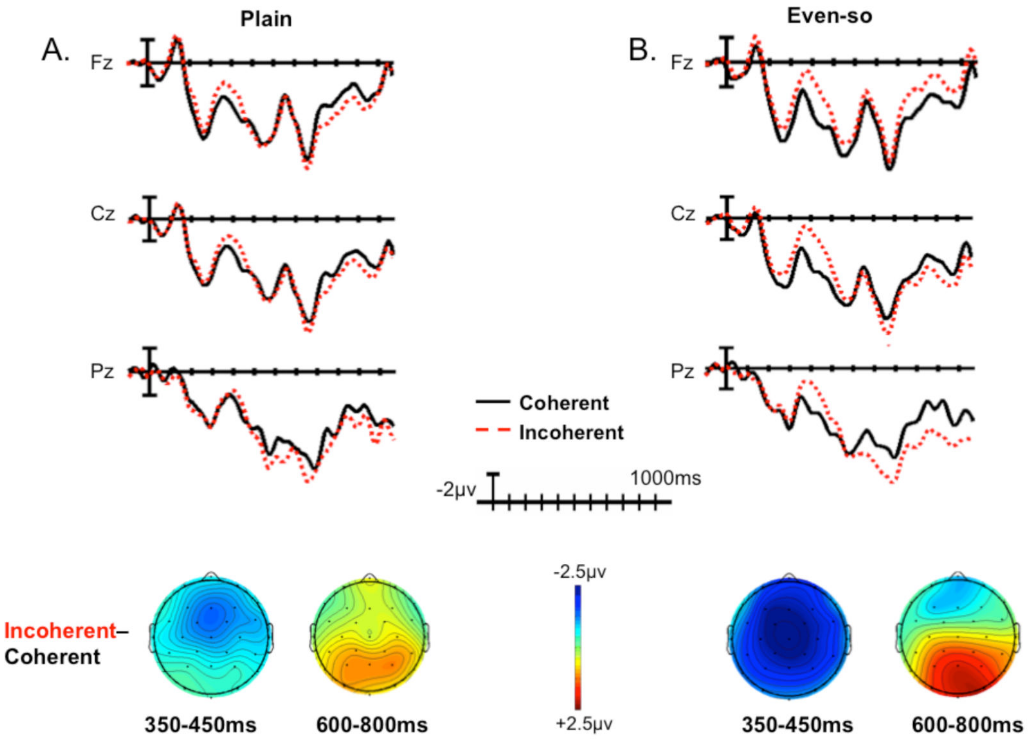

Figure 2.

Grand-averaged waveforms to critical words in Experiment 1 (coherence rating task) showing effects of Coherence at electrodes Fz, Cz and Pz.

Panel A: waveforms to critical words in plain coherent (black solid) and plain incoherent (red dotted) scenarios.

Panel B: waveforms to critical words in even-so coherent (black solid) and even-so incoherent (red dotted) scenarios.

Voltage maps show differences in ERPs between incoherent and coherent critical words (incoherent *minus* coherent) between 350-450ms (N400) and 600-800ms (P600).

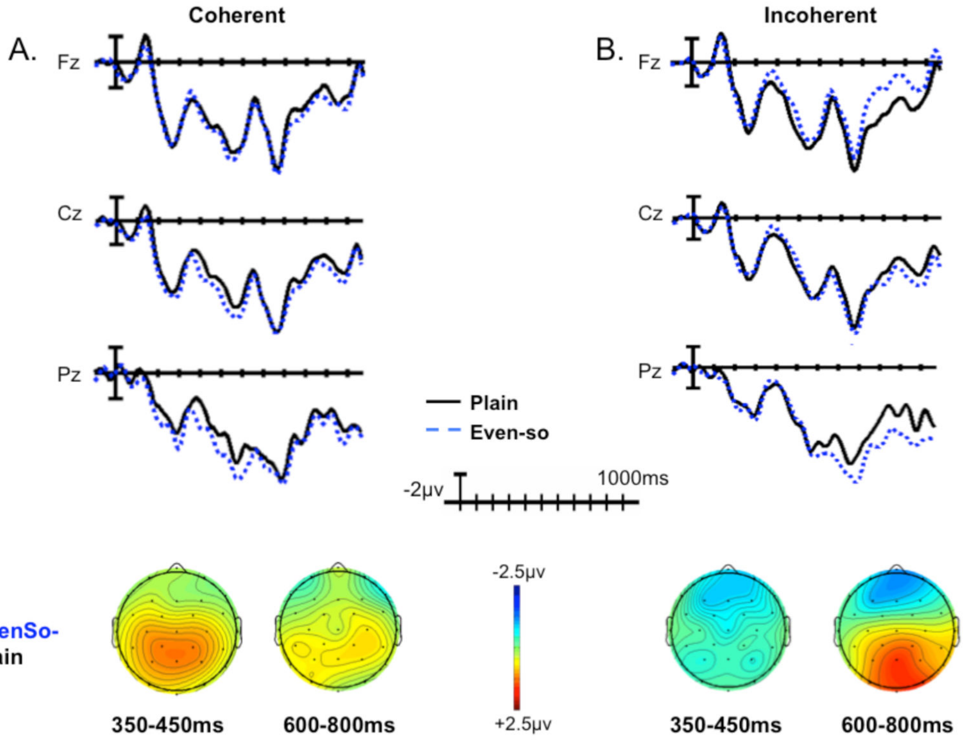

Figure 3.

Grand-averaged waveforms to critical words in Experiment 1 (coherence rating task) showing effects of Even-so at electrodes Fz, Cz and Pz.

Panel A: waveforms to critical words in coherent plain (black solid) and coherent even-so (blue dotted) scenarios.

Panel B: waveforms to critical words in incoherent plain (black solid) and incoherent even-so (blue dotted) scenarios.

Voltage maps show differences in ERPs between even-so and plain critical words (even-so *minus* plain) between 350-450ms and 600-800ms.

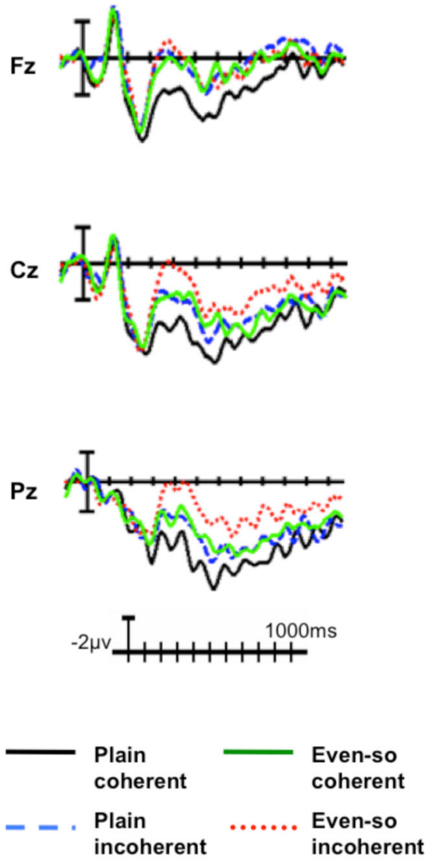

Figure 4.

Grand-averaged waveforms to sentence-final words in Experiment 1 (coherence rating task). ERPs to sentence-final words in all four conditions are shown at electrodes Fz, Cz and Pz. Plain coherent: black solid line; Plain incoherent: blue dashed line; Even-so coherent: green solid line; Even-so incoherent: red dotted line.

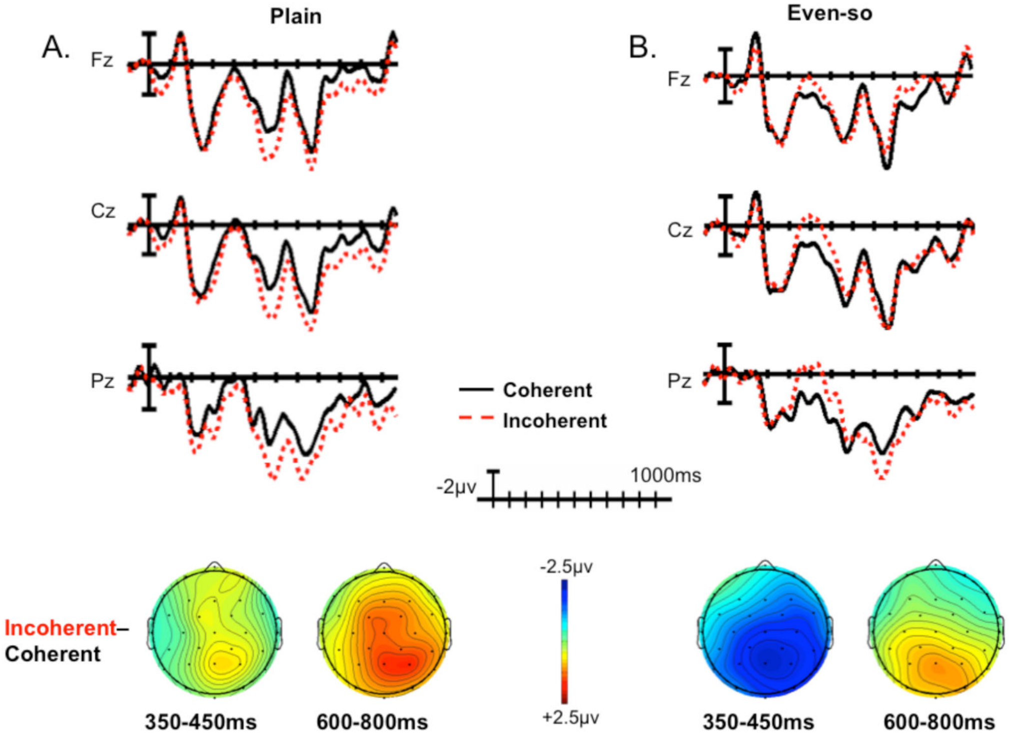

Figure 5.

Grand-averaged waveforms to critical words in Experiment 2 (comprehension task) showing effects of Coherence at electrodes Fz, Cz and Pz.

Panel A: waveforms to critical words in plain coherent (black solid) and plain incoherent (red dotted) scenarios.

Panel B: waveforms to critical words in even-so coherent (black solid) and even-so incoherent (red dotted) scenarios.

Voltage maps show differences in ERPs between incoherent and coherent critical words (incoherent *minus* coherent) between 350-450ms (N400) and 600-800ms (P600).

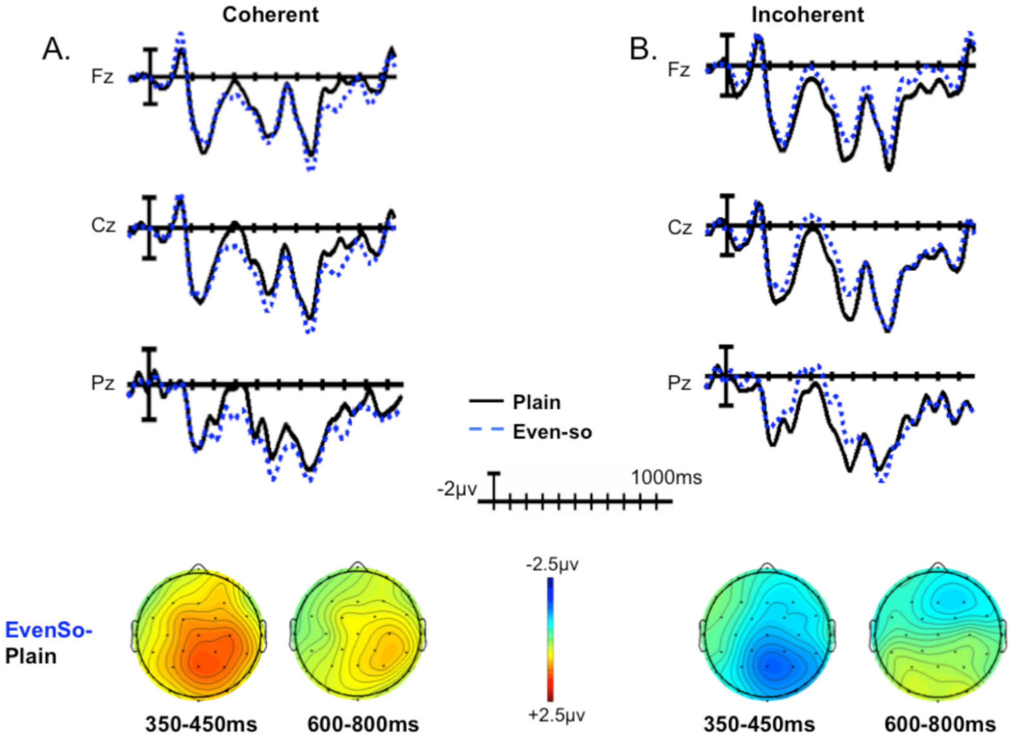

Figure 6.

Grand-averaged waveforms to critical words in Experiment 2 (comprehension task) showing effects of Even-so at electrodes Fz, Cz and Pz.

Panel A: waveforms to critical words in coherent plain (black solid) and coherent even-so (blue dotted) scenarios.

Panel B: waveforms to critical words in incoherent plain (black solid) and incoherent even-so (blue dotted) scenarios.

Voltage maps show differences in ERPs between even-so and plain critical words (even-so *minus* plain) between 350-450ms and 600-800ms.

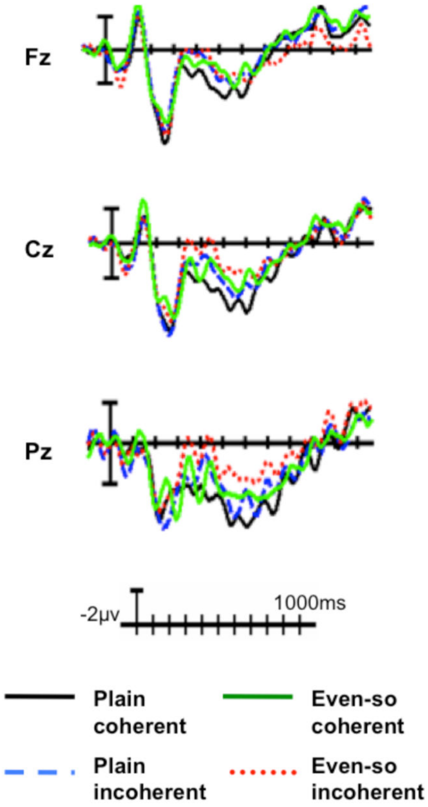

Figure 7.

Grand-averaged waveforms to sentence-final words in Experiment 2 (comprehension task). ERPs to sentence-final words in all four conditions are shown at electrodes Fz, Cz, and Pz. Plain coherent: black solid line; Plain incoherent: blue dashed line; Even-so coherent: green solid line; Even-so incoherent: red dotted line.

Table 1

Example stimuli and characteristics.

Scenario Type (n=45 per condition)	Example	SSV of critical word[‡]	Cloze[*]	Constraint[*]	Coherence ratings[^]
1. Coherent	Elizabeth had a history exam on Monday. She took the test and aced it. She went home and celebrated wildly.	0.179 [0.078]	0.42 [0.32]	0.52 [0.26]	4.8 [0.2]
2. Incoherent	Elizabeth had a history exam on Monday. She took the test and failed it. She went home and celebrated wildly.	0.174 [0.079]	0.03 [0.09]	0.40 [0.24]	1.7 [0.4]
3. Even-so Coherent	Elizabeth had a history exam on Monday. She took the test and failed it. Even so, she went home and celebrated wildly.	0.174 [0.079]	0.31 [0.28]	0.44 [0.25]	3.3. [1.0]
4. Even-so Incoherent	Elizabeth had a history exam on Monday. She took the test and aced it. Even so, she went home and celebrated wildly.	0.179 [0.078]	0.04 [0.11]	0.40 [0.24]	2.4 [1.0]

Means are shown with standard deviations in square parentheses. The critical word in each of the example sentences is underlined (although this was not the case in the experiment itself).

[‡] LSA was used to calculate Semantic Similarity Values (SSVs) between the critical word and its preceding content words.

[*] Cloze probability and constraint are represented as the proportion of total responses from 40 participants.

[^] Coherence ratings, on a 1-5 scale, were collected during the ERP recording session in Experiment 1. 5: very coherent; 1: incoherent.