Reversing Expectations: The Use of Concessive Connectives during Discourse Comprehension

Ming Xiang\textsuperscript{a} and Gina Kuperberg\textsuperscript{b,c}

\textsuperscript{a}. Language and Processing Lab, Linguistics Department, University of Chicago, IL, 60615.

\textsuperscript{b}. NeuroCognition Laboratory, Department of Psychology, Tufts University, Medford, MA, 02155.

\textsuperscript{c}. Department of Psychiatry, Massachusetts General Hospital, Bldg 149, 13th Street, Charlestown, MA, 02129

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
Abstract

In two ERP experiments, we investigated how people comprehend discourse describing events and states that are unexpected in the real world, through the use of the concessive connective, *even so*. Participants read short discourse scenarios like “The storm last night left only a dusting of snow. Even so, the school was open/closed…”, while they actively rated coherence (Experiment 1) or passively read for comprehension (Experiment 2). In both experiments, the N400 to coherent critical words (*closed*) was smaller than to incoherent critical words (*open*), indicating that the scalar-reversing function of *even so* was available immediately to set up *reversed expectations* that overrode stored real-world knowledge. These benefits of *even so*, however, did not come for free. We also observed a later sustained negativity effect on the sentence-final word of the even-so scenarios that may reflect costs of constructing and maintaining an alternative reversed world on which predictions were based.
General Introduction

Successful language comprehension draws heavily upon our experience in the real world. This real-world knowledge, stored in long-term memory, tells us whether what we hear or read is plausible, implausible, true or false. Moreover, during online language comprehension, we continually draw upon this knowledge to facilitate the processing of words within sentences that describe familiar states or events (Marslen-Wilson, Brown and Tyler, 1988; Rayner, Warren, Juhasz, and Liversedge 2004; Warren and McConnell 2007; McRae et al. 1997). We also use such knowledge to establish coherence relationships across multiple sentences in a larger discourse context (Singer et al. 1996; Bloom et al. 1990; Keenan et al. 1984; van Dijk and Kintsch, 1983; Myers et al., 1987).

At the same time, however, language gives us the remarkable ability to construct mental models that do not necessarily conform to our knowledge of the real world. For example, we are able to talk about events that have a very low probability of actually occurring, and we can carry out counterfactual reasoning in which we describe eventualities that are quite different to what actually happened. In such cases, the parser still seems able to construct a mental situation model quickly and seemingly effortlessly. In the current study, we use event-related potentials (ERPs)—an online neural measure with high temporal resolution—to examine when and how the brain accomplishes this task during real-time comprehension.

Concessive connectives

One way in which natural language can describe situations that are unexpected or improbable in the real world, is through the use of concessive connectives. Words such as but, however, and although function to link propositions that do not have any default
coherent relationships between them. Indeed, in some cases, the propositions connected by these connectives can actually contradict our real-world knowledge of how likely the event depicted in one proposition follows from the other. For example, the connective, “but”, has been characterized by Lakoff (1971) as a denial of the expectation built up from the previous discourse. This means that the two propositions linked by “but” would have actually been in conflict with each other, if not for its presence.

In this paper, we focus on the connective “even so”, which is used in discourse to link two propositions that would be unlikely to follow on from one another in the real world. At the core of this connective is the scalar operator “even” that marks a “reversal-to-expectation” contrast. Being a grammatical scalar operator, even 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). For example, in the sentence, “The school even closed”, the implication is that, under normal circumstances, the likelihood that the school is closed is very low, but nonetheless this event actually occurred.

Even so inherits this scale-reversing function at the discourse level. Consider the following scenario: “The storm last night left only a dusting of snow. Even so, the school was closed…”. This scenario is fully coherent, despite our knowledge that schools usually don’t close if there is only a dusting of snow on the ground. The use of even so acknowledges the default discrepancy between the two linked propositions but, at the same time, it asserts that the unlikely second event nevertheless occurred. Thus, in addition to its scalar reversal function, Even so also serves an important pragmatic communicative function. By explicitly signaling a peculiar property of the upcoming
proposition to the comprehender, it sends a clear signal that it is a special one and that it will contrast from what is expected based on real-world knowledge. This cue helps the comprehender infer the relevance of the upcoming proposition with respect to the current discourse, in the overall communicative context (Sperber and Wilson 1986).

Processing discourse scenarios that describe relationships between events or states that contradict our real-world knowledge poses a real challenge to the language processing system. This is because, as noted above, we constantly draw upon our stored real-world knowledge during normal language comprehension. This stored real-world knowledge is thought to interact, word by word, with the incrementally-constructed message-level representation to generate ‘expectations’ \(^1\) about upcoming words, or groups of words that share semantic features. When a word’s semantic features match such expectations, its processing is facilitated. *Even so*, however, tells us to construct a coherent discourse model that no longer maps directly on to these stored long-term memory representations. This raises important questions. When exactly does the meaning of *even so* first become available to reverse the default real-world relationships between the events being described? And what are the costs of overriding such default real-world knowledge?

---

\(^1\) Throughout this manuscript, we will use the general term ‘*expectations*’ to refer to a pre-activation of a semantic representation, regardless of mechanism. We will use the term *prediction* more specifically to refer to a more active process that commits to particular expectation(s), as discussed later in the manuscript. Of note, the process of generating an expectation, and matching this expectation to the semantic features of a target word, may not always be temporally distinct. This is because expectations can be further constrained by interactions with bottom-up perceptual information of the incoming word (e.g. Marslen-Wilson, 1987; Van Petten 1993; Holcomb, Grainger and O’Rourke 2002).
In theory, there are two possibilities for when *even so* becomes available to reverse the real-world relationships of the events being described. The first is that the integration of *even so* is delayed until the comprehender has computed the meaning of both propositions. This would imply that the parser links one proposition with another only *after* deriving a full representation of each one individually (through semantic-syntactic integration). In this situation, the incongruity between the events or states described and stored real-world knowledge would soon be detected. Only then might *even so* act to reverse the real-world likelihood scale such that, despite the mismatch with real-world knowledge, the input is considered coherent. This type of mechanism would be consistent with models holding that discourse-level pragmatic information asserts its influences relatively late during online processing (Fodor 1983; Forster 1979).

The second possibility is that the scale-reversing function of *even so* is available immediately to set up *reversed real-world expectations* for upcoming words that would otherwise be implausible continuations in a normal discourse. This mechanism has the advantage of paving the way for the smooth processing of incoming words into the current discourse context. It would be generally consistent with models of language comprehension which hold that all information, including pragmatic cues and the wider discourse context, is used as soon as it becomes available, to facilitate the processing of each incoming word (Altmann and Steedman, 1988; MacDonald, Pearlmutter, and Seidenberg, 1994; Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy, 1995; van Berkum, 2009). It is important to recognize that this second account does not imply that reversing the real-world likelihood scale under *even so* is cost-free. Setting up and maintaining mental situations models that are coherent under a set of likelihood relations,
which differ from real-world knowledge, may well come at a cost. This cost may not necessarily be seen at the point of the critical word. However, it may manifest later, at the sentence-final word, which is when so-called sentence-final wrap-up occurs and when the coherence relations between individual propositions are evaluated.

There has been some psycholinguistic work examining how connectives are processed (e.g. Millis and Just 1994; Townsend 1983; Sanders and Noordman 2000; Traxler, Bybee and Pickering 1997). Of particular relevance to the current discussion are findings reported by Murray (1994, 1997). Murray (1994, Experiments 1&2) contrasted coherent two-sentence narratives that were linked by different types of connectives, including concessive connectives like yet, nevertheless, however and but, e.g. “Rudy and Tom laughed with each other on the bus to the amusement park. Yet they didn’t speak to one another for the rest of the day”. Participants self-paced their way through each narrative, presented one sentence at a time. They were faster to read the second sentence when it followed a concessive connective, than when it did not follow any connective. Furthermore, this facilitation effect on the second sentence was observed for concessive connectives, but not for additives (e.g. also, and, furthermore, and moreover) or causal connectives (e.g. therefore, so, thus and consequently). In a follow-up study, Murray (1997) also looked at narratives that were made incoherent by inserting inappropriate connectives, as in “Manny informed his staff about the garage sale. However he arranged for the flyers to be made”. He found that, although all types of inappropriate connectives led to reading disruption, the inappropriate use of a concessive connective led to even more disruption, i.e. it took people still longer to read an incoherent second sentence
when it began with a concessive connective than when it began with other types of inappropriate connectives.

Murray’s observations indirectly support the second mechanism outlined above: that the semantic function of a concessive connective can come into play quite quickly to set up reversed expectations, thus facilitating the processing of upcoming information that is consistent with these expectations. Moreover, his findings of even more facilitation in the coherent concessive sentences (Murray 1994), and even more slow-down in the incoherent concessive sentences (Murray 1997), relative to no-connective discourse, suggest that such reverse expectations may have actually been stronger than expectations based only on default real-world knowledge. This may be because, as noted above, concessive connectives also serve a pragmatic function by sending the comprehender a clear signal that the upcoming proposition is a special one and that it will contrast from what is expected based on real-world knowledge. This pragmatic information may have cued the comprehender to generate more top-down expectations than usual, leading to more facilitation of upcoming words when these expectations were met (Murray 1994) and more costs when they were not met (Murray 1997).

These behavioral findings, however, are certainly not conclusive. Reaction times were measured to the whole second sentence of each two-sentence scenario, providing a relatively crude picture of how and when concessive connectives are used. Whole-sentence reading times reflect the net or cumulative costs of processing at different stages over multiple different words. Thus, facilitated processing of a whole sentence does not necessarily mean that there was no cost of using concessive connectives at any stage of
processing, and an overall slow-down does not rule out the possibility of temporary facilitation.

To gain a more precise picture of how concessive connectives are used at different time points during word-by-word incremental parsing, we used event-related potentials (ERPs)—a direct measure of neural processes with excellent temporal resolution, and which can be time-locked to individual words in a discourse. Distinct ERP components, distinguished by their latencies, polarity, topographical distribution, and morphology, can be mapped onto (at least partially) distinct neurocognitive mechanisms. In the present study, we focused on two ERP components: the N400 and the P600 or late positivity.

The N400 and real-world knowledge

The N400 is a negative deflection of the ERP waveform, peaking at around 400ms after the onset of a word stimulus. Its amplitude can reflect the degree to which that word’s semantic features match expectations that are generated through the interaction between context and different types of stored semantic material, including real-world knowledge (Kutas and Federmeier 2011; Paczynski & Kuperberg, in press). Expectations can be generated in different ways. A contextual representation may passively resonate with networks encoding semantic relationships between individual words to trigger a spread of activation (e.g., Myers and O’Brien, 1998; Gerrig and McKoon, 1998) that can facilitate processing to some degree, leading to some attenuation of the N400 (e.g. Federmeier and Kutas 1999; Otten and van Berkum 2007; Metusalem et al., in press; Paczynski and Kuperberg, in press; Lau et al., under review). However, most facilitation—and most N400 attenuation—is seen when expectations are generated
through more active *prediction* mechanisms, in which some commitment is made to a particular item or set of items, in advance of new input (DeLong et al., 2005; van Berkum et al., 2005; Federmeier et al., 2007; see Lau et al., under review, for discussion). There are many factors that can influence the generation of more active predictions, and hence N400 attenuation. These include the lexical predictability/constraint of the context itself (often operationalized using cloze tasks, e.g. Kutas and Hillyard 1984; Federmeier et al. 2007), the engagement of the parser in predictive strategies (Lau et al., under review), and, as discussed further below, the pragmatic relevance of the discourse context (Nieuwland and Kuperberg 2008; Nieuwland, Ditman and Kuperberg 2010; Hald, Steenbeek-Planting, and Hagoort 2007).

In many cases, expectations are based *directly* on real-world knowledge about actions, events or states, and the amplitude of the N400 therefore patterns with real-world plausibility/probability of the events/states described. This means that, during sentence processing, words that are congruous with real-world event/state knowledge evoke a smaller N400 than incongruous words (Kuperberg, Sitnikova, Caplan, and Holcomb, 2003; Hagoort, Hald, Bastiaansen, Marcel, & Petersson, 2004; Ferretti, Kutas and McRae, 2007; Bicknell, Elman, Hare, McRae, and Kutas, 2010). This is also true at the level of discourse where a smaller N400 is seen to words that conform to real-world expectations based on causal relationships between familiar events and states, than to words that are incongruous with such expectations (Yang et al. 2007; St. George et al. 1997), even when the semantic relatedness between individual words is matched across conditions (Kuperberg et al. 2011).
There are some instances, however, in which the N400 does not necessarily pattern with real-world plausibility/probability of the events or states described. For example, Hald, Steenbeek-Planting, and Hagoort (2007) showed that a critical word that mismatches our real-world knowledge (such as “The city of Venice has very many roundabouts” instead of “The city of Venice has very many canals”) can be made more expected, leading to a reduced N400, if it is embedded in a context that shifts the focus of the discourse topic, making the originally unexpected critical word more relevant, e.g. a context explaining that the city council of Venice introduced more road layouts to ease traffic flow. And Nieuwland et al. (2006, Experiment 2) reported that, when embedded in stories that supported a fictional reading, critical words in highly implausible sentences (e.g. “The peanut was in love...”) evoked a smaller N400 than in plausible sentences (“The peanut was salted...”). Finally, Nieuwland and Martin (2012) showed that when the discourse context invites speakers to entertain counterfactual worlds, the N400 was modulated by what happened in the imaginary world rather than in the real world, e.g. the N400 to Russia was smaller than to America in the counterfactual sentence, “If NASA had not developed its Apollo Project, the first country to land on the moon would have been Russia/America surely”.

In these cases the discourse context set up a fairly predictable and constraining alternative situation, and expectations were based on this ‘alternative world’ rather than on any stored, long-term real-world knowledge. For example, a hypothetical scenario in which Russia, rather than America, won the space race and landed first on the moon is associated with a highly salient historical event, and thus the consequence of this scenario is easy to fill in. This contrasts with another study examining counterfactuals, which used
less relevant and less predictable contexts: 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 carrots/fish...”. This type of context is less pragmatically relevant and less constraining: when constructing a hypothetical world of vegetarian cat, feeding it carrots is not the first thing to come to mind. Unlike Nieuwland and Martin (2012), in these scenarios, the inconsistent word (*fish*) elicited a smaller N400 than the consistent word (*carrot*), i.e. in this case, the N400 was modulated by real-world consistency, rather than by the counterfactual world. Similarly, eye-movement studies of counterfactuals using very similar material (Fergurson and Sanford 2008; Fergurson et al. 2008, Experiment 1) found no differences in early fixations measures between consistent and inconsistent critical words.

Together, these findings suggest that the pragmatic relevance and constraint of a discourse context may play an important role in influencing whether real-world knowledge can be overridden. This is consistent with previous evidence that pragmatic cues can be quickly integrated within a discourse model to influence the semantic processing of upcoming words (van Berkum 2009; van Berkum, van den Brink, Tesink, Kos and Hagoort, 2008), and that pragmatic relevance can influence how quickly negation operators (Nieuwland and Kuperberg 2008) and scalar implicatures (Nieuwland, Ditman and Kuperberg 2010) are calculated during online discourse processing.

**Late positivities/the P600**

The N400 reflects the match between the semantic features of each incoming word and prior expectations set up by the preceding context (Kutas & Federmeier, 2011; Paczynski & Kuperberg, In press). However, it does not necessarily reflect the final
stages of computing the plausibility or coherence of a sentence or discourse. There are some situations in which, in order to come to a full final representation of meaning, processing must continue and/or the context needs to be reanalyzed. These types of additional analyses are sometimes reflected by a group of posteriorly-distributed late positivities or P600s.

The P600 was first described to syntactic violations and ambiguities (Hagoort, Brown, & Groothusen, 1993; Osterhout & Holcomb, 1992), but has more recently been associated with semantic propositional incoherence where it has been descriptively termed a ‘semantic P600 effect’ (Kuperberg et al., 2003; Hoeks, Stowe, & Doedens, 2004; Kim & Osterhout, 2005; van de Meerendonk, Kolk, Vissers, & Chwilla, 2010). A semantic P600 is most likely to be triggered when propositional incoherence is detected by the parser (Kuperberg, 2007; Sanford et al. 2011) and when it conflicts with an alternative intermediate semantic representation. We have suggested that this conflicting intermediate representation is generated through a semantic memory-based analysis (Kuperberg, 2007), and more specifically through the (inaccurate) prediction of an alternative lexical item or group of items sharing common semantic features (see Paczynski & Kuperberg, in press for a recent discussion). A semantic P600 effect is most likely to be produced when comprehenders make explicit judgments about propositional coherence, increasing both the likelihood that the parser generates semantic predictions, and that it detects propositional incoherence (see Kuperberg, 2007 for

\[1\] There has been some debate about the nature of this intermediate representation. Some have suggested that it is constructed by integrating the incoming word with its preceding context through asyntactic combinatorial (Kim & Osterhout 2005) or heuristic (van de Meerendonk, Kolk, Chwilla, & Vissers, 2009) mechanisms, rather than simply reflecting the activation of stored information.
discussion). It may reflect additional analysis or reanalysis that is required to override this conflict and come to a final veridical representation of discourse meaning.

The present study

The present study employed ERPs to investigate the time course of resolving coherence relationships under *even so*, in comparison with plain scenarios without this connective. Participants read three-sentence scenarios like those in Table 1. The stimuli were constructed by crossing the coherence relationship between the two sentences (coherent or incoherent) and the presence of *even so* (plain scenarios or even-so scenarios in which *even so* appeared before the final sentence). This gave rise to four conditions. Referring to Table 1, the critical word *closed* was either coherent (1) or incoherent (2) in its discourse context. By adding “Even so” at the beginning of the final sentence of the same two scenarios, the content words remained the same, but the coherence relationships between the two clauses were reversed (3 and 4). This design allowed us to determine when *even so* became available to influence processing, distinguishing between the two general mechanisms described above.

We carried out two experiments. In Experiment 1, we asked participants to explicitly rate the coherence of each discourse situation, encouraging them to attend to the coherence relationships between the events and states described. In Experiment 2, participants answered intermittent comprehension questions, allowing us to determine when the scalar reversing function of *even so* became available during more passive discourse comprehension.

**Experiment 1**
Introduction

In this experiment, we recorded ERPs while participants read the four types of scenario described in Table 1. After each three-sentence scenario, participants gave a coherence rating about whether the last sentence in each scenario followed naturally from the first two. This rating task served two main purposes. First, it encouraged readers to try and make sense of the incoming material, and therefore to pay attention to the discourse context and the internal semantic relations between each proposition. We reasoned that this would make them maximally sensitive to the scalar reversing properties of even so. Second, the task yielded a complementary behavioral measure, allowing us to compare the coherence rating results with the pattern of ERP modulation.

We first considered the effects of Coherence on the two plain scenarios in (1 and 2 in Table 1). Based on our previous ERP study examining the effects of causal coherence on similar causally related and unrelated scenarios, using a similar judgment task (Kuperberg et al. 2011), we predicted a smaller N400 amplitude on critical words in the coherent scenarios (1) than in the incoherent scenarios (2). At least when the parser is actively engaged in making sense of the whole discourse, situation models built from the discourse context, coupled with stored real-world knowledge about relationships between states and events, will encourage the generation of expectations, leading to some attenuation of the N400 on critical words when such expectations are met.

Our main question was how Coherence would modulate the N400 and P600 on critical words in the even-so scenarios. Again, note that the even-so coherent scenarios (e.g. 3) paired together events/states that are unlikely to follow on from one another in the real world, while the even-so incoherent scenarios (e.g. 4) paired together events/states
that are likely to follow on from one another in the real world. The two potential mechanisms outlined in the General Introduction predict different patterns of modulation on both the N400 and P600 at the point of encountering the critical word.

The first overall possibility was that the scalar reversing function of *even so* is delayed. Readers would base any expectations about upcoming words on their real-world knowledge. This would predict a larger N400 to coherent than incoherent even-so scenarios, i.e. a reverse N400 effect, consistent with the ERP findings on counterfactuals by Fergurson et al. (2008). Only after some inconsistency between the current discourse and real-world knowledge is detected might the comprehender use the scalar reversal function of *even so* to reconcile this inconsistency. This might engage additional neural resources, which might manifest as an additional waveform (positivity or negativity), past the N400 time window.

The second overall possibility is that the scale-reversal function of *even so* is used immediately to reverse expectations that are generated prior to the onset of the critical word. This would predict that, just as in the plain scenarios, critical words should evoke a smaller N400 in the coherent than incoherent even-so scenarios. Such a finding would be consistent with ERP studies suggesting that, given a sufficiently constraining sentence or discourse context, expectations based purely on real-world knowledge can be overridden (e.g. Nieuwland et al. 2006; Hald et al., 2007; Nieuwland and Martin 2012). Indeed, the behavioral findings by Murray (1994) suggest that *even so* may lead to the generation of reverse expectations that are actually *stronger* than those based only on default real-world knowledge, perhaps because of its pragmatic communicative function. This would lead to further facilitation of semantic processing on incoming words whose semantic
properties match these top-down expectations, therefore leading to a smaller N400 on critical words in the coherent even-so than the coherent plain scenarios.

The immediate engagement of even so to actively reverse predictions might also lead to costs when such predictions are disconfirmed, particularly if the resulting interpretation is incoherent (see supporting behavioral evidence by Murray, 1997). As discussed in the General Introduction, the waveform most commonly associated with a constraining discourse context, in conjunction with sentence or discourse incoherence, is the posteriorly-distributed P600 effect, which is thought to reflect continued analysis or reanalysis in an attempt to establish coherence (Kuperberg, 2007; Paczynski & Kuperberg, In press). A P600 effect is most likely to be triggered when participants carry out active tasks (including judgment tasks), which encourage predictive processing, as well as the detection of overall discourse incoherence at the point of the critical word (see Kuperberg, 2007 for a discussion of mechanisms by which task might bias towards the production of a P600 effect on critical words in contextually-constrained incoherent versus coherent scenarios). An immediate engagement of even so to actively predict upcoming material would therefore lead to a larger P600 on critical words in the even-so incoherent scenarios than the plain incoherent scenarios.

Finally, we were interested in whether there were any costs of using even so, independent of when its meaning first became available. To address this question, we examined its effects on the final word of the discourse—the point at which the comprehender was required to explicitly evaluate coherence. Costs of evaluating coherence at sentence-final wrap-up are often manifested as a sustained negativity. For example, a sustained negativity effect is seen on sentence-final words of anomalous or
implausible (versus plausible) sentences (e.g. Osterhout and Holcomb 1992, 1993; Hagoort and Brown, 2000; Hagoort, 2003; Ditman, Holcomb and Kuperberg, 2007; De Grauwe, Swain, Holcomb, Ditman and Kuperberg, 2011). If the scale-reversing function of *even so* is also used at the sentence-final word for evaluating overall coherence, this might lead to additional processing costs, which might manifest as a larger sustained negativity in the even-so than the plain scenarios.

**Methods**

**Construction and characterization of stimuli**

One-hundred-and-eighty sets of three-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 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) and Coherence (the coherence of the critical word in relation to its preceding context: coherent or incoherent). In the plain coherent and plain incoherent conditions, the final sentence was identical, but differences in the first two sentences yielded different coherence. The addition of “Even so,” to the beginning of the final sentence of these two conditions effectively reversed the coherence of these scenarios. Hence, the original plain coherent condition became the even-so incoherent condition, and the original plain incoherent scenarios became the even-so coherent condition. The critical word was thus identical in all four conditions.

**Cloze ratings and contextual constraint**
The 180 sets of scenarios were counterbalanced across four lists using a Latin Square design. Each list contained 180 scenarios, consisting of only one condition from each scenario, but they were counterbalanced such that all four conditions in all scenarios appeared once across the four lists. For each scenario, the critical word and all words following it were removed and replaced by an ellipsis, e.g. “Carol listened to the radio. The storm last night had dumped several feet of snow. The schools were….”. 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 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 x 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 was larger for the plain than the even-so scenarios. Follow-up t-tests examining effects of Even-so at each level of Coherence showed that the cloze probability 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 degree of contextual constraint for each type of scenario stem by finding the most common completion across participants who saw that scenario stem, whether or not it matched the critical word, and tallying the number of subjects who provided this completion. For example, for a given scenario, if 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 subjects provided “confused” as their answer, then the constraint probability for this scenario stem would be .5. Contextual constraints for each type of scenario stem are also given in Table 1. A 2 x 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 contextual constraint of the plain coherent scenario stems was greater than the plain incoherent scenario stems (t(179)=5, p<0.001), while the difference
between the even-so coherent and incoherent scenario stems was only marginally significant (t(179)=1.9, p<0.06). In addition, the contextual constraint of the plain coherent scenario stems was greater than the even-so coherent scenario stems (t(179)=3.6, p<0.001), but there was no difference in contextual 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. 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 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, 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 each of these two sentences at their own pace, and pressed a button to move on to the second sentence. After the second sentence, they 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 and, 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**

Averaged ERPs, time-locked to critical words, were formed off-line from trials free of ocular and muscular artifact and were quantified by calculating the mean amplitude (relative to a 100ms pre-stimulus baseline). At the critical word, we carried out analyses across two time windows: 350-450ms, which encompassed the peak of the N400 in all four conditions, and 600-800ms, which encompassed the more broadly-distributed
P600 effect in all four conditions. 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). Follow-up analyses were constrained to the three regions where ERP modulation in the omnibus analyses was maximal (where higher-order interaction effects were most robust). In these follow-ups, interactions between Coherence and Even-so were parsed in two ways: by examining the effects of Coherence at each level of Even-so, and by examining the effects of Even-so at each level of Coherence.

All trials were included in the ERP analyses. 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 x 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). There was also 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 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 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 x 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**

Collapsed across the even-so and plain scenarios, there was a widespread N400 effect (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 effects of Coherence in each of these three regions (Fs>23, p<.001).

Over these frontal, central and parietal mid-regions (9 electrodes total), the effect of Coherence was further modulated by Even-so (Coherence x Even-so interaction: F(1,26)=5.4, p<.05). Follow-ups showed 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 scenarios (F(1, 26)=5.3, p<.05), 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 scenarios (F(1, 26)<1, p>.4), see Figure 3B.

**Late 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(1, 26)=22.3, p<.001, with follow-ups showing effects of Coherence in each of these four regions, all Fs > 6.8, ps < 0.05).

Over these parietal and occipital regions (12 electrodes), 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).

**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), due to slightly more rejected incoherent than coherent trials, and an effect of Even-so (F(1,26)=9.8, p<.01) due to slightly more rejected even-so than plain trials, but 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 larger sustained negativity to
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 to 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 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 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 were all larger than that the waveform evoked by sentence-final words of the coherent plain scenarios (Fs > 6, ps < .05).

**Discussion**

In this experiment, we examined the effects of Even-so as participants made explicit judgments about the coherence of three-sentence scenarios. In the plain
scenarios, we found both a small N400 and a small P600 effect of Coherence on the critical words. We also found N400 and P600 effects of Coherence in the even-so scenarios. In fact, both these effects were larger than the effects seen in the plain scenarios. However, whereas the larger N400 coherence effect in the even-so scenarios was driven by a smaller N400 to coherent even-so (versus plain) critical words, the larger P600 effect in the even-so scenarios was driven by a larger P600 to incoherent even-so (versus plain) critical words. Finally, we saw effects of both Coherence and Even-so at the sentence-final word.

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 discourse scenarios while participants carried out a coherence judgment task, similar to the task employed in the present study. As in this previous study, the N400 was attenuated to the coherent critical words, even though its semantic relatedness with the preceding context (as operationalized by LSA) was matched to the incoherent scenarios (see Table 1). We interpret this N400 attenuation as reflecting semantic facilitation that resulted from expectations based on stored real-world knowledge about the coherence relationships between familiar events and states. Of note, the degree of N400 attenuation to the coherent critical words in this study was less than in Kuperberg et al. (2011). This is probably because any specific predictions based on stored relationships between real-world events/states were weaker in this study than in Kuperberg et al. (2011) because the discourse context was less highly constraining. Also, unlike Kuperberg (2011), the small N400 effect was followed by a small P600 effect,
which may have reflected the detection of discourse incoherence at a slightly later stage of processing.

The even-so scenarios

The N400 attenuation to critical words in the coherent versus incoherent even-so scenarios clearly indicates that readers are able to use the scalar reversing function of even so immediately to reverse the likelihood scale on which they generate their predictions about the semantic properties of upcoming words. At the point of the critical word, default relationships between events and states, stored in memory, were overridden by these reverse expectations.

Our finding that this N400 Coherence effect was larger in the even-so than the plain scenarios, due to a smaller N400 to coherent even-so than plain critical words, further indicates that the reverse expectations set up by even so were actually stronger than expectations based only on default world knowledge: these reverse expectations led to more facilitation of processing semantically coherent critical words.

The generation of stronger expectations in the even-so than the plain scenarios cannot be attributed to differences in the lexical constraint of their contexts: the cloze probability of critical words the even-so coherent scenarios was not particularly high (see Table 1), and was, in fact, lower than in the plain coherent scenarios. Rather, we propose that encountering even so triggered readers to adopt a more predictive strategy of comprehension than that adopted to process the plain scenarios. This type of shift to a predictive mode of processing under even so would not be reflected by cloze ratings because, in an offline cloze task, raters have as long as they want to produce words: by definition, they are already adopting a highly predictive strategy. During online
incremental parsing, however, comprehenders have limited time to generate predictions; thus, any shift in predictive strategy will influence N400 modulation (see Lau, Holcomb & Kuperberg, Under review).

Exactly why even so triggered this type of shift to a more predictive mode of processing is an open question. However, we hypothesize that a key factor is its pragmatic function: encountering even so at the beginning of the final sentence provided a highly salient cue to the comprehender to constrain his/her expectations for upcoming words. We return to this idea that pragmatic factors are key for triggering shifts in predictive strategy in the General Discussion.

In addition to a larger N400 effect, critical words in the even-so scenarios also produced a larger P600 effect than in the plain scenarios. This larger P600 effect, however, was driven entirely by an enhanced P600 to incoherent critical words in the even-so (versus plain) scenarios. As discussed in the General Introduction, the amplitude of the P600 evoked by a semantically incoherent word is enhanced when discourse incoherence conflicts with an alternative semantic representation that is generated through semantic memory-based expectations (Kuperberg, 2007; Paczynski & Kuperberg, In press). In the present study, the even-so incoherent critical words were not more incoherent than the plain incoherent critical words (in fact, they were rated as more coherent). We therefore take the enhancement of the P600 in the even-so incoherent scenarios as further support for the hypothesis that readers were more strategically engaged in predictive processing in the even-so scenarios than in the plain scenarios. That is, under a more predictive parsing mode, they were more likely to generate alternative semantic representations of the upcoming words, which ended up conflicting with their
detection of an incoherent overall meaning upon integrating the actual critical word into its discourse context. We suggest that the enhanced P600 reflected a process of continued analysis or reanalysis of the incoherent critical word in an effort to come to a consistent final representation of discourse meaning (Kuperberg, 2007).

In sum, we interpret the larger N400 and P600 effects of Coherence in the even-so than the plain scenarios as both reflecting an increased engagement in predictive strategies under **even so**. This increased prediction led to both facilitated processing of coherent critical words (i.e. an attenuated N400 on critical words in coherent even-so versus plain scenarios), as well as to costs when predictions were unfilled and participants detected discourse incoherence.

**The sentence-final word**

Finally, by examining activity at the sentence-final word, we gained some insight into whether there were any global costs of using **even so** to reverse the real-world likelihood scale during sentence-final wrap-up. We found that there were. In addition to a prolonged sentence-final negativity effect in incoherent (versus coherent) plain scenarios, sentence-final words in the even-so (versus the plain) scenarios also produced a prolonged negativity effect, regardless of **even so**. At frontal sites, the negativities evoked by even-so coherent, even-so incoherent and plain incoherent scenarios were all of the same magnitude. At posterior sites, however, the effects of coherence and even-so appeared to be additive: the negativity produced by sentence-final words of even-so incoherent scenarios was larger than that produced by sentence-final words of either incoherent plain or coherent even-so scenarios. This suggests that the neurocognitive processes of assessing coherence and reversing the real-world likelihood scale were
somewhat distinct, although operating within the same time window. We develop these ideas further in the General Discussion.

**Experiment 2**

**Introduction**

Experiment 1’s findings suggest that, when participants make active coherence judgments, the brain can engage the concessive connective, *even so*, to both reverse and enhance expectations about the semantic features of upcoming words, leading to N400 attenuation. The aim of this second experiment was to determine whether *even so* is used in a similar way under more passive reading conditions. We therefore presented a new set of participants with the same experimental stimuli. However, rather than carrying out a coherence judgment task, they were asked to simply read the discourse scenarios and to answer some intermittent comprehension questions dispersed randomly throughout the experiment.

In Experiment 1, the N400 attenuation to critical words in the plain coherent (versus incoherent) scenarios was fairly small. We suggested that this was because any specific predictions that were based on stored relationships between familiar real-world events/states were fairly weak. Under more passive reading conditions, these types of predictions might be even weaker. This might lead to processing that is driven primarily by passive resonance with stored networks encoding the semantic relatedness between individual words (see Kuperberg et al., 2011 and Paczynski & Kuperberg, In press for discussion). Since the semantic relatedness between the critical word and its preceding context were controlled to be equal in the coherent and incoherent scenarios (the LSA measure in Table 1), this would predict no N400 difference between coherent and
incoherent plain scenarios. One relevant previous finding in this regard was reported by Burkhardt (2007) who found a P600 effect, but crucially no N400 effect as participants passively read discourse scenarios that required the generation of inferences.

Experiment 1 also showed that the N400 attenuation to coherent critical words in the even-so scenarios was greater than in the plain scenarios. We suggested that, under even so, comprehenders generated reverse predictions that were actually stronger than predictions generated in the plain scenarios. It is possible, however, that this increased N400 attenuation was not simply driven by even so, but rather by an interaction between even-so and the coherence judgment task. That is to say, when comprehenders were required to explicitly rate coherence, they became more sensitive to the discourse function of even so, and maximally incorporated the scale-reversing function of even so into their parsing strategy. If this were the case then, during more passive reading comprehension, one would predict no modulation on the N400 between the even-so and plain coherent scenarios: once again, processing might be driven by a more passive resonance between the context and semantic relatedness networks. If, on the other hand, even so alone is sufficient to trigger the generation of reverse predictions, this would predict more N400 attenuation to coherent critical words in the even-so than the plain scenarios, just as in Experiment 1.

In addition to the attenuation on the N400 to coherent critical words in the even-so scenarios, Experiment 1 demonstrated an enhancement of the P600 to the incoherent critical words in the even-so scenarios (each relative to the plain incoherent scenarios). We suggested that this enhanced P600 was driven by a conflict between two processes: (a) participants’ immediate detection of discourse incoherence on the critical word; and
(b) the generation of an alternative semantic representation from active predictions that were encouraged under *even so*. Both these processes are likely to be reduced under more passive reading conditions, and we therefore predicted less or no enhancement of P600 to critical words in the even-so incoherent, relative to the plain incoherent scenarios.

Finally, we were once again interested in whether there were any global costs of using *even so* to reverse the default likelihood scale at sentence-final wrap-up. We therefore examined the final word of these scenarios to see whether, as in Experiment 1, there was a larger sustained negativity to sentence-final words in the even-so than the plain scenarios.

**Methods**

**Participants and procedure**

Twenty-three undergraduates initially participated in Experiment 2. Three subjects were subsequently excluded from data analysis because of extensive ocular movements. 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 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, “Was it necessary for the schools to close?” The planned correct answer for this particular question was “yes” for the plain coherent and even-so coherent versions, and “no” for the plain incoherent and even-so incoherent
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 needed to simply read and understand each discourse scenario and answer questions when they came up.

**Results**

**Behavioral results**

Overall accuracy in answering the comprehension questions was 87% (SD: 4.41%). A 2 x 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<.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<.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<.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<.01).

**Event related potentials**

*Critical Word*
At the critical word, 17% of trials were rejected for artifact. A 2 x 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 these 9 parietal electrodes revealed no N400 effect of Coherence in the plain scenarios (F(1,19)<.1, p>.6; Figure 5A), 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 mainly driven by an attenuation of the N400 to critical words of the coherent even-so (vs. coherent plain) scenarios (F(1,19)=4.2, p<.05; see Figure 6A, again voltage map in Figure 6A shows a positivity between 350-450ms because the plain condition was subtracted from the even-so condition); there was no difference in N400 evoked by the critical words in the incoherent even-so (vs. 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 x 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)>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 plain
scenarios.

**Sentence-final word**

At the sentence-final word, 26% of trials were rejected for artifact. A 2 x 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.

**300-600ms**

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 (Fs≤3, ps≤1), but that did not differ from one another (F(1,19)<1, p>.9). In the anterior 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 viewed the same stimuli as in Experiment 1, but this time, they simply read the scenarios for comprehension rather than carrying out an explicit rating task. We saw some differences and some similarities between the two experiments in the pattern of results.

Unlike in Experiment 1, there was no N400 effect of Coherence in the plain scenarios—only a P600 effect was seen, with a larger P600 to incoherent than coherent critical words. As in Experiment 1, however, we did see an N400 effect of Coherence in the even-so scenarios, and this N400 effect was once again driven by a smaller N400 to coherent critical words in the even-so scenarios than in the plain scenarios. The N400 Coherence effect in the even-so scenarios was once again followed by a P600 effect.
However, unlike in Experiment 1, the P600 effect in Experiment 2 was not modulated by even-so, i.e. it was no larger in the even-so scenarios than in the plain scenarios. Finally, as in Experiment 1, we saw modulation across the four conditions on a sentence-final negativity. Once again, at frontal sites, the negativities evoked by even-so coherent, even-so incoherent and plain incoherent scenarios were larger than the negativity evoked by plain coherent scenarios; at posterior sites, however, the sentence-final negativity in the even-so incoherent scenarios was larger than that in either the plain incoherent or the even-so coherent scenarios, although these effects were weaker than in Experiment 1.

**Plain scenarios**

The absence of an N400 effect in the plain scenarios suggests that, during passive reading comprehension, the N400 was modulated primarily by the semantic relatedness between individual content words, which were matched between coherent and incoherent scenarios (through LSA). In the absence of any task pressure, and in a relatively non-lexically constraining context (as indexed by a relatively low cloze probability), comprehenders are less likely to have actively predicted upcoming words on the basis of stored real-world knowledge about relationships between familiar events and states, leaving processing to be driven primarily by passive resonance with stored relatedness networks. As in Experiment 1, there was, however, some registration of the incoherence of the plain scenarios, which manifest slightly later, during the P600 time window.

**Even-so scenarios**

As in Experiment 1, we interpret the reduction of the N400 to coherent critical words in the even-so versus the plain scenarios, as reflecting an increased engagement in predictive processing under *even so*. Again, we suggest that *even so* acted as a pragmatic
cue that triggered comprehenders to adopt a (reverse) predictive strategy about the upcoming words. This indicates that the N400 reduction to the coherent even-so (versus plain) critical words in Experiment 1 was not driven by the interaction between even so and the coherence judgment task. Rather, even so alone was sufficient to engage participants in reverse predictions, even under more passive reading conditions.

In contrast to Experiment 1, however, there was no increase in the P600 to the incoherent even-so (versus incoherent plain) critical words. This suggests that the enhancement of this effect in Experiment 1 was driven by the combination of even so and the coherence judgment task. We will return to the issue of how the task influenced both N400 and P600 modulation in the General Discussion.

**Sentence-final word**

Once again, we observed effects of both Even-so and Coherence on the sentence-final words. At posterior sites, these effects again seemed to be additive, with a larger negativity produced by sentence-final words of even-so incoherent scenarios, relative to both incoherent plain or coherent even-so scenarios. This suggests that, even under passive reading comprehension, the processes of reversing the real-world likelihood scale under even so and assessing final discourse coherence, were somewhat independent and additive.

**General Discussion**

In this study, we used the concessive connective even so to study the interaction between real-world knowledge and discourse information during online language processing. We aimed to distinguish between two possible mechanisms for how and when even so is applied during discourse processing. We also wanted to determine
whether the pragmatic communicative function of *even so* serves a purpose during comprehension. Finally, we were interested in whether the discourse function carried by *even so*, i.e. reversing default real-world likelihood relations, incurs any overall processing costs. In the discussion below we address each of these questions. We then discuss some general implications of our findings before offering some conclusions.

*Even so* is engaged immediately to reverse expectations during discourse processing

In the General Introduction, we outlined two possible mechanisms for how and when a concessive connective like *even so* is applied during discourse processing to reverse the default likelihood probabilities of two propositions describing sequential events or states. The first was that default real-world knowledge about the relationships between these events/states is used before *even so* is integrated. On this view, real-world knowledge is first used to facilitate the processing of the two propositions, and *even so* only operates to reverse real-world likelihood at a later stage, perhaps when the parser detects an incongruity. The second possible mechanism was that *even so* immediately interacts with real-world knowledge and the event or state described in one proposition to set up *reversed expectations* for upcoming words in the following proposition. On this account, both expectations and overall discourse coherence are evaluated based on this new set of reversed relationships between events/states, instead of the old default ones.

Data from both our experiments clearly support the second mechanism. In both experiments, processing of the two even-so scenarios was driven by the newly-established likelihood probabilities. In the even-so coherent scenarios, the relationship between the two propositions did not conform to stored real-world knowledge about likely relationships between familiar events or states. Nonetheless, critical words still
evoked a smaller N400 than in the even-so incoherent scenarios where the relationship between the propositions did conform to real-world knowledge. These findings are in line with a larger body of previous research showing that contextually-introduced coherence relations can, at least under some circumstances, override default real-world knowledge (Nieuwland et al., 2006; Hald, Steenbeek-Planting, and Hagoort, 2007; Nieuwland and Martin, 2012).

The mechanisms by which *even so* was used to set up these types of reverse expectations are unclear, but we hypothesize at least two operations. First, default real-world knowledge about the relevant events/states, and the relationships between them, must be activated from long term memory storage; second, their likelihood relations must be reversed to create a parallel *reversed world* of the events/states under discussion, which must be held within working memory. As discussed below, setting up such a reversed world under *even so* was not without costs—these were clearly manifest on the sentence-final word. However, we suggest that the immediate application of the scalar-reversal function of *even so* led to more efficient online sentence comprehension before this point: new incoming information interacted with the reversed world, rather than with stored real-world knowledge, to set up reverse expectations about the semantic features of upcoming words. When such expectations were met, the processing of upcoming words was facilitated, leading to a smaller N400.

*Even so* may function as a pragmatic cue to trigger more predictive processing

---

3 We use the terms *reversed world* or *alternative world* in this paper in an informal sense to mean that a separate set of likelihood relations need to be established under *even so* in some mental space that is different from the stored knowledge in long-term memory. We do not mean to apply the formal possible world semantics to the linguistic analysis of *even so*.
Beyond simply reversing the likelihood scale on which comprehenders set up expectations about upcoming events, *even so* seemed to lead to even more semantic facilitation of critical words than in the plain scenarios: in both experiments, the N400 coherence effect was actually larger in the even-so than in the plain scenarios and this larger N400 effect was driven entirely by a reduced N400 to critical words in the coherent even-so than in the coherent plain scenarios. This is in line with the behavioral results by Murray (1994) who found faster whole-sentence reading times in coherent scenarios with concessive connectives than coherent plain scenarios. The current study further demonstrates that this type of facilitation effect is immediately evident as language unfolds, word by word. We argue that this additional facilitation occurred because *even so* triggered more predictive processing than in the plain scenarios. That is, expectations set up on the basis of the reverse world established under *even so* were actually stronger than those set up on the basis of default real-world knowledge.

The link between the N400 and predictive processing has been discussed before, but mainly in relation to contextual constraint and cloze probabilities (Kutas and Hillyard 1984; Federmeier and Kutas 1999). Contexts that provide a higher cloze probability of upcoming words are said to be more semantically constrained, leading to more semantic facilitation (e.g. Federmeier et al. 2007). In the current study, however, offline cloze/constraint measures and N400 amplitude did not pattern together in the coherent scenarios (see Table 1): critical words in the even-so coherent scenarios had lower cloze and constraint ratings than in the plain scenarios, but still evoked a smaller amplitude N400. An account that only makes use of offline cloze or constraint measures to predict N400 amplitude falls short of explaining these observations.
We suggest that, while the cloze probability of a particular word is a good proxy for how likely it is that a context is predictive and how much the N400 will be attenuated, it is not the only factor to influence predictive processing and N400 amplitude. This is because cloze and constraint measures are collected offline when participants are explicitly instructed to make predictions and are given plenty of time to do so. They therefore do not necessarily reflect parsing strategies employed during incremental online processing, when participants’ conscious awareness of predictions is minimal and there is limited time before the next word is encountered. As a result, when contextual constraint is relatively low, the parser may be highly sensitive to any additional cues that encourage more predictive strategies during processing.

The sensitivity of the N400 to predictive strategies is illustrated in a recent semantic priming paradigm in which we manipulated the proportions of the same set of semantically associated prime-target pairs across experimental blocks (Lau, Holcomb & Kuperberg, under review). In the high relatedness proportion block, which encouraged a more predictive strategy, the N400 priming effect was larger than in the low relatedness proportion block and this larger N400 effect was once again, driven by an attenuation of the N400 to associated targets. Importantly, this N400 attenuation was seen despite the immediate context (the prime) being kept constant.

Similarly, in the present study, we suggest that the N400 attenuation on critical words in the even-so coherent (versus plain coherent) scenarios was driven by a strategic shift to a more active prediction mode of comprehension, which was triggered by the pragmatic function of *even so*. In the plain discourse scenarios, without explicit connectives, the parser did not know how the first proposition would be linked to the
second. In theory, they could have been connected along several possible dimensions—causal, spatial, temporal, premise-conclusion reasoning, intentionality, etc (Mandler, 1986; Speer and Zacks, 2005; Zwaan, 1996; Clark, 1972; Linde and Labov, 1975; Morrow, Greenspan, and Bower, 1987; Black and Bower, 1980; Keenan, Baillet, and Brown, 1984; Trabasso, Secco, and van den Broek, 1984). Because contextual constraint was not high, there was little incentive to actively predict upcoming words. Even so, however, provided an explicit communicative cue that signaled the relevance of the upcoming proposition with respect to the current one: it told the reader that the upcoming proposition would form a “contrary-to-the-expectation” relation to the current proposition, as opposed to any other coherence relation. We suggest that this led to more processing effort towards searching for the best alternative that contrasted with what default real-world knowledge would have predicted. That is, the parser actively searched the reversed world to generate more top-down predictions about the semantic features of incoming target words, leading to more attenuation of the N400 when the target’s features matched these predictions.

It is not totally clear why the even-so coherent scenarios received lower cloze/constraint ratings than the plain coherent scenarios, but one possibility is that these ratings reflected increased costs in processing the even-so scenarios as a whole. This interpretation is supported by the lower overall coherence ratings given to the even-so than the plain coherent scenarios in Experiment 1 (3.3 vs. 4.8 out of 5, see Table 1), as well as by the larger negativity on the sentence-final words of the even-so coherent than the plain coherent scenarios in both experiments. We discuss these findings in more detail below. At this point we note that they imply that, despite the even-so coherent scenarios
being more difficult to process at a global level than the plain coherent scenarios, participants used *even so* to maximally exploit what information they had to make *online* predictions.

The significance of pragmatic relevance on discourse processing has also been highlighted in some previous studies (Nieuwland and Kuperberg, 2008; Nieuwland, Ditman and Kuperberg 2010; Nieuwland and Martin 2012). However, in these studies, more predictive processing, triggered by pragmatic relevance, was conflated with the high cloze probability on the critical word. The present findings show that cloze probability does not always pattern with predictive processing, as indexed through N400 attenuation, and that the two constructs can be empirically dissociated.

**Immediate costs of detecting discourse incoherence**

In both experiments, a P600 effect was seen to incoherent (versus coherent) critical words, regardless of *even so*. Unlike the N400, which reflects the degree of match between discourse predictions and the semantic features of an incoming word, the P600 appears to be triggered by the detection of incoherence that results from a full integration of the critical word into its context (e.g. Kuperberg, 2007; van de Meerendonk et al. 2009; Paczynski and Kuperberg, In press). In this study, we suggest that the larger P600 evoked by incoherent than coherent critical words reflected additional efforts to reinterpret the input and arrive at a veridical representation of discourse meaning.

Importantly, however, there was a difference between the two experiments in whether the P600 was modulated by *even so*: in Experiment 1, when participants carried out explicit coherence judgments, the P600 to *incoherent* critical words was larger in the *even-so* than in the incoherent plain scenarios. We suggested that this was because
participants were more likely to detect discourse incoherence if it conflicted with an alternative predicted representation. In Experiment 2, however, there was no enhancement of the P600 by *even so*. This is not because participants failed to engage in predictive processing at all under *even so*; as discussed earlier, the N400 on critical words in the coherent even-so scenarios was still attenuated in Experiment 2. Rather, we argue that, during more passive comprehension, there was less predictive processing overall and participants were less likely to register discourse incoherence on the critical word itself. The combination of these two factors in Experiment 2 was not enough to reach a threshold for enhancing the P600 on incoherent critical words under *even so*.

Together, these findings indicate that, although a P600 effect can be generated by incoherent critical words, without particularly strong discourse predictions in a passive task, as in the plain incoherent scenarios, it is significantly larger when there is a strong contextually-driven prediction that is clearly flouted by incoherent input in an explicit judgment task (see Kuperberg, 2007).

Costs of maintaining and evaluating discourse input during sentence-final wrap-up

In addition to the ERP effects at the critical word itself, we also observed modulation of a sustained negativity on the sentence-final word—the point at which overall discourse evaluation or ‘wrap-up’ usually takes place. This sentence-final sustained negativity effect was driven by both Coherence and Even-so, and the pattern of modulation across the four experimental conditions was qualitatively similar between the two experiments, although effects were stronger in Experiment 1, probably because an explicit discourse evaluation was a necessary component of its task demands.
The effect of Coherence in the plain scenarios—the larger negativity on the final words of plain incoherent than plain coherent scenarios—is consistent with several previous studies describing a sustained negativity effect on sentence-final words that follow violations of real-world knowledge. In these studies, it has been interpreted as reflecting a mismatch between the meaning of the discourse as a whole and stored real-world knowledge (Hagoort and Brown, 2000; Hagoort, 2003; Ditman, Holcomb and Kuperberg, 2007; De Grauwe, Swain, Holcomb, Ditman and Kuperberg, 2011). What the present study further shows is that a sentence-final Coherence effect can also be seen when discourse coherence is not based on real-world knowledge: final words of the incoherent even-so scenarios evoked a larger sustained negativity than in the coherent even-so scenarios. This indicates that, at the sentence-final position, overall coherence of the incoherent scenarios was evaluated against the alternative world, rather than the real world. In other words, wrap-up costs are not only incurred when overall discourse coherence mismatches stored real-world knowledge, but also when it mismatches an alternative world that is previously set up by the discourse context.

In addition to the effect of Coherence, there was also a main effect of Even-so, with a larger sentence-final negativity in the even-so scenarios than the plain scenarios, regardless of coherence. As discussed above, we suggest that under even so, comprehenders constructed a ‘reverse world’ that was held in working memory, and was used to generate predictions about upcoming information. Here we suggest that, although this led to a facilitation of incoming words, there was some cost to setting up and maintaining this reversed world, even in coherent sentences, and that this cost manifest as a prolonged negativity effect during sentence-final wrap-up.
There was some evidence that the sentence-final negativity effects of Coherence and Even-so were somewhat independent of one another. In both experiments, at posterior regions, they were additive: the negativity effect was largest in the incoherent even-so scenarios, and smallest in the plain coherent scenarios, with the negativities in the incoherent plain and the coherent even-so scenarios falling in between. This additivity supports the idea that the two sentence-final negativities reflected somewhat distinct neurocognitive processes, with the effect Even-so reflecting maintenance of the reverse world within working memory, and the effect of Coherence reflecting an evaluation of the discourse against the real-world (for the plain scenarios) or the reverse world (for the even-so scenarios), as discussed above. At more frontal sites, however, final words of the coherent even-so, incoherent plain, and incoherent even-so scenarios all evoked a negativity effect of the same magnitude (relative to coherent plain scenarios). One possibility is that the more frontal component of the sentence-final negativity reflected a more general engagement of working memory resources that maintained multiple representations before passing them on to posterior sites for coherence evaluation. This would be broadly in line with previous studies reporting sustained anterior negativities in association with holding different types of constituents online within working memory before settling on a final representation of meaning (e.g. King & Kutas, 1995; Kluender and Kutas, M.1993; Nieuwland & van Berkum, 2008a; van Berkum et al., 2003).

General Implications of these findings

During both active coherence judgments as well as more passive reading, even so led to an immediate scalar reversal of the default relationships between the events/states on which comprehenders based their expectations, influencing the semantic processing of
incoming words. As we discuss below, these results have some general implications for when and how we use predictions during discourse processing, as well as for the functional interpretations of the N400 and P600 ERP components.

**Implications for mechanisms of discourse processing**

We have argued that discourse predictions draw not only upon stored long-term knowledge, as has been repeatedly demonstrated in the literature, but also upon discourse-internal information, even when this deviates from the real world. Establishing and maintaining a temporary alternative discourse world within working memory is also necessary for processing fictional or counterfactual discourse, as showed by Nieuwland et al. (2006) and Nieuwland and Martin (2012). However, in these other studies, the alternative world was set up and elaborated by the discourse context itself. In the current study, the parser had no explicit input as to what this alternative world should look like—it was set up solely under the abstract semantic function of *even so*. Understanding how other types of connectives and discourse devices interact with discourse context and real-world knowledge is an important direction for future studies.

We have also highlighted the importance of pragmatic communicative cues in discourse processing. Our findings are consistent with previous studies indicating that pragmatic information can used very quickly to guide online discourse processing (van Berkum 2009; van Berkum et al. 2008; Nieuwland and Kuperberg 2008; Nieuwland and Kuperberg 2010). Here, we have further argued that the pragmatic function of *even so* actually encouraged *more* predictive processing than in the plain scenarios. In previous studies, it has been suggested that a discourse context can only override real-world knowledge if it is sufficiently constraining, as indexed by cloze (see Nieuwland and
Martin 2012 versus Ferguson et al. 2008). Our findings, however, indicate that a highly constrained contextual content is not the only factor to trigger active prediction. Strong communicative cues—in the current case, even so—can put people in a “predictive mode”, even when the specific content itself is not highly constraining. In future research, it will be important to determine whether this effect generalizes beyond even so. Do other concessive connectives show the same effect? And, apart from concessive connectives, are there other pragmatic cues that can also enhance predictive strategies, even in a low-cloze discourse?

Finally, we discussed the idea that distinct “functional spaces” may be temporarily established during discourse processing and held within working memory, allowing for both online predictions as well as sentence-final coherence evaluations. In processing an isolated plain sentence, the parser can constantly evaluate the incremental message-level representation of meaning against activated stored long-term real-world knowledge. When the discourse gets more complex, however, as in the current study, the parser can build up an alternative world, which may be held within a distinct working space. Although establishing such a temporary alternative world can make online processing more efficient (i.e. triggering active prediction and facilitating subsequent processing), there is an ultimate cost associated with its maintenance, as we observed at sentence-final wrap-up. Future studies need to further define what types of processing give rise to such “working spaces”, how they are constrained, and whether individual differences in working memory or additional working memory load can influence the costs incurred through their maintenance.

Implications for the functional relevance of the N400 and P600 ERP components
We have previously argued that the N400 is sensitive to a balance between passive resonance mechanisms, driven by the semantic relatedness between individual words, and more active predictive processing that is based on the message-level meaning of a discourse context (Lau, Holcomb & Kuperberg, under review; Kuperberg et al., 2011; Paczynski and Kuperberg, in press; see also Otten and van Berkum, 2007). The present findings support this idea. In Experiment 1, there was some attenuation of the N400 evoked by the critical word in the plain coherent (versus plain incoherent) scenarios, suggesting that the active task encouraged some prediction, in this case based on real-world knowledge. In Experiment 2, however, with a more passive comprehension task, the amplitude of the N400 was the same between the plain incoherent and coherent scenarios, suggesting that its modulation was driven mainly by semantic relatedness between individual words, which was controlled to be the same between the two conditions (see LSA values in Table 1). In both experiments, however, the N400 evoked by critical words in the even-so coherent scenarios was attenuated relative to the even-so incoherent ones, which, as argued above, we think reflected more active predictive processing.

Our data also reveal an interesting asymmetry between N400 and P600 with respect to how they were modulated by experimental task. Overall, both effects were generally increased in the active judgment versus the passive reading task. However, whereas N400 was attenuated to critical words in the even-so coherent (versus the plain coherent) scenarios in both experimental tasks, the P600 was only enhanced in the even-so incoherent (versus plain incoherent) scenarios when participants were actively judging coherence. This is in line with the idea that N400 directly reflects the match between
semantic features of a word and semantic memory-based predictions, but that the P600 is triggered under a dynamic interaction between semantic memory-based predictions and incoherence detection, rather than either of these mechanisms alone. Most discussion of this dynamic interaction in the literature has been at the level of single sentences (e.g. Kuperberg 2007; see Paczynski & Kuperberg, In Press). Here we suggest that some of the same principles can translate and scale up to the level of whole discourse, and it will be important for future endeavors to explore these ideas further.

Conclusion

To conclude, we have shown that the concessive connective, *even so*, led to a reversal of expectations such that comprehenders were able to use discourse-internal information to override the effects of stored world knowledge. Moreover, comprehenders made maximal use of the pragmatic communicative function of *even so* to enhance predictions about the upcoming events, leading to facilitated processing of incoming words, even when the discourse constraint, indexed by cloze, was relatively low. This benefit of *even so*, however, did not come for free: we also observed at least some 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 integrate new information into an abstract mental model amazingly quickly, even when this model mismatches our real-world experience.
Acknowledgements

We thank Abigail Swain, Eric Fields, Kristina Fanucci for their assistance in data collection, and Ellen Lau for her helpful comments on an early version of the draft.
References


Academic Press.


Metusalem, R., Kutas M., Urbach, T.P., Hare, M., McRae, K. & Elman, J.L. (In press)

Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*.


concepts, Language and Cognitive Processes, 12, 2-3, 137-176.
Metusalem, R., Kutas M., Urbach, T.P., Hare, M., McRae, K. and Elman, J.L. In press.
Generalized event knowledge activation during online sentence comprehension.
*Journal of Memory and Language*.


Nieuwland, M.S. & Martin, A.E. (2012). If the real world were irrelevant, so to speak: The role of propositional truth-value in counterfactual sentence comprehension. *Cognition*, 122, 102-109


Table 1: Example stimuli and characteristics.

<table>
<thead>
<tr>
<th>Scenario Type (n=45 per condition)</th>
<th>Example</th>
<th>SSV of critical word‡</th>
<th>Cloze*</th>
<th>Constraint*</th>
<th>Coherence ratings^</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Coherent</td>
<td>Carol listened to the radio. The storm last night had dumped several feet of snow. The schools were closed that day.</td>
<td>0.179 [0.078]</td>
<td>0.42 [0.32]</td>
<td>0.52 [0.26]</td>
<td>4.8 [0.2]</td>
</tr>
<tr>
<td>2. Incoherent</td>
<td>Carol listened to the radio. The storm last night had only left a dusting of snow. The schools were closed that day.</td>
<td>0.174 [0.079]</td>
<td>0.03 [0.09]</td>
<td>0.40 [0.24]</td>
<td>1.7 [0.4]</td>
</tr>
<tr>
<td>3. Even-so Coherent</td>
<td>Carol listened to the radio. The storm last night had only left a dusting of snow. Even so, the schools were closed that day.</td>
<td>0.174 [0.079]</td>
<td>0.31 [0.28]</td>
<td>0.44 [0.25]</td>
<td>3.3 [1.0]</td>
</tr>
<tr>
<td>4. Even-so Incoherent</td>
<td>Carol listened to the radio. The storm last night had dumped several feet of snow. Even so, the schools were closed that day.</td>
<td>0.179 [0.078]</td>
<td>0.04 [0.11]</td>
<td>0.40 [0.24]</td>
<td>2.4 [1.0]</td>
</tr>
</tbody>
</table>

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.
Figure Captions:

**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.
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.