Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification

Stephen Politzer-Ahles [a], *Robert Fiorentino* [a], *Xiaoming Jiang* [b], *Xiaolin Zhou* [b,c]

[a] Neurolinguistics and Language Processing Laboratory, Department of Linguistics, University of Kansas, 1541 Lilac Lane, Blake Hall, Rm. 427, Lawrence, KS 66044, United States

[b] Center for Brain and Cognitive Sciences, Department of Psychology, Peking University, Beijing 100871, China

[c] Key Laboratory of Machine Perception and Intelligence (Ministry of Education), Peking University, Beijing 100871, China

Address correspondence to

Stephen Politzer-Ahles

Department of Linguistics

University of Kansas

1541 Lilac Lane

Blake Hall, Room 427

Lawrence, KS 66044-3177

Tel.: +1-717-580-7732

Fax: +1-785-864-5724

E-mail: sjpa@ku.edu

Additional e-mail addresses: fiorentino@ku.edu (R. Fiorentino), xmjiang1983@pku.edu.cn (X. Jiang), xz104@pku.edu.cn (X. Zhou)

Abstract

The present study examines the brain-level representation and composition of meaning in scalar quantifiers (e.g., *some*), which have both a semantic meaning (*at least one*) and a pragmatic meaning (*not all*). We adopted a picture-sentence verification design to examine event-related potential (ERP) effects of reading infelicitous quantifiers for which the semantic meaning was correct with respect to the context but the pragmatic meaning was not, compared to quantifiers for which the semantic meaning was inconsistent with the context and no additional pragmatic meaning is available. In the first experiment, only pragmatically inconsistent quantifiers, not semantically inconsistent quantifiers, elicited a sustained posterior negative component. This late negativity contrasts with the N400 effect typically elicited by nouns that are incongruent with their context, suggesting that the recognition of scalar implicature errors elicits a qualitatively different ERP signature than the recognition of lexico-semantic errors. We hypothesize that the sustained negativity reflects cancellation of the pragmatic inference and retrieval of the semantic meaning. In our second experiment, we found that the process of re-interpreting the quantifier was independent from lexico-semantic processing: the N400 elicited by lexico-semantic violations was not modulated by the presence of a pragmatic inconsistency. These findings suggest that inferential pragmatic aspects of meaning are processed using different mechanisms than lexical or combinatorial semantic aspects of meaning, that inferential pragmatic meaning can be realized rapidly, and that the computation of meaning involves continuous negotiation between different aspects of meaning.

Keywords: scalar implicature, pragmatics, ERP, sustained negativity, N400

1. Introduction

Understanding an utterance involves rapidly combining elements of its meaning from multiple sources, including the meanings of individual words (lexical semantics), the relationships between words (compositional semantics), and the relationship between the utterance and the context (pragmatics and discourse). Electrophysiology is particularly useful for investigating the interplay between these different aspects of meaning during language comprehension, since it offers both the high temporal resolution necessary to investigate the time course of meaning composition, and the ability to detect qualitative differences in effects elicited by different types of meaning composition. Many previous neurolinguistic studies examining pragmatic meaning have focused on real-world plausibility (e.g., Kuperberg et al. 2000; Hagoort et al., 2004; Filik & Leuthold, 2008), rather than aspects of meaning based on inferential pragmatics—i.e., meaning based on assumptions about the intentions of the speaker who makes an utterance and the context in which she utters it. The present study aims to investigate how the brain realizes linguistically-motivated distinctions between different aspects of meaning (semantically inherent versus pragmatically constructed) and how these aspects of meaning are composed online.

Information from the wider discourse and pragmatic context is used rapidly during sentence comprehension to make words easier or more difficult to integrate into the utterance meaning (Hagoort & Van Berkum, 2007; Van Berkum, 2009). Pragmatic and discursive information can guide comprehenders' predictions about upcoming words and thus, in event-related potential (ERP) studies, produce modulations in the N400, a negative-going ERP component emerging between about 200 and 500 ms after the presentation of a word and

showing a greater amplitude to words that are less expected and more difficult to retrieve or integrate (Kutas & Federmeier, 2000; Lau et al., 2008; Pylkkänen et al., 2011). Previous studies have shown that discourse context can override semantic constraints, making semantically appropriate but discursively inappropriate words elicit an increased N400, an effect normally elicited by semantically anomalous words (Nieuwland & Van Berkum, 2006; Filik & Leuthold, 2008). Language-external variables like the hearer's personal values or the speaker's gender, age, or class can make words easier or more difficult to retrieve from memory and integrate into a sentence and thus influence the N400 (Van Berkum, 2009) and brain activation in the medial prefrontal cortex (Tesink et al., 2009). N400-like ERP responses to pronouns are affected by the social status of their antecedents (Jiang et al., 2011) and gender stereotypes held by the comprehender (Osterhout et al., 1997). Pragmatic information can also play a role in semantic composition: there is evidence that negatives are not always rapidly integrated into the meaning of infelicitous sentences such as "A robin is not a bird" (Fischler et al., 1983; Wiswede et al., in press; but see Urbach & Kutas, 2010) but that they are when pragmatic context makes the sentence felicitous (Nieuwland & Kuperberg, 2008).

In contrast to these studies examining how pragmatic context influences retrieval and integration of a later word in the sentence, comparatively few have probed for ERP activity directly related to pragmatic inferencing or tested whether this activity is qualitatively distinct from that elicited by semantic retrieval and integration. Pragmatic inferencing may elicit sustained negativities rather than N400s. A sustained negativity known as the Nref, which begins at a latency of about 300ms in response to words with multiple or ambiguous referents as compared to words with unique referents (Van Berkum et al., 2007), has been suggested to be

related to computationally costly inference-making (Van Berkum, 2009). This hypothesis remains to be tested empirically. Crucially, similar sustained negativities have been observed for sentences in which the reader must re-compute a discourse model about whether or not an action was completed (Baggio et al., 2008) or revise a discursive inference that turns out to be incorrect (Pijnacker et al., 2011), although in the latter study the negativity had a more centro-parietal distribution.

In the present study, we examine the role *scalar implicatures* (Doran et al., 2009; Katsos & Cummins, 2010) play during processing. Scalar implicatures offer a clear distinction between semantic and pragmatic interpretations, which makes them ideal for investigation via methods with high temporal resolution like ERPs. Scalar implicature refers to the interpretation of a weak term as meaning that a stronger term is not true. Consider, for instance, the exchange in (1):

1) a. Are all of the students in your department hardworking?

   b. Some of them are.

In this context, because speaker B chose not to say "All of them are", a hearer often interprets the utterance "some of them are" as meaning "*not all of them* are". The interpretation *not all*, however, is not part of the inherent semantics of the quantifier *some*. Rather, it is generated through a pragmatic enrichment process (Noveck & Sperber, 2007), based on a hearer's expectation that a cooperative speaker will use the most informative term possible, and thus that the speaker's choice not to use the stronger term *all* must mean that the stronger term is not true (Grice, 1975; Horn, 1972). On the other hand, the inherent, semantic meaning (also known as the logical meaning) of the quantifier *some* is *at least one*, and could be consistent with *all*. The fact that the pragmatic meaning (*not all*) is context-based and not part of the semantic meaning (*at*

least one) is evident from the fact that this meaning can be revised or cancelled (e.g., in 2a) without resulting in a nonsensical sentence (Rullman & You, 2006; Doran et al., 2012), whereas the semantic meaning cannot (e.g., 2b, which is self-contradictory):

2) a. Some of the students in this department are hardworking. In fact, all of them are.

b. Some of the students in this department are hardworking. In fact, none of them are.

Since scalar implicatures have dissociable semantic and pragmatic meanings, they offer an ideal test case for examining the relationship between pragmatics and semantics online, including the time course of pragmatic inferencing and whether or not pragmatic and semantic processes elicit qualitatively different patterns of brain activity. Below we briefly review previous studies of scalar implicature and what they suggest about the interaction of semantic and pragmatic processing.

1.1 Previous investigations of scalar implicature

A number of recent psycholinguistic studies have investigated the speed at which pragmatic readings of scalar terms become available, the costs engendered by inferencing, and the role of context in scalar implicature processing (see, e.g., Noveck & Posada, 2003; Bott & Noveck, 2004; Feeney et al., 2004; Breheny et al., 2006; De Neys & Schaeken, 2007; Chevallier et al., 2008; Degen, 2009; Bott et al., 2012; Hartshorne & Snedeker, submitted). Many of these studies have used speeded verification or self-paced reading tasks. Response times in such tasks, however, may reflect not only processing related to implicature generation but also controlled decision-making components (Huang & Snedeker, 2009; Nieuwland et al., 2010; Tavano, 2010). This leaves open the question of what occurs before an overt response (or decision to move to

the next word) is made, and how implicature processing unfolds over time. Thus, it is worthwhile to investigate these questions using a methodology that both provides fine-grained temporal resolution and allows the researcher to track different processing stages prior to overt responses. One such methodology is event-related potentials (ERPs). In addition to offering high temporal resolution, ERPs have the potential to probe the extent to which the neural mechanisms of scalar implicature processing differ from those of other aspects of meaning composition, since ERP components may differ in terms of topography, polarity, and morphology, as well as latency (see, e.g., Kutas et al.,2006). To date, only two studies have investigated scalar implicature processing using ERPs.

Noveck and Posada (2003) measured ERPs while participants read and judged *underinformative* sentences such as "Some cats have ears." Such sentences are correct under a semantic interpretation (there do exist dogs that have ears) but incorrect under a pragmatic interpretation (it is not the case that "not all dogs have ears"). At the sentence-final critical word which determines the truth, falsehood, or underinformativeness of the sentence, the investigators found a decreased N400 for underinformative sentences relative to a control condition. The interpretation of this finding is complicated, however, by between-item differences in lexico-semantic relatedness between subjects and objects in their materials, the fact that critical words were not matched for any lexical properties (e.g., frequency), and the possible effect of global wrap-up processes that occur at the end of a sentence (for a review of these concerns, see Nieuwland et al., 2010; for a discussion of sentence wrap-up effects, see Hagoort, 2003). A later study by Nieuwland, Ditman, and Kuperberg (2010, Experiment 1) tested similar sentences while addressing these methodological factors. The authors found that participants with high

pragmatic ability (as measured by performance on the communication subscale of the Autism-Spectrum Quotient questionnaire) showed a greater N400 for underinformative than informative critical words. These results suggest that scalar implicatures can guide expectations about upcoming linguistic input and can override lexico-semantic influences on the N400.

These studies have provided the first insights into how scalar implicatures affect online processing as measured by ERPs. However, some open questions remain regarding the time course and neural instantiation of scalar implicature processing. These studies, like the other N400 studies summarized above, tested whether scalar implicatures can influence the processing of later words in the sentence after the scalar implicature has been computed. As acknowledged by Nieuwland and colleagues, the results of these studies do not "directly reflect full-fledged, online pragmatic inferencing, but rather ... reflect the semantic processing consequences of earlier and relatively implicit pragmatic inferencing" (Nieuwland et al., 2010, p. 341). Because violations in the previous studies only became detectable on words well downstream of the quantifier, these studies cannot make strong claims about how and when the scalar inference is realized. It remains to be seen what pattern of effects may be elicited by processing the scalar implicature itself; this is the question explored in the present study. Furthermore, these studies used critical words that differed lexico-semantically as well as pragmatically, in ways that overrode or may have hidden pragmatics-related ERP effects (e.g., the participants with low pragmatic ability in Nieuwland et al., 2010, whose N400s were driven by lexico-semantic relatedness rather than pragmatics). The truth or underinformativeness of sentences in the previous studies was also based on real-world knowledge, the processing of which may involve mechanisms distinct from the processing of linguistic information (Pylkännen et al., 2011) and

which may introduce various between-condition differences. For these reasons, it is worthwhile to further investigate scalar implicature processing using a design that dissociates semantic and pragmatic aspects of meaning and examines how each is processed. The present study does so by providing visual contexts against which participants judge sentences, as described below.

1.2 The present study

The present study, which was conducted in Mandarin Chinese, adopts a picture-sentence verification design (Wu & Tan, 2009; Tavano, 2010) to compare the neural responses to pragmatically underinformative versus informative sentences that are identical in lexico-semantic content. On each trial a participant is presented with a picture, followed by a sentence that correctly, incorrectly, or underinformatively describes it. Following a picture in which some of the characters are engaging in one activity and others in another (e.g., girls sitting on blankets or on chairs; the upper left portion of Figure 1), a sentence such as "Some of the girls are sitting on blankets" is acceptable, whereas the same sentence is underinformative if it follows a picture in which all of the characters are engaging in the same activity (upper right portion of Figure 1). In this way we strictly control the context in which the sentence is interpreted, keeping lexico-semantic content identical across conditions. Furthermore, inconsistency becomes detectable at the quantifier itself, allowing us to directly examine the response to underinformative quantifiers rather than the downstream effects of expectations generated by pragmatic inferencing.

The experiments reported here were conducted in Mandarin Chinese, whereas previous online studies of scalar implicature have all used western languages. The characteristics of Mandarin scalar implicature, however, are not different from those of English (see Chi, 2000;

Xie, 2003; Tsai, 2004; Rullman & You, 2006; Wu & Tan, 2009). The critical scalar quantifier in the present experiment is *yǒu de* (有的), which is partitive (Xie, 2003; Tsai, 2004) and has a strongly pragmatic interpretation (Wu and Tan's (2009) adult participants reported a pragmatic interpretation of *yǒu de* in 89% of trials). It is roughly equivalent in meaning to the English partitive *some of*, which robustly elicits a pragmatic interpretation (Grodner et al., 2010; Degen & Tanenhaus, 2011).

--- Insert Figure 1 about here ---

In Experiment 1 we factorially manipulate picture type (in *Some*-type pictures, some characters are engaging in one activity and some in another, whereas in *All*-type pictures all characters are engaging in the same activity) and the quantifier used in the sentence (*some of—yǒu de* 有的—versus *all of—suǒyǒu de* 所有的); see Figure 1 for example pictures and sentences. When used in a sentence following an *All*-type picture, the quantifier *some of* is semantically consistent but pragmatically inconsistent with the picture; when used in a sentence following a *Some*-type picture, the quantifier *all of* is semantically inconsistent with the picture (the inconsistency is due to the inherent semantics of *all*, not due to a pragmatically-enriched meaning).[1] Thus, the experiment has a 2 (Quantifier) × 2 (Consistency) design. Crucially, both

[1] Note that, at the position of the quantifier, participants could not be certain whether the inconsistent *all of* sentences were consistent or not with the picture. For instance, if a picture showed some girls sitting on chairs and some sitting on blankets, a sentence beginning "All of…" could be felicitously continued as "All of the girls are wearing hats" or "All of the chairs have girls sitting on them". A similar possibility exists for the *some of* sentences; for instance, a picture showing a group of girls all sitting on chairs could be felicitously continued as "Some of the

inconsistent conditions are compared with lexically matched controls: *some of* following a *Some*-type picture formed the control for the inconsistent *some of* condition, and *all of* following an *All*-type picture formed the control for the inconsistent *all of* condition. In this design, after seeing a picture the participant can form an expectation about the upcoming quantifier—in other words, she can verbally pre-encode the sets as *Some*-type or *All*-type sets (Huang, Hahn, & Snedeker, 2010; Hartshorne & Snedeker, submitted). Thus, both inconsistent *some of* and inconsistent *all of* are words that are unexpected in their context. Including the *all of* conditions allows us to examine the pragmatically inconsistent *some of* condition for effects that are unique to pragmatic processing, above and beyond the effect of seeing an unexpected word.

In Experiment 2, we test whether inferential processes involved in comprehending an underinformative sentence interact with lexico-semantic processes, by factorially manipulating the presence or absence of a pragmatic violation early in the sentence with the presence or absence of a lexico-semantic violation on a content word later in the sentence. We do so by using the same picture-sentence verification design as in Experiment 1, and additionally manipulating the lexical consistency between the picture and the sentence: lexically inconsistent sentences have objects (downstream of the quantifier) that do not match any of the objects portrayed in the preceding picture. Thus, Experiment 2 has a 2 (Pragmatic Consistency) × 2 (Lexical Consistency) design, in which sentences are lexically identical across conditions but the pictures preceding the sentences vary.

2. Results

girls are happy". None of these sentence types was included in the experiment; mismatches between picture and quantifier always led to sentences that were ultimately inconsistent.

2.1 Experiment 1

2.1.1 Behavioral results

Participants responded both to comprehension questions irrelevant to the interpretation of the quantifier and to acceptability judgment prompts during the course of the experiment. Behavioral data from one participant were lost due to a data logging error, leaving eighteen participants for the behavioral data analysis. In the comprehension task, mean accuracy rates were 86.1% for the pragmatically infelicitous condition (*some of* sentence following an *All*-type picture), 77.5% for consistent "some", 82.8% for semantically inconsistent (*all of* sentence following a *Some*-type picture), and 78.2% for consistent "all". A repeated measures ANOVA revealed no significant differences in mean accuracy across conditions ($F(3, 51) < 1$).

Acceptability judgments on the pragmatically underinformative sentences have no correct or incorrect answer, given that participants can interpret such sentences semantically or pragmatically. Across participants, 39.8% of pragmatically underinformative sentences were judged as correct, indicating a semantic judgment; in comparison, only 19.6% of semantically inconsistent sentences were judged as correct. The difference was significant by participants ($t(17) = -4.47$, $p < .001$), indicating that participants accepted pragmatically infelicitous sentences more often than semantically inconsistent sentences. As for the remaining conditions, which do have clear expected judgments, mean accuracy rates were 78.7% for the consistent "some" condition, 80.4% for the semantically inconsistent condition, and 85.5% for the consistent "all" condition. A repeated measures ANOVA revealed no significant differences across conditions ($F(2, 34) < 1$).

Several previous studies have distinguished pragmatic and semantic responders (Noveck & Posada, 2003; Bott & Noveck, 2004; Tavano, 2010). Thus, we divided our participants into groups using the following criteria: participants who made 5 or more semantic responses (1 or fewer pragmatic responses—each participant judged 6 underinformative sentences, see section 4.1.3) to the underinformative trials were classified as semantic responders, those who made 5 or more pragmatic responses (1 or fewer semantic responses) were classified as pragmatic responders, and those who made 2 to 4 semantic responses (no more than 4 responses of a given type) were classified as inconsistent responders. Five participants met the criteria to be considered semantic responders, while two were pragmatic responders and eleven inconsistent; there were not enough consistent responders to form participant groups for the ERP analysis.[2] There was a greater number of inconsistent responders in the present study than in some previous studies (Noveck & Posada, 2003; Tavano, 2010), which is consistent with Feeney and colleagues (2004), who found that participants tended to respond inconsistently to underinformative sentences when the variety of stimulus conditions is large (see the Experimental procedures for more information about the conditions included in the present experiment).

2.1.2 ERP results

--- Insert Figure 2 about here ---

--- Insert Figure 3 about here ---

[2] Using slightly more lax criteria (4 or more semantic responses for semantic responders, 4 or more pragmatic responses [2 or fewer semantic responses] for pragmatic responders, and 3 semantic [3 pragmatic] responses for inconsistent responders), 8 responders were classified as semantic responders, 3 as pragmatic, and 7 as inconsistent.

Visual inspection of the waveforms (Figures 2 and 3) suggests that semantically inconsistent *all of* elicited a less negative ERP than consistent *all of* from about 200 to 500 ms in the anterior and central regions, whereas pragmatically inconsistent *some of* elicited a sustained negative ERP compared to consistent *some of* in the right posterior regions. Thus, we conducted ANOVAs on the mean ERP amplitudes for the 200-500 ms and 500-1000 ms time windows; the omnibus ANOVA results are shown in Table 1.

--- Insert Table 1 about here ---

2.1.2.1 200-500 ms

The ANOVA revealed a significant interaction of Consistency and Region.[3] We resolved the interaction by testing the effect of Consistency at each region. Both types of inconsistent quantifier elicited significantly more positive ERPs than controls in the left anterior region ($F(1,18) = 4.52$, $p = .048$), marginally more positive in the midline anterior ($F(1,18) = 3.91$, $p = .063$) and left central ($F(1,18) = 3.21$, $p = .090$) regions, and marginally more negative ERPs in the right posterior region ($F(1,18) = 4.08$, $p = .059$); the simple effect of consistency did not reach significance in any other region (ps > .143).[4]

[3] There were also effects of Quantifier by Region in this time window and of Quantifier in the later time window. These, however, are not of theoretical interest since they involve direct comparison between different words, and thus we do not discuss them here. The significant main effects of Region are also not discussed since they do not reveal differences based on the experimental manipulation.

[4] Visual inspection of the waveforms and topographic plots (Figures 2 and 3) suggests that the posterior negativity revealed in the Consistency by Region interaction was due to the pragmatically inconsistent quantifiers, whereas the anterior positivity was present in both conditions—i.e., that semantically inconsistent quantifiers elicited an anterior

2.1.2.2 500-1000 ms

In the later time window there was a significant interaction of Quantifier, Consistency, and Region in the lateral ANOVA only. Resolving the interaction by Quantifier revealed that pragmatically inconsistent quantifiers elicited both a significant main effect of Consistency ($F(1,18) = 4.56$, $p = .047$) and a Consistency by Region interaction ($F(5,90) = 3.07$, $p = .039$), but neither an interaction nor a main effect of Consistency was observed for the semantically inconsistent quantifiers (Fs < 1). For the pragmatically inconsistent quantifiers, the main effect of Consistency was due to a more negative ERP for inconsistent than consistent quantifiers in this time window, and the interaction with Region was due to the fact that the simple effect of Consistency for *some of* reached significance at the right central ($F(1,18) = 7.09$, $p = .016$) and right posterior ($F(1,18) = 11.63$, $p = .003$) regions, but not at other regions (ps > .108).

2.1.3 Discussion

This experiment tested whether the pragmatic meaning of a scalar quantifier affects processing immediately when the quantifier itself is read, and how the detection of pragmatic implicature violations is manifested electrophysiologically when lexico-semantic differences are controlled for. Both quantifiers that were semantically inconsistent with a context and those that were pragmatically inconsistent elicited a less negative anterior ERP than controls in an earlier (200-500 ms) time window. This early effect indicates that the pragmatic interpretation of the

positivity only, whereas pragmatically inconsistent quantifiers elicited both an anterior positivity and posterior negativity. However, the interaction of Quantifier and Consistency in the omnibus ANOVA did not reach significance (see Table 1), providing no evidence for differential ERP responses to semantic and pragmatic inconsistencies in this time window.

scalar quantifier was used rapidly during processing, since the quantifier was only inconsistent with its context when interpreted pragmatically; this effect was not unique to scalar implicature processing, however, as it was also elicited by unexpected, semantically inconsistent quantifiers. Effects unique to scalar implicature processing were observed later in the epoch (500-1000 ms), at which time pragmatically inconsistent but not semantically inconsistent quantifiers elicited a sustained posterior negativity. While this negativity also appeared earlier in the epoch with a topography similar to an N400 effect, it is apparent from the waveforms that the effect is more likely the beginning of a sustained negativity lasting throughout the epoch; we note that Pijnacker and colleagues (2011) also found a dissociation between a transient N400 elicited by lexico-semantic violations, and a more long-lasting negativity elicited by discourse processing.

In experimental contexts like ours, rapid effects of pragmatic inconsistency could be due to participants' ability to verbally pre-encode the picture contexts as *Some*-type or *All*-type contexts, and then make a forward prediction about the quantifier that will appear in the sentence (Huang et al., 2010; Hartshorne & Snedeker, submitted). Indeed, the presence of an early effect is not surprising, as previous research has already shown that pragmatic expectations about upcoming words can modulate ERPs as early as the N400 (Van Berkum, 2009; Nieuwland et al., 2010). However, it is unlikely that the results of the present experiment are due only to effects of seeing an unexpected word. First of all, unexpected linguistic input typically elicits a N400 or P300/P600 effect (Lau et al., 2008; Bornkessel-Schlesewsky et al., 2011), whereas the topography and polarity of the early effect in the present experiment was consistent with neither of these. Rather, the effect is consistent in timing and topography with the Nref (Van Berkum et al., 2007); the smaller negativity for inconsistent quantifiers may reflect a decrease in effort

made to link *all of* or *some of* with an antecedent when the participant recognizes it to be pragmatically or semantically inconsistent with the context. More importantly, if participants were making predictions based on verbal pre-encoding, then *all of* and *some of* would both be unexpected; nevertheless *some of* elicited a qualitatively different effect later in the epoch. Sustained negativities have also been observed on sentences in which readers must re-compute a discourse model or revise a discursive inference (Baggio et al., 2008; Pijnacker et al., 2011). Thus, we propose that the sustained negativity in the present experiment reflects revision of the reader's interpretation of the quantifier's meaning (inhibition of the pragmatic reading and retrieval of the semantic reading) after the reader realizes that the pragmatic reading is inappropriate.

While psycholinguistic models assert that realizing and/or cancelling a pragmatic inference may involve processing costs (Katsos & Cummins, 2010, Hartshorne & Snedeker, submitted; see also Garrett & Harnish, 2007), they do not yet articulate precisely what sort of costs or mechanisms this computation entails (see Bott et al., 2011, for further discussion). Thus, in the next experiment we examine whether the canceling or suppression of a pragmatic inference (reflected by the sustained negativity in Experiment 1) interacts with semantic processing by modulating basic lexical processes downstream. We factorially manipulate the presence of a lexico-semantic violation (i.e., a sentence object that does not match the objects in the picture) and the felicity of the quantifier *some of* upstream of the violation; example pictures and sentences for Experiment 2 are shown in Figure 4. For example, the sentence "Some of the girls are sitting on blankets suntanning" is pragmatically and lexically correct when preceded by a sentence in which some girls are sitting on blankets and some sitting instead on couches

(depicted in the upper-left portion of Figure 4). The same sentence is pragmatically correct but lexically incorrect when none of the girls are sitting on blankets but not all the girls are sitting on the same thing (upper-right portion). The sentence is pragmatically incorrect but lexically correct when in fact all of the girls are sitting on blankets (lower-left portion). Finally, when all the girls are sitting on the same thing and that thing is not a blanket, the sentence is both pragmatically and lexico-semantically incorrect (lower-right portion), allowing us to examine how the neural response to lexico-semantic inconsistency at the object position interacts with the processing of the pragmatic inconsistency previously instantiated at the quantifier position.

--- Insert Figure 4 about here ---

Lexico-semantic picture-sentence mismatches have been shown to elicit robust N400s (Knoerfle et al., 2011). If the ongoing pragmatic revision process after encountering an infelicitous quantifier affects lexico-semantic processing, either by limiting the extent to which the parser commits to predictions about upcoming material or by using the same processing resources that would otherwise be used for lexico-semantic prediction and integration, then the N400 effect for lexico-semantic violations at the object position should be modulated. For instance, Panizza and colleagues (2011) found that participants in a visual world eye-tracking experiment were slower to look to an unambiguous target (e.g., slower to look to socks after *socks* had already been heard) when the target word was in a scalar implicature-supporting context than when it was not, suggesting that generating a scalar implicature may have interfered with their ability to use disambiguating phonological information for lexico-semantic integration.

In a similar vein, our Experiment 2 tests whether revising an underinformative scalar inference interferes with lexico-semantic integration between the picture and the sentential object. We also include the same Quantifier by Consistency manipulation at the quantifier position as we had in Experiment 1, in order to test whether the effect obtained in that experiment would be replicated. (The pragmatically inconsistent "some" and correct "some" conditions were included in the critical items; items corresponding to the semantically inconsistent "all" and correct "all" of Experiment 1 were included in the fillers for this experiment.) While the primary motivation for Experiment 2 was to examine the interaction of pragmatic and lexical processing rather than effects of modality, we found that auditory presentation of sentences was both comfortable for participants and reduced the duration of each trial. For this reason, sentence stimuli were presented auditorily rather than visually in Experiment 2.

2.2 Experiment 2

2.2.1 Behavioral results

The participants' task was to rate the consistency between the picture and the sentence using a 7-point scale. Average ratings were 6.3 for correct *some of* and 6.2 for correct *all of* sentences, 5.4 for pragmatic violations, 3.2 for lexical violations, 2.3 for double violations and 2.0 for semantically incorrect *all of* sentences. A repeated measures ANOVA on the four critical conditions (correct *some of*, pragmatic violation, lexical violation, and double violation) revealed significant effects of Pragmatic Consistency ($F(1,18) = 29.11$, $p < .001$) and of Lexical Consistency ($F(1,18) = 206.68$, $p < .001$), but no significant interaction ($F(1,18) = .03$, $p = .862$). Furthermore, pairwise *t*-tests between all six conditions, with the two-tailed alpha level

Bonferroni-corrected to $\alpha = .003$, revealed significant differences for every comparison except correct *some of* vs. correct *all of* ($p > .5$) and the double violation vs. semantically incorrect *all of* ($p = .32$). These results indicate that participants treated correct sentences, pragmatic violations, picture-sentence mismatches, and double violations as decreasingly acceptable, but they did not differentiate between the two correct conditions or between double violations (with both pragmatic violation and picture-sentence mismatch) and semantically incorrect "all". Because the present experiment used a gradient rating task rather than a categorical judgment task, it was not possible to classify participants as pragmatic or semantic responders using the same criteria as in Experiment 1 or previous studies (Noveck & Posada, 2003; Bott & Noveck, 2004; Feeney et al., 2004).[5]

--- Insert Figure 5 about here ---

2.2.2 ERP results

The waveforms time-locked to the quantifier position (Figure 5) show a sustained negativity for pragmatically inconsistent quantifiers, similar to the one obtained in Experiment 1

[5] Nevertheless, we assessed the number of pragmatic responders using one-tailed independent samples *t*-tests for each participant comparing ratings for pragmatic violations against ratings for correct sentences. Twelve participants reliably rated pragmatically inconsistent sentences lower than correct sentences (ps $< .05$), whereas eight did not. The former group may be considered pragmatic responders (those who interpreted *some* as meaning *not all*), whereas the latter group may be either semantic responders (those who interpreted *some* as meaning *at least one*) or inconsistent responders. Compared to the acceptability judgment task used in Experiment 1, in which most participants were inconsistent, the consistency rating in Experiment 2 yielded a greater number of pragmatic responders. Comparisons of the ERP responses between different groups of responders, however, are beyond the scope of this paper and therefore are not reported here.

but broader in distribution, and a sustained positivity for semantically inconsistent quantifiers. At the object position (Figure 6), both picture-sentence mismatches and double violations elicited broadly-distributed negativities from about 200-600 ms, whereas both types of objects following pragmatically inconsistent quantifiers elicited a sustained negativity from about 400-1000 ms. In this time window the sustained negativity appeared to be present for the objects following pragmatic violations and for the double violations, but not for the picture-sentence mismatches. These patterns of effects are examined statistically below; the omnibus ANOVA results for the quantifier and object positions are presented in Tables 2 and 3, respectively.

--- Insert Table 2 about here ---

--- Insert Figure 6 about here ---

2.2.2.1 Quantifier position

We quantified the effects of pragmatic and semantic inconsistency using the mean ERP amplitudes over the 300-1000 ms window. There was a significant interaction of Quantifier and Consistency, reflecting the fact that inconsistent *some of* elicited a negativity (lateral: $F(1,19) = 8.03$, $p = .011$; midline: $F(1,17) = 3.59$, $p = .073$) whereas inconsistent *all of* elicited a positivity (lateral: $F(1,19) = 7.72$, $p = .012$; midline: $F(1,17) = 5.63$, $p = .028$). There was also a significant interaction of Quantifier, Consistency, and Region in the lateral ANOVA only. The interaction was due to the fact that the negativity for the *some of* sentences was broadly distributed (the Consistency by Region interaction for *some of* did not reach significance, $F(5,95) < 1$), whereas the positivity for the *all of* sentences was somewhat left-posterior in distribution. Specifically, for

semantically inconsistent *all of* sentences, the Consistency by Region interaction was significant ($F(5,95) = 2.80$, $p = .033$); the simple effect of semantic Consistency was significant in the left posterior ($p = .001$), right posterior ($p = .005$), and left central ($p = .046$) regions, and marginal in the left anterior ($p = .063$) and right central ($p = .054$) regions.

--- Insert Table 3 about here ---

--- Insert Figure 7 about here ---

2.2.2.2 Object position

2.2.2.2.1 N400

We quantified the N400 using mean amplitudes in the 200-500 ms time window. In this window we observed a highly significant effect of Lexical Consistency, reflecting the fact that both lexically inconsistent conditions (picture-sentence mismatch and double violation) elicited more negative ERPs than lexically consistent conditions (correct object, and correct object following a pragmatically inconsistent quantifier). The effect was broadly distributed (it did not interact significantly with Region). The effect of Pragmatic Consistency was not significant. Crucially, no interactions of Pragmatic Consistency and Lexical Consistency were significant, indicating that the presence of a pragmatic violation did not modulate the lexico-semantic N400 effect.

2.2.2.2.2 Late negativity

We quantified the late ERP effect using the mean amplitudes in the 500-1000 ms window. In this window there was a significant main effect of Pragmatic Consistency, indicating that objects following pragmatic violations elicited more negative ERPs in the late window.

In the lateral ANOVA there was a marginal interaction between Pragmatic Consistency, Lexical Consistency, and Region, due to the fact that although the main effect of pragmatic inconsistency was significant for both lexically correct (i.e., correct objects following pragmatically inconsistent quantifiers) and lexically incorrect (i.e., double violations) sentences, it was somewhat broadly distributed for lexically correct sentences (the interaction of Pragmatic Consistency and Region did not reach significance in these sentences, $F(5,95) = 1.20$, $p = .320$), but was more limited to the anterior regions for the double violations. Specifically, for the double violations, the interaction of Pragmatic Consistency and Region was marginally significant ($F(5,95) = 2.23$, $p = .095$), and the Pragmatic Consistency effect was significant or marginal in the left anterior ($p = .017$), right anterior ($p = .034$), and left central region ($p = .070$), but not significant in the right central, left posterior, or right posterior regions ($ps > .190$).

To investigate whether the topographical difference was likely to be due to qualitatively different underlying sources or to quantitative differences in the signal, we performed a scaling analysis (Jing et al., 2006), which tests whether the signal in one effect has the same topography as the signal in another effect after being scaled based on a hypothetical scaling factor that represents the change in signal that would occur from a quantitative change in the strength of the underlying source. In this analysis, in which we directly compared the pragmatic effects for the double violation (formed by subtracting the ERP for the mismatching object condition from the

ERP for the double violation condition) and the pragmatic violation (subtracting the correct condition from the pragmatic violation), the interactions with region were not significant ($F(5,95) = 1.60$, $p = .204$; $F(5,95) = 1.85$, $p = .147$),[6] indicating that the topographic differences found in the raw analysis are not likely to result from different underlying generators.

2.2.3 Discussion

At the quantifier position, we partially replicated the finding of Experiment 1: pragmatic violations elicited a sustained negativity, albeit broader in distribution than the effect in Experiment 1. Unlike Experiment 1, semantically and pragmatically inconsistent quantifiers did not elicit similar effects in any time window; also unlike Experiment 1, we observed a sustained positivity for the semantically inconsistent quantifiers. The primary differences between the experiments were stimulus presentation modality (auditory in Experiment 2, visual in Experiment 1), task (consistency rating in Experiment 2, correctness judgments and comprehension questions in experiment 1), and composition of other sentences in the experiment (in particular, Experiment 1 did not include sentences with both pragmatic and lexico-semantic violations). Importantly, in both experiments semantically inconsistent quantifiers elicited a qualitatively different ERP pattern than the pragmatically inconsistent quantifiers, which provides evidence that the sustained negativity for pragmatically inconsistent quantifiers does not reflect a general reanalysis mechanism or a general response to unexpected input, but rather a process specific to the kinds of revision or inhibition processes that are necessary for

[6] In the procedure proposed by Jing and colleagues (2006), it is recommended to perform two comparisons: one between the raw Condition 1 and the scaled Condition 2, and one between the scaled Condition 1 and the raw Condition 2. Therefore, two *F*-tests are reported here.

revising/inhibiting the pragmatic interpretation of a quantifier and activating its semantic meaning.

The ERPs elicited at the object position showed evidence that pragmatic and lexico-semantic information were processed independently: the presence or absence of a pragmatic violation upstream did not modulate the lexico-semantic N400 effect elicited by picture-sentence mismatch. The lack of an interaction cannot be explained by assuming that pragmatic revision had already been completed by the time the object was heard, since the objects still elicited a sustained negativity associated with pragmatic revision. Rather, the finding suggests that the revision or inhibition of the pragmatic interpretation of scalar terms utilizes different processing resources than those used for lexico-semantic prediction and integration. The late time window on the ERPs time-locked to objects continued to show a sustained negativity in response to pragmatically inconsistent sentences, suggesting that pragmatic revision was not yet completed by the time the object was encountered (which was, on average, 1300 ms after the onset of the quantifier). Thus, our data seem to suggest that pragmatic and semantic aspects of meaning were processed in parallel and their respective effects were additive.

3. General discussion

The two experiments reported here examined the neural responses to pragmatic violations while controlling for lexico-semantic factors and allowing us to detect effects at the moment the critical quantifier is encountered. Perhaps most importantly, we found different ERP patterns for pragmatic and semantic violations: whereas lexico-semantic violations elicited an N400 and quantificational semantic violations elicited positivities, pragmatic violations consistently

elicited sustained negative components. The results suggest that the pragmatic reading of the quantifier is used rapidly during online processing and must be inhibited effortfully if it is inconsistent with the context. We also examined the interaction between pragmatic and lexico-semantic processing and found that pragmatic reanalysis did not modulate lexico-semantic processing downstream, suggesting that pragmatic and lexico-semantic aspects of meaning were processed independently. Below, we discuss each of these findings in turn.

3.1 The sustained negativity

At the quantifier position, in both experiments a sustained negativity was elicited by quantifiers that are pragmatically inconsistent with a context. This effect seems to be related to pragmatic processing in particular, as it was not elicited by quantifiers that were semantically inconsistent with a context. The effect could not be due only to processes related to seeing or hearing an unexpected word, since semantically inconsistent quantifiers and lexico-semantically inconsistent objects elicited qualitatively different effects even though they were also unexpected. The effect could also could not be due to revising expectations about what aspect of the picture will be pointed out later in the sentence, since this sort of revision is also possible in the semantically inconsistent *all of* sentences but did not elicit a sustained negativity. It is not likely to be due to generating or retrieving the pragmatic interpretation of the quantifier, since that process may have already been initiated during verbal pre-coding when the participant viewed each picture (Huang et al., 2010; Hartshorne & Snedeker, submitted). Rather, the sustained negativity is more likely to be related to effortful pragmatic reanalysis: specifically, inhibiting the pragmatic reading of *some of* and retrieving the semantic reading. This interpretation is

consistent with several recent studies (Baggio et al., 2008; Pijnacker et al., 2011) that have observed sustained negativities related to revising discourse models or discourse-based inferences. Further support for this interpretation comes from a study by Leuthold and colleagues (2012), who observed a sustained right-posterior negativity (and corresponding left-frontal positivity) in response to emotion words that were incongruent with a situation previously described (e.g., "The golf pro was distraught", after a context suggesting that the golf pro had a good chance to win a tournament). They speculated that this negativity may be due to suppressing the expected emotion words. It is possible that such an operation also involves reconsideration of the character's point of view, which is a hallmark of Gricean pragmatic processing. While the linguistic manipulation in the present study is different than those discussed above, pragmatic violations in the present study would have led participants to reanalyze the implicature-based meaning of *some*, similar to Pijnacker et al. (2011), and to re-consider the point of view of another speaker or character, as in Leuthold et al. (2012).

It should be noted that an alternate strategy participants could employ to interpret sentences with inconsistent quantifiers is to make no attempt to evaluate the meaning and reference of the quantifier whatsoever until more information becomes available later in the sentence. Recall that semantic violations consisted of *Some*-type pictures (e.g., several girls sitting on blankets and the rest sitting on chairs) followed by sentences beginning "In the picture, all…". Such a sentence could turn out to be correct (e.g., "…all the girls are wearing bathing suits"), and thus it is possible that participants waited until they had more information before attempting to further evaluate the consistency between the sentence and the picture. Crucially, however, pragmatically inconsistent quantifiers could also be followed by sentences that turn out

to be correct (e.g., an *All*-type picture could be followed by "…some of the girls are happy"). If participants employed such a processing strategy, we might expect effects to appear at the verb or object position, where the semantically incorrect sentences become unambiguously incorrect (e.g., at "…all the girls are sitting on…" or "…some of the girls are sitting on…", it becomes impossible to analyze the sentence as "…all the girls are wearing bathing suits" or "…some of the girls are happy"). Because the structure of the verbs used in the present study varied (verbs were presented simultaneously with aspect markers that preceded or followed them and differed in length and other properties) as did the point where the violation becomes unambiguous, such an analysis was not feasible with the present data, although the sustained negativity elicited by objects following pragmatic inconsistencies in Experiment 2 may be evidence for this sort of processing. Crucially, however, participants showed different ERP responses to the two types of inconsistency, even though this delayed interpretation strategy is available for both. Only the pragmatically inconsistent quantifiers can be reconciled with the context by reanalyzing the meaning of the quantifier (cancelling the implicature and retrieving the semantic meaning), and accordingly only the pragmatically inconsistent quantifiers showed the sustained negativity.

An alternative account of the sustained negativity observed in the present study is that it reflects truth-verificational processes initiated by the inconsistency between the quantifier and the context. Wiswede and colleagues (in press) found a late negativity elicited by nouns that make sentences untrue (e.g., "Africa is a planet"), but this negativity only occurred for participants who were performing a truth-value judgment task, not those who were performing a memory task. One might argue that pragmatically inconsistent *some of* in our study initiated this truth-verificational process, whereas semantically inconsistent *all of* did not since its

interpretation could be delayed until later in the sentence. Other aspects of our results, however, speak against this interpretation. In particular, no late negativity was elicited by objects that mismatched only the lexico-semantic content of the picture (e.g., the pure picture-sentence mismatch condition in Experiment 2, which only elicited an N400, as did the lexico-semantically mismatched objects in the Experiment 1 fillers in an exploratory analysis). Such words also introduce falsehood into the sentence, and are more similar to the words that elicited the late negativity in Wiswede and colleagues' (in press) study. Nevertheless, the sustained negativity in our study only occurred in conditions where the inconsistency was related to pragmatic meaning.

The fact that the responses to the pragmatic condition were characterized by early recognition of the inconsistency and revision of the inference has implications for both the theory of scalar implicature processing and for the cognitive neuroscience of language; these implications are discussed below.

3.2 The costs of scalar implicature processing

The present study was not designed to test the time course and processing costs of generating a pragmatic meaning (such a study would have to compare sentences in which *some of* will ultimately be interpreted pragmatically against those in which *some of* will be interpreted semantically, as in Breheny et al., 2006, and Hartshorne & Snedeker, submitted), but it does provide evidence about the time course and costs of adjudicating between the semantic and pragmatic readings. As noted above, the sustained negativity effect at the quantifier position for conditions in which the pragmatic reading of the quantifier was inconsistent with the context suggests that suppressing that aspect of meaning and accessing the semantic aspect was costly

and effortful. Data from a follow-up ERP experiment (Politzer-Ahles et al., forthcoming) further suggest that the sustained negativity appears mainly in participants who are poor at distinguishing between the pragmatic and semantic readings, as evaluated by an independent task; retrieving the semantic reading may take more effort for these participants, making the sustained negativity more prominent. Feeney and colleagues (2004), based on findings from a speeded verification task, also concluded that participants reading underinformative instances of *some* needed to suppress the pragmatic meaning and that this suppression is cognitively taxing. Garrett & Harnish (2007) provide evidence from another pragmatic phenomenon, *standardization implicitures* (e.g., "I've had breakfast" is interpreted as "I've had breakfast today"), that the pragmatically enriched reading is computed by default and the semantic reading can only be retrieved with effort—although we note that it is not necessarily the case that standardization-based implicitures are processed via the same mechanisms as scalar implicatures (see also Bezuidenhout & Cutting, 2002). On the other hand, a recent study in Mandarin suggests that the retrieval of the literal meanings of conventionalized lexical metaphors are not delayed relative to their metaphorical meanings (Lu & Zhang, 2012), raising the interesting possibility that pragmatic inferencing (at least scalar inference triggered by quantifiers) unfolds in a different manner than metaphor comprehension.

In sum, our results suggest that accessing the semantic reading of a scalar quantifier takes extra cognitive effort, eliciting a sustained negativity in the ERP. This is easy to reconcile with default models of scalar implicature processing (Levinson, 2000), which assume that implicatures are generated quickly and with little regard for whether the enriched pragmatic meaning makes the sentence more informative, and subsequently can only be revised or inhibited

with effort. It does not, however, preclude context-driven (Noveck & Sperber, 2007) or constraint-based models (Degen & Tanenhaus, 2011), since the possibility of verbal pre-encoding of our stimuli should have made the pragmatic reading easy to generate rapidly, and these models do not necessarily predict inhibition of pragmatic meaning to be effortless. Further study of the processing costs associated with both scalar implicature generation and scalar implicature reanalysis is needed to elucidate which cognitive resources are used for pragmatic processing and allow these models to become more explicit about this issue.

3.3 Neural correlates of different aspects of meaning

Much work on the processing of meaning in the brain has focused on the N400 ERP component and its sensitivity to manipulations of real-world plausibility (e.g., sentences such as "She spread her bread with socks"). Substantially fewer studies have examined how the brain processes compositional aspects of meaning (for reviews see Pylkkänen et al., 2011; Panizza, 2012) and how context and discourse interact with meaning (see Van Berkum, 2009). Scalar implicatures offer a promising test case for these issues, given that they represent an aspect of meaning that is composed in concert with semantic meaning and that the generation of scalar implicatures is strongly affected by context and expectations about speakers.

The present study offers converging evidence with other emerging work in neurosemantics suggesting that the mechanisms by which the brain composes meaning may not be the same as those by which it accesses words from the lexicon, notices associations between words, or evaluates real-world plausibility (i.e., several of the processes reflected by the N400). Recent investigations suggest that the patterns of brain activation elicited by violations of real-

world plausibility are not the same as those elicited by linguistically-motivated abstract operations such as semantic composition (Pylkkänen et al., 2011), licensing of negative polarity items (Steinhauer et al., 2010; Panizza, 2012) and semantic subcategorization (Kuperberg et al., 2000). In our experiments we found that quantifiers which were pragmatically inconsistent with a context elicited a qualitatively different ERP response than quantifiers which were semantically inconsistent, suggesting that they were processed by different mechanisms. We also found that costly pragmatic reanalysis of a quantifier's meaning did not modulate concurrent processing of lexico-semantic errors, providing further evidence that the processing of these two aspects of meaning are processed independently. We note, however, that while the qualitative differences in ERP responses found in the present study are consistent with distinct mechanisms of pragmatic and semantic meaning composition, it is difficult to infer the underlying sources of the ERP pattern. For this reason, localizing the neural generators of these effects using methods with high spatial resolution would be a valuable avenue for further research, and could provide additional evidence for a dissociation of pragmatic and combinatorial semantic meaning composition.

3.4 Conclusion

The present study examined responses to underinformative scalar quantifiers, manipulating pragmatic informativeness independently of lexico-semantic correctness and real-world knowledge. We found that pragmatic violations elicited qualitatively different ERP effects than lexico-semantic and quantificational semantic violations. The pragmatic meaning of the quantifier influenced processing rapidly and was later revised or inhibited if necessary. The electrophysiological response to underinformativeness observed at the quantifier position was

not due to lexico-semantic factors. Costly pragmatic revision of the quantifier's meaning continued through later portions of the sentence but remained independent from lexico-semantic processing, which unfolded in parallel. The results of the study suggest that the brain both integrates pragmatic and semantic aspects of quantifier meaning rapidly, and continually negotiates and updates these aspects of meaning if necessary. The present work both extends the cross-linguistic coverage of research on the online processing of scalar implicatures, and offers a novel approach to investigating the instantiation of scalar implicatures at the brain level.

4. Experimental procedures

4.1 Experiment 1

4.1.1 Participants

Data were collected from 23 right-handed Mandarin native speakers (10 females, age range 18-27, mean 20.8) from mainland China who were students at the University of Kansas. Four of these participants were excluded from the statistical analysis because of excessive artifacts in their recordings. All participants had normal or corrected-to-normal vision and were right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971). All participants provided their informed consent and received payment, and all methods for the study were approved by the Human Subjects Committee of Lawrence at the University of Kansas.

4.1.2 Materials

One hundred sixty sets of picture arrays were created for the critical trials (see Figure 1 for an example set). Each picture array included three to five actors or items. In the *All*-type picture array from each set, all of the actors were interacting with identical objects (for instance, four girls were all sitting on blankets, or five baskets were all holding pumpkins). In the *Some*-type picture array from each set, a subset of the actors was interacting with one type of object, and the rest were interacting with a different type of object (for instance, some girls were sitting on blankets and some on sofas, or some baskets were holding pumpkins and some holding bananas). The placement of the actors within the image and the relative locations of actors with different items in the *Some*-type pictures were allowed to vary randomly across sets. All picture arrays were black-and-white cartoons or line drawings, sized 1024 × 768 pixels, and with minimally complex backgrounds. Care was taken to limit pictures to those portraying plausible events. The base materials for the pictures were taken from freely available clipart from two published databases (Bonin et al., 2003; Szekely et al., 2004) and Google Images, and further edited using Adobe Photoshop, the GNU Image Manipulation Program, and Microsoft Paint by two paid graphic arts students from Peking University and the first author.

For each set of picture arrays, *some of* and *all of* sentences were written to match the *All*- and *Some*-type arrays (see Figure 1). Each sentence began with "图片里" ("in this picture"), followed by a subject quantified by either "有的" (*yǒu de*, *some of*), or "所有的" (*suǒyǒu de*, *all of*), followed by a verb and aspect marker, object, and an additional phrase to separate the object from the end of the sentence. Verbs in the critical sentences were marked for progressive, perfective, or prospective aspect. *All of* sentences included the mandatory adverbial 都 *dōu*

before the verb (see Li & Thompson, 1981; Jiang et al., 2009). The sentences were written with the help of a paid linguistics student from Peking University who was a native speaker of Mandarin.

Additionally, 148 picture-sentence pairs were created for use as fillers. The filler picture arrays met the same criteria as the critical trials, except that some of them depicted intransitive events. Thirty-seven of these fillers were *Some*-type pictures paired with matching, felicitous *some of* sentences, and thirty-seven were *All*-type pictures paired with matching, correct *all of* sentences. The other seventy-four pictures were paired with sentences that had appropriate quantifiers but either an object that did not match any of the objects in the picture of a verb that did not match the activity shown. Several of these included verbs that yielded semantically anomalous sentences (e.g., "all the scientists are planting squirrels"), whereas most had verbs that were semantically plausible but not congruous with the picture (e.g., "all the boys are going for a walk with their classmates", after a picture in which all the boys are wrestling with their classmates). The filler sentences all included quantifiers that were not used in the critical sentences but were similar in meaning to *all of* or *some of*, or classifier phrases in place of quantifiers. None of the filler sentences used numbers in the place of quantifiers (for discussion of how the presence/absence of numbers and quantifiers in the experimental context may affect the perception of scalar implicature, see Degen, 2009; Grodner et al., 2010; Huang et al., 2010; and references therein). The set of fillers with mismatching pictures and sentences was included to distract participants from the quantifier manipulation in the critical sentences, and the remaining matching fillers were included to maintain a proportion of acceptable sentences of at

least 50% during the experiment, assuming that pragmatically infelicitous stimuli are judged as unacceptable.

4.1.3 Procedure

Participants were seated in a dimly-lit room about 1 meter in front of a 41-cm CRT monitor. Stimuli were presented at the center of the screen using the Presentation software package (Neurobehavioral Systems). Each trial began with a fixation point presented for 500 ms, followed by a picture which remained on the screen for 4000 ms. The picture was followed by a fixation point of random duration (between 500 and 1500 ms), after which the sentence was presented region by region using the serial visual presentation paradigm. Regions were presented using a variable presentation procedure (see, e.g., Nieuwland et al., 2010), whereby each region was presented at a base duration of 425 ms per region, plus 80 ms for each character more than 3 in the region; because the critical quantifiers were all three characters or less, their presentation durations do not differ across conditions. The interstimulus interval was 400 ms for all regions.[7] Twenty percent of trials were followed by comprehension questions or acceptability judgments (see below), which were presented on the screen for 5000 ms or until the participant's response. Each trial was followed by a blank screen for 1500 ms before the start of the next trial. The experiment was divided into six blocks of approximately 50 sentences each, and participants were given short breaks between the blocks. Participants were instructed not to blink during the presentation of the sentences.

[7] An 800-ms stimulus onset asynchrony (400-ms word presentation, 400-ms interstimulus interval) has been found to be natural and comfortable for Chinese readers in previous studies (e.g., Jiang et al., 2009), but the regions used in the present study tended to be longer than the regions used in those studies, and pilot participants reported the variable presentation rate described above to be the most comfortable.

Participants performed a mixture of acceptability judgments and comprehension questions. On ten percent of trials, after the sentence ended, a question that probed information about the picture and was irrelevant to the sentence was presented (e.g., after the sentence "In this picture, some of the girls are sitting on blankets", the comprehension question "Are the girls wearing swimsuits?" appeared). In an additional ten percent of trials, the sentence was followed instead by an acceptability judgment (the question "对不对," "Is that correct?"). Participants were not given explicit instructions about what criteria to consider in judging the sentences, unless they asked for clarification; if they asked, they were instructed to judge, based on their own intuition, whether the sentence was consistent with the picture and described it appropriately. The experimenter stressed that some sentences had no right or wrong answer and that the experiment was meant to measure the participant's own language intuitions. The comprehension questions were included to prevent participants from being able to adopt a strategy of only paying attention to the quantifiers and the number of objects in a picture, and the acceptability judgments were included to ensure that participants pay attention to the sentence rather than just try to remember the picture. Acceptability judgment prompts were allotted to six of the forty pragmatically infelicitous sentences for each participant, allowing us to determine whether participants accepted or rejected these sentences when making an explicit judgment. Participants responded to both the comprehension questions and acceptability judgment prompts using the left and right buttons on a mouse.

The experimental sentences were divided into four lists according to a Latin square design, such that every sentence appeared once in each condition across lists but no sentence or picture was repeated within a list. The item order in the list was fully randomized for each

participant. The first block of the experiment was preceded by a practice block of seven trials which followed the same presentation procedure as the main experiment but did not include any quantifier-related violations. The practice sentences included some sentences with existential quantifiers (e.g., "图片里有。。。," "in the picture there are") and some without quantifiers (e.g. "图片里的小狗," "the dogs in the picture are…"). Feedback was given for behavioral responses in the practice block, but not in the main experiment. The recording itself took 70 to 80 minutes.

4.1.4 Data acquisition and analysis

The EEG was continuously recorded using an elastic electrode cap (Electro-Cap International, Inc.) containing 32 Ag/AgCl scalp electrodes organized in a modified 10-20 layout (midline: FPZ, FZ, FCZ, CZ, CPZ, PZ, OZ; lateral: FP1/2, F7/8, F3/4, FT7/8, FC3/4, T3/4, C3/4, TP7/8, CP3/4, T5/6, P3/4, O1/2). Polygraphic channels were placed at the left and right outer canthi for monitoring horizontal eye movements, above and below each eye for monitoring blinks, and on the left and right mastoids. The left mastoid served as a reference during data acquisition and AFz served as the ground. Impedances for scalp electrodes and mastoids were kept below 5 kΩ. The recordings were amplified by a Neuroscan Synamps2 amplifier (Compumedics Neuroscan, Inc.) with a bandpass of 0.01 to 200 Hz, and digitized at a sampling rate of 1000 Hz.

The continuous EEG was re-referenced to the average of both mastoids and segmented into epochs from 1 second before to 2 seconds after the presentation of the critical word. Based on visual inspection, trials containing excessive muscle artifact or alpha activity within the epoch

of 200 ms before to 1200 ms after the onset of the stimulus were excluded from the analysis. Following artifact rejection, the data were demeaned using the mean amplitude of the each epoch (Groppe et al., 2009), and an independent components (ICA) decomposition algorithm (Makeig et al., 1996) was applied to remove ocular artifacts. After artifact correction, the EEG was visually inspected again to remove trials in which any artifact remained. A total of 18.8% of trials was rejected in this way (18.9% of pragmatically inconsistent *some of* trials; 16.2% of correct *some of* trials; 20% of semantically inconsistent *all of* trials; and 20.1% of consistent *all of* trials); a repeated measures ANOVA revealed that marginally more *some of* than *all of* trials were kept in the analysis ($F(1,18) = 3.49$, $p = .078$) and that there was no significant effect of consistency or interaction between quantifier or consistency in terms of trials kept ($ps > .16$). Participants with fewer than 25 trials remaining for any condition after artifact rejection were excluded from the analysis. Subsequently, data epochs were baseline-corrected using a 200-ms pre-stimulus baseline and averaged to calculate ERPs.

Time windows for analysis were chosen based on visual inspection of the data, and mean ERP voltage amplitudes were compared using repeated measures ANOVAs involving the factors Consistency (consistent, inconsistent), Quantifier (*some of, all of*), and the topographical factor Region. Midline and lateral regions were analyzed separately. For the lateral ANOVA, regions were defined by averaging within the following electrode groups: left anterior (F7, F3, FC3), left central (T3, C3, CP3), left posterior (T5, P3, O1), right anterior (F4, F8, FC4), right central (C4, T4, CP4), and right posterior (P4, T6, OZ). For the midline ANOVA, regions were defined as follows: anterior (FZ, FCZ), central (CZ, CPZ), and posterior (PZ, OZ). The Huynh-Feldt correction was applied to *F*-tests with more than one degree of freedom in the numerator.

4.2 Experiment 2

4.2.1 Participants

Twenty-three Peking University students (9 females; mean age 22.5 years, range 18-26) who were native speakers of Mandarin participated in the study. Three were excluded from the statistical analysis due to excessive artifacts in their recordings, leaving a total of 20 participants in the final analysis. All participants had normal or corrected-to-normal vision and were right-handed according to the Chinese Handedness Survey (Li, 1983). All participants provided their informed consent and received payment, and all methods for the study were approved by the Ethics Committee of the Department of Psychology, Peking University, and the Human Subjects Committee of Lawrence at the University of Kansas.

4.2.2 Materials

Two hundred and sixty sets of picture arrays were designed according to the same criteria as in Experiment 1. Each *Some*- and *All*-type picture array had two versions, such that in the first version the object being interacted with by some or all of the characters matched the object mentioned in the associated sentence, and in the second version it mismatched. At the object position, this formed a 2 (Lexical Consistency) × 2 (Pragmatic Consistency) design: sentences with correct objects, sentences with lexical violations at the object position, sentences with correct objects but a pragmatic violation upstream, and sentences with both a pragmatic violation upstream and a lexically incorrect object. It formed a one-factor design at the quantifier position: sentences with consistent quantifiers and those with pragmatically inconsistent quantifiers (each

of these conditions collapsed across lexically consistent and inconsistent sentences, since at the quantifier position the lexical violation has not yet been encountered). A sample stimulus set is shown in Figure 4. Critical sentences were written so that none of the critical objects were at the end of the sentence. All the critical objects used were either 2 or 3 syllables long.[8]

Two hundred forty filler sentences were prepared, using picture-sentence pairs that had not been chosen for the critical items as well as new picture-sentence pairs. Eighty were used to test the semantic violation at quantifier position (forty correct *all* and forty semantically inconsistent *all of* sentences, counterbalanced across participants); these sentences, together with the critical sentences, allowed us to test whether the Consistency by Quantifier interaction reported in Experiment 1 could be replicated. Of the remaining fillers, eighty were correct *all of* sentences that were not analyzed, and the last 80 were sentences using other quantifiers. Of those 80, 40 used *some*-like quantifiers (e.g. 有一些 *a few*) and 40 used *all*-like quantifiers (e.g. 每个 *every*). None included quantifier-related violations; 40 were entirely correct, 20 mismatched with the picture at the object position, and 20 mismatched at the verb position. (Out of each of these types, half of the items used *all*-like quantifiers and half used *some*-like.)

Auditory stimuli were read by a female native speaker from the Peking University Chinese department, who was instructed to avoid placing contrastive stress on the quantifiers and

[8] The 200 plausible most plausible all-type pictures were normed with a sentence completion task to select pictures in which the objects were most identifiable. Twenty-eight students from Beijing Union University participated in the task. Participants were presented with the pictures along with sentence fragments up to but not including the objects (e.g. "图片里，所有的女孩都坐在。。。", "In the picture, all the girls are sitting on…") and asked to complete the sentence. For critical stimuli for the ERP experiment we chose the 160 sentence-picture pairs whose objects had the highest cloze probability, with the condition that a pair was not chosen if any identical objects were given in response to both pictures. All sentences chosen had an object cloze probability above 46% (mean 81%). Due to reorganization of target and filler stimuli to avoid repetition of target objects, two picture/sentence pairs that had not been cloze tested were later introduced into the critical stimuli.

objects. The recordings were digitized at 22050 Hz using CoolEdit Pro (Syntrillium Software) and segmented using Praat (Boersma & Weenik, 2012), and the onset latencies of the quantifiers and objects were measured. The onset of the quantifier *yǒu de* (*some of*) was defined as the point of lowest intensity between the preceding syllable *lǐ* and the *yǒu*, which in most tokens also coincided with a perceptible change in phoneme quality and preceded, by 10-20 ms, a 200-400 Hz drop in frequency of the second through fourth formants. The onset of the quantifier *suǒyǒu de* (*all of*) was defined as the onset of high-frequency energy in the spectrogram. Onsets of objects were measured as the audible onset of the first consonant of the word (plosives were measured at the burst), except in two cases where the onset of the first consonant of the second syllable was measured since this was the point of disambiguation for the critical word. The latency between quantifier onsets and object onsets in the critical sentences was 1309 ms on average (*sd* = 203 ms, range 832-2127 ms).

The 400 trials (160 critical *some of* sentences, 80 *all of* fillers, and 160 other fillers) were arranged into four lists in a Latin square design. Each list contained 40 trials per object condition. For the *all of* sentences tested, each list contained 40 trials per condition (correct "all", semantically inconsistent "all").

Each list was divided into five blocks of 80 trials each, such that the first trial in each block was a filler sentence. Each block was pseudorandomized according to the following criteria: no more than three trials of the same condition could appear consecutively, no more than four correct or incorrect trials could appear consecutively, no more than six *Some*-type or *All*-type pictures could appear consecutively, and no more than six *some of* or *all of* sentences could appear consecutively. The order of trials was kept the same for each list, such that a given item

appeared at the same position (but in different conditions) in every list, and each of the lists adhered to the above constraints.

4.2.3 Procedure

Participants were seated comfortably in a dimly lit and electromagnetically shielded room, about 80 cm in front of a 51-cm CRT monitor. Pictures were presented on the monitor and sentences were presented through tube earphones (Etymotic Research, Inc.). Stimulus presentation and recording of behavioral responses was controlled using Presentation software (Neurobehavioral Systems). Each trial began with a fixation point presented at the center of the screen for 500 ms, followed by the picture, which was presented at the center of the screen for 4000 ms. After this time the picture disappeared and was immediately replaced by a fixation point at the center of the screen, which remained on the screen throughout the presentation of the auditory sentence. The sentence began between 500 and 1500 ms after the appearance of the fixation point. After the end of the sentence, a 1-7 scale appeared on the screen.

The participants' task was to rate how consistent the sentence was with the preceding picture within 3000 ms. The rating task was chosen to encourage participants to pay attention to the entire sentence and thus reduce the possibility that they could complete the task strategically simply by matching numbers of items in the picture with quantifiers in the sentence; rating tasks have been used in previous online studies investigating quantification (Urbach & Kutas, 2010) and scalar implicature (Foppolo, 2007). After the rating task was complete, the trial was followed by a 2500 ms blank screen before the fixation point signaling the beginning of the next trial.

After every 80 trials the participants were given a break. In addition, after every 20 trials they were given a 15-second break, during which time a message appeared on the screen asking them to relax briefly. The formal experiment was preceded by a practice session consisting of 10 trials. The trial structure and picture formats were identical to those used in the main experiment, but no violations involving picture-object mismatch or pragmatic underinformativeness were included. The recording took about 100 minutes.

4.2.4 Data acquisition and analysis

The EEG was continuously recorded using an elastic electrode cap (Brain Products, Munchen, Germany) containing 64 tin electrodes organized according to the 10-20 system. Additional channels were placed above the right eye and at the outer canthus of the left eye for monitoring vertical and horizontal electro-oculograms (EOGs), respectively. An electrode placed on the tip of the nose served as the reference during data acquisition, and AFz served as the ground. Impedances were kept below 10 kΩ. The recordings were amplified using a Brain Products Brainamp amplifier with a bandpass from 0.016 to 100Hz, and digitized at a sampling rate of 500 Hz.

The raw EEG was segmented into epochs from 1000 ms before to 4250 ms after the quantifiers (this epoch ensured at least 2000 ms after each critical object). Data were then demeaned using the mean amplitude of each epoch (Groppe et al., 2009), decomposed with an ICA algorithm (Makeig et al., 1996) to remove ocular artifacts, and re-segmented into two separate datasets (one consisting of -200 to 1000 ms epochs time-locked to the quantifiers, and one consisting of -200 to 1000 ms epochs time-locked to the objects). Artifact rejection was

performed separately for the quantifier and object data, and ERPs time-locked to the object used a 100-ms post-stimulus baseline rather than a 200-ms pre-stimulus baseline, since the pre-stimulus interval contained sustained effects of processing violations at the quantifier. 11.7% of trials were rejected (9.8% of epochs time-locked to the objects, and 13% of epochs time-locked to the quantifiers); all subjects included in the analysis had at least 29 trials per condition in the object analysis and 25 per condition in the quantifier analysis. The proportion of trials rejected did not differ between conditions in either analysis (objects: Fs < 1; quantifiers: Fs < 1.06, ps > .315).

The following electrode regions were defined on this cap: left anterior (F1, F3, F5, FC1, FC3, FC5), right anterior (F2, F4, F6, FC2, FC4, FC6), left central (C1, C3, C5, CP1, CP3, CP5), right central (C2, C4, C6, CP2, CP4, CP6), left posterior (P1, P3, P5, PO3, PO7, O1), right posterior (P2, P4, P6, PO4, PO8, O2), midline anterior (Fz, FPz), midline central (Cz, CPz), midline posterior (POz, Oz). For the quantifier position, the analysis used the factors Consistency (consistent, inconsistent), Quantifier (*some of, all of*), and Region (6 levels for the lateral ANOVA, 3 for the midline ANOVA). For the object position, the factors were Pragmatic Consistency (consistent, inconsistent), Lexical Consistency (consistent, inconsistent), and Region. The Huynh-Feldt correction was applied to F-tests with more than one degree of freedom in the numerator.

Acknowledgements

This research was supported by the National Science Foundation East Asia and Pacific Summer Institutes (award ID #1015160) to SPA, the China Post-Doctoral Science Foundation (award IDs

#20100480150, #2012T50005) to XJ, and the Natural Science Foundation of China (award ID #30970889) and Ministry of Science and Technology of China (award ID# 2010CB833904) to XZ. Experiment design, data analysis, and preparation of this manuscript was completed by the authors. The authors thank Liang Yan, Wu Chunping, and Wu Yue, Luo Yingyi, Zhu Mengyan, Wu Junru, and Lamar Hunt III for assistance in the construction of materials; Wu Yin and Zhou Yuqin for assistance with data collection; and Jamie Bost and Natalie Pak for assistance in the preparation of this manuscript.

References

Baggio, G., van Lambalgen, M., & Hagoort, P. (2008). Computing and recomputing discourse models: An ERP study. *J Mem Lang, 59,* 36-53.

Bezuidenhout, A., & Cutting, J. (2002). Literal meaning, minimal propositions and pragmatic processing. *J Pragmat, 34,* 433-456.

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer [Computer program]. http://www.praat.org/

Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behav Res Methods Instrum Comput, 35,* 158-167.

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genç, S., Philipp, M., … Schlesewsky, M. (2011). Think globally: Cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain Lang, 117,* 133-152.

Bott, L., Bailey, T., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *J Mem Lang, 66,* 123-142.

Bott, L., & Noveck, I. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *J Mem Lang, 51,* 437-457.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalized scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition, 100,* 434-463.Chevallier, C., Noveck, I., Nazir, T., Bott, L., Lanzetti, V., & Sperber, D. (2008). Making disjunctions exclusive. *Q J Exp Psychol, 61,* 1741-1760.

Chi, W. (2000). "Bùfen", "yǒu de" zhī luójì biànxī. [Towards a logical differentiation between "part" and "some"]. *Shāndōng Shīdà Xuébào (Shèhuì Kēxué Bǎn) [Shandong Normal University Journal (Social Science)], 169,* 91-103.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Exp Psychol, 54,* 128-133.

Degen, J. (2009). *Processing scalar implicatures: An eye-tracking study* (Master's thesis, University of Osnabrück, Osnabrück, Germany).

Degen, J., & Tanenhaus, M. (2011). Making inferences; the case of scalar implicature processing. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3299-3304).

Degen, J., & Tanenhaus, M. (2010). When contrast is salient, pragmatic "some" precedes logical "some". *Poster presented at the 23rd CUNY Conference on Human Sentence Processing,* New York, NY.

Doran, R., Baker, R., McNabb, Y., Larson, M., & Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *International Review of Pragmatics, 1,* 211-248.

Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language, 88,* 124-154.

Filik, R, & Leuthold, H. (2008). Processing local pragmatic anomalies in fictional contexts: evidence from the N400. *Psychophysiology, 45,* 554-558.

Fischler, I., Bloom, P., Childers, D., Roucos, S., & Perry, N. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology, 20,* 400-409.

Feeney, A., Scafton, S., Duckworth, A., & Handley, S. J. (2004). The story of some: Everyday pragmatic inferences by children and adults. *Can J Exp Psychol, 54,* 128-133.

Foppolo, F. (2007). Between 'cost' and 'default': a new approach to scalar implicature. *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue,* 125-132.

Garrett, M, & Harnish, R. (2007). Experimental pragmatics: testing for implicitures. *Pragmatics and Cognition, 15,* 65-90.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41-58). New York: Academic Press.

Grodner, D., Klein, N., Carbary, K., & Tanenhaus, M. (2010). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition, 116,* 42-55.

Groppe, D., Makeig, S., & Kutas, M. (2009). Identifying reliable independent components via split-half comparisons. *NeuroImage, 45,* 1199-1211.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304,* 438-41.

Hagoort, P., & Van Berkum, J. (2007). Beyond the sentence given. *Philo. Trans R Soc B, 362,* 801-811.

Hagoort, P.. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *J Cogn Neurosci, 16,* 883-899.

Hartshorne, J., & Snedeker, J. (submitted). The speed of inference: Evidence against rapid use of context in calculation of scalar implicatures.

Horn, L. (1972). *On the semantic properties of logical operators in English* (Doctoral dissertation, University of California, Los Angeles).

Huang, Y., Hahn, N., & Snedeker, J. (2010). Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures. *Poster presented at the 23rd CUNY Conference on Human Sentence Processing*, New York, NY.

Huang, Y., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cogn Psychol, 58*, 376-415.

Jiang, X., Li, Y., & Zhou, X. (2011). Is it over-respectful or disrespectful? Differential brain activities in perceiving pragmatic violation of social status information during utterance comprehension. *Poster presented at the 3rd Neurobiology of Language Conference*, Annapolis, MD.

Jiang, X., Tan, Y., & Zhou, X. (2009). Processing the universal quantifier during sentence comprehension: ERP evidence. *Neuropsychologia, 47*, 1799-1815.

Jing, H., Pivik, R., & Dykman, R. (2006). A new scaling method for topographical comparisons of event-related potentials. *J Neurosci Methods*, 151, 239-249.

Katsos, N., & Cummins, C. (2010). Pragmatics: from theory to experiment and back again. *Lang Linguist Compass, 4/5*, 282-295.

Knoeferle, P., Urbach, T., & Kutas, M. (2011). Comprehending how visual context influences incremental sentence processing: insights from ERPs and picture-sentence verification. *Psychophysiology, 48*, 495-506.

Kuperberg, G., McGuire, P., Bullmore, E., Brammer, M., Rabe-Hesketh, S., Wright, I., … David, A. (2000). Common and distinct neural substrates for pragmatic, semantic, and syntactic processing of spoken sentences: an fMRI study. *J Cogn Neurosci, 12*, 321-341.

Kutas, M., & Federmeier, K. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends Cogn Sci, 4*, 463-470.

Kutas, M., Van Petten, C., & Kluender, R. (2006). Psycholinguistics electrified II: 1994-2005. In M. Traxler & M.A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 659-724). New York: Elsevier.

Lau, E., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nat Rev Neurosci, 9*, 920-933.

Leuthold, H., Filik, R., Murphy, K., & Mackenzie, I. (2012). The on-line processing of socio-emotional information in prototypical scenarios: inferences from brain potentials. *Soc Cogn Affect Neurosci, 7*, 457-466.

Li, C., & Thompson, S. A. (1981). *Mandarin Chinese: A Functional Reference Grammar*. Los Angeles: University of California Press.

Li, T. (1983). Distribution of left/right handedness among Chinese people. *Xin Li Xue Bao, 15*, 268-276.

Lu, A., & Zhang, J. (2012). Event-related potential evidence for the early activation of literal meaning during comprehension of conventional lexical metaphors. *Neuropsychologia, 50*, 1730-1738.

Luck, S. (2005). *An Introduction to the Event-Related Potential Technique.* MIT Press.

Makeig, S., Bell, A., Jung, T., & Sejnowski, T. (1996). Independent component analysis of electroencephalographic data. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems 8* (pp. 145-151). Cambridge: MIT Press.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods, 164*, 177-190.

Nieuwland, M., Ditman, T., & Kuperberg, G. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *J Mem Lang, 63*, 324-346.

Nieuwland, M., & Kuperberg, G. (2008). When the truth is not too hard to handle: an event-related potential study on the pragmatics of negation. *Psychol Sci, 19*, 1213-1218.

Nieuwland, M. & Van Berkum, J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *J Cogn Neurosci, 18*, 1098-1111.

Noveck, I., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain Res, 85*, 203-210.

Noveck, I., & Sperber, D. (2007). The why and how of experimental pragmatics: The case of 'scalar inferences'. In N. Burton-Roberts (Ed.), *Advances in Pragmatics* (pp. 184-212). Basingstoke: Palgrave.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*, 97-113.

Osterhout, L., Bersick, M. & McLaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Mem Cognit, 25*, 273-285.

Panizza, D. (2012). Formal neurosemantics: Logic, meaning and composition in the brain. *J Neurolinguistics, 25*, 460-488.

Panizza, D., Huang, Y., Chierchia, G., & Snedeker, J. (2011). Relevance of polarity for the online interpretation of scalar terms. *Proceedings of the 19th Semantics and Linguistic Theory Conference (April 2009)*, 360-378.

Pijnacker, J., Geurts, B., van Lambalgen, M., Buitelaar, J., & Hagoort, P. (2011). Reasoning with exceptions: an event-related brain potentials study. *J Cogn Neurosci, 23*, 471-480.

Politzer-Ahles, S., Jiang, X., Fiorentino, R., & Zhou, X. (forthcoming). Individual differences in logical ability predict ERP responses to underinformative sentences. *Poster to be presented at the 4th Neurobiology of Language Conference*, San Sebastian, Spain.

Pylkkänen, L., Brennan, J., Bemis, D. K. (2011). Grounding the cognitive neuroscience of semantics in linguistic theory. *Lang Cogn Process, 26*, 1317-1337.

Rullman, H., & You, A. (2006). General number and the semantics and pragmatics of indefinite bare nouns in Mandarin Chinese. In K. von Heusinger & K. P. Turner (Eds.), *Where semantics meets pragmatics* (pp. 175-196). Amsterdam: Elsevier.

Steinhauer, K., Drury, J., Portner, P., Walenski, M., Ullman, M. T. (2010). Syntax, concepts, and logic in the temporal dynamics of language comprehension: Evidence from event-related potentials. *Neuropsychologia, 48*, 1525-1542.

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., … Bates, E. (2004). A new on-line resource for psycholinguistic studies. *J Mem Lang, 51*, 247-250.

Tavano, E. (2010). *The balance of scalar implicature* (Doctoral dissertation, University of Southern California, Los Angeles, CA).

Tesink, C., Buitelaar, J., Petersson, K., van der Gaag, R., Kan, C., Tendolkar, I., & Hagoort, P. (2009). Neural correlates of pragmatic language comprehension in autism spectrum disorders. *Brain, 2009*, 1941-1952.

Tsai, W.-T. (2004). Tán "yǒu rén," "yǒu de rén," hé "yǒu xiē rén" [On "yǒu rén," "yǒu de rén," and "yǒu xiē rén"]. *Hànyǔ Xuébào [Chinese Linguistics], 8*(2), 16-25.

Urbach, T., & Kutas, M. (2010). Quantifiers qualify more or less online: ERP evidence for partial incremental interpretation. *J Mem Lang, 63*, 158-179.

Van Berkum, J. (2009). The neuropragmatics of 'simple' utterance comprehension: An ERP review. In U. Sauerland & K. Yatsushiro (Eds.), *Semantics and Pragmatics: From Experiment to Theory* (pp. 276-316). Basingstoke: Palgrave Macmillan.

Van Berkum, J., Koornneef, A., Otten, M., & Nieuwland, M. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Res, 1146*, 158-171.

Wiswede, D., Koranyi, N., Müller, F., Langner, O., & Rothermund, K. (in press). Validating the truth of propositions: behavioral and ERP indicators of truth evaluation processes. *Soc Cogn Affect Neurosci.*

Wu, Z., & Tan, J. (2009). Hànyǔ értóng yǔyán zhōng de děngjí hányì—yí xiàng shíyàn yánjiū [Scalar implicature in Chinese child language: An experimental study]. *Wàiguóyǔ [Journal of Foreign Languages], 32*, 69-75.

Xie, Y. (2003). Guānyú "yǒu de+VP" [On the construction of "yǒu de+VP"]. *Yǔyán Yánjiū [Studies in Language and Linguistics], 23*, 37-4.

Figure captions

Figure 1 Sample pictures and sentences used in Experiment 1. Upper portion: *some of* sentences preceded by pictures that render them correct (left) or pragmatically incorrect (right). Lower portion: *all of* sentences preceded by pictures that render them semantically incorrect (left) or correct (right).

Figure 2 Effect of pragmatic inconsistency in Experiment 1. Upper portion: Grand average ERPs (a 30 Hz low-pass filter was applied for plotting) at nine scalp regions. Lower portion: Topographic maps formed by subtracting the correct *some of* condition from the pragmatically incorrect condition over two time windows.

Figure 3 Effect of semantic inconsistency in Experiment 1. Upper portion: Grand average ERPs (a 30 Hz low-pass filter was applied for plotting) at nine scalp regions. Lower portion: Topographic maps formed by subtracting the correct *all of* condition from the semantically inconsistent condition over two time windows.

Figure 4 Sample pictures used in Experiment 2; in this sample, all pictures were followed by the sentence 图片里，有的女孩坐在毯子上晒太阳 ("In the picture, some of the girls are sitting on blankets suntanning"). In a given trial, only one of the pictures was shown before the sentence. The condition labels on the picture are for expository purposes only and were not included in the experiment.

Figure 5 Effects of pragmatic and semantic inconsistency at the quantifier in Experiment 2. Upper portion: Grand average ERPs at the midline central region. Lower portion: Topographic maps formed by subtracting the correct quantifier condition from the corresponding inconsistent quantifier conditions.

Figure 6 Effects of lexical and pragmatic inconsistency in Experiment 2. Upper portion: Grand average ERPs (a 30 Hz low-pass filter was applied for plotting) at nine scalp regions. Lower portion: Topographic maps of difference waves.

Table captions

Table 1 Results of the lateral and midline omnibus ANOVAs in Experiment 1 at two time windows, with each cell showing the lateral ANOVA result above and the midline ANOVA result below. *.05 < p < .1; **p < .05; ***p < .005; ****p < .001

Table 2 Results of the lateral and midline omnibus ANOVAs at the quantifier position in Experiment 2, with each cell showing the lateral ANOVA result above and the midline ANOVA result below. *.05 < p < .1; **p < .05; ***p < .005; ****p < .001

Table 3 Results of the lateral and midline omnibus ANOVAs at the object position at two time windows in Experiment 2, with each cell showing the lateral ANOVA result above and the midline ANOVA result below. *.05 < p < .1; **p < .05; ***p < .005; ****p < .001

Figure 1 (black-and-white, no color version)
Click here to download high resolution image

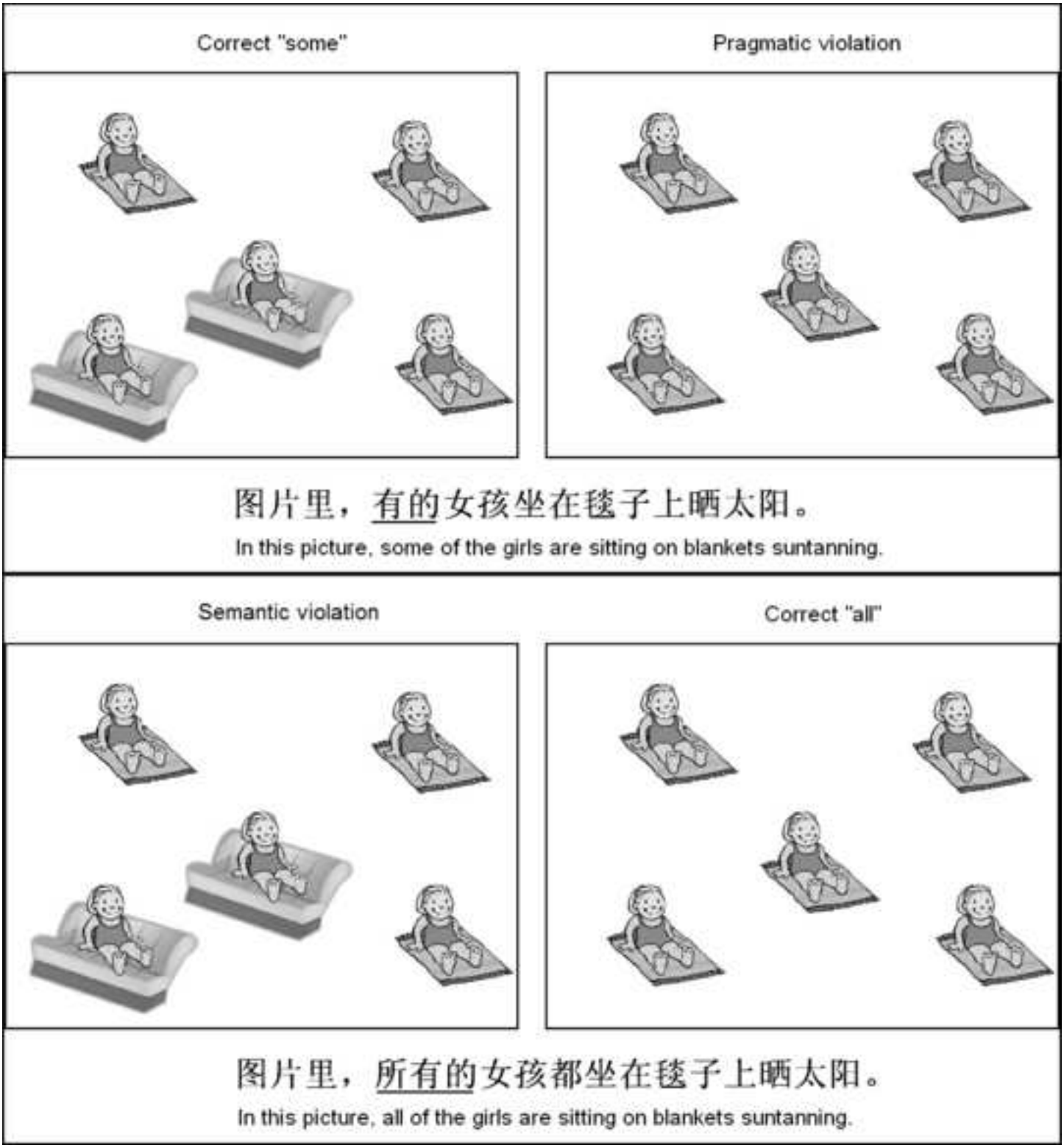

Figure 2 (color)

Click here to download high resolution image

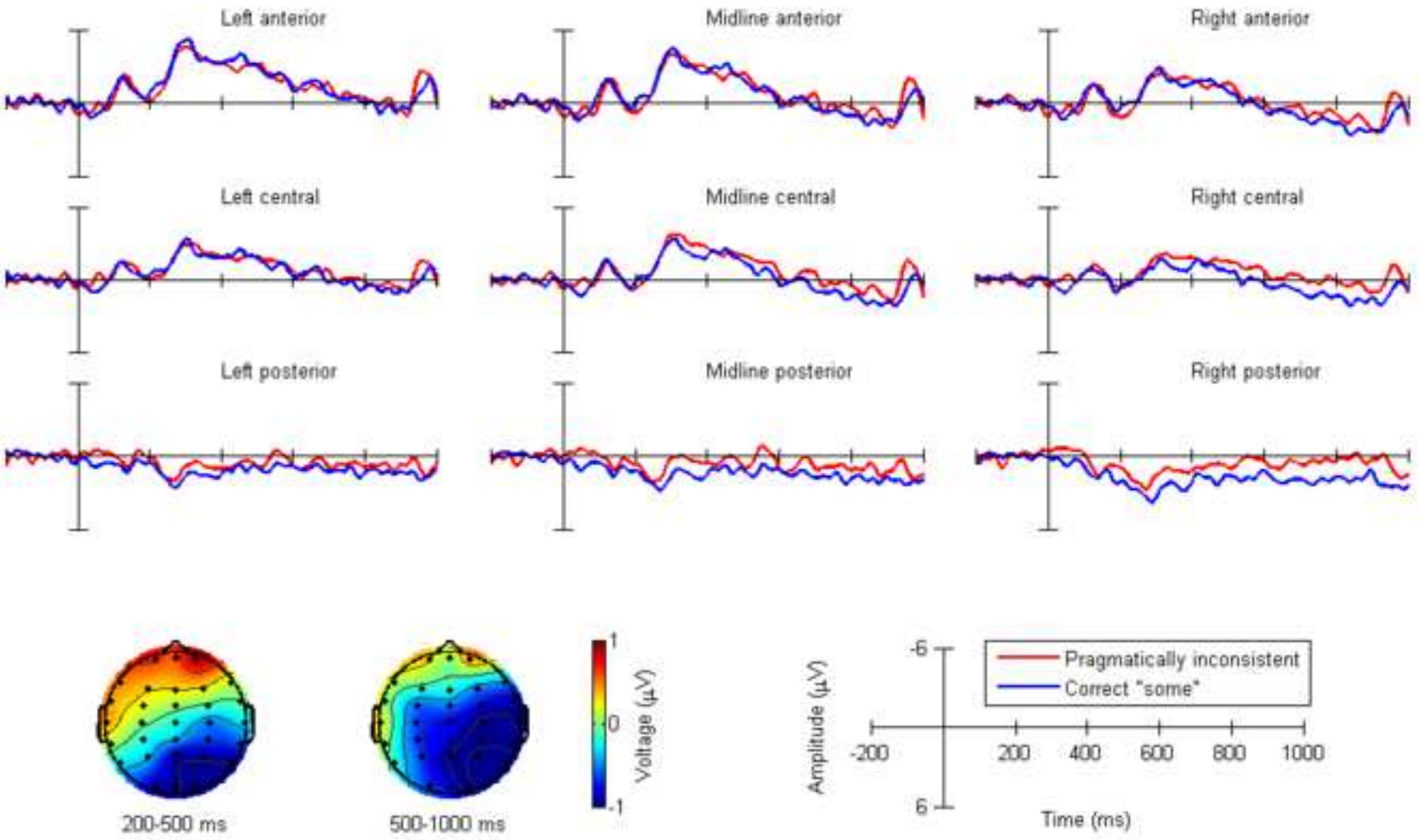

Figure 3 (color)

Click here to download high resolution image

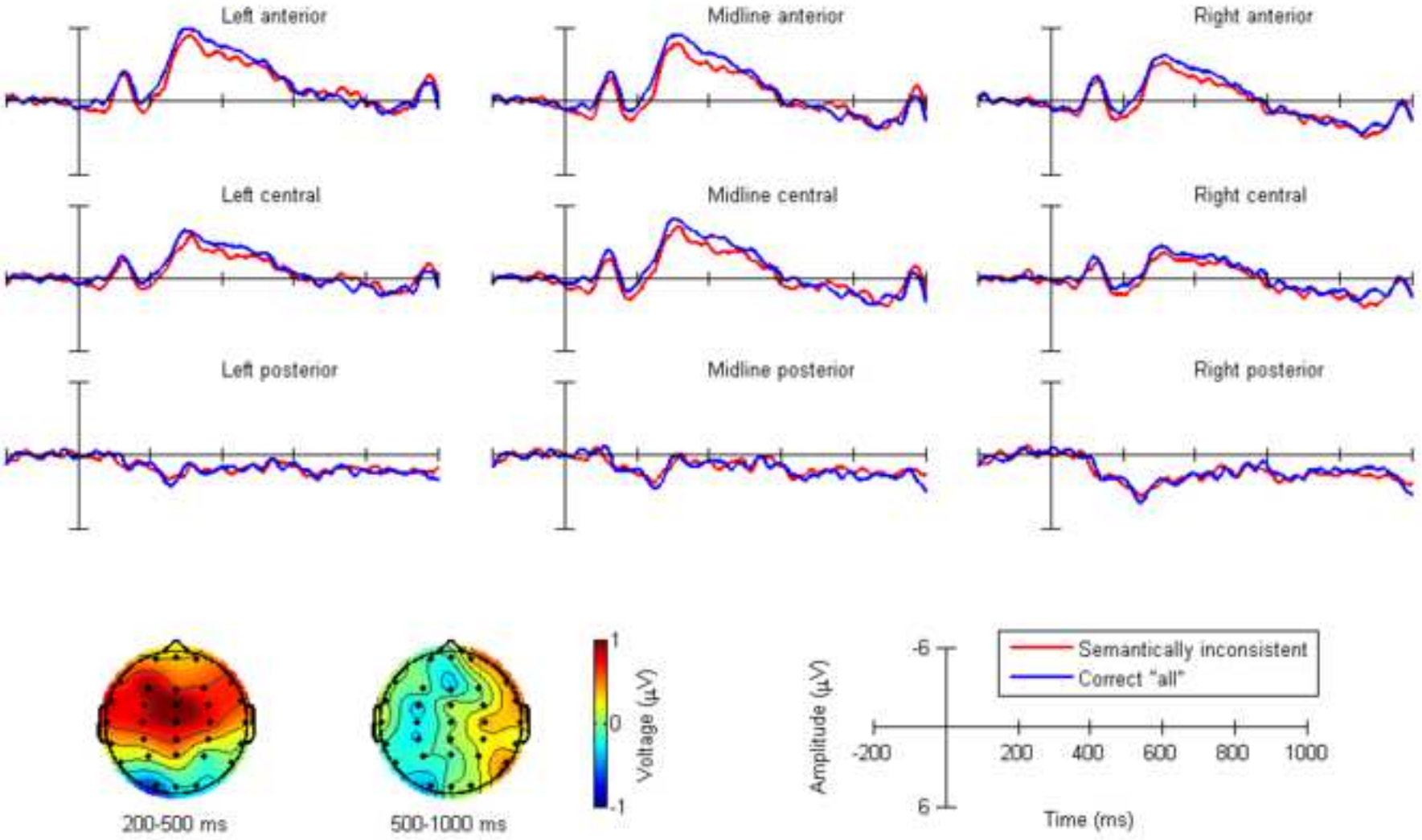

Figure 4 (black-and-white, no color version)
Click here to download high resolution image

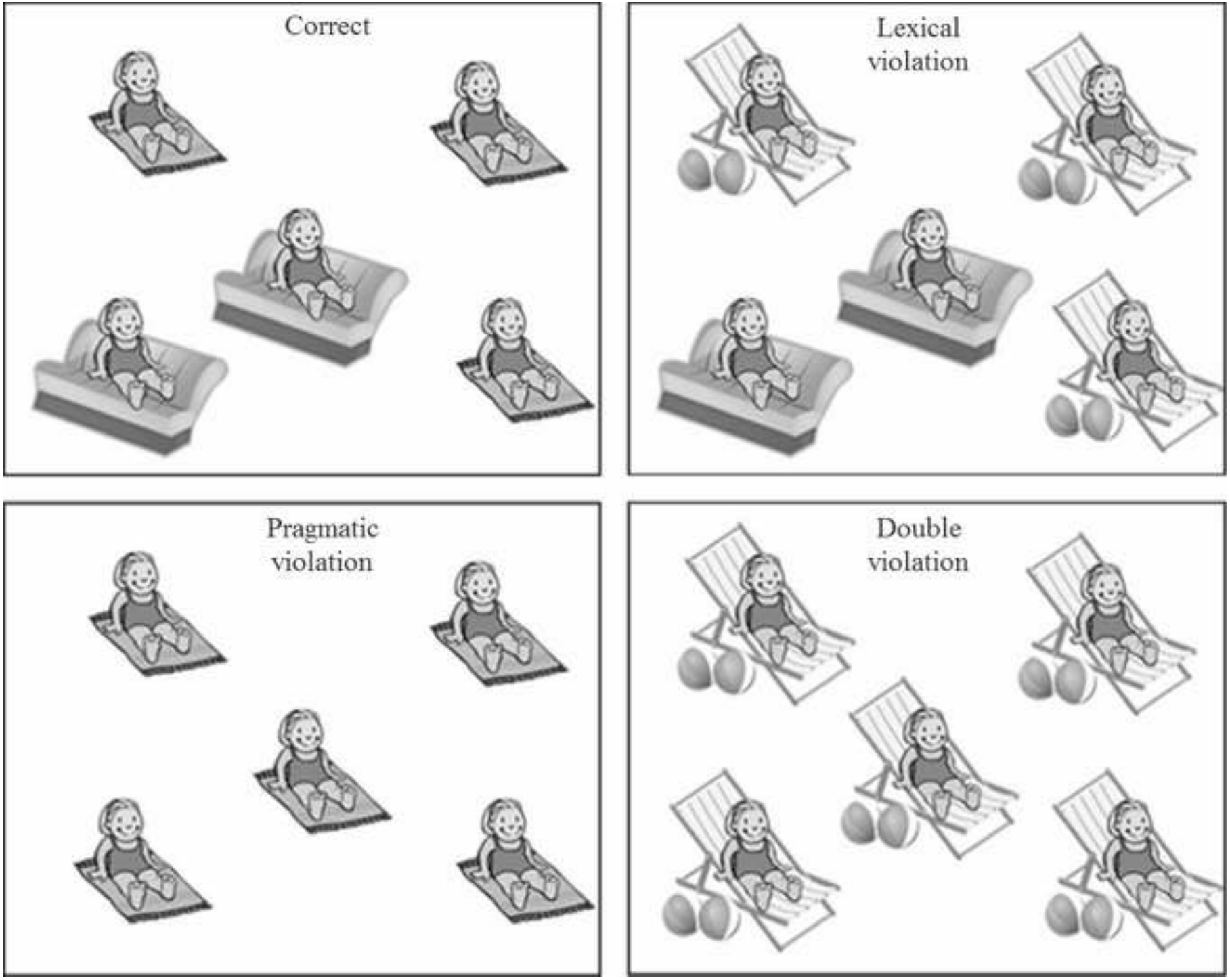

Figure 5 (color)

Click here to download high resolution image

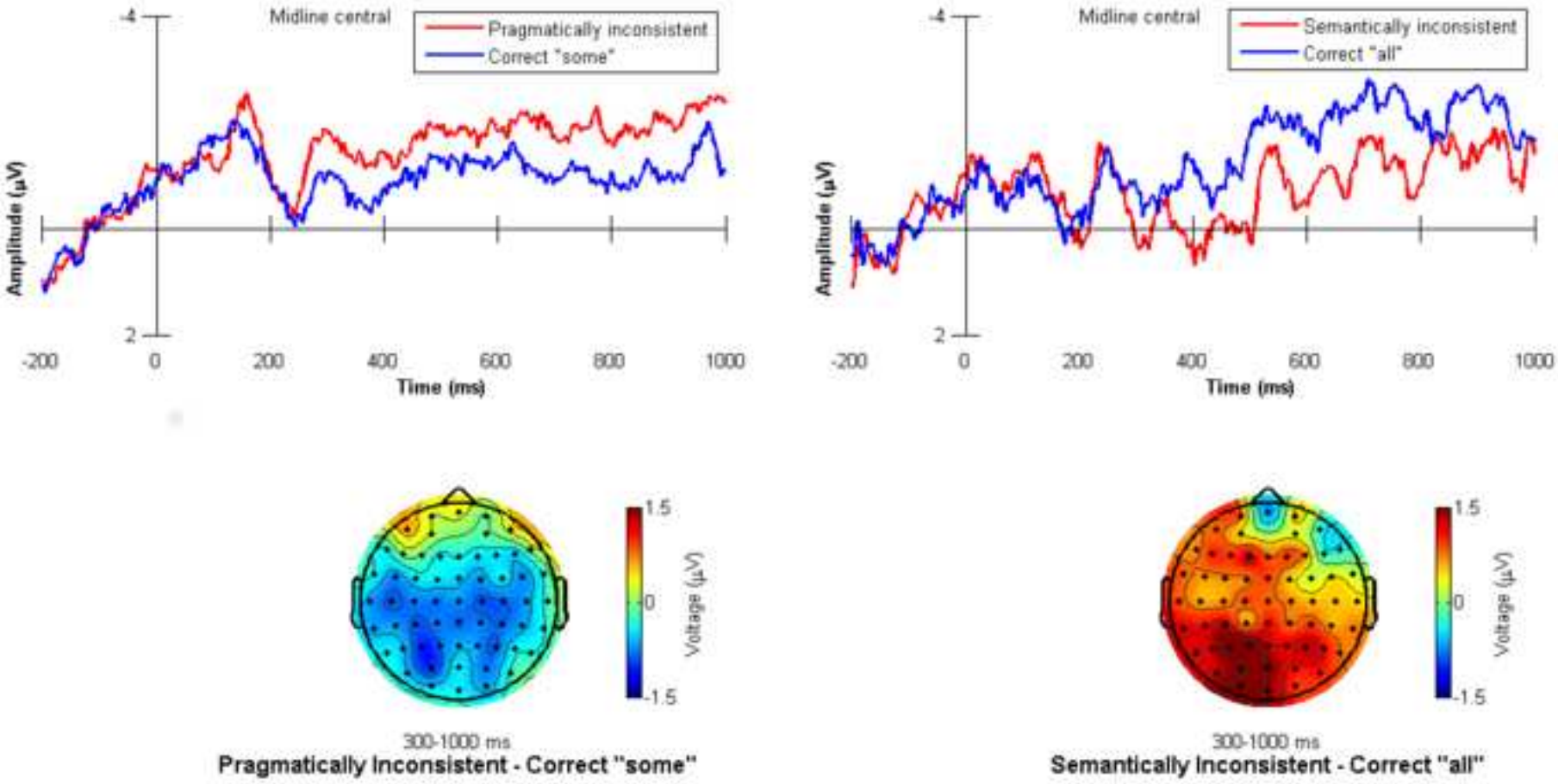

Figure 6 (color)

Click here to download high resolution image

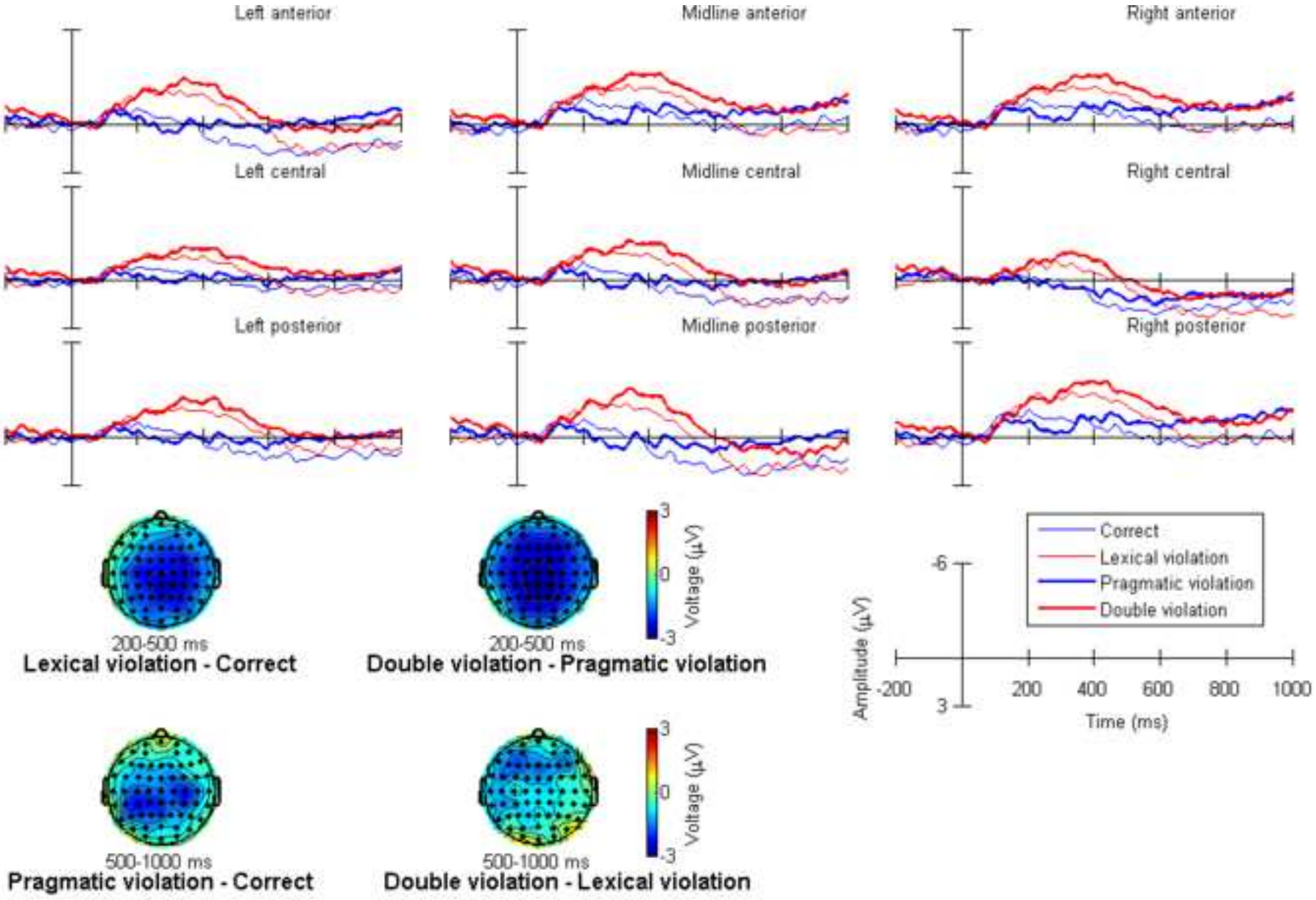

Table 1

Experiment 1

Effect	200-500 ms	500-1000 ms
Quantifier	$F(1,18) = 1.07$	$F(1,18) = 4.04$*
	$F(1,18) = 2.42$	$F(1,18) = 1.07$
Consistency	$F(1,18) = 0.15$	$F(1,18) = 2.08$
	$F(1,18) = 0.18$	$F(1,18) = 2.34$
Region	$F(5,90) = 49.19$****	$F(5,90) = 20.67$****
	$F(2,36) = 38.60$****	$F(2,36) = 11.12$***
Quantifier × Consistency	$F(1,18) = 1.92$	$F(1,18) = 2.63$
	$F(1,18) = 2.46$	$F(1,18) = 1.04$
Quantifier × Region	$F(5,90) = 2.98$**	$F(5,90) = 0.05$
	$F(2,36) = 1.90$	$F(2,36) = 0.48$
Consistency × Region	$F(5,90) = 6.73$***	$F(5,90) = 0.65$
	$F(2,36) = 7.25$***	$F(2,36) = 0.64$
Quantifier × Consistency × Region	$F(5,90) = 0.31$	$F(5,90) = 3.06$**
	$F(2,36) = 0.14$	$F(2,35) = 0.50$

Table 2

Experiment 2 – quantifiers

Effect	300-1000 ms
Quantifier	$F(1,19) = 0.06$
	$F(1,19) = 0.08$
Consistency	$F(1,19) = 0.85$
	$F(1,19) = 1.70$
Region	$F(5,95) = 71.96$****
	$F(2,38) = 34.36$****
Quantifier × Consistency	$F(1,19) = 10.92$**
	$F(1,19) = 6.10$**
Quantifier × Region	$F(5,95) = 1.30$
	$F(2,38) = 0.75$
Consistency × Region	$F(5,95) = 1.51$
	$F(2,38) = 0.98$
Quantifier × Consistency × Region	$F(5,95) = 2.83$**
	$F(2,38) = 0.82$

Table 3

Experiment 2 – objects

Effect	200-500 ms	500-1000 ms
Pragmatic Consistency	$F(1,19) = 0.36$	$F(1,19) = 22.96$****
	$F(1,19) = 0.45$	$F(1,19) = 23.76$****
Lexical Consistency	$F(1,19) = 58.82$****	$F(1,19) = 0.21$
	$F(1,19) = 53.15$****	$F(1,19) = 0.06$
Region	$F(5,95) = 60.48$****	$F(5,95) = 29.46$****
	$F(2,38) = 54.18$****	$F(2,38) = 27.64$****
Pragmatic Consistency × Lexical Consistency	$F(1,19) = 0.60$	$F(1,18) = 0.27$
	$F(1,19) = 0.60$	$F(1,19) = 0.19$
Pragmatic Consistency × Region	$F(5,95) = 0.57$	$F(5,95) = 1.24$
	$F(2,38) = 1.88$	$F(2,38) = 0.48$
Lexical Consistency × Region	$F(5,90) = 1.38$	$F(5,95) = 1.05$
	$F(2,38) = 2.59$	$F(2,38) = 0.59$
Pragmatic Consistency × Lexical Consistency × Region	$F(5,90) = 0.30$	$F(5,95) = 2.26$*
	$F(2,38) = 0.15$	$F(2,38) = 1.73$