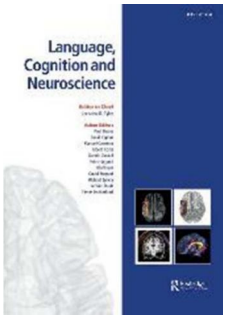


Involvement of prefrontal cortex in scalar implicatures: evidence from magnetoencephalography

Journal:	*Language, Cognition and Neuroscience*
Manuscript ID:	PLCP-2014-OP-9658.R2
Manuscript Type:	Original Paper
Date Submitted by the Author:	23-Feb-2015
Complete List of Authors:	Politzer-Ahles, Stephen; New York University, Abu Dhabi, NYUAD Institute Gwilliams, Laura; New York University, Psycholinguistics
Keywords:	scalar implicature, pragmatics, magnetoencephalography, prefrontal cortex

Involvement of prefrontal cortex in scalar implicatures: evidence from magnetoencephalography

Stephen Politzer-Ahles*, Laura Gwilliams

NYU Abu Dhabi Institute, New York University, Abu Dhabi, United Arab Emirates

**Address correspondence to*
Stephen Politzer-Ahles
NYUAD
PO Box 129188
Abu Dhabi
UAE
Tel.: +971-56-689-9497
E-mail: spa268@nyu.edu

Additional contacts
Laura Gwilliams
NYUAD
PO Box 129188
Abu Dhabi
UAE
Tel.: +971 563190446
E-mail: laura.gwilliams@nyu.edu

Involvement of prefrontal cortex in scalar implicatures: evidence from magnetoencephalography

While many recent studies have investigated scalar inferences (i.e., the interpretation of *some* as meaning *some but not all*) using offline as well as online behavioural methods, little is known about the neural correlates of scalar inference realisation. The present study used magnetoencephalography, which has high temporal resolution, to measure neural activity while participants heard stories that included the scalar inference trigger *some* in contexts that either provide strong cues for a scalar inference or provide weaker cues. The middle portion of the lateral prefrontal cortex (Brodmann area 46) showed an increased response to *some* in contexts with fewer cues to the inference, suggesting that this condition elicited greater effort. While the results are not predicted by traditional all-or-nothing accounts of scalar inferencing that assume the process is always automatic or always effortful, they are consistent with more recent gradient accounts which predict that the speed and effort of scalar inferences is strongly modulated by numerous contextual factors.

Keywords: scalar implicature; pragmatics; magnetoencephalography; prefrontal cortex

Introduction

Comprehending a natural utterance generally involves inferring messages that were not explicitly expressed. For example, the literal meaning of (1) is (1a), but in many contexts a competent listener will interpret (1) as meaning (1b).

(1) Some of the flowers have bloomed.

(a) At least one of the flowers has bloomed.

(b) Some, *but not all*, of the flowers have bloomed.

Meaning (1b), the *pragmatic interpretation*, is presumably realised by generating a set of stronger alternative sentences that were not spoken (e.g., "All of the flowers have bloomed") and inferring that a competent speaker choosing not to utter a stronger

sentence must have intended to convey that the stronger sentence is not true, e.g., that not all of the flowers had bloomed (Chemla & Singh, 2014; Horn, 1972; Katsos & Cummins, 2010; Noveck & Sperber, 2007; Sauerland, 2012). This process is known as a *scalar inference*. The pragmatic interpretation (1b) can be cancelled without yielding a self-contradictory sentence, as shown in (2); furthermore, in certain contexts the pragmatic interpretation may not arise at all, such as in semantic contexts where the alternative sentence is weaker rather than stronger (3) and epistemic contexts where the speaker does not have enough knowledge of the situation to license the scalar inference (4). Meaning (1a), the lexical interpretation, does not have these properties.

(2) Some of the flowers have bloomed; in fact, all of them have.

(3) If some of the flowers bloom, the garden will look beautiful.

(4) I took a quick glance at the garden and saw that some of the flowers had bloomed.

A major question of interest in psycholinguistics is how scalar inferences are cognitively realised. In recent years there has been substantial debate over whether the mechanisms for scalar inferencing are pragmatic (per Horn, 1972) or grammatical (per Chierchia, Fox, & Spector, 2012), and whether the pragmatic meaning is realised immediately and automatically (per Chierchia, 2004, and Levinson, 2000) or with a processing cost (per e.g. Sperber & Wilson, 1995). The debate over whether inference realisation is pragmatic or grammatical is mainly a question of listeners' competence and the unit of analysis is often overt judgments of whether or not enriched meanings are realised in various types of complex sentences (see e.g. Chemla & Singh, 2014; Sauerland, 2012), although the answer to this debate could also have implications for processing. On the other hand, the debate over the speed and context-dependence of scalar inferencing (see e.g. Katsos & Cummins, 2010; Noveck & Reboul, 2008) is

concerned mainly with psycholinguistic measures, but the specific grammatical or pragmatic mechanisms underlying putative processing costs (or lack thereof) are not as clearly spelled out. Both the question of pragmatic vs. grammatical processing and that of cost vs. automaticity have been addressed in numerous behavioural experiments using methods including online forced-choice paradigms (Bott, Bailey, & Grodner, 2012; Bott & Noveck, 2004; Chevallier et al., 2008; Feeney et al., 2004), offline forced-choice paradigms (Chemla & Spector, 2011; Degen & Tanenhaus, 2014; Geurts & Pouscoulous, 2009), self-paced reading (Bergen & Grodner, 2012; Breheny, Katsos, & Williams, 2006; Hartshorne & Snedeker, submitted; Lewis, 2013; Politzer-Ahles & Fiorentino, 2013), dual-task (De Neys & Schaeken, 2007; Dieussaert, Verkerk, Gillard, & Shaeken, 2011; Marty & Chemla, 2013; Marty, Chemla, & Spector, 2013), and visual world eye-tracking (Breheny, Ferguson, & Katsos, 2012, 2013; Degen & Tanenhaus, 2014; Grodner, Klein, Carbary, & Tanenhaus, 2010; Huang & Snedeker, 2009).

Less attention has been given, however, to the neural correlates of scalar implicature. Thus far, only one experiment has investigated the neural substrates of scalar inferencing using a method with high spatial resolution. Shetreet, Chierchia, and Gaab (2014) used functional magnetic resonance imaging (fMRI) to examine brain regions activated in *some*-sentences (e.g., "Some mice have grapes") and in *every*-sentences (e.g., "Every penguin is on the bus") during a picture-sentence verification task. They found that left inferior frontal gyrus (LIFG; Brodmann area 47) was activated more for *some*-sentences, which evoked scalar inferences, than for *every*-sentences, which did not. The authors argue that LIFG may thus be the locus of semantic aspects of inference computation, such as the generation of relevant alternatives and enrichment of the quantifier's interpretation via negation of the alternatives (see Chierchia, 2004; Chierchia et al., 2012).

This finding is somewhat complicated, however, by several aspects of the study's methods and design. While fMRI has excellent spatial resolution, it has poor temporal resolution, making it difficult to know at what point in the sentence the LIFG effect was elicited (an issue of importance, since several accounts of scalar inferencing argue that it occurs immediately when the scalar expression, e.g. *some*, is encountered). On the other hand, methods with high temporal resolution, such as magnetoencephalography and electroencephalography, would allow researchers to disentangle inference-related neural activity elicited at the moment a scalar expression is presented, from other processes (such as verification of the upper-bounded meaning relative to the context) that may be elicited later as the sentence unfolds. Furthermore, the critical comparison in the experiment was between different words, *some* and *every*. The effects observed could reflect downstream verification strategies rather than inferencing itself. Crucially, the denotations of *some* and *every*, and thus the verification strategies they may induce, are different. Verifying *every* requires only checking for one counterexample; the same is true for verifying the semantic interpretation of *some*, as it requires only finding one instance of e.g. a mouse with grapes. Verifying *some [but not all]*, on the other hand, requires identifying two different subsets (e.g., mice that have grapes and mice that do not). Thus, the effects could be due to processing different denotations rather than to the actual process of realising the enriched meaning.[1] To rule out such differences, it would be valuable to test scalar inferencing in a paradigm that 1) does not require an explicit verification task; 2) compares *some* in an inference-triggering context to *some* in a non-inference-triggering context, rather than comparing *some* to a different word; and 3) includes controls that replicate the differences in denotation between lexical *some* ("at least one") and pragmatic *some* ("at least one, but not all") but do not involve scalar inferences. Observing similar frontal activation in

such a design would add converging evidence that this region is involved in some aspect of scalar inferencing.

The present study

The present study used magnetoencephalography (MEG), which has the requisite combination of good spatial and temporal resolution, to investigate the neural substrates of realising scalar inferences. While several previous studies have used high-temporal-resolution techniques to examine inferencing (Chevallier, Bonnefond, Van der Henst, & Noveck, 2010; Hartshorne, Liem Azar, Snedeker, & Kim, in press; Hunt, Politzer-Ahles, Gibson, Minai, & Fiorentino, 2013; Nieuwland, Ditman, & Kuperberg, 2010; Noveck & Posada, 2003; Politzer-Ahles, Fiorentino, Jiang, & Zhou, 2013; Sikos, Tomlinson, Traut, & Grodner, 2013; Zhao, Liu, Chen, & Chen, 2015), they have all used electroencephalography (EEG), which has poorer spatial resolution; furthermore, other than Hartshorne et al. (in press) and Sikos et al. (2013), these studies have all used violation paradigms and/or examined words downstream of the quantifier. The present study is the first study with high spatial resolution to examine successful scalar inferencing while also controlling for the lexical issues described in the previous section.

We adopted a paradigm that has been widely used in reading time studies (Bergen & Grodner, 2012; Breheny et al., 2006; Hartshorne & Snedeker, submitted; Lewis, 2013; Politzer-Ahles & Fiorentino, 2013) and one EEG study (Hartshorne et al., in press), in order to contrast *some* in contexts that strongly support a scalar inference and in contexts that do not. An example is shown in (5). The first two sentences establish a context that introduces a salient set of referents and asks a question about either *all* of them (creating an upper-bounded context, 5a,c) or about *any* of them

(creating a lower-bounded context, 5b,d). *Some* is more likely to be interpreted pragmatically in an upper-bound context, where there is an explicit stronger alternative, than in a lower-bound context. Thus, brain regions associated with making scalar inferences may be expected to show greater activation in response to the word *some* in the upper-bound items.

(5)

(a) **Upper-bound *some***: Mary was preparing to throw a party for John's relatives. She asked John whether all of them were staying in his apartment. John said that some of them were.

(b) **Lower-bound *some***: Mary was preparing to throw a party for John's relatives. She asked John whether any of them were staying in his apartment. John said that some of them were.

(c) **Upper-bound *only some***: Mary was preparing to throw a party for John's relatives. She asked John whether all of them were staying in his apartment. John said that only some of them were.

(d) **Lower-bound *only some***: Mary was preparing to throw a party for John's relatives. She asked John whether any of them were staying in his apartment. John said that only some of them were.

While a lower-bound context like (5b) is often treated as an instance in which the scalar inference is not licensed (see, e.g., Breheny et al., 2006), and this context manipulation has been shown to modulate downstream reading times for words whose interpretation depends on the inference (Breheny et al., 2006; Lewis, 2013; Politzer-Ahles & Fiorentino, 2013), offline ratings we conducted (see Methods) showed that participants still incorporated the implicature into their ultimate interpretations of these sentences in this set of stimuli. Thus, we refer to these conditions as contexts that

provide stronger (5a, upper-bound) or weaker (5b, lower-bound) cues supporting the eventual scalar inference, rather than as conditions that do or do not license the inference.

We also included a pair of control conditions (5c,d) where the *not all* interpretation is made explicit via the addition of *only*. Unlike bare *some*, *only some* is semantically specified to mean *some but not all* and this meaning cannot be cancelled (Minai & Fiorentino, 2010). These conditions have the same context manipulation as the critical conditions (5a,b) but do not involve scalar inferencing. Thus, brain regions involved in scalar inferencing can be limited to those showing an interaction such that there is a context effect in the *some* conditions (5a,b) and not the *only some* conditions (5c,d).[2]

The direction of the context effect to be expected is an empirical question. Under accounts that assume a processing cost for scalar inferencing across the board (e.g. Noveck & Sperber, 2007), one would expect to see greater activation in the more strongly inference-supporting (upper-bounded) context. This may reflect extra effort involved in perspective-taking to make primary implicatures, under a pragmatic account, or extra effort involved in applying semantic operations to parse the strengthened interpretation, under a semantic account; some of these putative operations are likely to be required under both pragmatic and semantic accounts (see Chemla & Singh, 2014). On the other hand, under accounts that assume scalar inferences are made by default (e.g. Levinson, 2000), no difference in neural activation would be predicted between the two conditions (given that scalar inferences were realised at similar rates in both conditions). Results from reading time experiments using this paradigm are mixed; some find slowdowns for *some* in the more strongly inference-supporting context (Bergen & Grodner, 2012; Breheny et al., 2006; Hartshorne & Snedeker, submitted,

Experiment 1), consistent with accounts assuming extra processing effort for inferencing, whereas others find no slowdown for *some* (Hartshorne & Snedeker, submitted, Experiment 2; Lewis, 2013; Politzer-Ahles & Fiorentino, 2013), and none find slowdowns for *some* in the less strongly inference-supporting (lower-bounded) context. In the only electrophysiological study using this paradigm, Hartshorne and colleagues (in press) found no difference in scalp EEG for *some* (although they did, like all of the reading-time studies, find effects downstream confirming that scalar inferences were more available in the upper-bounded than lower-bounded contexts). In the present study we focus on source-level analysis, but in order to compare the results to those of Hartshorne and colleagues (in press) we also report sensor-level findings.

Methods

Participants

Eleven native English speakers (9 female, mean age = 23, SD = 5.39) took part in the experiment; four additional volunteers participated but were not included in the data analysis because of excessive artifacts in their recordings. All had normal or corrected-to-normal vision and hearing and were right-handed as assessed by the Edinburgh Handedness Questionnaire (Oldfield, 1971). All participants provided their informed consent and were paid for their participation, and experimental procedures were approved by the Institutional Review Board of New York University Abu Dhabi.

Materials

One hundred and twenty-eight vignettes were created for the critical trials, according to the template shown in (5). The first sentence of each vignette introduced a set of

referents (e.g., John's relatives) to the discourse, and the second established an upper- or lower-bounded question under discussion using either the quantifier *all* or the quantifier *any*. The only difference between contexts is the use of *all* or *any* in the second sentence. Finally, in the implicit upper bound (*some*) conditions (5a,b), the third sentence responded to the question using the phrase *some of them*, which relies on a scalar inference to establish the "some but not all" interpretation. In the explicit upper bound (*only some*) conditions (5c,d), the response sentence uses the phrase *only some of them*, the "not all" interpretation of which is semantically specified without any scalar inference, and which does not rely on contextual support from the discourse cues. This yielded a 2 (CONTEXT: upper- vs. lower-bounded) × 2 (QUANTIFIER: *some* vs. *only some*) design.

An additional 128 vignettes were created to serve as fillers. Sixty-four introduced upper-bounded (*all*) questions under discussion but included answers with the quantifier *all* in the position where *some* or *only some* occurred in the critical items. The other 64 introduced lower-bounded (*any*) questions under discussion but included answers with the quantifier *none*. These fillers served to keep the critical quantifiers from being wholly predictable, and to reinforce the contrast between *some* and *all* in the upper-bounded conditions and the contrast between *some* and *none* in the lower-bounded conditions.

The stimuli were read aloud by a female native speaker of American English in an anechoic chamber at the University of Kansas, who was instructed to avoid placing contrastive stress on the quantifiers. For the critical items, the context sentences of the vignettes were read separately from the critical sentences.[3] Filler items were read as complete vignettes. The recordings were digitised at 44100 Hz and later segmented and intensity-normalised using Praat (Boersma & Weenik, 2014), and the four combinations

of context and critical sentences were then spliced together to create the items. The onset latencies of the quantifiers were measured by hand.

Offline questionnaire

To evaluate the likelihood of making an upper-bounded interpretation of *some* in the stimuli used in this experiment, a separate offline judgment task was conducted (after the MEG experiment, and with different participants) in which participants answered, for each vignette, whether or not an "all" interpretation was possible. For example, participants heard vignette (6) and were then asked whether or not it was possible that all of John's relatives were staying in his apartment.

(6) Mary was preparing to throw a party for John's relatives. She asked John whether all of them were staying in his apartment. John said that some of them were.

The question probed the "all" interpretation, rather than the "not all" interpretation (as was done by Degen, in press), because the presence of an "all" interpretation guarantees that a scalar inference *was not* computed, whereas the presence of a "not all" interpretation does not guarantee that an inference *was* computed—the lower-bounded, semantic interpretation of some may still be consistent with "not all".

The 128 critical items and 128 filler items from the MEG experiment were divided into eight sub-experiments, with 16 critical and 16 filler items each. Within each sub-experiment, the 16 critical items were organised into four versions in a Latin square design, such that each version had four critical items per condition. This resulted in 32 lists for the offline experiment. The lists were administered over the Internet using Qualtrics, and each was completed by three participants, recruited via Amazon

Mechanical Turk. A given participant was allowed to complete multiple sub-experiments, with the constraint that they could not complete more than one Latin-square list within the same sub-experiment (which were clearly indicated as such). A total of 28 unique participants completed the task, with each participant completing between one and eight sub-experiments. Completion of each took about ten minutes.

MEG procedure

The vignettes were presented binaurally to participants over tube earphones (Aearo Technologies), using the Presentation stimulus delivery software (Neurobehavioral Systems). The participants' task was to listen to each vignette for comprehension. Each trial began with the presentation of a fixation cross, which remained on the screen while the vignette was playing. The vignette began playing 250-750 ms after the appearance of the cross, and the cross remained on the screen for an additional 250-750 ms after the sentence offset. Thirty-three percent of trials were followed by a comprehension question presented visually on the screen. The comprehension questions probed various portions of the vignettes, but on critical trials the comprehension questions never asked questions that depended on the interpretation of the critical quantifier (e.g., questions such as "How many of John's relatives were staying?" were not asked). Each question was presented along with two possible answers, and the participant indicated her choice using a button box placed below her left index and middle fingers. The next trial began after the participant made her response (in the case of trials with comprehension questions) or after the participant pressed either button (in the case of trials with no comprehension question).

The stimuli were organised into four lists in a Latin square design. The item order was randomised at runtime for each participant, with the restriction that the first

five trials were always fillers. The recording took about 50 minutes to complete.

Prior to the recording, the participant's head shape was digitised using a dual source handheld FastSCAN laser scanner (Polhemus, VT, USA). Three fiducial points and the position of five marker coils placed around the participant's face were also digitised. This was in order to localise the position of the participant's head in relation to the MEG sensors to allow for source reconstruction.

Data acquisition and preprocessing

MEG recordings used a whole-head 208 channel axial gradiometer system (Kanazawa Institute of Technology, Kanazawa, Japan) with a sampling frequency of 1000 Hz; data were filtered online with a low-pass of 200 Hz and a high-pass of 0.1 Hz. Data were epoched from -200 to 1200 ms, relative to the acoustic onset of the quantifier *some*.[4] Artifact rejection consisted of manual inspection of the data to remove blinks, and amplitude cut-offs at 3000 fT. This resulted in removal of 15% of the trials, leaving an average of 26 trials per condition for each subject. Participants with fewer than 15 trials in any condition were excluded from further analysis. Preprocessed data were averaged for each condition and subject, using the prestimulus interval for baseline correction.

L2 minimum norm estimates of source activity were calculated using BESA 6.0 (MEGIS Software GmbH). Unsigned current estimates were created for each condition and low-pass filtered at 40 Hz. The source space was parcellated into Brodmann areas by converting sources to Talairach space using the Talairach Daemon (www.talairach.org) and matching each point to the nearest labelled Brodmann area in Talairach space. For each condition and each participant, the current estimate for each region was found by averaging the current estimates for all source vertices within that region. Source-space analysis was restricted to the following regions in both

hemispheres: ventromedial prefrontal cortex (BA 11) and temporal pole (BA 38), which have been implicated in basic phrasal composition in MEG (e.g., Bemis & Pylkkänen, 2011); temporo-parietal junction (BA 39), which has been implicated in speech act processing and theory of mind (e.g. Egorova, Shtyrov, & Pulvermüller, 2013); temporal regions implicated in lexical semantics and auditory processing (BAs 21, 22, and 42); and frontal regions (inferior frontal gyrus: BAs 44-45[5]; middle prefrontal cortex: BA 46; ventral orbitofrontal cortex: BA 47) implicated in aspects of structural and grammatical processing and recently implicated in scalar inference generation (Shetreet et al., 2014). These regions were chosen based on *a priori* predictions that they may be involved in some aspect of pragmatic processing or semantic composition.

Statistical analysis

For the sensor-level statistical analysis, we conducted non-parametric spatiotemporal clustering (Maris & Oostenveld, 2007). While the experiment used a 2×2 design (see Materials), the 2×2 test was computationally implemented using a series of t-tests on selected subsets of the data, which is conceptually comparable to running factorial analyses of variance when the factors only involve two levels (see, e.g., http://mailman.science.ru.nl/pipermail/fieldtrip/2011-January/003447.html). To test the main effect of CONTEXT, two datasets were created by averaging, respectively, the upper-bounded *some* and *only some* datasets, and the lower-bounded *some* and *only some* datasets; these two datasets were then compared using paired *t*-tests. The main effect of QUANTIFIER was tested in a similar way. The interaction between CONTEXT and QUANTIFIER (i.e., the difference of differences) was tested by creating two datasets representing the two context effects (subtracting the lower-bounded *some* from the upper-bounded *some* dataset, and the lower-bounded *only some* from the upper-bounded *only some* dataset), and comparing these context effects using paired *t*-tests.

Source-level statistics were performed by conducting nonparametric temporal clustering (Maris & Oostenveld, 2007) on the estimates from 200-1000 ms post stimulus onset within each region, and correcting for false discovery rate (Benjamini & Yekutieli, 2001) across regions. F-tests were used to measure the main effects of CONTEXT and QUANTIFIER, and a custom statistic was used to measure the crucial interaction. Nonparametric temporal clustering allows for the use of user-defined test statistics, and in this case we constructed a test statistic that quantifies the presence of a CONTEXT × QUANTIFIER interaction such that there is a large effect of CONTEXT in either direction for *some* sentences, but no effect of CONTEXT for *only some* sentences; see equation (1). This statistic is large when such an interaction is present, and small otherwise.

$$\left|t_{(\text{upper-bound } some)} - (\text{lower-bound } some)\right| - \left|t_{(\text{upper-bound } only\ some)} - (\text{lower-bound } only\ some)\right| \quad (1)$$

For the largest cluster (any series of 10 or more contiguous timepoints with *p*-values[6] of .3 or less) in each region, a cluster test statistic was computed by summing the sample test statistics (the F-tests of the main effects, or the statistic in (1) for the interaction) over all samples in the cluster, and then compared to 10,000 permutation test statistics each calculated by permuting the condition labels of the data and re-calculating the test statistic. The proportion of permutation test statistics greater than the original cluster test statistic was the *p*-value for that region. False discovery rate adjustment was then applied to the *p*-values of the largest clusters of the regions.

Results

Offline questionnaire

On the filler items, participants correctly answered "yes" (i.e., indicated that an "all"

interpretation was possible) to 98.70% (SD=3.39%) of *all*-quantifier fillers and only answered "yes" to 0.52% (SD=3.33%) of *no*-quantifier fillers, indicating that participants were attentive to the final sentences of the vignettes. One of the critical items was incorrectly coded in the web version of the experiment and was removed from subsequent analyses. The mean percentages of "yes" responses (indicating upper-bounded or non-pragmatic readings) for the rest of the critical items are shown in the left portion of Figure 1 below. For statistical analysis, responses were treated as a categorical variable and were modelled with generalised linear mixed models, with fixed effects of QUANTIFIER, CONTEXT, and the QUANTIFIER × CONTEXT interaction, and maximal random effects structures for SUBJECT, ITEM, and Latin square LIST.

--Figure 1 about here--

A slight numerical effect of Quantifier is evident, with *some* vignettes more likely than *only some* vignettes to be consistent with an "all" interpretation. This effect failed to reach significance, however, in the mixed model with maximal random effects ($\chi^2(1) = 0.08$, $p = .776$), although it did reach significance in a model with only random intercepts ($\chi^2(1) = 38.52$, $p < .001$) and in repeated-measures ANOVA ($F(1,27) = 5.19$, $p = .019$). The effect of Context was not significant in any analysis (maximal mixed model: $\chi^2(1) = 0.07$, $p = .787$; intercept-only mixed model: $\chi^2(1) = 1.03$, $p = .31$; ANOVA: $F(1,27) = 2.69$, $p = .113$), nor did the interaction (maximal mixed model: $\chi^2(1) = 0.19$, $p = .663$; intercept-only mixed model: $\chi^2(1) = 0.32$, $p = .571$; ANOVA: $F(1,27) = 0.44$, $p = .513$)

As is evident from Figure 1, the variance between participants was much greater for *some*, which is ambiguous, than for *only some*, which is unambiguous. This is due to a split in the some responses between pragmatic responders, who tend to interpret *some* as upper-bounded, and semantic responders, who do not; such group differences are

common in the experimental literature on scalar inferences (Noveck & Posada, 2003; Bott & Noveck, 2004; Hunt et al., 2013). The higher average rate of "all" interpretations for *some* is apparently driven by the "tail" of a few semantic responders. Histograms of the subject means for *some* and *only some* items (averaged across upper-bound and lower-bound contexts) are shown in the right-hand portion of Figure 1.

Behavioural

All participants kept in the electrophysiological analysis responded with greater than 80% accuracy on critical trials. Comprehension accuracy was 93.3% for SOME_all items (items with the quantifier *some*, in the upper-bounded *all* context), 86.2% for SOME_any items, 94.1% in ONLYSOME_all items, and 94.1% in ONLYSOME_any items. Accuracy was measured using generalised linear mixed models comprising fixed effects of CONTEXT, QUANTIFIER, the CONTEXT × QUANTIFIER interaction, and TRIAL NUMBER, and random effects (including both random intercepts and maximal random slope structure, per Barr, Levy, Scheepers & Tily, 2013) of PARTICIPANT, ITEM, and LIST. Model comparisons using log-likelihood tests showed no significant fixed effects (χ^2s < 2.59, *p*s > .1).

Sensor space

Event-related fields evoked by *some* are shown in Supplementary Figure 1. The spatiotemporal clustering analysis in sensor space revealed no significant clusters for either the main effect of CONTEXT, the main effect of QUANTIFIER, or the CONTEXT × QUANTIFIER interaction.

Source space

One of the regions tested, left BA 46 (corresponding to the middle prefrontal cortex,

MPFC; $p_{unc.}$ = .002, p_{FDR} = .04), showed an interaction in the 571-711 ms time window that reached significance after FDR adjustment. Another region, left BAs 44-45 (corresponding to the inferior frontal gyrus; $p_{unc.}$ = .01, p_{FDR} = .1) showed a very marginal interaction in the 561-669 ms time window. The current estimates for these regions are displayed in Figure 2. Current estimates for the rest of the regions tested are shown in Supplementary Figure 2 in the online version of the article, and whole-brain maps of the current estimates are shown in Supplementary Figure 3. The results of the statistical analysis are shown in Table 1.

--Figure 2 about here--

--Table 1 about here--

In each of the significant regions, the FDR-significant interaction was due to greater activity evoked by bare *some* in the lower-bounded context with weaker cues for the inference than upper-bounded context with stronger cues (BA 46: $t(10) = 3.41$, 95% CI = 1.02–4.87; BAs 44-45: $t(10) = 4.44$, 95% CI = 1.65–4.98), and no substantial context effect evoked by *only some* (BA 46: $t(10) = -0.15$, 95% CI = -1.34–1.17; BAs 44-45: $t(10) = -0.66$, 95% CI = -1.55–0.84).

No regions showed FDR-significant main effects of either CONTEXT ($p_{unc.}$s > .018, p_{FDR}s > .38) or QUANTIFIER ($p_{unc.}$s > .125, p_{FDR}s > .999).

Discussion

The present study tested the neural-level computation of scalar inferences using the high temporal and spatial resolution of magnetoencephalography, and using a paradigm which allows for a direct comparison of the same word (*some*) occurring either in contexts that provide strong cues for incorporating a scalar implicature into its meaning, or contexts that do not. At the sensor level we observed no effect of scalar inferencing, consistent with Hartshorne and colleagues (in press), who only observed EEG effects of

inferencing well after the triggering expression. On the other hand, at the source level, we found that the MPFC (BA 46) was selectively sensitive to scalar inferencing: specifically, greater activation was evoked by *some* in context that provided *less* support for the inference, compared to the contexts that provided more support. The location of this activation is consistent with, although somewhat dorsal of, the region implicated in scalar inferencing by the fMRI study of Shetreet and colleagues (2014); nonetheless, the nature of the effect may be different.

Neural correlates of scalar inferencing

The present study observed activation related to scalar inferences in the left MPFC (BA 46), as well as a similar trend that did not reach significance in left inferior frontal gyrus (BAs 44-45). Of these two regions, the latter is spatially quite close to the activation that Shetreet and colleagues (2014) observed in BA 47 (MNI coordinates: -36, 17, -22) for the comparison between *some* sentences and *every* sentences (i.e., their "implicature generation") comparison—the region the authors implicated in the realisation of scalar inferences. On the other hand, the most significant region identified in the present study, BA 46, is closer to activation Shetreet and colleagues observed in BA 10 (middle frontal gyrus; MNI coordinates: -27, 44, -2) for the comparison between felicitous *some* (i.e., *some* describing pictures in which some but not all of the elements have the property described) versus infelicitous *some* (i.e., *some* describing pictures in which all of the elements have the property described), that is, the "implicature mismatch" comparison. They argued that this region is involved in managing the conflict between the inference-based meaning and the pragmatically mismatching context. In short, the present study observed activation that partially overlaps with the LIFG activation in Shetreet et al. (2014) but also extends to slightly more dorsal and anterior regions, and may overlap

with the region implicated in violation sentences as well as the region implicated in basic inference realisation.

The present results offer converging evidence that these regions are indeed involved in comprehending scalar inferences, rather than, for instance, verifying different denotations or performing different kinds of quantification. These results also extend the previous finding by showing that this activation occurs rapidly, within 600 ms after the realisation of the quantifier (adding to previous evidence that scalar implicatures, rather than being postponed until the end of a sentence or proposition as traditional Gricean accounts assume, are realised immediately and *in situ* during incremental sentence processing; see Nieuwland et al., 2010; Politzer-Ahles et al., 2013, Sikos et al., 2013; see Geurts, 2010, for discussion of how immediate incremental computation of implicatures can be reconciled with a Gricean view; see also Chemla & Spector, 2011, Chemla & Singh, 2014, Geurts & Pouscoulous, 2009, among others, for further discussion of local vs. global computation of scalar inferences). The fact that this activation was strongest in MPFC is interesting because the lateral prefrontal cortex has been suggested to play a role in working memory (e.g. Barbey, Koenigs, & Grafman, 2013, among others), and there is also some evidence that scalar inference realisation may be modulated by working memory (Dieussaert et al., 2011; Feeney et al., 2004; Marty & Chemla, 2013; Politzer-Ahles, Fiorentino, Durbin, & Li, 2014). This region, however, is also implicated in many other aspects of cognition, such as planning and cognitive control (Caplan, 2006), comprehension of syntactic structure (Hashimoto & Sakai, 2002), and lexical retrieval (Chee, Hon, Caplan, Lee, & Goh, 2002); therefore, the relationship between working memory and scalar inferencing at the neural level requires further investigation before strong conclusions regarding this topic can be made.

While the present experiment supports the conclusions of Shetreet and colleagues (2014) that frontal areas are involved in scalar inferencing, there are some important differences between the findings. Perhaps most noticeably, the present study observed the greatest activation for a condition that provides fewer cues for an inference (*some* in a context with a lower-bounded question under discussion), whereas Shetreet and colleagues observed the greatest activation in an inference condition (*some* sentences) compared to a no-inference condition (*every* sentences). The numerous differences between the studies in terms of design, critical comparisons, and methodologies used make it difficult to determine why the activation pattern is reversed—for example, fMRI activation observed in these brain regions in the previous study may reflect processes occurring at very different times than what was measured in the present study. The following section will discuss in more detail the potential implications of the direction of the effect observed in the present study. Another important difference between the present study and that of Shetreet and colleagues is that the previous study attributed two different loci of activation (inferior frontal and prefrontal) to two different cognitive processes (realising inferences, or "implicature generation", and comprehending infelicitous inferences, or "implicature mismatch"), whereas the present experiment observed activation overlapping with both of these areas in the same contrast. Again, the differences in contrasts and methods between these two studies make it premature to draw strong conclusions; in particular, more research into the neural correlates of implicature *violations* using techniques like MEG is needed, as the present study did not test violations and previous electrophysiological studies using violation paradigms (e.g., Hunt et al., 2013; Nieuwland et al., 2010; Noveck & Posada, 2003; Panizza et al., 2014; Politzer-Ahles et al., 2013; Zhao et al., 2015) did not perform source localisation. As the spatial resolution of fMRI is superior

to that of MEG, and the different regions identified in Shetreet et al. (2014) are close together, care should be taken in interpreting apparent differences between the results of the present study and of that study.

A final point to take note of is that in the present study, effects of scalar inferencing were only observed in the source analysis, not in the sensor analysis. This is consistent with an EEG study by Hartshorne and colleagues (in press), using a similar design as the present study's, which did not observe an effect of scalar inferencing at the quantifier, but confirmed through analysis of downstream words that the context did indeed modulate comprehension of *some*. Thus, it appears that the neural correlates of making scalar inferences that are observed at the source level may be difficult to detect at the scalp level in EEG or MEG.

Implications for processing models

Before addressing what the brain data can tell us about how scalar inferences are realised, an important question is whether the participants in the experiment even did realise them at all. While previous reading time evidence suggests that participants differentially realise or fail to realise scalar inferences as a function of upper- versus lower-bounded context with similar stimulus manipulations (Lewis, 2013; Politzer-Ahles & Fiorentino, 2013) and intuition suggests that the inference is less available in the nonsupporting lower-bounded context (Katsos & Cummins, 2010), our post-hoc offline questionnaire found that these participants were about equally likely to make scalar inferences in both upper- and lower-bound contexts with the present set of materials . This could be because *some* is such a strong cue for the upper-bounded interpretation, especially in an experiment where so many similar vignettes are presented together that this interpretation is generated even when it is not relevant. That indeed is the prediction of defaultist processing models (Levinson, 2000), but it could

also be modelled in Bayesian frameworks (Goodman & Stuhlmüller, 2013; Grodner & Russell, 2013), where in the lower-bounded context that makes the inference less relevant to the question under discussion, the probability of the speaker uttering *some* given the intended message "more than none" is still low. While our offline data have limitations (they are from a separate group of participants as the MEG experiment, and rather than using online measures of scalar inference realisation they only probe the eventual outcome of the inference), they do suggest that the context manipulation in the present experiment may have affected the ease of realising inferences, rather than the ultimate interpretation. In the future, experiments testing this question using MEG data from a downstream anaphorical expression (e.g., following the design of Hartshorne and colleagues, in press) would be valuable in further investigating this issue. [7]

As for the measures at the quantifier itself, traditional processing accounts of scalar inferencing make distinct predictions about the direction of context effects in behavioural measures and, by extension, neural activity. Context-driven processing accounts of scalar implicature, including that of Relevance theory (e.g. Noveck & Sperber, 2007) predict that realising scalar inferences always engenders a processing cost, and therefore there should be greater cognitive effort (which may be reflected by increased reading times, delayed eye movements, and possibly greater neural activity) when comprehending expressions that trigger an inference, compared to expressions that do not. This account does not necessarily predict that the amount of effort should be the same across contexts that do trigger inferences; it is feasible that more effort could be required in a context that provides less support for the inference. Such an account is thus consistent with the present results, although this pattern of results can be accounted for more explicitly by a constraint-based formulation of context-driven processing (see below).

Defaultist processing models (e.g., Levinson, 2000), on the other hand, suppose that scalar inferences of the sort tested here are realised automatically by default, and that realising the lower-bounded interpretation (i.e., realising that *some* may be consistent with *all* when its meaning is not enriched to include "and not all") is what takes more effort. Since our offline survey suggested that context did not affect the extent to which participants cancelled inferences and realised the lower-bounded interpretation of *some*, such models would have difficulty explaining why the lower-bounded condition in the present study engendered greater MEG activity: if inferences were realised automatically and with equal effort in both conditions, and did not require more cancellation in either condition, it is not clear what other mechanism the additional activation could be reflecting. Even if we assume that only the participants in the MEG study performed more inference cancellation in the lower-bounded than upper-bounded contexts, and that the participants in the offline questionnaire did not, there are conceptual challenges to attributing the observed MEG effect to cancellation. First of all, the prediction of greater processing cost in an inference-nonsupporting context only makes sense along with the assumption that the participant *must* actively cancel illicit inferences. In the paradigm used here, however, it is not clear why a participant would need to cancel even an illicit inference, as the inference does not conflict with anything else in the utterance, and the exchange can still be fully understood even with the unnecessary information added by the inference (see Politzer-Ahles & Fiorentino, 2013, for further discussion). Furthermore, some formal accounts of scalar implicature are not conducive to mechanisms for cancellation. For example, while early formal accounts included mechanisms for cancellation (e.g. Chierchia, 2004; Levinson, 2000), grammatical or lexicalist accounts of inferencing could also be formulated as involving competition between two readings (the enriched and non-

enriched) which are both realised before one is ultimately selected (e.g. Chemla & Singh, 2014); under such a view, it is not clear why a cancellation mechanism would be necessary, or why selecting one reading would be more computationally costly than selecting another reading.

While the psycholinguistic literature on scalar implicatures has traditionally been framed in terms of the competing context-driven and defaultist accounts, other formulations of scalar inferencing may better account for the present results. For instance, Degen and Tanenhaus (2011, 2014) have proposed a constraint-based account of inferencing in which inferences are always derived via the same mechanisms, but may appear context-based (i.e., slow and effortful) or default (i.e., rapid and easy) depending on the strength of numerous cues, such as context and prosody, that can interact to facilitate or inhibit the inference. The results of the present study could easily be explained under such an account: participants made the inferences in both contexts that strongly supported the inference and contexts that only weakly supported it (perhaps because the global experimental context introduced so many sentences of the type in (5)), but the inference was more difficult to make in the context with weak support, which provided fewer cues to facilitate the inference. While a computationally explicit account of how the facilitation occurs is beyond the scope of the present paper, one might imagine that the upper-bounded and lower-bounded items in the present experiment include many of the same cues (the word *some*, uttered with the same prosody and in the same experimental context—i.e., a context that did not include numerals and other lexical alternatives to *some*), but the upper-bounded items include an additional cue that the lower-bounded items do not: the explicit upper-bounded question under discussion (whether "all" was true) is a much stronger cue in favour of realising a scalar inference. Similarly, Huang and Snedeker (2011) propose that there

are both bottom-up and top-down routes to realising scalar inferences, and that inferences may be realised rapidly and effortlessly via top-down mechanisms when, for example, the information structure of the context makes the inference readily available even before the triggering expression. Such an account would also predict greater activation in the less supportive context, which might require a more bottom-up mechanism to derive the inference, compared to the more supportive context which explicitly introduces the quantifier *all* and thus allows the comprehender to verbally pre-encode the alternatives (thus giving top-down support to the derivation of alternative messages, which is a crucial part of making scalar inferences, see Chemla & Singh, 2014).

The final alternative we will discuss is a lexical/grammatical ambiguity account, whereby the enriched reading of *some* is not itself derived by an inference but is underlyingly specified in either the lexical entry for *some* (which may have both the lower-bounded and upper-bounded readings listed in the lexicon), or the grammatical parse of the utterance (which may yield parses both with and without an "Exhaustification" operator responsible for triggering the enriched meaning; see Chierchia et al., 2012). Under such accounts, the role of pragmatics in scalar inference is not to actually realise the enriched meaning, but rather to choose between the two competing meanings. In this view, the increased activation observed in the lower-bounded context may reflect greater ambiguity, as inferior frontal gyrus is in some studies more strongly activated by ambiguous words than nonambiguous words, or by subordinate as compared to dominant readings of ambiguous words (Bilenko, Grindrod, Myers, & Blumstein, 2009; Fiebach, Vos, & Friderici, 2004; c.f. Copland et al., 2007). Specifically, in the more supportive context, the salient upper-bounded question under discussion strongly biases the hearer towards realising a "some but not all"

interpretation; on the other hand, the less supportive context does not introduce this sort of bias. Arguably, however, the lower-bounded context may bias the reader just as strongly *against* realising a "some but not all" interpretation. An ambiguity-based account of the present findings would require a more detailed, and perhaps quantitative, account of how the questions under discussion create biases regarding the reading of *some*, as well as an account of the locus of the ambiguity.

In short, further research is required to disentangle several competing accounts with different explanations of the pattern of results observed in the present study. The pattern of results is potentially consistent with gradient constraint-based processing and with an ambiguity account of scalar inferences. One important conclusion that can be made is that the neural findings suggest a more detailed processing account is needed compared to the traditional defaultist vs. context-driven dichotomy that has been drawn in most of the psycholinguistic literature until recently. For example, under the context-driven view which has seen much support from psycholinguistic data, it would have been easy to predict that there is an EEG/MEG component associated with making scalar inferences, that such a component appears in sentences where inferences are made and not sentences where inferences are not made, and that neurolinguistic experiments simply need to identify when and where the component surfaces. The present results suggest that the actual situation is much more complicated, and that the search for scalar implicatures in the brain needs to be informed by more explicit models of what specific operations (such as constraint evaluation or ambiguity resolution strategies) feed into the derivation of inferences, as well as by gradient data on how strongly a given sentence supports an inference rather than just all-or-nothing judgments of whether it does or does not trigger an inference (as in Degen, in press). For future research, it will likely not be enough to just compare inference-supporting and

inference-nonsupporting stimuli, but to also develop ways to test more specific mechanisms such as those described above.

Conclusion

The present study, combining for the first time a design that is linguistically motivated and controlled for all relevant variables with neurolinguistic techniques that provide good temporal and spatial resolution, showed that the prefrontal cortex is involved in the comprehension of scalar inferences. The specific pattern of activation shown, with greater activity elicited in contexts that provide fewer cues to support a scalar inference rather than contexts that provide more cues, is in line with both constraint-based and lexical/grammatical ambiguity-based proposals for how scalar inferences are realised. The data suggest that inference realisation is not a monolithic process, but rather that it may involve multiple more fine-grained mechanisms or cues. The present results offer the first brain-level evidence for an electrophysiological component associated with the realisation of scalar inferences, while also highlighting the need for a more spelled-out model of the specific cognitive operations that contribute to inferential processing.

Acknowledgements

This research was funded by grant G1001 from the NYUAD Institute, New York University Abu Dhabi. We also thank Liina Pylkkänen, Diogo Almeida, and two anonymous reviewers for feedback on previous versions of this paper; Estibaliz Blanco Elorrieta for assistance in data collection; and Kelly Berkson, Natalie Pak, and Kate Coughlin for assistance in stimulus development. The authors are responsible for any errors in the manuscript.

References

Barbey, A., Koenigs, M., & Grafman, J. (2013). Dorsolateral prefrontal contributions to human working memory. *Cortex, 49*, 1195-1205.

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255-278.

Bemis, D., & Pylkkänen, L. (2011). Simple composition: a magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of Neuroscience, 31*, 2801-2814.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*, 1165-1188.

Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1450.

Bilenko, N., Grindrod, C., Myers, E., & Blumstein, S. (2009). Neural correlates of semantic competition during processing of ambiguous words. *Journal of Cognitive Neuroscience, 21*, 960-975.

Boersma, P., & Weenink, D (2014). Praat: doing phonetics by computer [Computer program]. http://www.praat.org/.

Bott, L., Bailey, T., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language, 66*, 123-142.

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

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition, 100*, 434-463.

Breheny, R., Ferguson, H., Katsos, N. (2012). Investigating the time-course of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes, 28*, 443-467.

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Taking the epistemic step: toward a model of on-line access to conversational implicatures. *Cognition, 126*, 423-440.

Caplan, D. (2006). Why is Broca's area involved in syntax? *Cortex, 42*, 46-471.

Chee, M. W., Hon, N. H., Caplan, D., Lee, H. L., & Goh, J. (2002). Frequency of concrete words modulates prefrontal activation during semantic judgments. *Neuroimage, 16*, 259-268.

Chemla, E. & Singh, R. (2014). Remarks on the experimental turn in the study of scalar implicature. *Language and Linguistics Compass, 8*, 373-386.

Chemla, E., & Spector, B. (2011). Experimental evidence for embedded scalar implicatures. *Journal of Semantics, 28*, 359-400.

Chevallier, C., Bonnefond, M., Van der Henst, J-B, & Noveck, I (2010). Using ERPs to capture inferential processes guided by prosodic cues. *Italian Journal of Linguistics, 22*, 125-152.

Chevallier, C., Noveck, I., Nazir, T., Bott, L., Lanzetti, V., & Sperber, D. (2008). Making disjunctions exclusive. *Quarterly Journal of Experimental Psychology, 61*, 1741-1760.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Structures and Beyond, 3*, 39-103.

Chierchia, G., Fox, D., & Spector, B. (2012). Scalar implicature as a grammatical phenomenon. In Maienborn, von Heusinger & Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning, Vol. 3* (pp. 2297-2331). Berlin: Mouton de Gruyter.

Copland, D., Zubicaray, G., McMahon, K., & Eastburn, M. (2007). Neural correlates of semantic priming for ambiguous words: an event-related fMRI study. *Brain Research, 1131*, 163-172.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load. *Experimental Psychology, 54*, 128-133.

Degen, J. (in press). Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. *Semantics and Pragmatics*.

Degen, J. and Tanenhaus, M. (in press). Processing scalar implicature: a constraint-based approach. *Cognitive Science*. doi: 10.1111/cogs.12171

Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: further evidence that scalar implicatures are effortful. *Quarterly Journal of Experimental Psychology, 64*, 2352-2367.

Egorova, N., Shtyrov, Y., & Pulvermüller, F. (2013). Early and parallel processing of pragmatic and semantic information in speech acts: neurophysiological evidence. *Frontiers in Human Neuroscience*, *7*.

Feeney, A., Scafton, S., Duckworth, A., & Handley, S. (2004). The story of some: everyday pragmatic inferences by children and adults. *Canadian Journal of Experimental Psychology, 54*, 128-133.

Fiebach, C., Vos, S., & Friderici, A. (2004). Neural correlates of syntactic ambiguity in sentence comprehension for low and high span readers. *Journal of Cognitive Neuroscience, 16*, 1562-1575.

Geurts, B. (2010). *Quantity Implicatures*. Cambridge: Cambridge University Press.

Geurts, B., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics and Pragmatics*, *2*, 4-1.

Goodman, N., & Stuhlmüller, A. (2013). Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science, 5*, 173-184.

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.

Grodner, D., & Russell, B. (2013). Evidence for a rational probabilistic account of Gricean implicatures. *Poster presented at 26th CUNY Conference on Human Sentence Processing.*

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

Hartshorne, J., Liem Azar, S.,Snedeker, J., & Kim, A. (in press). The neural computation of scalar implicature. *Language, Cognition and Neuroscience*.

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

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

Huang, Y., & Snedeker, J. (2011). Logic and conversation revisited: Evidence for a division between semantic and pragmatic content in real-time language comprehension. *Language and Cognitive Processes, 26*, 1161-1172.

Hunt, L., Politzer-Ahles, S., Gibson, L., Minai, U., & Fiorentino, R. (2013). Pragmatic inferences modulate N400 during sentence comprehension: Evidence from picture-sentence verification. *Neuroscience Letters, 534*, 246-251.

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

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge: MIT press.

Lewis, S. (2013). *Pragmatic enrichment in language processing and development*. PhD dissertation, University of Maryland.

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

Marty, P., & Chemla, E. (2013). Scalar implicatures: working memory and a comparison with only. *Frontiers in Psychology, 4*.

Marty, P., Chemla, E., & Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua, 133*, 152-163.

Minai, U., & Fiorentino, R. (2010). The role of the focus operator only in children's computation of sentence meaning. *Language Acquisition, 17*, 183-190.

Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language, 63*, 324-346.

Noveck, I., & Reboul, A. (2008). Experimental pragmatics: a Gricean turn in the study of language. *Trends in Cognitive Sciences, 12*, 425-431.

Noveck, I., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language, 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*. Basingstoke: Palgrave.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97-113.

Panizza, D., & Onea, E. (2014). Some implicatures take their time: an ERP study on scalar implicatures with 'sentence-picture vs. picture-sentence' verification task. Talk presented at *27th CUNY Conference on Human Sentence Processing*. Columbus, OH.

Politzer-Ahles, S., & Fiorentino, R. (2013). The realization of scalar inferences: context sensitivity without processing cost. *PloS ONE*, *8*(5), e63943.

Politzer-Ahles, S., Fiorentino, R., Durbin, J., & Li, L. (2014). The role of working memory in the online realization of scalar inferences. Poster presented at *27th CUNY Conference on Human Sentence Processing*. Columbus, OH.

Politzer-Ahles, S., Fiorentino, R., Jiang, X., & Zhou, X. (2013). Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification. *Brain Research, 1490*, 134-152.

Sauerland, U. (2012). The Computation of scalar implicatures: Pragmatic, lexical or grammatical? *Language and Linguistics Compass, 6*, 36-49.

Shetreet, E., Chierchia, G., & Gaab, N. (2014). When some is not every: Dissociating scalar implicature generation and mismatch. *Human Brain Mapping, 35*, 1503-1514.

Sikos, L., Tomlinson, S., Traut, H., & Grodner, D. (2013). Incremental computation of scalar implicatures: An ERP study.Poster presented at *26th CUNY Conference on Human Sentence Processing*. Columbia, SC.

Sperber, D., & Wilson, C. (1995). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Yekutieli, D., & Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference, 82*, 171-196.

Zhao, M., Liu, T., Chen, G., & Chen, F. (2015). Are scalar implictures automatically processed and different for each individual? A mismatch negativity (MMN) study. *Brain Research, 1599*, 137-149.

Table 1. Results of the statistical analysis of current estimates. Each row gives uncorrected *p*-value (controlled for multiple comparisons over time, but not for false discovery rate across regions), FDR-adjusted *p*-value (Yekutieli & Benjamini, 1999), and time window (begin and end times in ms) of the largest interaction cluster in that region. See Methods for a description of the test statistic used to quantify the interaction for the purpose of defining clusters. The final column lists the MNI coordinates for the center of the region.

ROI	$p_{unc.}$	p_{FDR}	Cluster times	Coordinates (x, y, z)
Left SFG (BA 10)	.479	.798	[933 951]	-21, 60, 8
Left MeFG (BA 11)	.404	.808	[927 960]	-13, 38, -16
Left MTG (BA 21)	>.99	>.99	[161 182]	-62, -18, -11
Left STG (BA 22)	.235	.783	[800 839]	-63, -27, 8
Left ATL (BA 38)	.567	.872	[118 141]	-36, -11, -32
Left TPJ (BA 39)	.737	>.99	[816 835]	-51, -67, 25
Left TTG (BA 42)	.067	.335	[773 851]	-66, -24, 11
Left IFG (BA 44-45)	.01	.1	[563 668]	-55, 24, 13
Left MPFC (BA 46)	.002	.040	[572 709]	-48, 40, 18
Left vOFC (BA 47)	.212	.848	[116 143]	-37, 27, -15
Right SFG (BA 10)	.807	>.99	[493 523]	25, 60, 8
Right MeFG (BA 11)	.440	.800	[972 1000]	16, 36, -17

Right MTG (BA 21)	.241	.689	[808 839]	62, -22, -9
Right STG (BA 22)	.271	.602	[161 207]	63, -33, 10
Right ATL (BA 38)	.035	.233	[769 902]	35, 10, -32
Right TPJ (BA 39)	.902	>.99	[965 982]	51, -66, 28
Right TTG (BA 42)	.254	.635	[503 571]	65, -20, 11
Right IFG (BA 44-45)	>.99	>.99	[110 250]	54, -26, 6
Right MPFC (BA 46)	>.99	>.99	[110 343]	49, -40, 17
Right vOFC (BA 47)	>.99	>.99	[137 197]	40, 27, -13

Figure captions

Figure 1. Left: mean percentages of "yes" responses (indicating lower-bounded or non-pragmatic readings); error bars represent ±2×SE (the standard error of the by-subject means). Right: histograms of the subject means for *some* and *only some* items (averaged across upper-bounded and lower-bounded contexts)

Figure 2. Current estimates for regions showing FDR-significant or -marginal interactions. The time window of the significant cluster is highlighted in gray, and the upper left portion of each plot shows the spatial location of the vertices comprising that region. SOME_all: *some* in the "all" (upper-bounded, inference-supporting) context; SOME_any: *some* in the "any" (lower-bounded, inference-nonsupporting) context; ONLYSOME_all: *only some* in the "all" (upper-bounded) context; ONLYSOME_any: *only some* in the "any" (lower-bounded) context.

[1] The authors did include baseline conjunction analyses to argue against this possibility. These analyses involved comparing activations for *some*-sentences in various conditions to activations for correct *every*-sentences (that is, those accompanying pictures in which every element has the property described), and showed that processing the lexical meaning of *some* either did not elicit greater activation than the lexical meaning of *every*, or only did so in other brain regions. This analysis may not have ruled out potential differences due to lexical meanings, however, because it included *some*-NONE items—that is, *some*-sentences accompanying pictures in which none of the elements have the property described. For example, if verification of *some but not all* followed a two-step procedure (first verify whether any Xs have the property, then check whether all the Xs have the property), then *some*-NONE items would not elicit the second step, and therefore might not have shown greater verification costs than *every*-sentences. Crucially, *some*-NONE sentences were included in this baseline analysis but not in the critical analysis, which may have caused lexical differences to emerge in only the critical analysis.

[2] An anonymous reviewer notes that "only some" is somewhat infelicitous in the lower-bounded context (i.e., "She asked John whether any of them were staying his apartment. John said that only some of them were"). While this may indeed introduce additional differences in activity between the two "only some" conditions, the statistical analysis conducted in the present study (see Methods) used a test statistic designed to identify brain regions that showed an effect of Context in the "some" sentences and not the "only some" sentences. Therefore, while this aspect of the control items may have caused our analysis to miss some regions that could have been involved in scalar inferencing, it would not have caused us to falsely detect any regions.

[3] That is to say, the following four sentences corresponding to the possible versions of the item shown in (5) were each read during different blocks of the recording, with multiple other sentences in between:

- Mary was preparing to throw a party for John's relatives. She asked John whether all of them were staying in his apartment.
- Mary was preparing to throw a party for John's relatives. She asked John whether any of them were staying in his apartment.
- John said that some of them were.
- John said that only some of them were.

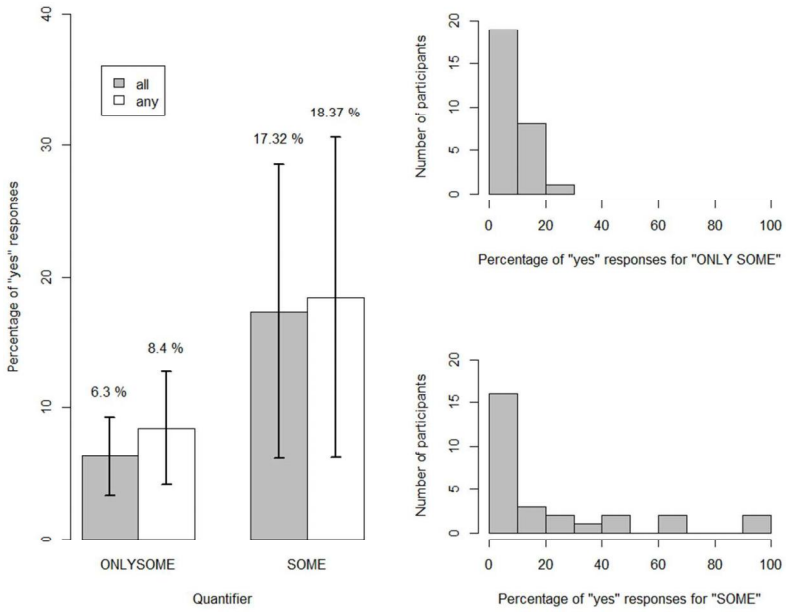
[4] The data also showed the same pattern of results when the *only some* conditions were time-locked to the onset of *only* rather than the onset of *some*.

[5] The small sizes of BAs 44 and 45, along with the coarse-grained nature of the decimation of the BESA cortex into sources, meant that each of these BAs contained a very small number of source vertices. Therefore, in the parcellation used for our analysis, these two BAs are combined.

[6] Because the test statistic used for identifying clusters was not a true *t*-statistic, but rather a difference of *t*-statistics, it does not necessarily fit a *t* distribution, and thus *p*-values based on the *t* distribution are not accurate. In other words, $p < .3$ in the present context does not necessarily mean a test-wise Type I error rate of less than .3; rather, it is a somewhat arbitrary value. This does not matter for a permutation test, however, as the threshold used for creating clusters does not affect the familywise error rate of the overall analysis (see Maris & Oostenveld, 2007).

[7] An additional question raised by the offline data is as follows: why is it the case that both the MEG data at the quantifier and the offline data at the end of the sentence suggest that that inference is realised in both contexts, whereas reading time and EEG data at a downstream expression (e.g. Breheny et al., 2006; Bergen & Grodner, 2012; Hartshorne et al., in press; Lewis, 2013; Politzer-Ahles & Fiorentino, 2013) suggest that the inference is more likely in

the upper-bounded context? It is difficult to compare these findings, given that they come from different measures, different samples of participants, and different experimental designs. One possible explanation is that, even if participants realise the inference in both contexts, only in the upper-bounded context do they commit to the inference so fully as to make a strong forward prediction that "the rest" will appear later in the sentence. This explanation is ad-hoc, however, and this topic requires further investigation.



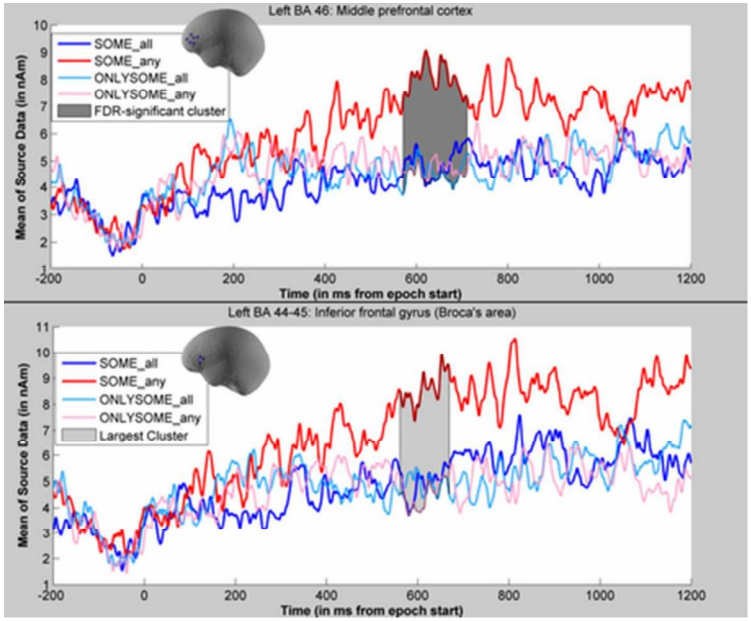


Figure 2. Current estimates for regions showing FDR-significant or -marginal interactions. The temporal window of the significant cluster is highlighted in gray, and the upper left portion of each plot shows the spatial location of the vertices comprising that region. SOME_all: some in the "all" (upper-bounded, inference-supporting) context; SOME_any: some in the "any" (lower-bounded, inference-nonsupporting) context; ONLYSOME_all: only some in the "all" (upper-bounded) context; ONLYSOME_any: only some in the "any" (lower-bounded) context.
19x15mm (600 x 600 DPI)

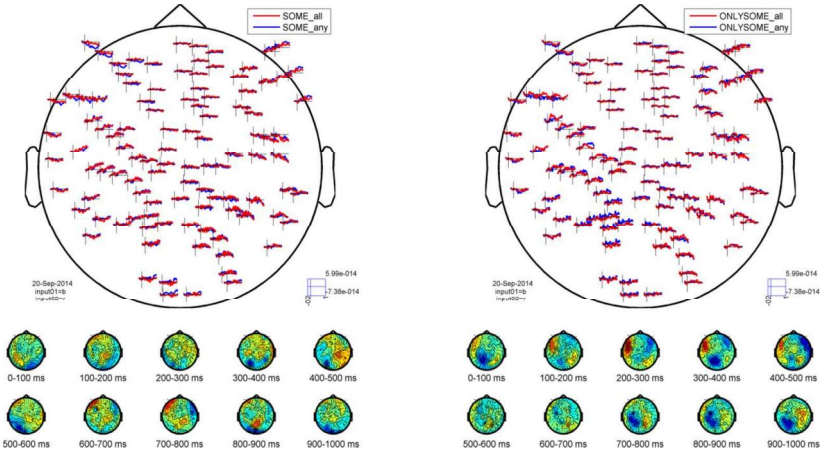
Supplementary figure captions

Supplementary figure 1. Top: Waveforms showing the event-related fields evoked by the critical word. SOME_all: *some* in the "all" (upper-bounded, inference-supporting) context; SOME_any: *some* in the "any" (lower-bounded, inference-nonsupporting) context; ONLYSOME_all: *only some* in the "all" (upper-bounded) context; ONLYSOME_any: *only some* in the "any" (lower-bounded) context. Waveforms represent a subset of the channels in the sensor space. Bottom: Topographic plots of the difference waves for the context effect at *some* (SOME_all – SOME_any) and at *only some* (ONLYSOME_all – ONLYSOME_any).

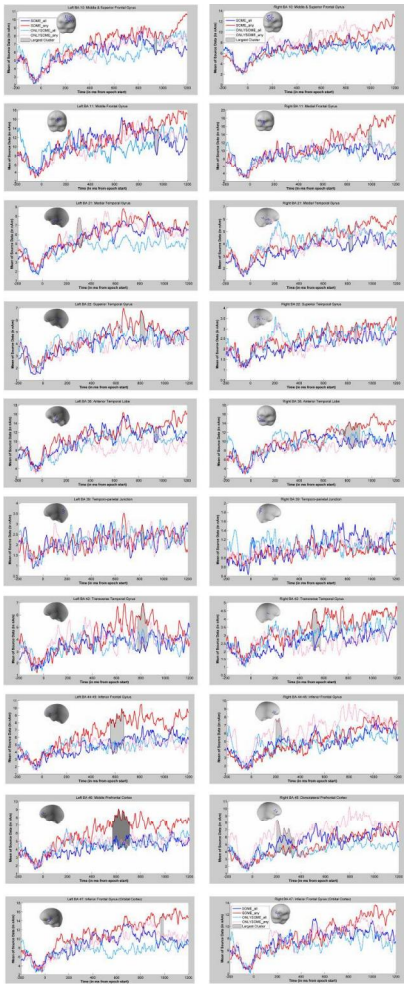
Supplementary figure 2. Current estimates for all regions tested (except BA 46 and BAs 44-45, which are shown in Figure 1). For each region, the temporal window of the largest cluster is highlighted in gray, and the upper left portion of each plot shows the spatial location of the vertices comprising that region. SOME_all: *some* in the "all" (upper-bounded, inference-supporting) context; SOME_any: *some* in the "any" (lower-bounded, inference-nonsupporting) context; ONLYSOME_all: *only some* in the "all" (upper-bounded) context; ONLYSOME_any: *only some* in the "any" (lower-bounded) context.

Supplementary figure 3. Maps of current estimates overlain on the whole brain. The difference between SOME_all and SOME_any is shown in the upper portion (A), and the difference between ONLYSOME_all and ONLYSOME_any in the lower portion (B). For each contrast, the plots are masked at $p = .01$ (paired t-tests). Each row shows a different view (left, right, front, back), and each column shows a different time window.

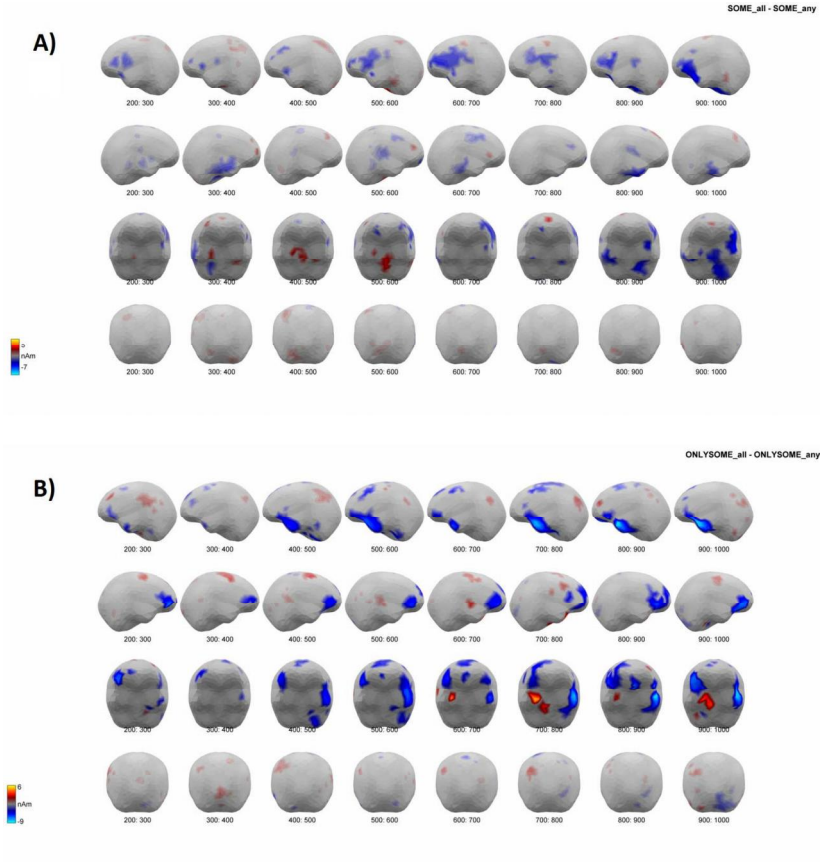
Supplementary file 1. Description of the post-hoc offline survey carried out on Mechanical Turk.



Waveforms showing the event-related fields evoked by the critical word. SOME_all: some in the "all" (upper-bounded, inference-supporting) context; SOME_any: some in the "any" (lower-bounded, inference-nonsupporting) context; ONLYSOME_all: only some in the "all" (upper-bounded) context; ONLYSOME_any: only some in the "any" (lower-bounded) context. Waveforms represent a subset of the channels in the sensor space. Bottom: Topographic plots of the difference waves for the context effect at some (SOME_all – SOME_any) and at only some (ONLYSOME_all – ONLYSOME_any).
322x174mm (116 x 116 DPI)



Supplementary Figure 2. Current estimates for all regions tested (except BA 46 and BAs 44-45, which are shown in Figure 1). For each region, the temporal window of the largest cluster is highlighted in gray, and the upper left portion of each plot shows the spatial location of the vertices comprising that region. SOME_all: some in the "all" (upper-bounded, inference-supporting) context; SOME_any: some in the "any" (lower-bounded, inference-nonsupporting) context; ONLYSOME_all: only some in the "all" (upper-bounded) context; ONLYSOME_any: only some in the "any" (lower-bounded) context.
150x369mm (300 x 300 DPI)



Supplementary Figure 3. Maps of current estimates overlain on the whole brain. The difference between SOME_all and SOME_any is shown in the upper portion (A), and the difference between ONLYSOME_all and ONLYSOME_any in the lower portion (B). For each contrast, the plots are masked at p = .01 (paired t-tests). Each row shows a different view (left, right, front, back), and each column shows a different time window. 402x423mm (120 x 120 DPI)