Journal of Memory and Language 83 (2015) 79–96

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Quantifiers are incrementally interpreted in context, more than less

Thomas P. Urbach [a,*], Katherine A. DeLong [a], Marta Kutas [a,b]

[a] Department of Cognitive Science, University of California, San Diego, United States
[b] Department of Neurosciences, University of California, San Diego, United States

ARTICLE INFO

Article history:
Received 11 October 2014
revision received 27 March 2015

Keywords:
Quantifier
Incremental, shallow, partial, interpretation
Brain potential
ERP
N400
Language comprehension

ABSTRACT

Language interpretation is often assumed to be incremental. However, our studies of quantifier expressions in isolated sentences found N400 event-related brain potential (ERP) evidence for partial but not full immediate quantifier interpretation (Urbach & Kutas, 2010). Here we tested similar quantifier expressions in pragmatically supporting discourse contexts (*Alex was an unusual toddler. Most/Few kids prefer sweets/vegetables*...) while participants made plausibility judgments (Experiment 1) or read for comprehension (Experiment 2). Control Experiments 3A (plausibility) and 3B (comprehension) removed the discourse contexts. Quantifiers always modulated typical and/or atypical word N400 amplitudes. However, the real-time N400 effects only in Experiment 2 mirrored offline quantifier and typicality crossover interaction effects for plausibility ratings and cloze probabilities. We conclude that quantifier expressions *can* be interpreted fully and immediately, though pragmatic and task variables appear to impact the speed and/or depth of quantifier interpretation.

Introduction

This report describes a series of experiments that investigate when and to what extent the meanings of natural language quantifier expressions like *most kids* and *few kids* are interpreted as sentences unfold over time. These experiments extend our previous investigations of the time course of quantifier interpretation (Urbach & Kutas, 2010).

When all goes well in verbal communication, comprehenders reflexively respond to a sequence of linguistic tokens—spoken or written words, signed gestures—by constructing an interpretation of what was meant. There is considerable consensus among language researchers on the coarse-grained principle that interpretation is incremental, i.e., that representations of structural form and semantic content are typically constructed word by word rather than being deferred until additional, potentially informative words are encountered (see Just & Carpenter, 1980 for an influential early account and overviews in, e.g., Altmann & Mirkovic, 2009; Hagoort & van Berkum, 2007; Rayner & Clifton, 2009). This principle of incremental interpretation is characteristic of theoretical accounts of language comprehension that differ in other important ways. These include "syntax first" models that postulate a modular, serial, processing architecture, such as the garden-path model (e.g., Frazier, 1987; see also Friederici, 2002 for application to speech), "interactive" or "constraint based" models with interconnected network architectures that do not privilege syntactic or any other type of information (e.g. Bates & MacWhinney, 1989; Macdonald, Pearlmutter, & Seidenberg, 1994; Marslen-Wilson & Tyler, 1975; McRae, Spivey-Knowlton, & Tanenhaus, 1998), and "multi-stream" views on which

* Corresponding author at: Department of Cognitive Science, Mail Code 0515, University of California, San Diego, La Jolla, CA 92093-0515, United States. Fax: +1 858 534 1128.
E-mail address: turbach@ucsd.edu (T.P. Urbach).

http://dx.doi.org/10.1016/j.jml.2015.03.010
0749-596X/

syntactic and semantic analyses are rapidly constructed in parallel (e.g., Bornkessel & Schlesewsky, 2006; Kim & Osterhout, 2005; Kos, Vosse, van den Brink, & Hagoort, 2010; Kuperberg, 2007; van Herten, Kolk, & Chwilla, 2005). Still other approaches aim to explain sentence comprehension phenomena within the constraints of general principles of human cognitive processing (e.g., Lewis & Vasishth, 2005). Notwithstanding their considerable differences, each of these frameworks is committed to some form of incremental interpretation.

At the same time, there is a growing appreciation of the wide range of phenomena indicating that lexical and propositional information readily available to the comprehender nonetheless may not always make its way into the semantic representations constructed in real-time (“shallow”, “underspecified”, “just good enough” interpretation, for overviews see, e.g., Frisson, 2009; Sanford & Graesser, 2006). Notable laboratory examples include so-called semantic illusions wherein descriptions of patent errors and contradictions go unnoticed as in Moses rather than Noah taking animals on the ark (Erickson & Mattson, 1981), survivors rather than victims of a plane crash being buried (Barton & Sanford, 1993) and kids giving out rather than getting candy on Halloween, (Reder & Kusbit, 1991). The interpretation of such cases is that comprehenders' semantic representations are incomplete or partial or underspecified with respect to crucial information. Special cases abound: factual errors are noticed less often when they occur outside of discourse focus (Baker & Wagner, 1987) and in passives rather than actives (Ferreira, 2003), see also the reports collected in “Shallow Processing and Underspecification,” (2006). However, relatively little is known about general principles governing what information is and is not represented and when. Few studies have probed the time course of partial or underspecified interpretation construction (though see, e.g., self-paced reading in Reder and Kusbit (1991), eye movements in Daneman, Lennertz, and Hannon (2007), and event-related brain potentials (ERPs) (Sanford, Leuthold, Bohan, & Sanford, 2011; Tune et al., 2014).

Evidence that the comprehension system is interpretively lazy at times challenges the generality of the strong principle of incremental interpretation. Since the inventory of expressive devices in natural language is quite large, delimiting the scope of the strong incremental interpretation principle requires us to examine real-time interpretation in a wide range of linguistic constructions. For example, ERPs have proved useful in shifting the theoretical landscape toward incremental interpretation and this trajectory is particularly clear in regard to negation. In an influential early study, Fischler, Bloom, Childers, Roucos, and Perry (1983) found that negation did not modulate N400 ERP amplitudes at the predicate word in simple subject–predicate sentences, *A robin* [*is*/*is not*] *a* [*bird*/*tree*] but did modulate a later potential. This pattern was interpreted as evidence that the reference and predication are first composed to form the propositional content, e.g., ISA(robin, bird) with the negation operator subsequently applied to the result, i.e., later, despite appearing before the predicate in the surface form. Similar ERP findings (Kounios & Holcomb, 1992) were observed for categories and exemplars (*No rubies are gems*/*spruces*) and also interpreted in line with delayed processing of negation. However more recent investigations of the time-course of negation interpretation have taken a cue from behavioral research and consider not just the semantic (truth-functional) content of negative propositions but the circumstances in which they are appropriate to use. In seminal work (Wason, 1965) manipulated visual displays and found that with truth value held constant, negative sentence verification times could be reduced by “contexts of plausible negation”. Following this logic, recent sentence processing work has investigated the on-line processing of negative sentences under conditions that better conform to pragmatic principles. In her dissertation work Staab (2007) constructed (counterbalanced pairs of) scenarios (*During his long flight Joe needed a snack. The flight attendant could only offer him pretzels and cookies. Joe wanted something salty*/*sweet*) and found N400 effects of negation on the critical words in sentence continuations, (*So he bought the pretzels*/*cookies* vs. *So he didn't buy the pretzels*/*cookies*). Independently, Nieuwland and Kuperberg (2008) also tested negation in pragmatically licensing contexts, *With proper equipment, scuba-diving is*/*isn't very safe*/*dangerous*) and out. They also found N400 effects on the critical word including, critically, a clear cross-over interaction where negation fully reverses the N400 for words that are vs. are not compatible with world knowledge, i.e., about the hazards of scuba diving. So despite the early reports, there is evidence that negation can be incrementally interpreted under some conditions, presumed to better approximate ordinary language use than isolated negative sentences.

Quantifier interpretation: what and when?

Among the expressive devices that augment reference and predication in natural language are expressions of quantity, e.g., in English words and expressions such as *all*, *some*, *none*, *many*, *most*, *half of*, *exactly three*, *always*, *never*, *often*, and *rarely* that allow speakers to specify, with more or less precision, how many, how much, and how frequently. Comprehenders, for their part, must make sense of quantifiers along with reference and predication to arrive at an interpretation of the propositional content of, e.g., *Most birds can fly*. The project for theories of real-time language comprehension is to determine what quantifier interpretations are constructed and there is growing interest a variety of quantifier types: a selective sample of topics and reports includes investigations of bare cardinal quantifiers (e.g., Frazier et al., 2005; Kaan, Dallas, & Barkley, 2007; Wijnen & Kaan, 2006); existential quantifiers and their scalar implicatures (e.g., Breheny, Katsos, & Williams, 2006; Huang & Snedeker, 2009; Politzer-Ahles, Fiorentino, Jiang, & Zhou, 2013); consequences of quantifier interpretations for discourse processing (e.g., Paterson, Filik, & Moxey, 2009; Sanford, Dawydiak, & Moxey, 2007); multiple quantification and scope ambiguities (e.g., Dwivedi, 2013; Filik, Paterson, & Liversedge, 2004; Kurtzman & Macdonald, 1993, and quantifiers and long-distance dependencies, e.g., Hackl, Koster-Hale, & Varvoutis, 2012).

Even though it is clear (to linguistic intuition) that the meanings of quantifiers are fully interpreted – eventually – it is by no means clear when (and under what circumstances) this occurs. In particular, it is not at all obvious that it routinely happens fully and at the earliest opportunity. In previous studies (Urbach & Kutas, 2010, hereafter U&K 2010) we probed *when* – immediately vs. delayed – and to what extent – fully vs. partially – quantifier interpretations are incorporated into message-level representations in real-time. The experimental design pitted the meaning of quantifier expressions such as *most* and *few* against general world knowledge of what is typical and atypical as expressed in simple subject–verb–object sentences, e.g., *Most* vs. *few farmers grow crops* vs. *worms* and *Farmers often* vs. *rarely grow crops* vs. *worms*. We recorded scalp EEG while participants read these sentences RSVP and rated their plausibility, testing whether the meanings of quantifier expressions are processed fully at the earliest opportunity by measuring N400 event-related brain potential (ERP) responses elicited by the critical typical and atypical words. As expected, when judged for plausibility after the sentence, those quantified propositions that were consistent with general world knowledge, e.g., *Most farmers grow crops*... were rated more plausible than those that were not, e.g., *Few farmers grow crops*... Critically, when the object noun referred to something atypical, e.g., *worms*, the quantifiers reversed the pattern of plausibility judgments, i.e., sentences such as *Most farmers grow worms*... were rated less plausible than *Few farmers grow worms*... We interpreted these plausibility judgments as evidence that the quantifiers were indeed fully interpreted by the time the judgment was made and in a way that was consistent with linguistic intuition about their meaning in conjunction with general world knowledge. However, N400 amplitudes elicited by the typical and atypical object nouns (*crops* vs. *worms*) as these were encountered during real-time processing told a different story. *Most*- and *few*-type quantifiers (U&K 2010, Experiment 2) and the adverbs *often* and *rarely* (U&K 2010, Experiment 3) did modulate the ERPs in the expected direction, i.e., *few*-type and *rarely* reliably increased N400 amplitude for typical nouns and marginally decreased it for atypical nouns in the context of most-type quantifiers, providing evidence of registration of some difference between *most*- and *few*-type quantifiers by the time the critical word was encountered. However, across all conditions, the N400 modulations were well short of the crossover interaction observed for the plausibility judgments. The on-line N400 effects at the critical word did not mirror the offline plausibility judgments as predicted by full and immediate incrementality. So, for models of real-time sentence comprehension that have parameters for speed and depth of interpretation (e.g., immediate vs. delayed and full vs. partial respectively), the data from our initial studies are most consistent with immediate partial but delayed full interpretation of quantifier semantics.

Our previous quantifier studies left open a number of questions. On the one hand whereas N400 amplitude modulations evidenced registration of some difference between the quantifiers for typical words, it did not for atypical words. This asymmetry may reflect a theoretically uninteresting lack of power or quirks about the experimental stimuli or something systematic about the way quantifiers impact processing when language makes contact with typical vs. atypical world knowledge.

There is also a potential concern about the pragmatic felicity of isolated quantified sentences, c.f., negation. In isolation, statements of shared general knowledge, e.g., *most birds can fly*, though patently true are uninformative and thus pragmatically infelicitous. However, by analogy with negation, there are "contexts of plausible quantification" in which true generalizations can be informative, e.g., by dint of contrast with contextually salient exceptions or special cases: *Penguins are unusual birds. Most birds can fly but penguins cannot*. Even if the comprehender already knows full well that, *Most birds can fly* is true, it is pragmatically felicitous in this context because it adds information: it specifies the particular respect in which penguins are unusual. We thus asked whether embedding quantified propositions in pragmatically appropriate contexts could substantially increase the speed and/or depth of quantifier interpretation as has been reported for negation.

We also consider the potential impact of the plausibility rating task on the time course of quantifier interpretation. Evaluating the plausibility of sentences is a commonplace and natural adjunct to language comprehension and part of what comprehenders do (or perhaps should do) while reading or listening to, e.g., explanations of teenager's post-curfew arrivals, political debates, and scientific research reports. The impact of this task on the pattern of ERP effects could be argued both ways. Rating each sentence for plausibility may have focused attention on the quantifiers and encouraged unusually rapid and/or deep interpretation. Alternatively, making numerical judgments and executing responses on every trial may impose a greater cognitive load than just reading or listening and thus could work against full and immediate quantifier interpretation if it competes for the same resources.

Finally, in our previous report, we proposed that if quantifiers are interpreted fully and immediately, the impact of their semantics as inferred from offline measures, e.g., linguistic intuition and plausibility ratings, should be evident in on-line measures sensitive to semantic processing, e.g., N400 ERPs. We operationalized this idea by testing whether the crossover interaction effect observed in plausibility judgments was also observed in the N400 amplitudes elicited by the critical words (it was not). However full crossover interactions can occur even when scores for one variable do not differ reliably across levels of the other, e.g., if the different quantifiers do not modulate N400 amplitude for one type of critical word. To rule out this case, we formulate a stricter test with four individually necessary and jointly sufficient criteria that make explicit the patterns of N400 effects (Fig. 1) that we propose constitute evidence of full and immediate quantifier interpretation.

1. This criterion requires evidence of an N400 typicality effect following *most*-type quantifiers and in a direction consistent with general world knowledge and the

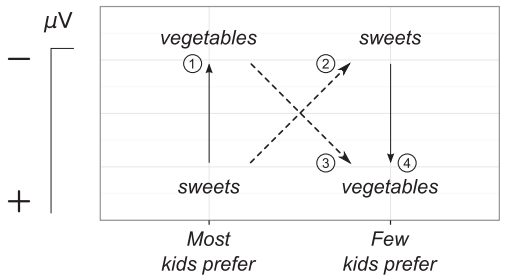

Fig. 1. Criterial N400 amplitude effects at critical words. Schematic diagram of the N400 pair-wise effects corresponding to the four individually necessary and jointly sufficient criteria for evidence of full incremental quantifier interpretation. Circled numbers indicate the criteria, arrows indicate the direction of the N400 effect (negative is plotted up).

compositional semantics of the sentence. For instance, in *Most kids prefer* sweets/vegetables, the N400 elicited by words denoting atypical objects (e.g., *vegetables*) must be relatively larger than that elicited by words denoting typical objects (e.g., *sweets*). This canonical N400 effect establishes that the experimental materials are well-behaved at the critical word.

2. This criterion requires that the *most*- and *few*-type quantifiers differentially impact the real-time processing of the subsequent critical typical word and do so in a manner consistent with the (offline, intuitive) quantifier semantics, critical word meaning, and general world knowledge. Since, e.g., it is plausible that kids generally prefer sweets to other kinds of food, this criterion requires larger N400s for *sweets* in *Few kids prefer* sweets ... than in *Most kids prefer* sweets ...
3. This criterion is the same as Criterion 2 except that it requires quantifiers to have the right sort of impact on processing on the atypical words. This requires a larger N400 for *vegetables* in *Most kids prefer* vegetables than in *Few kids prefer* vegetables. Note that the N400 effects required by Criterion 2 and Criterion 3 go in opposite directions.
4. Finally, there also must be a typicality effect in the context of the *few*-type quantifiers, again consistent with world knowledge, which entails that it is the reverse of the canonical typicality effect in the context of *most*-type quantifiers. For instance, the N400 elicited by the typical word *sweets* must be relatively larger than that elicited by the atypical word (e.g., *vegetables*) in a pair like, *Few kids prefer* sweets/vegetables.

The present studies

The present series of studies aimed to investigate the generality of our previous results and again test the hypothesis that quantifiers are fully and incrementally interpreted. The experimental design is similar. Two types of quantified noun phrases (*most*- and *few*-type) were pitted against general world knowledge, e.g., what kids prefer to eat (typical, atypical), in a fully crossed design, yielding four types of quantified sentences. For these experiments, a new set of stimulus materials was developed. Sentences containing the quantifiers and critical typical and atypical words were constructed with more structural variety than in U&K 2010. Crucially, these new RSVP sentences were preceded by discourse contexts introducing scenarios for which further elaboration by a quantified generalization would add information and thus be more pragmatically felicitous. The scenarios described in these discourse contexts reinforced and/or established the consistency of the "*Most ... typical*" and "*Few ... atypical*" forms of the quantified sentences with world knowledge, and thereby also, the inconsistency of the other two forms, i.e., "*Most ... atypical*" and "*Few ... typical*". In this design, the incompatibility becomes evident at the critical atypical or typical word.

In Experiment 1, the discourse context was read, and then followed by the RSVP quantifier sentence which was rated for plausibility as in U&K 2010. Experiment 2 tested the impact of the plausibility rating task using these same materials with a new group of participants who read for comprehension and answered content questions on a random 25% of the trials. Finally, to assess the impact of the (presumably) supporting discourse context on the offline and ERP measures of quantifier interpretation, control Experiments 3A and 3B were conducted as in Experiments 1 and 2, respectively, except without the preceding discourse context.

Predictions

As in U&K 2010, we suppose that if the comprehension system is strongly incremental it should fully and immediately interpret quantifiers, e.g., *most*, *few*, along with (presumably) open class content words, e.g., *kids*, *prefer* and integrate them all into the evolving interpretation of the sentence and broader discourse context. If so, then within a few hundred milliseconds after encountering the last word of *Most kids prefer* and *Few kids prefer* the message-level representation under construction should be systematically different as a function of the different meanings of the quantifiers. It is worth noting that this need not be so. If quantifier interpretation is deferred or substantially slower than the interpretation of content words, it may be that different interpretations of these initial quantified sentence fragments that are so patent in linguistic intuition have not (yet) been assembled by the time the critical word is encountered. In this case, despite the obvious difference in the surface forms of the expressions and despite the obvious differences in their eventual interpretation, at this moment, i.e., just before the critical word appears, the semantic representation of these fragments may not differ. To dismiss this possibility is to assume that quantifier interpretation is incremental without actually testing it empirically – which is what we did.

Our experiments aim to answer this question by measuring brain responses emitted when the comprehension system is probed on the fly with critical words that, in virtue of their meaning, tap into the comprehender's general knowledge. Based on prior ERP research showing rapid effects of world knowledge (e.g., Hagoort, Hald, Bastiaansen, & Petersson, 2004; Hald, Steenbeek-Planting,

& Hagoort, 2007) and discourse context (e.g., Hagoort & van Berkum, 2007; Nieuwland & Van Berkum, 2006; van Berkum, Hagoort, & Brown, 1999) on N400 amplitude, critical word continuations of the quantified sentence fragment that are inconsistent with the current discourse context and general world knowledge should elicit relatively larger N400 amplitude than continuations that are consistent. Critically, the hypothesis that quantifier interpretation is full and immediate in pragmatically supporting discourse contexts predicts that all four criteria for strong incremental quantifier interpretation outlined above should be satisfied. This interpretation of the ERP results leans heavily on empirically established relationships between experimental variables and N400 ERPs during RSVP reading, i.e., our assumptions about the impact of world knowledge and discourse context on the polarity and amplitude of potentials elicited by the critical words. However, our conclusions about the time course of quantifier interpretation are compatible with, and thus cannot distinguish between, competing fine-grained views about the functional significance of scalp potentials observed around 300–500 ms poststimulus, a controversial topic of independent theoretical interest (for discussion see, e.g., Brouwer, Fitz, & Hoeks, 2012).

Methods

The following experimental methods and procedures were the same in each experiment except for the different participant groups and the stimulus and task variables as detailed separately. All experiments reported here were conducted according to a research protocol approved by the Institutional Review Board of the University of California, San Diego Human Research Protection Program. Participants were volunteers who provided their informed consent in writing prior to enrolling in the study.

Participants

In each of the ERP experiments, a different group of 16 volunteers (8 female) were recruited from the University of California, San Diego campus community and could elect to receive course credit or $7 per hour for participating. All participants were right-handed, native English speakers with normal or corrected-to-normal vision and no self-reported history of neurocognitive impairment. In Experiment 1 the mean age was 22.7 years, range [19, 26] and 7 participants (2 female) reported left-handed first degree relatives. Data from all 16 participants initially recruited were included in the analysis. In Experiment 2, the mean age was 22.6 years, range [18, 27] and 7 participants (4 female) reported left-handed first degree relatives. Data from two participants were excluded for excessive EEG artifacts and replaced. In Experiment 3A the mean age was 20.6 years, range [18, 28] and 4 participants (2 female) reported left-handed first degree relatives. Data from 7 participants were excluded and replaced, 5 for excessive EEG artifacts, 2 because of research staff EEG data acquisition errors. In Experiment 3B the mean age was 20.1 years, range [18, 29] and 4 participants (2 female) reported left-handed first degree relatives. Data from one participant was excluded for excessive EEG artifacts and replaced.

Materials

Stimuli (e.g., Table 1; see Supplementary Material for a complete list) consisted of a single discourse context followed by one of four target sentences constructed by crossing two types of quantifier (*most* vs. *few*) with typical and atypical critical words. The discourse contexts draw on world knowledge and introduce information about individuals or a specific scenario, often involving an exception or departure from the norm, e.g., *an unusual toddler*. The experimental materials crossed the quantifier type (*most vs. few*) with typicality (typical vs. atypical) relative to general world knowledge. Two of the four resulting combinations differ in quantificational form but are both plausible (*Most kids prefer sweets, Few kids prefer vegetables*); the other two continuations also differ in quantificational form but are implausible (*Most kids prefer vegetables, Few kids prefer sweets*).

One hundred and forty such sets of four were constructed with context sentences of various lengths and grammatical structures and target sentences with variants of the quantifiers, *most* and *few*. The 140 pairs of *most*- and *few*-type quantifier expressions were approximately matched for length in number of words: 126 quantifier pairs were the same length, 8 of the *most*-type quantifiers were one word longer than the *few*-type quantifiers, 5 were one word shorter, and one was two words longer. The critical typical and atypical words were controlled for several variables known to modulate N400 ERPs and were approximately matched on average across the 140 items for length, frequency, and orthographic neighborhood. The length of typical words ($M = 6.3$ characters, $SD = 2.18$, range [3, 13]) did not differ reliably from the length of atypical words ($M = 6.5$ characters, $SD = 2.19$, range [2, 12]), Welch's $t(278) = -0.903$, $p = .367$ (Welch, 1947). Log frequency of the typical critical words ($M = 1.96$, $SD = 1.62$, range [0, 6.66]) did not differ reliably from log frequency of the atypical words ($M = 1.95$, $SD = 1.39$, range [0, 6.74]), Welch's $t(161) = 0.08$, $p = .94$. For the 82 typical and 91 atypical critical words appearing in the Brown corpus (Francis & Kucera, 1979), the size of the orthographic neighborhood of the typical words ($M = 4.11$, $SD = 5.56$, range [0,24]) did not differ reliably from the size of the orthographic neighborhood of atypical

Table 1
Example discourse context and corresponding quantified sentences. *Alex was an unusual toddler in that he favored broccoli and peas over cookies and candy.*

Quantifier	Typicality	RSVP sentence
Most	Typical	*Most kids prefer sweets to other foods.*
Most	Atypical	*Most kids prefer vegetables to other foods.*
Few	Typical	*Few kids prefer sweets to other foods.*
Few	Atypical	*Few kids prefer vegetables to other foods.*

Note: The critical word is underlined for display here; stimuli presented during the experiment were not underlined.

words ($M = 3.66$, $SD = 4.91$, range [0,22]), Welch's $t(274) = 0.73$, $p = .47$. Since only the quantifier expressions and critical typical and atypical words are varied in this design, low-level lexical properties and relations (e.g., lexical and semantic association, semantic feature overlap, as well as co-occurrence and other distributional relations among words in the discourse context, target sentence, and critical words) are held constant. Consequently, any differences in the typicality effect observed at the critical words may be unequivocally attributed to the impact of the preceding quantifier expressions.

Two normative studies (Fig. 2, Panel A) were conducted to determine the predictability of the critical typical and atypical words in the target sentences when these sentences were preceded by the discourse contexts or presented in isolation (methods described in the Supplementary Material). The predictability of the typical and atypical critical words was operationalized as cloze probability, i.e., the proportion of responses in a fill-in-the-blank sentence completion task (c.f., Taylor, 1953). For the version in which discourse contexts preceded the *most*- and *few*-type quantifier sentence fragments, cloze probabilities ranged between .01 and .36 in the four conditions. The typical critical words (*sweets*) were moderately predictable as completions of the *most*-type sentence fragments (*Most kids prefer* ___) and the cloze probability was substantially higher ($M = 0.36$, $SD = 0.31$) than the cloze probability of atypical critical words (*vegetables*) which was quite low ($M = 0.01$, $SD = 0.04$). This cloze probability typicality effect was reversed for the *few*-type quantifier sentence fragments (e.g., *Few kids prefer* ___) where the cloze probability of atypical critical words was higher ($M = 0.36$, $SD = 0.28$) than that of typical critical words which was again quite low ($M = 0.04$, $SD = 0.06$). These effects resulted in a robust and nearly symmetrical cross-over interaction effect between quantifier and typicality, $F(1,139) = 297.24$, $p < .001$, $\eta_p^2 = .68$, (Fig. 2; Panel A, top row). In the no context version, cloze probability of the critical targets in all conditions was generally low, ranging between .04 and .14 (Fig. 2, Panel A, bottom row). In the context of the *most*-type quantifiers, e.g., *Most kids prefer* ___, the typical words (*sweets*), were more predictable ($M = 0.14$, $SD = 0.23$) than the atypical words (*vegetables*) ($M = 0.04$, $SD = 0.11$). This pattern was reversed in the context of the *few*-type quantifiers, e.g., *Few kids prefer* ___, where the atypical word was more predictable ($M = 0.12$, $SD = 0.19$) than the typical word ($M = 0.05$, $SD = 0.09$). Even within this reduced range of cloze probabilities, the

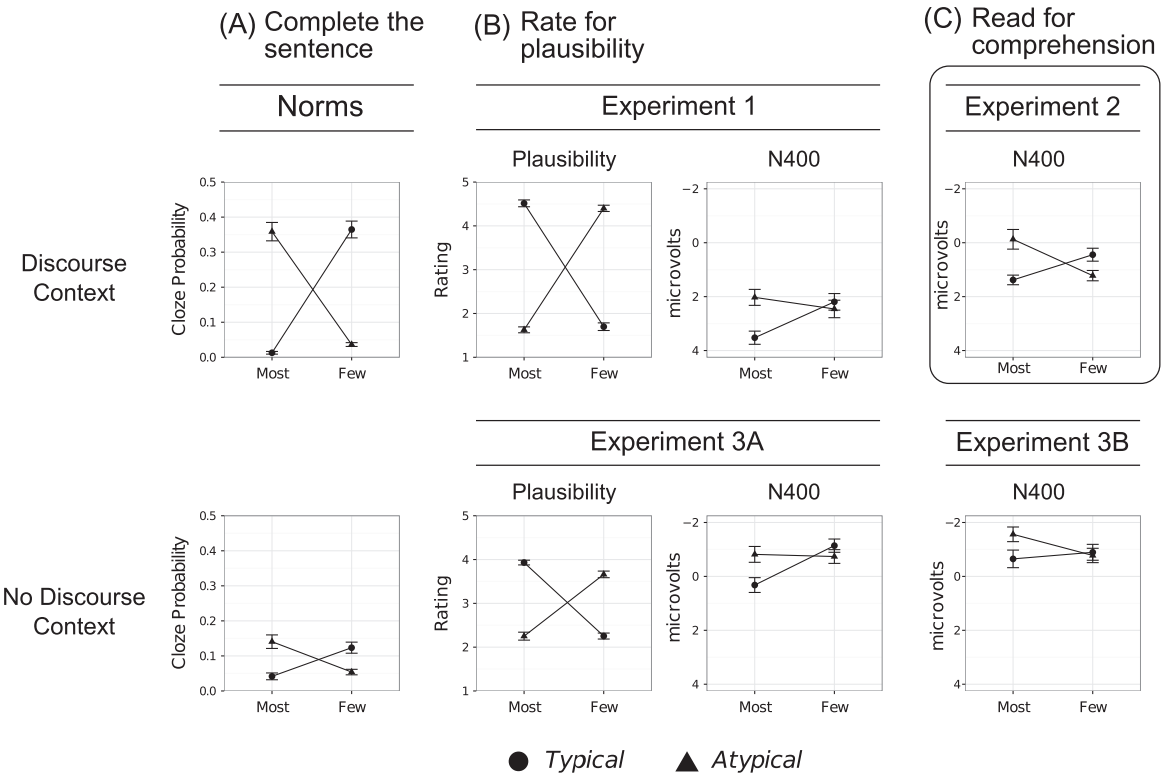

Fig. 2. Quantifier by typicality interaction effects. Panel (A) Critical word cloze probability (=proportion of responses) in quantified sentence fragments following supporting discourse contexts (top row) and without the preceding discourse contexts (bottom row). Panel (B) Sentence plausibility ratings on a five-point scale (left column) and N400 ERP amplitudes (right column) at the cloze-normed critical word for these same quantified sentences. The quantified sentences were presented following supporting discourse contexts in Experiment 1 (top row, $N = 16$), and without the discourse contexts in Experiment 3A (bottom row, $N = 16$). Panel (C) N400 ERP amplitudes in microvolts at the cloze-normed critical word for these same quantified sentences read for comprehension. The quantified sentences were presented following supporting discourse contexts in Experiment 2 (top row, $N = 16$) and without the discourse contexts in Experiment 3B (bottom row, $N = 16$). All four criteria for strong incremental quantifier interpretation are satisfied only in Experiment 2 (boxed). Whiskers indicate ±1 *SE*; in Panels (B and C) the within-participants *SE* is calculated according to Morey (2008).

crossover interaction effect between the quantifier and typicality factors was nearly symmetrical and statistically reliable, $F(1,139) = 42.91$, $p < .001$, $\eta_p^2 = 0.23$).

Thus, in off-line cloze testing, the most salient consequence of the preceding discourse context was to increase the predictability of the typical critical words in the context of *most*-type quantifiers and of atypical critical words in the context of *few*-type quantifiers, and by comparable amounts. To a lesser extent, the discourse context also reduced the predictability of atypical words in the context of *most*-type quantifiers and typical words in the context of *few*-type quantifiers (cloze probabilities for both were already near floor for the no-context sentence fragments). These normative data show that the *most*- and *few*-type quantifiers had the expected crossover interaction effect on the predictability (cloze probability) of the typical and atypical words, and further, the crossover interactions were approximately symmetric. Consequently, the predictability was higher and well-matched for typical words in the context of *most*-type quantifiers and atypical words in the context of *few*-type quantifiers. The predictability was lower and also well-matched for typical words in the context of *few*-type quantifiers and atypical-words in the context of *most*-type quantifiers.

Procedure

Participants were seated in a comfortable chair in a dimly lit electrically shielded, sound attenuating testing chamber (Industrial Acoustics Co.). Stimuli were presented under computer control on a 21" VGA monitor in a white Arial font against a dark background at a viewing distance of about 120 cm. Discourse contexts were presented in their entirety on a single screen. For the RSVP quantifier sentences, a fixation target (++++) subtending about 2 degrees of visual angle was presented briefly at the center of the screen, followed by presentation of the sentence one word at a time at an SOA of 500 ms (=30 monitor refresh cycles at 60 Hz). Each word appeared centered on the screen for approximately 200 ms (=12 refresh cycles). Stimuli were presented in blocks of 20 trials followed by a brief break. Participants were instructed that they would be reading sentences on the computer screen while their brainwaves were recorded. They were encouraged to minimize eye-movements and blinks while the sentences were presented in order to reduce artifacts in the EEG. The instructions were followed by a brief practice session to familiarize participants with the stimulus presentation and task using sentences unrelated to the experimental materials.

Secondary tasks: rating plausibility and reading for comprehension

Two types of secondary tasks were employed. In Experiments 1 and 3A, 2.3 s after each RSVP quantified sentence, an on-screen cue appeared ("How plausible?"), prompting participants to rate the plausibility of the sentence on a five-point scale (1 = highly implausible, 5 = highly plausible). Responses were registered on a 5-button keypad actuated by the thumb and fingers of the right hand. In Experiments 2 and 3B, participants were instructed to read the sentences for meaning and told that they would occasionally be asked to answer questions about what they had read. Comprehension questions were presented in a random order and in equal numbers for the *most*- and *few*-type quantifiers on 25% of the trials For the discourse context version (Experiment 2), the comprehension question could be answered in reference to the quantified sentence along with the scenario described, e.g., *In mountainous areas where the terrain is uneven and it's easier to go around obstacles than forge direct routes through them . . . few roads are curved due to engineering limitations. Are there many winding roads in mountainous areas?* For the no context version of the materials (Experiment 3B), the comprehension questions could be answered in reference to the quantified sentence alone. Participants responded by making a button-press with response switches held in the left and right hands.

EEG recording and analysis

Scalp ERPs were recorded from 26 tin electrodes embedded in an elastic cap as described in Ganis, Kutas, and Sereno (1996), arrayed in a laterally symmetric quasi-geodesic pattern of triangles approximately 4 cm on a side (see U&K 2010, Fig. 1). An additional electrode was located over the right mastoid (A2); eye movements and blinks were monitored by recording the electro-oculogram (EOG) via four electrodes, one located adjacent to the outer canthus of and one below each eye. Potentials at all locations were recorded against a common reference electrode located over the left mastoid (A1), amplified with Grass Model 12 Neurodata Acquisition System (20 K gain except for 10 K gain for EOG and prefrontal locations, high pass filter 0.01 Hz, low pass filter 100 Hz), and digitally sampled (12-bits, 250 samples/s). Recordings were re-referenced offline to the mathematical average of the potentials over left and right mastoid. Single trial epochs spanning the interval from 500 ms pre-stimulus to 1500 ms post-stimulus were extracted from the continuous EEG and screened for artifacts by computer algorithm and confirmed by visual inspection. On average across participants, artifact rejection rates were approximately balanced across conditions in each experiment: 4–5% of the trials were rejected in Experiment 1; 7–8% were rejected in Experiment 2; 7–8% were rejected in Experiment 3A; 8–11% were rejected in Experiment 3B.

ERP analyses at midline and mediolateral electrodes were conducted as described in U&K 2010. Time-domain average ERPs at the critical target word position were computed for each participant. Mean ERP amplitude relative to a 200 ms prestimulus baseline was calculated at the following latencies: N400, 300–500 ms post-stimulus; late positivity (LP) 500–800 ms post-stimulus; and slow wave (SW) 800–1300 ms post-stimulus. Mean potentials were analyzed separately for the midline electrodes and for sixteen of the remaining electrodes at locations distributed across the scalp in a laterally symmetrical array. For the midline electrodes we conducted fully crossed repeated measures ANOVAs with stimulus factors of quantifier type (two levels: *most*-type, *few*-type), critical word typicality (two levels: typical, atypical), and an electrode location factor. Following U&K 2010, for the N400 window, midline

central and posterior electrodes were analyzed (Ce, Pa, Oc). For the LP and SW windows, the analysis included the prefrontal electrode as well (Pf, Ce, Pa, Oc). To characterize scalp distribution of the effects, we conducted a $2 \times 2 \times 2 \times 4$ ANOVA crossing the quantifier type and critical word typicality factors with electrode location factors: two levels of hemisphere (left, right), two levels of laterality (lateral, medial), and four levels of anteriority, prefrontal (Pf), frontal (Fr), temporo-central (TC), parieto-occipital (PO). For *F* tests involving more than one degree of freedom in the numerator, we report *p* values for Greenhouse–Geisser epsilon-adjusted degrees of freedom (Greenhouse & Geisser, 1959), the value of epsilon, and the original (unadjusted) degrees of freedom. R (R Development Core Team, 2014; ggplot2 Wickham, 2009) and Inkscape (Bah, 2007) were used for statistical analyses and figure construction.

Results

We present here planned comparisons for the critical hypothesis tests and ANOVAs for main effects of quantifier, typicality, and interactions involving these factors and electrode location for the midline electrodes. For completeness, ANOVAs for the mediolateral electrodes are reported in the Supplementary Material.

Experiment 1: supporting discourse context and rating for plausibility

Quantified sentences were read and rated for plausibility following brief discourse contexts that introduced exceptional and/or specific situations.

Plausibility judgments

In Experiment 1, the mean plausibility ratings made after each sentence ranged from a low of 1.6 to a high of 4.5 on a five-point scale from 1 to 5 (Table 2; Fig. 2, Panel B, top row, left). For items containing *most*-type quantifier sentences, plausibility ratings were higher for those containing typical critical words ($M = 4.5$, $SD = 0.29$) than for those containing atypical critical word ($M = 1.6$, $SD = 0.29$). This pattern was reversed for the *few*-type quantifier sentences where plausibility ratings for items with typical critical words ($M = 1.7$, $SD = 0.38$) were lower than for those with atypical critical words ($M = 4.4$, $SD = 0.28$). The resulting nearly symmetric crossover interaction effect for the quantifier and typicality factors was reliable, $F(1,15) = 421.3$, $p < .001$, $\eta_p^2 = .97$.

Table 2
Plausibility judgments for Experiments 1 and 3A.

Quantifier	Typicality	Experiment 1 discourse context	Experiment 3A no discourse context
Most	Typical	4.5 (0.29)	3.93 (0.27)
Most	Atypical	1.6 (0.29)	2.25 (0.40)
Few	Typical	1.7 (0.38)	2.25 (0.32)
Few	Atypical	4.4 (0.28)	3.66 (0.32)

Note: Mean values, *SD* in parentheses on scale from 1 to 5.

ERP results

The ERP morphology at the critical words was unexceptional for the 500 ms SOA RSVP paradigm (Fig. 3, Panel A; Fig. 4, Panel A). P1–N1–P2 potentials were observed over lateral occipital scalp evolving between about 50 and 200 ms poststimulus, followed by a large fronto-central positive-going deflection (P2). Following the P2, a broadly distributed negative-going deflection (N400) was observed, largest over medial central and parietal scalp, peaking between about 300 to 400 ms poststimulus and varying by experimental condition. In the context of *most*-type quantifiers, there was a typicality effect with larger N400s (i.e., greater relative negativity) for atypical words (*Most kids prefer vegetables*) than for typical critical words (*Most kids prefer sweets*). This effect was observed over midline central and parietal, and to a lesser extent occipital, scalp. In the LPC analysis window (500–800 ms), differences between conditions are small and variable across the scalp and there is little indication of systematic effects of the experimental manipulations. In the SW window (800–1300 ms) a typicality effect appears for both *most*- and *few*-type quantifiers, though reversed in polarity in comparison with the N400 typicality effect. That is, in the context of the *most*-type quantifiers the atypical words (*Most kids prefer vegetables*) elicited a greater positivity than did the typical critical words (*Most kids prefer sweets*) whereas in the context of *few*-type quantifiers where the typical critical words (*Most kids prefer sweets*) elicited a greater positivity than the atypical words (*Most kids prefer vegetables*).

N400 300–500 ms. Midline analysis. ANOVA found that quantifier type interacted with critical word typicality, $F(1,15) = 6.92$, $MSE = 5.38$, $p = .019$, $\eta_p^2 = .32$; no other main effects or interactions involving the quantifier, typicality, and electrode factors were reliable. *Hypothesis tests* (Fig. 2, Panel B, top row, N400 effects): Criterion 1 was satisfied, i.e., in the context of *most*-type quantifiers, an N400 effect in the predicted direction was found, with atypical critical words ($M = 2.03$, $SD = 2.87$) eliciting an N400 that was 1.50 μV greater (relatively more negative) than typical words ($M = 3.52$, $SD = 2.43$), $t(15) = 3.66$, $p_{1\text{-tailed}} = .001$, $d = .915$. Criterion 2 was also satisfied. For typical critical words, the 1.33 μV effect of quantifier on N400 amplitude was in the predicted direction, i.e., larger (more relatively negative) in the context of few-type quantifiers ($M = 2.19$, $SD = 2.43$) than *most*-type quantifiers ($M = 3.52$ μV, $SD = 2.43$ μV) and reliable, $t(1,15) = 3.09$, $p_{1\text{-tailed}} = .004$, $d = .773$. Criterion 3 was not satisfied, i.e., there was no evidence that the *most*- and *few*-type quantifiers had a differential impact on processing the atypical words which elicited relatively large N400s in the context of both *few*-type quantifiers ($M = 2.46$ μV, $SD = 2.20$ μV) and *most*-type quantifiers ($M = 2.03$ μV, $SD = 2.87$ μV). The small numerical difference between them was not reliable ($p_{2\text{-tailed}} > .41$). Nor was Criterion 4 satisfied. In the context of the *few*-type quantifiers the typicality effect was negligible ($p_{2\text{-tailed}} > .61$) and hence, there was no reversal of the

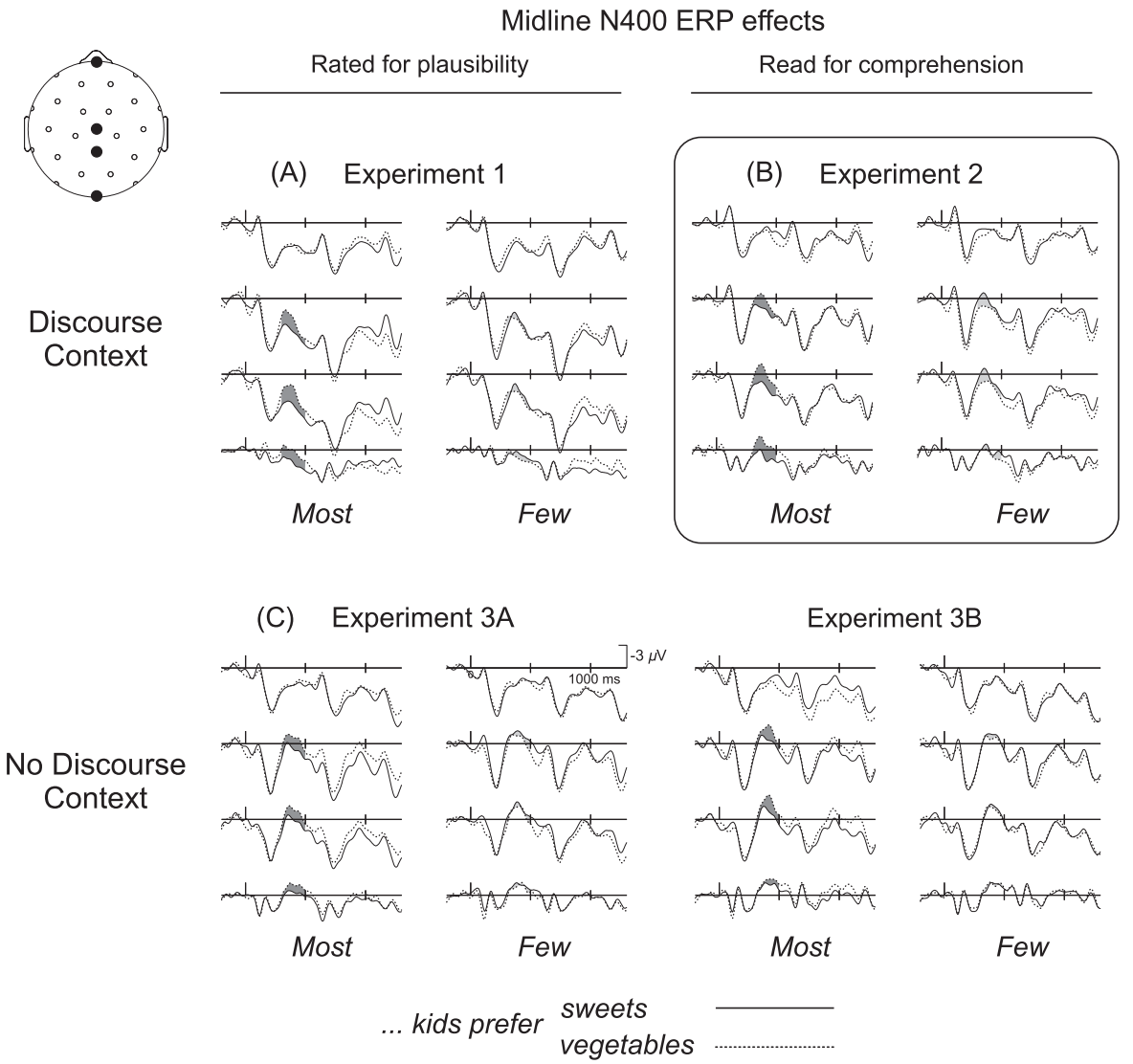

Fig. 3. Grand average ERP potentials at four midline scalp locations (prefrontal, central, parietal, occipital). Panel (A) Experiment 1 ($N = 16$) with quantified sentences presented following supporting discourse and rated for plausibility. Panel (B) Experiment 2 ($N = 16$) with the same quantified sentences and discourse contexts read for comprehension. Panel (C) Control experiments with the quantified sentences presented without the discourse contexts and rated for plausibility in Experiment 3A (left column, $N = 16$) or read for comprehension in Experiment 3B (right column, $N = 16$). Shading indicates the N400 mean potential analysis window for incremental quantifier interpretation hypothesis testing; negative is plotted up.

typicality effect that was observed in the context of the *most*-type quantifiers.

LPC 500–800 ms. Midline analysis. In the analysis of the LPC amplitude at midline electrodes, quantifier and typicality did not reliably interact, $F(1,15) < 1$.

SW 800–1300 ms. Midline analysis. There was a quantifier and typicality interaction effect for the SW amplitude, $F(1,15) = 5.06$, $MSE = 6.12$, $p = .040$, $\eta_p^2 = .25$. Pair-wise comparisons uncorrected for Type I error-rate inflation, found a marginal typicality effect in the context of *few*-type quantifiers where the SW for typical words was more positive ($M = 4.33$, $SD = 2.42$), than for atypical words ($M = 3.62$, $SD = 1.81$), $t(15) = 2.07$, $p_{2\text{-tailed}} = .056$, $d = 0.519$. A quantifier effect was also found for atypical words which were more positive in the context of *most*-type quantifiers ($M = 4.36$, $SD = 2.42$) than in the context of *few*-type quantifiers ($M = 3.62$, $SD = 1.81$), $t(15) = 2.52$, $p_{2\text{-tailed}} = .023$, $d = 0.631$. The sizes of these effects varied by location. In the context of the *most*-type quantifiers, the typicality effect was largest over central and parietal scalp where the SW elicited by atypical words was generally more positive than the SW elicited by typical words. This effect was smaller over occipital scalp and reversed over prefrontal scalp where the SW elicited by typical words was more positive than the SW elicited by atypical words. In the context of the *few*-type quantifiers, typical words elicited a

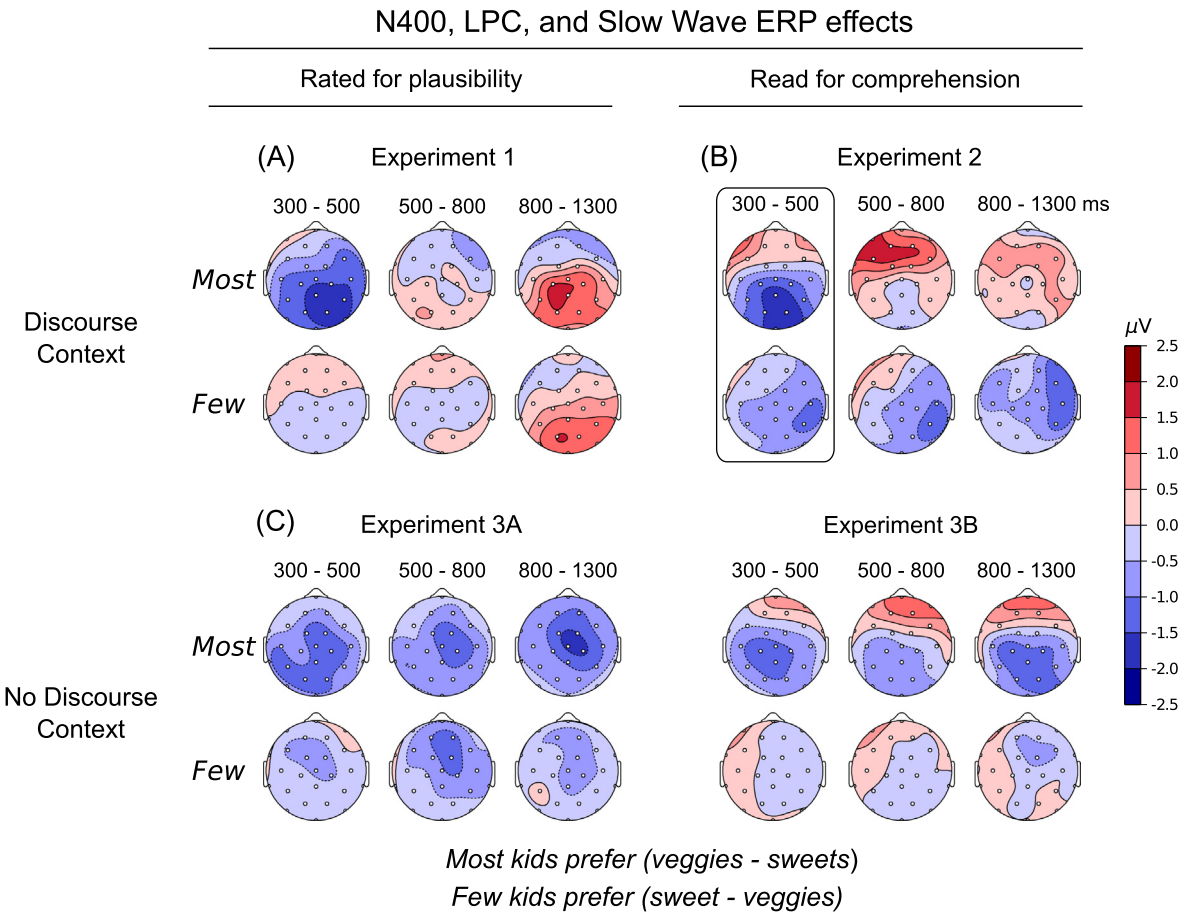

Fig. 4. Spline-interpolated scalp distribution plots of N400 (300–500 ms), LPC (500–800 ms) and slow wave (800–1300 ms) ERP amplitude effects. Panel (A) Experiment 1 ($N = 16$) with quantified sentences presented following supporting discourse and rated for plausibility. Panel (B) Experiment 2 ($N = 16$) with the same quantified sentences and discourse contexts read for comprehension. Panel (C) Control experiments with the quantified sentences only presented without the discourse contexts and rated for plausibility in Experiment 3A (left column, $N = 16$) or read for comprehension in Experiment 3B (right column, $N = 16$). For the *most*-type quantifiers, negative values 300–500 ms poststimulus indicate greater N400 amplitude for atypical–typical words, i.e., the canonical typicality effect (Criterion 1). *Note:* For the *few*-type quantifiers, the typicality effect calculation is reversed so negative values 300–500 ms indicate a reversal of the canonical typicality effect, i.e., satisfaction of Criterion 4. Contour lines indicate 0.5 μV intervals.

more positive SW than atypical words; this effect was largest over central and parietal scalp, smaller over occipital scalp and negligible over prefrontal scalp. These differences in the polarity and distribution of the typicality effect along the midline for the *most*- vs. *few*-type quantifiers resulted in an interaction effect between quantifier, typicality, and electrode location, $F(1, 15) = 4.98$, $MSE = 1.69$, $p = .014$, $\eta_p^2 = .25$.

Experiment 1 discussion

As expected, the plausibility ratings exhibited a large, approximately symmetric, crossover interaction effect. This pattern aligns with the cloze probability effects and anchors our assumption that the quantifier expressions and critical words in these materials were interpreted in accord with their usual meanings by the time the plausibility judgment was rendered. Furthermore, the two plausible conditions (*most + typical*, *few + atypical*), both have high and well-matched plausibility ratings and the two implausible conditions both have low and well-matched plausibility ratings. In this respect, these materials mark a substantial improvement over U&K 2010 (see Figs. 2 and 3 therein) where the *few + atypical* sentences (*e.g., Few farmers grow worms*) were not rated as plausible as the *most + typical* sentences (e.g., *Most farmers grow crops*). Consequently, in U&K 2010, the quantifier effect (*most*- vs. *few*-type) on the plausibility ratings of the atypical words was smaller than for typical words and may have contributed to the failure to find a corresponding quantifier effect on the N400. The large plausibility effects and more nearly symmetric quantifier by typicality crossover interaction observed here in Experiment 1 militate against this concern.

A key ERP finding in Experiment 1 is that, as expected, in the context of *most*-type quantifiers, the atypical words elicited larger N400s than typical words (Criterion 1). In contrast with the offline (non-speeded) measure of cloze probability and the post-sentential plausibility judgments.,

this canonical typicality effect occurs in real-time as the critical words (typical or atypical) are first encountered, i.e., incrementally. We take the fact that the *most*- and *few*-type quantifiers modulated typical word N400 amplitudes reliably in the expected direction to show that the meaning of the different quantifiers types is appreciated rapidly enough and to a sufficient degree to impact processing by the time critical words are encountered. However, as in U&K 2010, the N400 data provide no evidence that these same quantifiers impact real-time processing of the atypical words. Even though the quantifier by typicality interaction effect is reliable and (numerically) in the predicted direction of a crossover interaction, this interaction effect is primarily a consequence of the increased N400s to typical words following *few*-type quantifiers.

A somewhat different pattern is evident for the LPC and slow wave potentials following the N400. In these intervals, the potentials over central and parietal scalp in all conditions are markedly positive in comparison to those in the other experiments in this report as well as the isolated sentences judged for plausibility in Experiments 2 & 3 in U&K 2010. The LPC elicited by the critical words in Experiment 1 differed little between any of the conditions but in the slow wave interval 800–1300 ms post-stimulus, a typicality effect in the context of *most*-type quantifiers emerges, opposite in polarity from the N400 typicality effect in the context of most-type quantifiers and there is a reversed SW typicality effect for *few*-type quantifiers as well. These SW effects in Experiment 1 are notable in that they are the first approximately symmetrical crossover interaction observed for potentials at any latency in this series or Experiments 2 and 3 in U&K 2010. One of the ways to improve upon the common characterizations of incremental interpretation as "rapid" or "not substantially delayed" is by specifying upper bounds on processing times. Though there are other stimulus differences besides the addition of the supporting discourse context, this finding initially suggests that the impact of the context may occur downstream of the processing reflected by the N400 amplitude modulations, i.e., primarily after about 800 ms.

In comparison with U&K 2010, Experiment 1 was designed to employ more natural and varied stimuli and use a discourse context to increase the pragmatic felicity of the quantified propositions. However, in contrast with what has been reported for pragmatically felicitous negation (discussed above), the pattern of N400 responses to the critical typical and atypical words do not yet indicate full incremental interpretation. Rather, in satisfying only the first two of the requisite four criteria, the N400 results in Experiment 1 are evidence that quantifier semantics are rapidly incorporated into the sentence interpretation in some comparisons, but do not impact incremental processing as fully as they could have in other cases. In this respect, the online processing consequences of the quantifier semantics appear to dissociate from the end product of this processing as inferred from slower, downstream cloze task word continuations and plausibility judgments.

Experiment 2: supporting discourse context and reading for comprehension

Experiment 1 was designed to address the issue of pragmatic infelicity in U&K 2010 by testing for full incremental quantifier interpretation in supporting discourse contexts. Experiment 2 extends this series by addressing the potential impact of plausibility judgments, with a new group of participants whose secondary task was to read for comprehension and answer occasional questions about the scenario described.

ERP results

In Experiment 2 (Fig. 3, Panel B; Fig. 4, Panel B), potentials did not exhibit the pronounced broadly distributed slow positive shift observed in Experiment 1. As in Experiment 1, there was a typicality effect in the context of *most*-type quantifiers, with larger N400s (i.e., greater relative negativity) for atypical than for typical critical words. This effect was largest over bilateral medial parietal and occipital scalp, and also evident to a lesser extent over central scalp. In the context of *few*-type quantifiers (e.g., *Few kids prefer*...) and, for the first time in this series of experiments, this N400 pattern was reversed: the typical critical words (*sweets*) elicited a larger N400 than the atypical words, (*vegetables*).

In the LPC window (500–800 ms) the typicality effect in the context of the *few*-type quantifiers was little changed from the pattern in the N400 window. In the context of the *most*-type quantifiers, the posterior relative negativity for atypical *vs.* typical words was small and confined to the parietal and occipital midline whereas the magnitude of the anterior relative positivity effect was larger in comparison with 300–500 ms. In the SW analysis window (800–1300 ms), the typicality effect in the context of *most*-type quantifiers is negligible and in the context of the *few*-type quantifiers, it appears as a small somewhat right lateralized effect, potentials elicited by atypical words slightly more positive than those elicited by typical words.

N400 (300–500 ms). Midline analysis. ANOVA for data recorded at the three midline electrodes found a reliable crossover interaction effect between quantifier type and critical word typicality, $F(1,15) = 19.75$, $MSE = 3.17$, $p < .001$, $\eta_p^2 = .57$. *Hypothesis tests* (Fig. 2, Panel C, top row, N400 effects): Importantly, all four criteria for evidence of full immediate quantifier interpretation were satisfied. Criterion 1: following *most*-type quantifiers, there was a reliable 1.51 μV typicality effect in the expected direction with the atypical critical words, ($M = -0.13$, $SD = 1.47$) more relatively negative than typical words, ($M = 1.38$, $SD = 2.04$), $t(1,15) = 3.31$, $p_{1\text{-}tailed} = .002$, $d = 0.83$. Criterion 2: For typical critical words, there was a reliable 0.94 μV quantifier effect in the expected direction, i.e., relatively greater N400 amplitude ($M = 0.44$ μV, $SD = 2.01$) following *few*-type quantifiers (e.g., *Few kids prefer sweets*) than following *most*-type quantifiers (e.g., *Most kids prefer sweets*) ($M = 1.38$ μV, $SD = 2.04$), $t(1,15) = 3.07$, $p_{1\text{-}tailed} = .004$, $d = 0.77$. Criterion 3: For atypical critical words, there was a reliable 1.35 μV N400 quantifier effect in the expected

direction with N400 amplitude in the context of *most*-type quantifiers (e.g., *Most kids prefer vegetables*), more relatively negative ($M = -0.13$ μV, $SD = 1.47$) than in the context of *few*-type quantifiers ($M = 1.22$ μV, $SD = 1.81$), $t(1,15) = 2.73$, $p_{1\text{-tailed}} = .008$, $d = 0.68$. Criterion 4. Crucially, in the context of the *few*-type quantifiers, there was a reliable 0.78 μV N400 typicality effect in the expected direction, i.e., the **reverse** of the *most*-type typicality effect, with the N400 elicited by typical critical words ($M = 0.44$ μV, $SD = 2.01$ μV) more relatively negative than atypical critical words ($M = 1.22$ μV, $SD = 1.81$ μV), $t(1,15) = 2.84$, $p_{1\text{-tailed}} = .006$, $d = 0.71$.

LPC (500–800 ms). Midline. At the four midline electrodes there were no reliable LPC effects involving the quantifier and typicality factors.

Slow wave (800–1300 ms). Midline. At the four midline electrodes there were no reliable SW effects involving the quantifier and typicality factors.

Experiment 2 discussion

In Experiment 2 we find, for the first time in this series, and to our knowledge, anywhere, that all four criteria for evidence of full incremental quantifier interpretation are satisfied. Quantifiers modulated N400 amplitude for both typical and atypical words, in the predicted directions, and with effects large enough that the canonical typicality effect for *most*-type quantifiers was reversed by *few*-type quantifiers. We conclude that when reading such sentences for comprehension in pragmatically supporting contexts, the interpretations of the *most* and *few*-type quantifiers are fully processed and incorporated without significant delay as meaning is constructed in real-time.

Control Experiments 3A and 3B: no discourse context

Although the key findings in Experiment 2 make a strong case for full incremental quantifier interpretation, an open question remains. Since Experiment 2 is the first in this series (including U&K 2010) that uses a reading for comprehension task instead of plausibility judgment, we cannot say for sure whether the N400 cross-over effect at the critical word is a consequence of the pragmatically licensing discourse context or the reading task or the combination. Though it seems unlikely at this juncture, the plausibility rating may tax the cognitive system in ways that disrupts incremental interpretation. If so, quantifiers may be interpreted fully and immediately even in isolated sentences.

To resolve this uncertainty we conducted two additional control experiments using the same quantified sentences without the preceding discourse context. In Experiment 3A, the sentences were rated for plausibility as in Experiment 1; in the critical Experiment 3B, the sentences were read for comprehension as in Experiment 2. Comparing Experiment 1 with Experiment 3A tests the impact of the discourse context when rating the single quantified sentences for plausibility. Based on the other experiments in this series, there is no reason to predict full and immediate quantifier interpretation in Experiment 3A. The crucial test of the contribution of discourse context is in Experiment 3B. If the reading task rather than discourse context is driving full incremental quantifier interpretation, the four criteria should be satisfied when isolated sentences are read for comprehension.

Experiment 3A plausibility ratings

In Experiment 3A the quantifier sentences were presented to new group of participants word by word, without a preceding discourse context, and rated for plausibility. Mean plausibility ratings ranged from a low of 2.25 to a high of 3.93 (Table 2, Fig. 2, Panel B, bottom row). For items containing *most*-type quantifiers, plausibility ratings were higher when the sentence contained a typical critical word ($M = 3.93$, $SD = 0.27$) than atypical critical word ($M = 2.25$, $SD = 0.40$). This pattern was reversed for the *few*-type quantifiers where plausibility ratings for items with typical critical words ($M = 2.25$, $SD = 0.32$) were lower than for those with atypical critical words ($M = 3.66$, $SD = 0.32$). The resulting nearly symmetric crossover interaction effect for the quantifier and typicality factors was reliable, $F(1,15) = 337.31$, $MSE = 0.11$, $p < .001$, $\eta_p^2 = .96$.

Experiment 3A ERP results

When following *most*-type quantifiers, N400s were generally larger for atypical critical words and smaller for typical critical words. This pattern persists throughout the LPC (500–800 ms) and SW (800–1300 ms) analysis windows, increasing slightly in the latter (Fig. 3, Panel C, left column; Fig. 4, Panel C, left column). By contrast, in the context of *few*-type quantifiers the pattern of typicality effects differs both in time course and polarity. There is little evidence of an N400 typicality effect; both typical and atypical critical words elicit similar N400s comparable in amplitude to atypical words in the context of *most*-type quantifiers. However, in the LPC window following the N400, the typicality effect reverses in comparison with *most*-type quantifiers, i.e., the potentials elicited by the typical words are relatively more negative than atypical words. This cross-over interaction effect of quantifier and typicality in the LPC window is also evident, though smaller in magnitude in the SW window. This pattern of typicality effects was broadly distributed across the scalp though larger over centro-posterior scalp in the context of *most*-type quantifiers, larger over fronto-central scalp in the context of *few*-type quantifiers, and larger over medial than lateral scalp for both.

N400 300–500 ms. Midline analysis. Quantifier type interacted with critical word typicality, $F(1,15) = 8.62$, $MSE = 3.30$, $p = .010$, $\eta_p^2 = .36$; no other main effects or interactions involving the quantifier or typicality factors were reliable. *Hypothesis tests* (Fig. 2, Panel B, bottom row, N400 effects): Criterion 1 was satisfied: in the context of the *most*-type quantifiers, the 1.14 μV canonical N400 typicality effect was reliable and in the expected direction with N400 amplitude more relatively negative for atypical critical words ($M = -0.82$ μV, $SD = 1.72$) than for typical critical words ($M = 0.32$ μV, $SD = 2.02$), $t(15) = -2.65$), $p_{1\text{-tailed}} = .009$, $d = 0.66$. Criterion 2 was satisfied: for the typical critical words, there was a reliable 1.46 μV

quantifier effect in the expected direction with N400 amplitude relatively more negative in the context of the *few*-type quantifiers ($M = -1.14$ μV, $SD = 1.31$) than in the context of the *most*-type quantifiers ($M = 0.32$ μV, $SD = 2.02$), $t(15) = 3.51$, $p_{1\text{-tailed}} = .001$, $d = 0.88$. However, Criterion 3 was not satisfied: we did not find a reliable quantifier effect for atypical words; N400 amplitude in the context of *few*-type quantifiers ($M = -0.74$ μV, $SD = 1.31$) was little different than in the context of *most*-type quantifiers ($M = -0.82$ μV, $SD = 1.72$ μV), $p_{1\text{-tailed}} = .43$. Nor was Criterion 4 satisfied: the 0.4 μV difference between typical and atypical words in the context of *few*-type quantifiers was small and not reliable, $p_{1\text{-tailed}} = .12$. Since there was no clear typicality effect for *few*-type quantifiers at all, trivially, it was not reversed relative to the canonical typicality N400 effect following *most*-type quantifiers.

Late Positive Complex (LPC): 500–800 ms. Midline analysis. At the four midline electrodes, there was a reliable interaction effect between quantifier type and typicality for LPC amplitude, $F(1,15) = 6.49$, $MSE = 4.45$, $p = .022$, $\eta_p^2 = .30$. Pair-wise comparisons found a reliable quantifier effect for typical words which were more positive in the context of *most*-type quantifiers ($M = 3.29$ μV, $SD = 2.03$) than in the context of *few*-type quantifiers ($M = 1.99$ μV, $SD = 1.74$), $t(15) = 3.29$, $p_{2\text{-tailed}} = .005$, $d = 0.82$. There was no evidence of a quantifier effect on LPC amplitude elicited by atypical words ($p > .92$). There were weak trends toward typicality effects in the context of the *most*-type quantifiers ($p_{2\text{-tailed}} = .12$) and *few*-type quantifiers ($p_{2\text{-tailed}} = .09$) and these effects were in the same direction as those observed for the N400 analysis window though largest over frontal rather than parietal scalp.

Slow wave 800–1300 ms. Midline electrodes. The pattern of slow wave effects at the midline electrodes was generally similar to the N400 and LPC, though the interaction effect between quantifier type and typicality was marginal, $F(1,15) = 4.19$, $MSE = 6.32$, $p = .059$, $\eta_p^2 = .22$.

Experiment 3B ERP results

In Experiment 3B the quantifier sentences were presented to a new group of participants word by word, without preceding discourse context, and read for comprehension with probe questions presented at random on 25% of the trials. The critical N400 effects in Experiment 3B (Fig. 3, Panel C, right column; Fig. 4, Panel C, right column) were all smaller than in Experiment 2. In the context of *most*-type quantifiers, there was a canonical typicality effect with larger N400s (i.e., greater relative negativity) for atypical words than for typical critical words. This effect was observed over midline central and parietal, and to a lesser extent occipital, scalp. At the midline prefrontal electrode, the typicality effect reversed in polarity, with atypical critical words more positive than typical words. This effect was broadly distributed, and for the most part laterally symmetric, with the posterior relative negativity slightly larger over the left than right medial scalp and the prefrontal positivity evident at right, but not left, lateral prefrontal scalp. This effect persisted throughout the LPC and SW analysis windows. In the LPC interval the relative negativity over central and posterior scalp was somewhat reduced and the prefrontal positivity was more pronounced, evident over bilateral frontal and prefrontal scalp, still slightly right lateralized. In the SW interval, the posterior relative negativity increased slightly in comparison with the LPC interval and the SW prefrontal positivity was generally comparable to that in the LPC interval, though somewhat more bilaterally symmetric. By contrast, typicality effects in the context of the *few*-type quantifiers are negligible in the N400 and LPC intervals, with only a small relative positivity over fronto-central scalp emerging in the SW interval.

N400 (300–500 ms). Midline. ANOVA for data recorded at the three midline electrodes found a marginal interaction effect between quantifier type and critical word typicality, $F(1,15) = 4.21$, $MSE = 3.02$, $p = .058$ and a three-way interaction between quantifier type, typicality, and electrode location, $F(1,15) = 4.78$, $MSE = 0.32$, $p = .022$, $\eta_p^2 = .24$. The canonical N400 typicality effect in the context of *most*-type quantifiers was larger over central and parietal scalp than occipital, whereas the typicality effect in the context of *few*-type quantifiers was slightly positive over central scalp and negligible elsewhere. No other main effects or interactions involving the quantifier or typicality factors were reliable. *Hypothesis tests* (Fig. 2, Panel C, right column N400 effects): Criterion 1 was satisfied: In the context of *most*-type quantifiers, a 0.91 μV N400 typicality effect in the expected direction was observed, with atypical words more negative ($M = -1.56$ μV, $SD = 1.80$ μV) than typical words ($M = -0.65$ μV, $SD = 1.87$ μV), $t(15) = 1.87$, $p_{1\text{-tailed}} = .040$, $d = 0.47$. Criterion 2 was not satisfied: Quantifier type did not have a reliable impact on the N400 amplitudes of typical critical words which differed by less than 0.25 μV in the context of *most*-type quantifiers ($M = -0.65$, $SD = 1.87$) and *few*-type quantifiers ($M = -0.89$, $SD = 2.02$), $t(15) = 0.53$, $p_{1\text{-tailed}} > .3$. Criterion 3 was satisfied: quantifier type reliably modulated N400 amplitude for the atypical words in the expected direction, i.e., in the context of *few*-type quantifiers, N400 amplitude for atypical words was smaller ($M = -0.78$ μV, $SD = 1.40$ μV) than in the context of *most*-type quantifiers ($M = -1.56$ μV, $SD = 1.80$ μV), $t(15) = -2.27$, $p_{1\text{-tailed}} = .019$, $d = 0.57$. Criterion 4 was not satisfied: there was no evidence of a typicality effect in the context of *few*-type quantifiers, where the N400s elicited by typical and atypical words differed by less than 0.13 μV, $p_{1\text{-tailed}} > .39$.

LPC 500–800 ms. Midline. ANOVA conducted on LPC amplitudes at the four midline electrodes found no main effects of quantifier or typicality. There was no interaction effect between quantifier and typicality. The typicality effect in the context of *few*-type quantifiers was negligible at all the midline electrodes and in the context of the *most*-type quantifiers, the relative negativity for atypical *vs.* typical words (ranging between −0.52 μV and −0.80 μV on average) at the three electrodes over central and posterior scalp was offset by a substantial relative positivity (1.41 μV) at the midline prefrontal electrode. These different anterior-to-posterior scalp distributions of the typicality effect for

the two types of quantifier resulted in a reliable three-way interaction between quantifier type, typicality, and electrode location, $F(3,45) = 6.55$, $MSE = 1.01$, $p = .009$, $\eta_p^2 = 0.304$.

Slow wave (800–1300 ms). Midline analysis. The pattern of quantifier and typicality slow wave effects was similar to that observed for the LPC, albeit slightly larger in magnitude. In the context of the *few*-type quantifiers there was still little indication of any typicality effect. In the context of the *most*-type quantifiers, the relative negativity elicited by atypical words in comparison with typical words ranged between −0.76 μV and −1.33 μV on average over central and posterior scalp; at the prefrontal electrode, this typicality effect was a relative positivity of 1.45 μV. ANOVA found that the quantifier and typicality factors did not reliably interact, again because of the prefrontal relative positivity offsetting the central and posterior relative negativity. Consequently, as also observed for the LPC, there was a three-way interaction between quantifier, typicality, and electrode location, $F(3,45) = 5.88$, $MSE = 1.27$, $p = .012$, $\eta_p^2 = .28$, $\epsilon_{GG} = 0.531$.

Experiment 3A and Experiment 3B discussion

Experiment 3A served as the no-discourse-context control for Experiment 1. Without discourse context, the N400 effects for the quantifier and typicality variables were generally similar to those observed in Experiment 1 in that the same two (and only two) criteria were satisfied. The pattern of plausibility ratings also was similar in the two experiments. Removing the discourse context had only a modest quantitative impact on the patterns of effects in the plausibility ratings and N400 measures, generally reducing their sizes. The most salient impact of removing the discourse was on the cloze probabilities of the critical words in both plausible conditions, i.e., the predictability of typical words following *most*-type quantifiers and atypical words following *few*-type quantifiers both dropped to very low levels. Overall, the plausibility ratings and N400 results replicate the findings in U&K 2010 where different RSVP quantified sentences also were rated for plausibility without supporting context. We again interpret these results as evidence of partial incremental interpretation. In U&K 2010 (Experiment 2 and Experiment 3) a typicality effect was observed as a frontal positivity in the SW interval following few-type quantifiers (atypical more positive than typical); no such late frontal positivity was evident here in Experiment 3A. Since the procedures and secondary task were held constant by design, the difference seems most likely attributable to the new stimulus materials though participant variables cannot be ruled out.

Experiment 3B served as the no-discourse-context control for Experiment 2, where the critical evidence for strong incremental quantifier interpretation was observed. In Experiment 3B we found (as in all the other experiments herein) a canonical typicality effect wherein atypical critical words elicited a larger N400 than typical words in the context of *most*-type quantifiers (Criterion 1) even without a supporting discourse context or the plausibility rating task. As for the question of whether *most*- and *few*-type quantifiers modulate N400 for the critical words, we find a different pattern in Experiment 3B than in any of our previous experiments in this series (Experiments 1, 2, 3A; also U&K 2010 Experiments 2 and 3). We have previously found in every case that the different quantifiers reliably modulate N400 amplitude of typical words which is larger in the context of *few*-type quantifiers than *most*-type quantifiers. And, with the exception of Experiment 2 in this report, the different quantifiers did not modulate N400 amplitude for the atypical words. Here in Experiment 3B we find the opposite pattern where the quantifiers modulate N400 for atypical but not typical words. We return to this unexpected result briefly in the general discussion. The key result from Experiment 3B is that criteria for full and immediate quantifier interpretation were not all satisfied. We conclude that reading the supporting discourse context in this design plays a role in the full and immediate quantifier interpretation observed in Experiment 2 and that this effect is not merely a consequence of changing to the reading for comprehension task.

General discussion

The current experiments were conducted to probe the speed and depth of noun phrase quantifier interpretation. The primary aim was to test a strong form of the principle of incremental interpretation which predicts that quantifier expressions, like other words, should be processed immediately and fully. To test this prediction we measured N400 amplitudes elicited by critical test words in quantified sentences where two types of quantifier expressions (*most*/*few*), were crossed with typicality (typical, atypical). In the four types of quantified sentences that result, two are consistent with general world knowledge (*Most kids prefer* sweets/vegetables…) and the two that are not (*Most kids prefer* vegetables/sweets…) become so upon encountering the underlined critical word. This feature of the experimental design is the same as in (U&K 2010) where N400 amplitude modulations at the critical test word indicated that quantifier expressions are interpreted incrementally, i.e., their meaning had some impact on processing the critical and atypical critical words, albeit not fully at that critical word. The experiments reported here extend this series by embedding a new set of quantifier sentences in "contexts of plausible quantification".

Normative testing showed that preceding the isolated sentences with the discourse context had a substantial impact on the predictability (cloze probability) of the critical word and a modest impact on the plausibility ratings. Both these offline measures exhibited a robust and approximately symmetric quantifier-by-typicality crossover interaction effect that was smaller, but qualitatively similar without the discourse context. The largest effect of removing the discourse for either offline measure was to decrease the cloze probability of the critical target words in the two more plausible conditions, i.e., typical words following *most*-type quantifiers and atypical words following *few*-type quantifiers.

N400 evidence for strong incremental quantifier interpretation

The strong incremental quantifier interpretation hypothesis—that quantifiers are fully and immediately interpreted in real-time—predicts that quantifier's consequences as sentences unfold mirror their consequences in offline processing. In the present studies this means that the on-line N400 effects elicited by the critical words must mirror the crossover interaction effects observed for the off-line plausibility ratings and critical cloze probabilities. To test this, we proposed a decision rule based on four individually necessary and jointly sufficient criteria that observed N400 effects must satisfy to constitute positive evidence of strong incremental quantifier interpretation. The results across all four experiments can be summarized as follows:

Criterion 1. Is there a typicality effect in the right direction in the context of most-type quantifiers? All the experiments satisfied this criterion. Regardless of discourse context or secondary task, atypical critical words elicited a larger N400 than typical words (*Most kids prefer sweets/vegetables*). This effect is expected since many variables known to modulate N400 amplitude consistently pull in the same direction in this comparison. This N400 typicality effect aligns with pre-theoretic intuitions about the degree of "fit" or "congruity" of the atypical vs. typical critical based on general world knowledge as well as the normative cloze probability ratings. This N400 effect also patterns with the post-sentence plausibility judgments (Experiments 1 and 3A) of those individuals whose brains are generating these N400 effects at the critical word. Finding these N400 typicality effects confirms that the experimental materials behave as intended for the experimental design and that there is sufficient statistical power to detect N400 effects of this magnitude – two key assumptions in our subsequent inferences about the time course of quantifier interpretation.

Criteria 2 and 3. Are there quantifier effects in the right direction on both typical and atypical words? The answer requires the semantics of quantifiers to express: sometimes yes and sometimes No. The different quantifier types modulated N400 amplitude for typical critical words (Experiments 1, 2, and 3A herein; Experiments 2 and 3 in U&K 2010), atypical critical words (Experiments 2 and 3B), or both (Experiment 2). Since only the quantifiers differ and the direction of the N400 amplitude modulation is consistent with the quantifier's meaning in conjunction with the compositional semantics of the sentence and real-world knowledge, we conclude that in each experiment, at least some relevant information about the meaning of the quantifiers is incorporated into the evolving semantic representation of propositional content prior to encountering the critical word rather than being significantly delayed. Lexical properties, e.g., frequency or familiarity, and relations among lexical items, e.g., the semantic relatedness of or associations among *kids*, *prefer*, and *sweets* vs. *vegetables* also likely contribute to the N400 effects at the critical word but in this experimental design these factors are held constant while only the quantifiers vary. So whatever drives the typicality N400 effect of the critical typical and atypical words when they follow the *most*-type quantifiers, if the N400 differs in the context of *few*-type quantifiers, the difference may be attributed to the quantifiers. Finding at least some evidence that the two types of quantifiers impact processing at the critical word position is consistent with our previous findings (U&K 2010, Experiment 2 and Experiment 3), and we again interpret them as evidence of at least partial or underspecified incremental quantifier interpretation.

Criterion 4. Can the canonical atypical vs. typical N400 effect be reversed with few-type quantifiers? This criterion, along with the other three, was satisfied in Experiment 2 (only) where the test materials were read for comprehension and the quantified sentences were presented following pragmatically supportive discourse context. We interpret this N400 cross-over as compelling evidence that the semantic representations of the initial fragment of *few*-type quantified sentences, e.g., *Few kids prefer*..., can be constructed rapidly enough to make processing of what is canonically typical *more* difficult to process than what is canonically atypical. Since typicality is a consequence of general world knowledge and the opposite pattern was observed in the context of *most*-type quantifiers, we conclude that both types of quantifiers were interpreted (1) *incrementally* because the N400 effect is elicited by the critical word when it might have been deferred to a later time and (2) *fully* because in this experimental design, the reversal of the canonical N400 typicality effect can be attributed to the quantifier. To our knowledge the N400 quantifier by typicality interaction effect in Experiment 2 is the first on-line measure that exhibits the full symmetric crossover quantifier interaction pattern that we have consistently observed in the offline plausibility and cloze measures.

Post-N400 discourse and task effects

Across the series of studies, discourse context had a substantial impact on the pattern of relative positivities in the LPC and SW intervals and these differed by task. There were two general tendencies. First, in the absence of discourse context (Experiments 3A and 3B), the general pattern of quantifier and typicality effects evident in the N400 interval (300–500 ms) tended to persist throughout the subsequent LPC (500–800) and SW (800–1300 ms) intervals. Second, when preceded by supporting discourse context (Experiments 1 and 2), the pattern of quantifier and typicality effects was more variable over time. In Experiment 1 with the plausibility judgments, a posterior positivity effect emerged in the SW window. In Experiment 2 with reading for comprehension, a short-lived frontally positive typicality effect emerged in the LPC interval following the *most*-type quantifiers. Positivities following an N400 are widely reported in the literature and have been descriptively labeled "post-N400 positivity" or PNP effects. Despite recent attention, the functional significance of PNPs is not fully understood (for some candidate interpretations see e.g., Brouwer et al., 2012; Gouvea, Phillips, Kazanina, & Poeppel, 2010; Kolk & Chwilla, 2007; Kuperberg, 2007; Van Petten & Luka, 2012). Systematic individual differences have been

proposed to account for some of the wide range of PNP effects (e.g., Kos, van den Brink, & Hagoort, 2012). Our findings suggest that empirically adequate accounts of PNPs must account for discourse and task variables.

Incremental interpretation of logical operators: negation and quantification

Our findings align with recent ERP work on negation and the historical trajectories are strikingly similar. Kounios and Holcomb (1992) manipulated truth value with logical quantifiers, *some* and *all* in addition to negation (*All*/Some/*No rubies are gems*/*spruces*). The quantifiers were not found to have an impact on N400 amplitude in the predicate segment (though they did affect RTs in the sentence verification task). U&K 2010 revisited the time course of quantifier interpretation with a somewhat wider variety of non-logical quantified noun phrases and adverbs of quantification and sentences intended to tap general though not specifically categorical world knowledge. Those experiments marked some progress by finding that quantifier semantics could be appreciated rapidly enough to impact N400 for the typical critical words. We generally replicated those results here in Experiment 1 and, crucially, extended them in Experiment 2 with the full N400 crossover interaction between quantifiers and general world knowledge of the sort reported for negation by Nieuwland and Kuperberg (2008) and Staab (2007).

Two lessons of general interest emerge from this parallel. First, it might have turned out that the time course of interpreting logical function words, including but not limited to negation and quantification, is fundamentally different than the interpretation of open-class content words. That is they could have proven, in time, to be in-principle exceptions to strong formulations of full and immediate incremental interpretation hypotheses. This does not now appear to be the case which is an important fact about the time course of meaning construction during comprehension. The second lesson is a cautionary tale. A few early empirical ERP findings provided evidence that interpretation of negation and quantification is delayed or not fully incremental under some conditions. A few recent studies have found evidence of incremental negation and quantifier interpretation under other conditions. These later studies highlight the risks of overgeneralizing from the earlier ones. However, recency alone does not mitigate the risks of overgeneralizing from a few studies.

Incremental interpretation and underspecification revisited

Affirmations of incremental interpretation are widespread in the psycholinguistics literature. For example, in an influential article (Just & Carpenter, 1980) this theoretical commitment is formulated as an assumption: "The immediacy assumption posits that the interpretations at all levels of processing are not deferred; they occur as soon as possible, a qualification that will be clarified later (p. 330)." They go on to illustrate that delayed processing in contravention of the immediacy assumption may be brief, e.g., the interpretation of the word *large* awaits the appearance of what it modifies (*large insect* vs. *large building*, (p. 341) or potentially longer, e.g., when integrating information across clause boundaries in longer stretches of text. Just and Carpenter are also alive to the role of individual differences and the content and emphasize the role of the comprehender's goals: "There is no single mode of reading. Reading varies as a function of who is reading, what they are reading, and why they are reading it. . . . The reader's goals are perhaps the most important determinant of the reading process (Just & Carpenter, 1980, p. 350)". For some in the field, it appears that subsequent research has promoted incremental interpretation from an assumption to a conclusion. For example, a lucid expression appears in Altmann and Mirkovic (2009, p. 604), where, after reviewing a range of experimental evidence, the authors write, "The view we are left with is of a comprehension system that is "maximally incremental"; it develops the fullest interpretation of a sentence fragment at each moment of the fragment's unfolding." This principle of maximally incremental interpretation is also quickly and explicitly qualified (Altmann & Mirkovic, 2009): "Of course, conversational goals (including participants' goals while engaged in psycholinguistic studies, as well as other nonlinguistic goals) will necessarily change the attentional state of the system . . . leading to changes in what constitutes the fullest possible interpretation of a sentence . . . The "maximal" in "maximal incrementality" is thus situation dependent." Both these passages, separated by some three decades, illustrate a general, and in our view, central feature of current thinking about the time course of language comprehension: as soon as a strong principle of incremental interpretation is articulated, it is immediately qualified, "as incremental as possible". Without elaboration this qualification pushes the principle toward triviality: interpretation is incremental except when it is not. Consequently systematic investigations of when interpretation is incremental and when it is not have an important theoretical role to play in determining the scope of the principle of incremental interpretation in language comprehension.

Future work

The previous observation leads directly to the next one. The results of this series of studies marks progress in understanding the timecourse of quantifier interpretation but there are clearly many open questions. Linguists distinguish different types of quantifier expressions on syntactic and semantic grounds, e.g., some quantifiers license negative polarity items (*Most doctors are not criminals*), and others do not (**Few doctors are not criminals*). Some but not all quantifiers are semantically (truth-functionally) equivalent to others in combination with negation (*a few* vs. *not many*). Some quantifiers are more vague (*many*, *few*) while others are more precise (*at least one*, *exactly three*, *half of*, *all*). Quantified noun phrases constructed from such expressions can occur in syntactic argument positions with thematic roles (*Many kids like sports*) and in adjuncts, e.g., prepositional phrases (*Although nearly hunted to extinction, wild turkeys are now found in every state in the continental U.S.*). Quantified noun phrases can be combined within and across clauses that result in well-known interpretive ambiguities: *Many kids like a few*

sports can be true if many kids like the same few sports or each of the many kids likes a different few sports. These illustrative quantifier phenomena occur within a single sentence of clause. Language comprehension in the general case involves constructing interpretations of information spanning multiple sentences. In a discourse about a birthday party, *There were a bunch of kids at Joanna's birthday party*. *Some boys ate all the candy*, the comprehender must work out that there is a definite group of boys, definite stock of candy, and an eating event completed in the past that exhausted the candy supply. Furthermore, she must also work out whether the group of boys denoted in the second sentence was among the group of kids denoted in the first or whether this is some new group, e.g., the bad boys from down the street. We have seen that discourse contexts can impact the fine-grained time course of quantified subject noun phrases. We have not yet probed the different kinds of discourse information that might have such an effect nor whether or to what extent similar results hold for the many other types of quantifiers.

Conclusion

Language researchers work at putting together the puzzle of how the comprehension system maps verbal input to meaning on the fly. Expressions of quantity are an integral part of meaning in natural languages and appear in a wide range of forms and constructions. This series of studies continued the line of work reported in U&K 2010 by investigating quantified noun phrases, testing them in and out of discourse contexts and under two task conditions with the aim of determining whether the real-time interpretations of quantifiers is best characterized as full and immediate or partial and delayed. One important result is the evidence from Experiment 2 that under the right conditions – in supporting discourse context and while reading for comprehension – quantified noun phrases can be interpreted fully and incrementally, in so far as this can be inferred from what is known about the sensitivity of N400 brain potentials to experimental manipulations of meaning. Since these N400 results parallel those from recent investigations of negation, further study may allow both to be subsumed under a general regularity governing the processing of logical semantic elements in message-level interpretations. A second important result is that across the studies, the patterns of offline behavior, i.e., plausibility judgments, cloze probabilities, and patterns of real-time brain activity at the critical word often align but sometimes dissociate. These dissociations crucially mean that inferences from any single measure alone to conclusions about the general operation of the system are incomplete at best. For while the plausibility judgments and cloze probabilities consistently exhibit strong effects of quantifiers and typicality that are amplified by supporting discourse context, the brain potentials show that the time course of processing on the way to these behavioral responses is more variable across tasks and discourse contexts. So looking just at these real-time brain potentials would miss the consistency evident in the end state of the interpretive processing and looking just at the end state would miss the fine-grained differences in time-course evident in the brain potentials. These complex patterns of associations and dissociations across systematic experimental manipulations serve as a reminder of how complex language comprehension truly is, while at the same time highlighting the value of using different measures on different time scales to constrain conclusions about the nature of the processing.

Acknowledgment

This research was supported by NIH grants HD-22614 and AG-08313 to Marta Kutas.

A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jml.2015.03.010.

References

Altmann, G. T. M., & Mirkovic, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33*, 583–609.
Bah, T. (2007). *Inkscape: Guide to a vector drawing program*. Prentice Hall Press.
Baker, L., & Wagner, J. L. (1987). Evaluating information for truthfulness – The effects of logical subordination. *Memory & Cognition, 15*, 247–255.
Barton, S. B., & Sanford, A. J. (1993). A case-study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & Cognition, 21*, 477–487.
Bates, E., & MacWhinney, B. (1989). The competition model. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of sentence processing*. Cambridge, UK: Cambridge University Press.
Bornkessel, I., & Schlesewsky, M. (2006). The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological Review, 113*, 787–821.
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.
Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research, 1446*, 127–143.
Daneman, M., Lennertz, T., & Hannon, B. (2007). Shallow semantic processing of text: Evidence from eye movements. *Language and Cognitive Processes, 22*, 83–105.
Dwivedi, V. D. (2013). Interpreting quantifier scope ambiguity: Evidence of heuristic first, algorithmic second processing. *PLoS ONE, 8*.
Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior, 20*, 540–551.
Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology, 47*, 164–203.
Filik, R., Paterson, K. B., & Liversedge, S. P. (2004). Processing doubly quantified sentences: Evidence from eye movements. *Psychonomic Bulletin & Review, 11*, 953–959.
Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry, N. W. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology, 20*, 400–409.
Francis, W. N., & Kucera, H. (1979). *Brown corpus manual*. Brown University, Department of Linguistics.
Frazier, L., Clifton, C., Rayner, K., Deevy, P., Koh, S., & Bader, M. (2005). Interface problems: Structural constraints on interpretation? *Journal of Psycholinguistic Research, 34*, 201–231.
Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart (Ed.), *Attention and performance 12: The psychology of reading* (pp. 559–586). Hove, England: Lawrence Erlbaum Associates.
Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences, 6*, 78–84.
Frisson, S. (2009). Semantic underspecification in language processing. *Language and Linguistics Compass, 3*, 111–127.

Ganis, G., Kutas, M., & Sereno, M. I. (1996). The search for "common sense": An electrophysiological study of the comprehension of words and pictures in reading. *Journal of Cognitive Neuroscience, 8*, 89–106.
Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes, 25*, 149–188.
Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95–112.
Hackl, M., Koster-Hale, J., & Varvoutis, J. (2012). Quantification and ACD: Evidence from real-time sentence processing. *Journal of Semantics, 29*, 145–206.
Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*, 438–441.
Hagoort, P., & van Berkum, J. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society B – Biological Sciences, 362*, 801–811.
Hald, L. A., Steenbeek-Planting, E. G., & Hagoort, P. (2007). The interaction of discourse context and world knowledge in online sentence comprehension. Evidence from the N400. *Brain Research, 1*, 210–218.
Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology, 58*, 376–415.
Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*, 329–354.
Kaan, E., Dallas, A. C., & Barkley, C. M. (2007). Processing bare quantifiers in discourse. *Brain Research, 1146*, 199–209.
Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language, 52*, 205–225.
Kolk, H., & Chwilla, D. (2007). Late positivities in unusual situations. *Brain and Language, 100*, 257–261.
Kos, M., van den Brink, D., & Hagoort, P. (2012). Individual variation in the late positive complex to semantic anomalies. *Frontiers in Psychology, 3*, 1–10.
Kos, M., Vosse, T., van den Brink, D., & Hagoort, P. (2010). About edible restaurants: Conflicts between syntax and semantics as revealed by ERPs. *Frontiers in Psychology, 1*, 1–11.
Kounios, J., & Holcomb, P. J. (1992). Structure and process in semantic memory: Evidence from event-related brain potentials and reaction-times. *Journal of Experimental Psychology – General, 121*, 459–479.
Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146*, 23–49.
Kurtzman, H. S., & Macdonald, M. C. (1993). Resolution of quantifier scope ambiguities. *Cognition, 48*, 243–279.
Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*, 375–419.
Macdonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676–703.
Marslen-Wilson, W., & Tyler, L. K. (1975). Processing structure of sentence perception. *Nature, 257*, 784–786.
McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38*, 283–312.
Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology, 4*, 61–64.
Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science, 19*, 1213–1218.
Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience, 18*, 1098–1111.
Paterson, K. B., Filik, R., & Moxey, L. M. (2009). Quantifiers and discourse processing. *Language and Linguistics Compass, 3*, 1390–1402.
Politzer-Ahles, S., Fiorentino, R., Jiang, X. M., & Zhou, X. L. (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.
R Development Core Team (2014). *R: A language and environment for statistical computing (version 3.1.0 (2014-04-10) – "Spring Dance")*. Vienna, Austria: R Foundation for Statistical Computing.
Rayner, K., & Clifton, C. (2009). Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research. *Biological Psychology, 80*, 4–9.
Reder, L. M., & Kusbit, G. W. (1991). Locus of the Moses illusion: Imperfect encoding, retrieval, or match. *Journal of Memory and Language, 30*, 385–406.
Sanford, A. J., Dawydiak, E. J., & Moxey, L. M. (2007). A unified account of quantifier perspective effects in discourse. *Discourse Processes, 44*, 1–32.
Sanford, A. J., & Graesser, A. C. (2006). Shallow processing and underspecification. *Discourse Processes, 42*, 99–108.
Sanford, A. J., Leuthold, H., Bohan, J., & Sanford, A. J. S. (2011). Anomalies at the borderline of awareness: An ERP study. *Journal of Cognitive Neuroscience, 23*, 514–523.
Shallow Processing and Underspecification (2006). [Special Issue]. *Discourse Processes, 42*(2).
Staab, J. (2007). *Negation in context: Electrophysiological and behavioral investigations of negation effects in discourse processing*. San Diego State, La Jolla: University of California, San Diego.
Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*, 415–433.
Tune, S., Schlesewsky, M., Small, S. L., Sanford, A. J., Bohan, J., Sassenhagen, J., et al. (2014). Cross-linguistic variation in the neurophysiological response to semantic processing: Evidence from anomalies at the borderline of awareness. *Neuropsychologia, 56*, 147–166.
Urbach, T. P., & Kutas, M. (2010). Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language, 63*, 158–179.
van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (1999). *Semantic integration in sentences and discourse: Evidence from the N400*.
van Herten, M., Kolk, H. H., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Brain Research. Cognitive Brain Research, 22*, 241–255.
Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*, 176–190.
Wason, P. C. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior, 4*, 7–11.
Welch, B. L. (1947). The generalization of students problem when several different population variances are involved. *Biometrika, 34*, 28–35.
Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. New York: Springer.
Wijnen, F., & Kaan, E. (2006). Dynamics of semantic processing: The interpretation of bare quantifiers. *Language and Cognitive Processes, 21*, 684–720.