Investigating scalar implicatures in a truth-value judgement task: evidence from event-related brain potentials

Maria Spychalska, Jarmo Kontinen and Markus Werning
Faculty of Philosophy and Educational Research, Institute of Philosophy II, Ruhr University Bochum, Bochum, Germany

It is considered underinformative to say *Some As are B* when it is known that all As are B. Such underinformative sentences receive divergent truth-value judgements: whereas so-called *logical responders* evaluate them to be true, *pragmatic responders* reject them as false. In a sentence-picture verification experiment, we found that the split in the behavioural responses correlates with the difference in the event-related potentials (ERP) signal (N400 and P600) recorded for underinformative and for unambiguously true or false sentences with *some*. However, the ERP patterns for sentences with *all* are similar for both groups. In contrast to previous findings, the effect is independent of the subjects’ autistic spectrum quotients. Assuming that the N400 amplitude is inversely correlated with the expected probability of the critical word we argue that the observed between-group difference in the ERP pattern can be explained by the hypothesis that ‘logicians’ and ‘pragmatists’ use distinct verification strategies in evaluating sentences with the quantifier *some*.

Introduction

Implicatures are contents that are *suggested* in utterances, but they are neither directly expressed nor strictly entailed by them. They result from *pragmatic* inferences, which unlike logical inferences, are based not only on the literal meaning of a given linguistic expression, but also on certain tacit principles assumed to govern any cooperative communication. One of such principles, first formulated by Grice (1975), is known as the *Maxim of Quantity*. It states that a speaker should contribute to a conversation by providing an appropriate amount of information. For instance, from statement (1) one can infer that not all of A’s students have passed the exam; otherwise A’s utterance would be considered underinformativ. 

(1) A: *Some of my students have passed the exam.*

This implicature has been described as *scalar implicature*, since in its analysis one assumes that *some* is a semantically weaker term than *all* on the same linguistic scale (Horn, 1972; Levinson, 1983). Semantically, the truth of *Some As are B* is compatible with the truth of *All As are B*. This standard meaning of *some*, informally paraphrased as *some and possibly all*, is usually referred to as its *logical or semantic meaning*. Yet, if the speaker uses the weak scalar quantifier *some*, as in the above example (1), the addressee is in a position to infer that a sentence of similar content with the stronger scalar quantifier *all* is false. This inference is based on the reasoning that the speaker, who is assumed to be obeying the *Maxim of Quantity*, is expected to utter a more informative statement with the quantifier *all* if she knows that such a statement is true. The meaning of *some*, when enriched with the implicature *not all*, is often referred to as its *pragmatic meaning*. However, the semantic meaning and the pragmatic one are considered unequal in their status. Implicatures, unlike semantic meanings are defeasible, i.e. they can be canceled without leading to self-contradictory sentences (Levinson, 1983). Compare the intuitively correct sentence (2(a)), in which the implicature of the first clause was canceled by the second clause, with (2(b)), which is self-contradictory.

(2) (a) *Some students passed the exam, in fact all of them passed.*
(b) *Some students passed the exam, in fact none of them passed.*

Although it is generally accepted that the phenomenon of scalar implicatures cannot be explained without reference to some pragmatic principles, there is still much controversy about the precise division of labour between semantics and pragmatics. One particularly often addressed question regarding scalar
Implicatures are the default vs. context-based controversy. According to the default view, the implicature not all is triggered by the lexical item some by default, i.e. locally and more or less automatically, although it may be canceled in special circumstances (Chierchia, 2004; Chierchia, Fox, & Spector, 2012; Horn, 1984; Levinson, 1983, 2000). The proponents of the other approach (Bott & Noveck, 2004; Breheny, Katsos, & Williams, 2006) postulate that scalar implicatures result from complex and global reasoning processes that are based on context or assumptions regarding the speaker’s intentions (Carston, 1998; Sperber and Wilson, 1986).

Somewhat orthogonal to this debate, the emphasis in the philosophy of language concerning the semantics-pragmatics distinction has recently shifted towards the role pragmatic processes play in establishing the intuitive truth-conditions of a sentence. Minimal semanticists (Borg, 2007, 2012; Cappelen and Lepore, 2005) argue that the truth-conditions of a sentence are determined in a solely compositional way (Janssen, 1996; Werning, 2005; Werning, Hinzen, & Machery, 2012), i.e. the semantic value of a sentence is a function only of the stable semantic values (literal meanings) of the sentence’s constituents and the way they are syntactically combined. This picture has recently been challenged. For instance, Recanati’s truth-conditional pragmatics (Recanati, 2010) allows a contribution of implicatures into the intuitive truth-conditions of a sentence, therefore questioning the classical semantics-pragmatics divide. Arguably, the issue whether implicatures contribute to the truth-conditional content is independent of their alleged default or non-default character. The linguistic literature provides approaches which in various ways reconcile the poles of these two debates (Chierchia, 2004; Chierchia et al., 2012; King and Stanley, 2005; Levinson, 2000). Thus, understanding the phenomenon of scalar implicatures requires not only providing an account of how they are generated, but also explaining their role in constituting sentence intuitive truth-conditions.

**Experimental investigation of scalar implicatures**

Various experimental studies have been conducted to shed light on the theoretical debate concerning the nature of scalar implicatures, even though none of the discussed theories makes an explicit experimental predictions. Most empirical work has focused on the default- vs. context-based controversy and the emphasis has been put on the question of whether implicatures are cognitively costly compared to the semantic meaning. The processing costs of implicatures have been then considered to be a proxy of their default vs. non-default character.

It is already a well-established result that people have divergent intuitions regarding the truth of so-called underinformative sentences with some. These are sentences such as Some people have lungs, which are semantically true but pragmatically infelicitous. When people are asked to evaluate the truth-value of such sentences they are more or less evenly divided into so-called “pragmatic” responders, i.e. those who reject pragmatically infelicitous sentences as false, and “logical” responders, i.e. those who accept such sentences as true. Along with this finding it has also been shown that pragmatic responders usually take more time than logical responders to evaluate underinformative sentences (Bott, Bailey, & Grodner, 2012; Bott and Noveck, 2004). The higher processing cost of the pragmatic interpretation of some compared to its semantic interpretation is usually interpreted as strong evidence that scalar implicatures cannot result from any automatic processes and hence are not default inferences.

However, the empirical data are far from being conclusive and a number of experiments on the time-course of the scalar implicature processing led to contradictory results. For instance, in their eye-tracking study, Huang and Snedeker (2009) demonstrated that adult subjects showed a preference for a target that was consistent with the scalar implicature prior to the phrasal completion, which indicates that scalar inferences can already occur as the utterance unfolds. Nevertheless, consistently with the reaction time results, subjects’ looks to the target were substantially delayed when the implicature was generated relative to trials with semantically unambiguous quantifiers. Thus, there seemed to be a lag between semantic processing and calculating the implicature. Yet, in a different eye-tracking study, Grodner, Klein, Carbary, and Tanenhaus (2010) obtained contrastive results. In their experiment, subjects’ reactions for some were as fast as for unambiguous quantifiers such as all or none. Accordingly, the authors argued that the scalar inference was computed immediately and was not delayed relative to the literal interpretation of some. An interesting contribution to this debate was provided by Tomlinson, Bailey, and Bott (2013). In a mouse-tracking experiment they found that when participants made pragmatic interpretations of some, their mouse movements first deviated towards the logical response option before targeting at the pragmatic response option. However, when participants made logical interpretations their mouse movements went directly towards the target. These results support the hypothesis that participants interpret the upper
bounded pragmatic meaning in two steps but the lower bounded logical meaning in a single step.

A good method to investigate the time-course of linguistic processing is electroencephalography (EEG). It is a technique of a high temporal resolution that allows investigating language comprehension by measuring so-called event-related potentials (ERPs), i.e. direct brain responses time-locked to specific events, e.g. visual or auditory stimuli. Whereas an ERP component is a scalp-recorded voltage change that is considered to reflect a specific neural process, an ERP effect is an amplitude difference between the values of a given ERP component observed in two compared conditions.

Two ERP components that have received special attention over the last few decades in language-oriented research are of particular relevance for our study: the N400 and P600. The N400 is a negative-going shift in an ERP waveform, maximal over the centro-parietal scalp sites, of a latency ranging between 200 and 600 ms, and peaking roughly 400 ms post-stimulus onset (Kutas & Federmeier, 2011; Swaab, Ledoux, Camblin, & Bouwewyn, 2012). It is, in principle, elicited by every content stimulus, however, its amplitude depends on the “semantic” expectancy of this stimulus. This expectancy is modulated by various factors, such as well-formedness of a sentence given inherent language rules of semantic composition, semantic fit of the stimulus to the background context, or its plausibility given our world-knowledge (Kutas & Federmeier, 2000; Kutas & Hillyard, 1980; Nieuwland & Van Berkum, 2006; Van Berkum, Hagoort, & Brown, 1999). Upon its discovery (Kutas and Hillyard, 1980), the N400 effect has been labelled to be a signature of widely concei ded “semantic incongruence” in language (Kutas & Federmeier, 2000). One of the classical and most often cited examples where the N400 effect was found is a study in which ERPs triggered by sentence-final semantically expected verbs were compared with ERPs triggered by less expected verbs, or unrelated verbs in sentences such as The pizza was too hot to eat/drink/cry (Kutas & Van Petten, 1994). Although the precise functional role of the N400 is still debated, it is thought to indicate semantic retrieval processes that are sensitive to the degree of lexical predication and facilitation based on the context.

The P600 is a slow late positive shift in an ERP waveform with an onset around 500 ms, reaching its maximum around 600 ms post-stimulus onset. It lasts several hundred milliseconds and appears mostly on the posterior sites, although an anterior distribution of the P600 has also been observed (Swaab et al., 2012). Traditionally, the P600 has been related to syntactic errors (Hagoort, Brown, & Groothuizen, 1993; Osterhout & Holcomb, 1992); however, recently it has also been observed for semantic anomalies in the context of sentences containing no syntactic problems (Kim & Osterhout, 2005; Kuperberg, Caplan, Sitnikova, Eddy, & Holcomb, 2006; Kuperberg, Sitnikova, Caplan, & Holcomb, 2003; Van Herten, Kolk, & Chwilla, 2005). Thus, the functional interpretation of the P600 does not seem to be straightforward and it has been suggested that the P600 may reflect more general or combinatorial repair processes that are aimed at resolving inconsistency in the input (Hagoort, 2003; Kolk and Chwilla, 2007; Kuperberg, 2007; Van Herten et al., 2005).

The first EEG studies on scalar inferences (Nieuwland, Ditman, & Kuperberg, 2010; Noveck and Posada, 2003) focused on comparing the modulation of N400 ERPs by sentence-final words in underinformative (pragmatically infelicitous) sentences with some, e.g. Some people have lungs, and in informative sentences (pragmatically felicitous), e.g. Some people have pets. The truth-value of such sentences can be evaluated by reference to a person’s world-knowledge: It is a sort of encyclopedic fact that all people have lungs, and we know from our experience that only some people have pets. Violations of real-world knowledge have consistently been associated with larger N400 responses (Hagoort, Hald, Bastiaansen, & Peterson, 2004; Hald, Steenbeek-Planting, & Hagoort, 2007; Nieuwland, & Kuperberg, 2008). Therefore, if the interpretation of some includes its scalar implicature, critical predicates in infelicitous sentences violate our world-knowledge-based expectations and should trigger larger N400 ERPs than critical predicates in felicitous sentences. An occurrence of the N400 effect in this case could then be interpreted as evidence of an early incremental integration of the implicature into the sentence meaning. However, the N400 is also modulated by the lexico-semantic relationship between the critical word and the main noun phrase in the sentence. The lexico-semantic relationship between two words is usually measured with the frequency of their co-occurrence in contexts, that is known as the latent semantic analysis (LSA) value (Landauer, Foltz, & Laham, 1998). The modulation of the N400 by the LSA value is systematic: The higher the LSA value, the smaller the N400 (Kutas, Van Petten, & Kluender, 2006; Van Petten, 1993). Underinformative sentences, unlike informative ones, often refer to our encyclopedic knowledge. Thus, it is expected that the lexico-semantic relationship of critical predicates to the main noun phrases should be stronger in the former type of sentences compared to the latter one. Therefore, although due to the scalar inference one can expect a larger N400 amplitude for critical predicates in underinformative relative to informative sentences, this effect might be balanced out by a relatively stronger lexico-semantic relationship between words in the
underinformative sentences, as was shown by Noveck and Posada (2003).

In their pioneering experiment, Noveck and Posada (2003) observed that critical predicates in underinformative sentences elicited flatter N400 ERPs than critical words in informative sentences or in false sentences (e.g. Some crows have radios). Moreover, the N400 ERPs for the underinformative sentences were not modulated by the subjects’ truth-value judgements that were recorded together with the EEG data. The authors interpreted these results as indicating that the scalar implicature is a result of a post-semantic decision process that is related to the truth-value evaluation, whereas the initial stage of semantic processing is linked only to the lexico-semantic relationship.

These conclusions were challenged by Nieuwland et al. (2010). By systematically controlling the LSA values within the tested sentences as well as for the personality traits of their subjects, they were able to demonstrate that people differ in the way they process the implicature, depending on their score in the Autism Spectrum Quotient Questionnaire (AQ) (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001). The AQ test is a self-assessment questionnaire measuring traits of autistic spectrum disorder (ASD) in healthy adults with normal IQ. It consists of five subscales in which the following personal traits are tested: social skills, communication skills, imagination, attention to detail and attention switching. Although it is not a diagnostic tool, a score above 32 points in the AQ test has been shown to correlate with the likelihood of being diagnosed with autism. It is particularly noteworthy that people with ASD, especially high-functioning adults with autistic disorder high-functioning autistic (HFA), seem to have problems with various linguistic skills including the processing of linguistic information in contexts (Pijnacker, Geurts, van Lambalgen, Buitelaar, & Hagoort, 2010), defeasible reasoning (Pijnacker, Geurts et al., 2009), and pragmatic inferencing (Pijnacker, Hagoort, Buitelaar, Teunisse, & Geurts, 2009). Although some results suggest that a significant impairment of pragmatic skills occurs only in a clinical group of patients (HFA) but not in all individuals who show only some autistic traits (Pijnacker, Hagoort et al., 2009), it is still a reasonable assumption that people’s general pragmatic abilities and possibly their processing of pragmatic information could be predicted based on their AQ scores. In the experiment by Nieuwland et al. (2010) participants with AQ scores lower than the group median had larger N400 ERPs when presented with underinformative rather than with informative statements, independently of the lexico-semantic relationship between the words in those sentences (LSA values). In contrast, subjects with AQ scores higher than the group median had larger N400 ERPs when presented with informative rather than with underinformative sentences, and these ERPs were correlated with higher LSA values.

A different approach to investigating implicature processing was chosen by Politzer-Ahles, Fiorentino, Jiang, and Zhou (2012). The authors applied a paradigm in which sentences with some and all were used to describe picture-scenarios, e.g. pictures in which all agents were engaged in the same activity (all-scenarios), or pictures in which only some of the agents were engaged in one activity and the rest in another activity (some-scenarios). They showed that those quantifiers that were used in a pragmatically inconsistent way (some in the case of all-scenarios), but not in a semantically inconsistent way (e.g. all in the case of some-scenarios), were associated with a sustained negativity effect, i.e. a prolonged negativity starting ca. 300 ms post-stimulus onset, that had earlier been observed in response to ambiguous words (Van Berkum, Koornneef, Otten, & Nieuwland, 2007). The authors proposed that in their study this sustained negativity effect reflected the process of implicature cancellation and the retrieval of the semantic meaning. However, it should be noted that in their experiment the ERPs were measured on the onset of the quantifier, i.e. at the beginning of the sentence whose semantic status was not yet decided. Taking into account the fact that the authors did not observe a significant N400 effect for semantically inconsistent quantifiers, one can hypothesize that the observed sustained negativity reflected an attempt to assign a proper reference to the ambiguously used quantifier phrase rather than the violation of the scalar implicature.

It is, however, noteworthy that the P600 effect in a combination with the left anterior negativity (LAN) effect (early negativity over anterior electrodes lateralised to the left side and typically related to morphosyntactic violations (Swaab et al., 2012) has been also recorded in response to a pragmatic violation of a similar sort, namely a violation of the exclusive reading of the contrastive or (Chevallier, Bonnefond, Van der Henst, & Noveck, 2010). Disjunctive sentences, such as A or B, are considered to give rise to the scalar implicature not both A and B, especially if the connective or is stressed. Chevallier et al. (2010) showed that the presence of a prosodic cue on or led to an increase in the P600 amplitude as well as to the LAN effect whenever the exclusive (i.e. consistent with the scalar implicature) interpretation of the connective or was applied. The authors interpreted the observed P600 effect as reflecting increased processing efforts and argued in favour of the context-based approaches to scalar implicatures.
The current study

It is remarkable that, given such a rich empirical literature, there is no definite view on the nature of scalar implicatures. Although relatively good evidence has been provided regarding the default vs. context-based controversy, where perhaps the results have more convincingly been used against the strong default view, little attention has so far been devoted to the role scalar implicatures play in constituting intuitive truth-conditions. Our current study aims at tackling this question. We investigate how intuitive truth-value judgements modulate the N400 evoked by sentence-final predicates in underinformative sentences, independently of the lexico-semantic constraints in the tested sentences or the subjects’ autistic spectrum quotients. In order to dissociate the process of calculating the implicature from the process of sentence evaluation that is based on world-knowledge, we use a sentence-picture verification paradigm and record ERPs elicited by pragmatic violations that are based on short-term memory.

It had previously been demonstrated that the propositional truth-value influences sentence processing, which results in larger N400 ERPs for false compared to true sentences (Nieuwland & Kuperberg, 2008; Nieuwland & Martin, 2012). Thus, one can ask to what extent the intuitive truth-value judgements influence the ERP components elicited by pragmatically infelicitous sentences. Nieuwland et al. (2010) did not provide any differential analysis of this type, whereas in the study by Noveck and Posada (2003) no modulation of the ERPs by the truth-value judgements was found. It is sometimes claimed that truth-value judgments are not appropriate in the context of pragmatic infelicity, since they are assumed to reflect semantic evaluation, whereas pragmatically infelicitous statements cannot be judged in semantic terms. However, this claim presupposes a classical stand in the dispute regarding the role pragmatic processes play in determining the truth-conditional content. Thus, if we aim to tackle this question, we need to investigate how the intuitive truth-value evaluation modulates the processing of underinformative sentences. We can then compare two levels: the neural level, where pragmatic infelicity can be detected by means of an ERP signature and the behavioural level, where the acceptability judgements given by the subjects indicate their intuitive truth-value evaluation. Therefore, the main research question of our study is whether these two levels correlate.

In a recent study, Hunt, Politzer-Ahles, Gibson, Minai, & Fiorentino (2013) also performed an ERP experiment focusing on similar questions. They presented scenarios in which an action was performed by an agent on the whole set of given objects or just on a subset, and asked their subjects whether sentences with some match those scenarios. Their results indicated a trend where the N400 effect was dependent on the participants’ evaluation of underinformative sentences: The comparison between the ERPs elicited in the underinformative and the true condition resulted in a marginally significant N400 effect only for the so-called pragmatic responders, but not for the semantic responders. It is surprising that this interesting study does not provide any between-group comparison of the ERPs in the critical underinformative condition, which could shed more light on the underlying verification strategies. Moreover, due to the lack of controlling for the AQ scores of the participants it is not possible to compare this study with the results of Nieuwland et al. (2010). For instance, it cannot be excluded that the pragmatic responders had significantly larger AQ scores and thereby obtained the N400 effect. It seems also of high relevance to contrast sentences containing the weak scalar quantifier some with control sentences containing the strong quantifier all. First, independently of whether some occurs in a felicitous or infelicitous sentence, as a weak and underinformative scalar term, it might generally trigger processes different from those triggered by an unambiguous quantifier such as all. Second, having experimental conditions with all would allow controlling for additional factors. For instance, unlike true or infelicitous sentences, false sentences with some are existentially void: Some As are B is false if there are no As that are B. In contrast, All As are B is false in two cases: when there are no As that are B and when there are As that are B, but not all As are B. Thus, to have a better understanding of the processing of the not all implicature one should contrast not only infelicitous some-sentences with true or false some-sentences, but also infelicitous some-sentences with false, but existentially not vacuous all-sentences.

In our present study we controlled for all those factors. We tested sentences with both the weak scalar quantifier some, which is considered to give rise to the scalar implicature, and the strong quantifier all, which does not trigger any scalar inferences. Truth-value judgements were gathered along with the EEG data in order to test whether the semantic evaluation of the underinformative sentences correlates with the elicited ERP components. Unlike Hunt et al. (2013) we did not ask whether the sentences matched the scenarios, but we explicitly asked for intuitive truth-value judgements. We also analysed how these truth-value judgements modulated the verification strategies and shaped the ERPs evoked in both the underinformative and unambiguous conditions. Additionally, we tested our
participants with three parts of the Wechsler Intelligence Test (WAIS-IV) to measure their non-verbal abstract reasoning skills (matrix reasoning test), linguistic skills (the vocabulary test), and working memory (digit span memory test). Our motivation was the following: Context-driven approaches assume that generating scalar implicatures involves complex inferential processes. Such processes might be facilitated for individuals with better working memory, reasoning or linguistic abilities. Therefore, it is possible that individual differences in these factors would account for the differences in the implicature processing. Based on the results by Nieuwland et al. (2010) all our subjects were also screened with the Autism Spectrum Quotient Questionnaire in order to examine whether the modulation of the N400 is primarily triggered by the differences in their AQ scores or by the differences in their truth-value evaluation of the underinformative sentences.

Method

Design

All target sentences that were used in the experiment were of form (3), where X denotes the critical noun.

(3) Einige / Alle Bilder enthalten Xs.
Some / All pictures contain Xs.

The sentences were evaluated with respect to visual scenarios, consisting of five pictures. Subjects were first presented with a quantifier phrase and the verb, i.e. some/all pictures contain. Afterwards a scenario was presented and finally a critical noun X. The critical noun determined whether the sentence was true and pragmatically felicitous with respect to the pictures. The reason why the presentation of the sentence was interrupted by the presentation of the scenario was to allow the subjects to create expectations regarding the critical noun while they were inspecting the pictures. In each scenario two different categories of objects were presented: one occurring in all of the pictures, the other occurring in only two or three pictures. There were three evaluation conditions for each of the two quantifiers. For the quantifier some (S-conditions) these were (i) true and felicitous, in short Some-True (ST); (ii) true and infelicitous, in short Some-Infelicitous (SI); and (iii) a false condition Some-False (SF). For the quantifier all (A-conditions) there was (i) one true condition, i.e. All-True (AT); and two false conditions: (ii) All-False-Primed (AFP), where the critical noun denoted one of the object categories presented in the pictures; and (iii) All-False-Non-primed (AFN), where the critical noun denoted an object category that was not displayed in the pictures. The conditions Some-Infelicitous and All-True corresponded to the case where X denoted the object category that was contained in each of the pictures, Some-True and All-False-Primed corresponded to the case where X denoted the object category that was contained in only a subset of the pictures, and finally Some-False and All-False-Non-primed corresponded to the case where X denoted an object category that was not displayed in any of the pictures.

The detailed structure of the experimental trial is illustrated in Figure 1, whereas Table 1 presents the evaluation conditions for each of the critical words in the provided example. The subjects were asked to give truth-value judgements at the end of each trial, after the critical word disappeared. “Yes” and “No” appeared on the screen in a pseudo-random manner to indicate the meaning of the buttons and there was 4000 ms to respond. The ERPs were measured at the onset of the critical word, which was presented for a relatively long time of 1300 ms in order to record ERPs (in both the N400 and P600 time-windows) being undistorted by the processes associated with selecting and pressing the response button. Based on the literature (Bott & Noveck, 2004), we expected that our subjects would be divided into those who evaluate sentences with some in condition Some-Infelicitous as true (“logical” response) and those who evaluate them as false (“pragmatic” response).

If the scalar implicate is generated by default and integrated into the compositional process of building-up sentence meaning, then it should significantly modulate subjects’ expectations regarding the upcoming noun in the trials with the quantifier some. Consequently, the N400 ERPs elicited by critical nouns in the Some-Infelicitous condition should be larger than in the Some-True condition, but smaller than in the Some-False condition. However, based on the results by Chevallier et al. (2010), we also considered a possibility that comparing pragmatically infelicitous with true sentences might result in a late positivity (P600) effect rather than the N400 effect, which could be interpreted in favour of a post-semantic character of the scalar implicature. Furthermore, we aimed to test whether the sizes of these ERP effects depend on the participants’ truth-value evaluation of the infelicitous sentences: The pragmatic evaluation should result in larger N400/P600 ERPs for condition Some-Infelicitous relative to Some-True, whereas the logical evaluation should result in larger N400/P600 ERPs for condition Some-False relative to Some-Infelicitous. Such a correlation would support a truth-conditional character of the scalar implicate.
Whether these effects would be additionally modulated by the subjects’ AQ scores was our additional research question.

Let us also pay attention to predictions associated with the differences in priming of critical words across conditions. In condition Some-False critical words denoted objects that were not depicted, whereas in conditions Some-Infelicitous and Some-True critical words denoted depicted objects. Since priming generally decreases the N400 amplitude (Kutas & Federmeier, 2000), any difference in the N400 EPRs between conditions Some-False and Some-Infelicitous, or Some-False and Some-True, that potentially would be triggered by the difference in the semantic status of the compared sentences, would be additionally increased by the difference in priming of the critical words. This effect could be later controlled by comparing the two false conditions for the quantifier all i.e. All-False-Primed and All-False-Non-primed, which differed only with respect to priming, but not with respect to the semantic value of the compared sentences.

**Materials**

For the preparation of stimuli, we constructed a list of 80 ordered triples of nouns \( N_1, N_2, N_3 \), which gave us three sets of unique words (240 words in total). All nouns denoted concrete objects, that are easy to identify in a picture and are well known to an average German speaker. For each triple we created two individual pictures showing \( N_1 \) and \( N_2 \) as single objects, and one picture containing both an image of \( N_1 \) and an image of \( N_2 \). Pictures were either searched using free clipart platform or edited with Adobe Photoshop. The database of nouns and pictures was then used to create a unique stimuli list for each participant in a pseudo-random manner.

In order to control the lexical factors that are known to modulate the ERPs, we prepared the stimuli in the following way: all critical words were used in their plural form, were two-syllabic and had a length of 4–9 characters; compound nouns were excluded. The logarithmic word frequency value was checked in the Wortschatz Leipzig corpus (http://wortschatz.uni-leipzig.de/, 2011), and was kept between 8 and 17 (moderate frequent words). In order to keep the three sets of words comparable, for each triple the words were matched with respect to their length (maximal character difference was 4) and log-frequency (maximal value difference was 2). In the end, there were no significant differences between the three sets \( N_1, N_2, N_3 \) with respect to the words’ log-frequencies (One-way ANOVA \( F(2, 237) = .359, p = .7, \) all pairwise comparisons \( p = 1.0 \), or lengths \( F(2, 237) = .651, p = .522, \) all pairwise comparisons: \( p > .9 \). The words’ mean log-frequencies and mean lengths with standard deviations for each set are reported in Table 2.

Furthermore, the assignment of nouns to conditions was done in a pseudo-random way. Half of the triples were randomly associated with the quantifier some and
half with the quantifier all. Then, the assignment of \( N_1 \) or \( N_2 \) to either the first or to the second condition of the given quantifier was also randomly determined, whereas the remaining word from the triple (\( N_3 \)) was used in the third condition. The reason for having a separate list of words for SF and AFN conditions was practical: In these conditions the critical word is not depicted. To have a perfect match between the word and its visual representation and avoid any confusion caused by potential difficulties in recognizing objects as well as to facilitate the process of stimuli preparation, we selected for visualization only those words that allowed for the best clipart-form representation. For instance, whereas the word Seifen (soaps) is as frequent as Gabeln (forks) or Kannen (jugs) in German, it was much easier to find a good clipart-style picture of a fork and of a jug than of a soap.

The visual scenarios were generated from the triples as 5-element pseudo-random combinations of pictures, in such a way that in each trial exactly two object categories were shown, namely corresponding to nouns \( N_1 \) and \( N_2 \): one of the shown object categories occurred in each of the pictures, the other in three or two of the pictures. The same pseudo-random combination of pictures was used for each of the three conditions for the quantifier to which it was assigned, but each time the pictures were randomly sorted when presented on the screen. Depending on the evaluation condition a different noun (\( N_1 \), \( N_2 \) or \( N_3 \)) was displayed at the end of the trial. This means that each combination of pictures was seen three times, but each critical word only once. Since the last noun, \( N_3 \), served as a critical word in conditions Some-False or All-False-Non-primed, it was never depicted. In sum, for each participant each of the conditions ST, SI, AT, AFP used a different, randomly selected, subset of nouns from the set of \( N_1 \) plus \( N_2 \), whereas each of the conditions AFN and SF used a randomly selected subset of the words from the list \( N_3 \). The proportion of the occurrences of the two displayed object categories in the whole scenario was either 5/3 or 5/2, which was balanced evenly per condition. It was also balanced out across conditions but kept consistent within a trial whether the critical object was displayed on the right-hand side or the left-hand side of the pictures. Thus, in each condition in half of the trials the critical noun denoted the object category that was shown on the left-hand side of all five pictures, and in half of the trials the critical noun denoted the object category that was shown on the right-hand side.

There were 40 trials per condition, which yields a total of 240 experimental trials, plus 60 filler trials, with quantifiers: no (keine), most (die meisten), two (zwei), three (drei), four (vier), five (fünf). The filler trials varied between participants only with respect to the positioning of particular pictures on the screen. The nouns used in the filler trials were different from those used for the test trials and were less strictly selected with respect to their length or frequency.²

### Participants

Fifty-seven (twenty-nine women) members of the Ruhr-University Bochum were recruited for the experiment (age: 18–44, mean: 24.2, SD: 4.4). They were reimbursed for their participation. All participants spoke German as their only mother tongue, had at least a secondary degree (German Abitur), normal or corrected to normal vision, no history of psychological or neurological problems, and were right-handed. Three people were excluded from the analysis: either due to technical problems that occurred during the recording (one participant), or because of their lack of attention during the experiment, i.e. one participant had over 50% of missed responses and one had over 50% of erroneous responses in one of the control conditions.

### Procedure

Upon arrival all our participants signed a written consent of participation including a statement concerning their vision, medication, neurological or psychiatric history. They filled in the Edinburgh Handedness Inventory test and were screened using the three parts of WAIS and the AQ Questionnaire. The measurement was conducted in a dim, electrically and acoustically isolated cabin. Subjects were seated in front of a computer screen and a keyboard with two buttons. For presenting the stimuli we used the Presentation® Software. The experiment started with a short instruction followed by an exercise session consisting of five example trials. No feedback was given throughout the experiment and subjects were asked to follow their intuition in the truth-value judgement task.

The experiment was divided into five blocks of 60 trials with optional breaks in between. The net measurement time (excluding breaks) was on average 46 minutes. The presentation times for the sentences and the scenarios were based on the standard reading vs.

| sub-list | Mean log-freq/SD | Mean length/SD |
|----------|------------------|----------------|
| \( N_1 \) | 13.13/1.84 | 5.96/1.17 |
| \( N_2 \) | 13.25/1.73 | 6.12/1.12 |
| \( N_3 \) | 13.36/1.75 | 5.95/0.94 |

---

² The occurrence of erroneous responses was assessed for each participant and each condition, that is, it was counted whether any of the 60 trials was missed or had over 50% misses. Only if all responses were correct, both for the test trials and for the filler trials, the participant was included in the analysis. If there were less than 50% correct responses or if erroneous responses were found, they were excluded from the analysis.
object recognition/counting times as well as on few pilot sessions.

**EEG recording and data processing**

EEG was recorded from 64 active electrodes held on the scalp by an elastic cap, with a BrainAmp acticap EEG recording system. AFz served as the ground electrode and FCz as the physical reference. Four electrodes (FT9, FT10, P09, PO10) were reprogrammed and used for controlling both vertical (above and below the right eye) and horizontal (on the right and left temple) eye-movements (EOG electrodes). The EEG was recorded with a sampling rate of 500 Hz and a band-pass filter of 0.53 (the time constant of 0.3 s was used as the low cut-off) – 70 Hz. Impedance was kept below 5 kΩ for the EOG electrodes and below 10 kΩ for the EEG electrodes.

The EEG data were processed using Brain Vision Analyzer 2.0 software. We applied an off-line high cut-off filter at 40 Hz, 12 dB/oct. Automatic raw data inspection rejected all trials with the absolute amplitude difference over 200 μV/200 ms, or with the activity lower than 0.5 μV in intervals of at least 100 ms. The maximal voltage step allowed was 50 μV/ms. For seven participants we disabled 1–4 channels due to technical problems or excessive artefacts (i.e. Fp1, Fp2, AF7, AF8), and for one subject we had to disable FC2 due to a technical recording error. These channels were subsequently excluded from the statistical analysis. Eye blinks were corrected using an independent component analysis. The data were off-line re-referenced to the linked mastoids comprising of TP9 and TP10. Segments from 200 ms pre-target onset until 1000 ms post-onset were separately extracted and averaged for each subject and every condition (2 quantifiers × 3 evaluation conditions). Baseline correction used the 200 ms interval preceding the onset of the stimulus. All segments with any remaining physical artefacts (including the amplitude lower than −90 μV or higher than 90 μV) were excluded before averaging. The minimal number of segments that was preserved in each condition was 26 out of 40 (60%).

**Behavioural results**

**Pragmatic vs. logical response: group division**

The analysis of the truth-value judgements revealed that our subjects were generally consistent in their choice of either the pragmatic or the logical interpretation of the quantifier some. Accordingly, we divided them into two groups based on their responses in condition Some-Infelicitous. People who had at least 70% of pragmatic responses were called “pragmatists” (N = 26), whereas those who had at least 70% of logical responses were called “logicians” (N = 28). Applying the threshold of 70% resulted in an exhaustive division (Figure 2).

Accuracy in all control conditions, i.e. those unambiguously true or false, was at ceiling level. The mean accuracy (Table 3) varied between 94.64% (All-False-Primed, “logicians”) and 99.64% (All-False-Non-primed, “logicians”). In condition Some-Infelicitous accuracy was defined with respect to group, i.e. the pragmatic response was considered correct for a “pragmatist” and the logical response was considered correct for a “logician”.

There was no significant difference with respect to some-defined accuracy between “logicians” and “pragmatists” neither in condition SI (“pragmatists” median = 97.5%, “logicians” median = 97.5%, Mann–Whitney U = 282, z = −1.473, p = .141, r = −.02), nor in any other condition (Pairwise comparisons after Bonferroni correction are reported at 0.05/6 = .0083 level of significance; p > .03 for the between-group comparison in condition AFP, p > .6 for all remaining conditions). Non-parametric Levene’s tests (Nordstokke and Zumbo, 2010; Nordstokke, Zumbo, Cairns, & Saklofske, 2011) revealed no statistically significant between-group differences in variances in any of the conditions (p = .035 for condition ET, p > .1 for all remaining conditions).

**Reaction time for correct responses: a delay in calculating the implicature**

In order to investigate whether subjects’ mean reaction time for correct responses was dependent on the quantifier (all vs some) or the evaluation condition (ST/SI/SF and AT/AFP/AFN), we first computed the mean response times by taking into account only those trials in which subjects gave a correct response (Table 4 and Figure 3). Subsequently, we conducted a mixed Repeated Measures ANOVA with Quantifier and Truth (evaluation condition) as within-subject factors and Group as a between-subject factor. In condition Some-Infelicitous, the accuracy was defined with respect to group. To reduce skewness we performed a logarithmic transformation of the data, i.e. we computed natural logarithm of each dependent variable, and we performed the analysis on the transformed data.3 Greenhouse-Geisser was applied whenever the assumption of sphericity was not met and the p-values of all pairwise comparisons were Bonferroni corrected. The Levene’s tests for each of the repeated measures variables were not significant, thus the variances are assumed to be homogeneous for all the compared levels.

The full-factorial analysis (Table 5) proved the significance of both factors: Truth (p = .001) and
Quantiﬁer (p = .033), with the quantiﬁer all receiving on average faster correct responses than the quantiﬁer some (584.57 ms vs. 596.75 ms). There was a significant interaction between Quantiﬁer and Truth (p < .001) as well as the three-way interaction Quantiﬁer × Truth × Group (p = .012), but the Truth × Group and the Quantiﬁer × Group interactions were not signiﬁcant. There was also no signiﬁcant effect of Group.

Since the correspondence between the evaluation conditions for both quantiﬁers was only partial (whereas condition Some-True can be considered corresponding to All-True, and Some-False to All-False-Non-primed, Some-Infelicitous and All-False-Primed lacked their corresponding conditions), in order to investigate the effect of Truth as well as the Truth × Quantiﬁer × Group interaction, we separately analysed the response times for each of the quantiﬁers. For each quantiﬁer Truth was a signiﬁcant factor (p ≤ .001), whereas there were no between-group differences, and the Group × Truth interaction was signiﬁcant (p = .003) only for the quantiﬁer some (Table 6). Pairwise comparisons for the quantiﬁer all showed that correct responses were given signiﬁcantly faster in condition All-True than in All-False-Non-primed (p < .001), but not signiﬁcantly faster than in All-False-Primed (p = .152), whereas the two All-False conditions did not differ signiﬁcantly with respect to the mean response times (p = .078). In the case of the quantiﬁer some pairwise comparisons revealed that the participants gave signiﬁcantly slower responses in the Some-Infelicitous condition relative to the Some-True condition (p = .006), and relative to the Some-False condition (p = .008), but there was no signiﬁcant difference between conditions Some-True and Some-False (p = 1.0). The analysis of contrasts proved that this effect was mainly due to the fact that “pragmatists” responded slower in condition Some-Infelicitous: the difference in the mean reaction time between conditions Some-Infelicitous and Some-True as well as between Some-Infelicitous and Some-False was signiﬁcantly larger for “pragmatists” than for “logicians”: F(1.52) = 6.757, p = .012, η² = .115 for the contrast Some-True vs. Some-Infelicitous and F(1.52) = 10.267, p = .002, η² = .165 for the contrast Some-Infelicitous

### Table 3. The mean accuracies and the standard deviations M(SD), in percentages, for all evaluation conditions.

|        | some          | all           |
|--------|---------------|---------------|
|        | ST            | SI            | SF            |                       |
|        | LogResp       | PragmResp     |               |                        |
| Logicians |               |               |               |                        |
|         | 95.71 (6.30)  | 96.52 (4.68)  | 3.39 (4.73)   | 99.20 (1.67)          |
| Pragmatists | 97.21 (2.86)  | 6.54 (8.16)   | 93.46 (8.16)  | 99.42 (1.29)          |
| All subjects | 96.44 (4.97)  | 53.19 (45.85) | 46.76 (45.89) | 99.31 (1.49)          |
|         |               |               |               |                       |
|         | 98.48 (2.39)  | 94.64 (4.45)  | 99.64 (0.89)  |                        |
|         | 99.04 (2.13)  | 97.02 (3.74)  | 99.52 (1.00)  |                        |
|         | 98.75 (2.26)  | 95.79 (4.26)  | 99.58 (0.94)  |                        |

Note: In condition Some-Infelicitous: the average percentage of pragmatic responses vs. the average percentage of logical responses after excluding the missing responses.
Table 4. The mean response times (correct responses) and the standard deviations \(M(SD)\) for all evaluation conditions.

|                | Logicians |               |               | Pragmatists |               |               | All subjects |               |               |
|----------------|-----------|---------------|---------------|-------------|---------------|---------------|--------------|---------------|---------------|
|                | (ms)      | (ln)          | (ms)          | (ln)        | (ms)          | (ln)          | (ms)         | (ln)          | (ms)         |
| **Logicians**  | 591.23(118.78) | 594.69(107.81)| 598.48(127.58)| 566.34(95.73)| 582.89(110.71)| 593.40(91.83)|             |               |               |
| **Pragmatists**| 579.83(99.85) | 639.24(151.47)| 577.01(92.20) | 566.64(92.65)| 587.47(127.96)| 610.68(119.34)|             |               |               |
| **All subjects**| 585.74(109.19)| 616.14(131.33)| 588.14(111.44)| 566.48(93.35)| 585.09(118.20)| 601.72(105.31)|             |               |               |

Notes: The accuracy in condition \textit{Some-infelicitous} is defined with respect to the group. For ease of interpretation, we present the untransformed values in milliseconds. Below the means for the log-transformed values (natural logarithm).

vs. \textit{Some-False}. Figure 3 graphically presents the mean times of correct responses for each group and condition.

It had earlier been observed that evaluating false sentences usually takes a longer time than evaluating true sentences (Carpenter and Just, 1975; Clark and Chase, 1972). Therefore, one could argue that the delay in responding pragmatically in the \textit{Some-Infelicitous} condition was at least partially an artefact of giving a “no” response in the truth-value judgement task. However, there is enough evidence that this delay cannot exclusively be explained by this effect, since “pragmatists” responded in the \textit{Some-Infelicitous} condition not only slower than in the \textit{Some-True} condition, but also slower than in the \textit{Some-False} condition. Admittedly, the relatively fast responses in the \textit{Some-False} condition compared to \textit{Some-Infelicitous}, could partially be related to differences in short-term memory load between \textit{Some-False} and the other \textit{S}-conditions: In the \textit{Some-False} condition the critical noun referred to an object that was not depicted in the trial, which might have facilitated the rejection of the sentence. In contrast, in order to reject – under a pragmatic interpretation – a sentence in the \textit{Some-Infelicitous} condition, it was necessary to recollect the information of which of the shown object categories occurred in all and which only in some of the pictures. However, this difference in memory demands between conditions \textit{Some-False} and \textit{Some-Infelicitous} cannot be the only reason for the difference in the response times. Firstly, there was no significant difference in the subjects’ response times for the two \textit{All-False} conditions, although these conditions only differed with respect to whether or not the critical word referred to one of the depicted objects. Secondly, although conditions \textit{Some-Infelicitous} and \textit{All-False-Primed} did not strictly correspond to each other, they could be regarded as corresponding conditions in the case of “pragmatists”: In both conditions the critical noun denoted one of the depicted object categories, and in both conditions the sentences were evaluated by the pragmatic responders.

Figure 3. The mean reaction times for correct responses in each evaluation-condition and for each group. For ease of interpretation we plot the untransformed values in milliseconds.
to be false. A dependent t-test comparing subjects’ reaction times in these two conditions proved that “pragmatists” responses in condition All-False-Primed were significantly faster than their responses in condition Some-Infelicitous ($t(25) = 3.229$, $p = .003$), but there was no significant difference in the case of “logicians” ($t(27) = 1.146$, $p = .262$). To summarize, the above-reported data can overall be interpreted as evidence that calculating scalar implicatures came with a significant delay. This is an important result supporting some of the previously reported data that the pragmatic interpretation of the quantifier some involves a longer processing time relative to the semantic interpretation (Bott and Noveck, 2004; Bott et al., 2012).

### The choice between the pragmatic and the logical interpretation of some does not depend on the measured personality traits or cognitive abilities

The mean values in all measured cognitive or personality tests, i.e. working memory span, linguistic skills, logical intelligence, or AQ score, are reported in Tables 7 and 8. Consistent with the literature (Baron-Cohen et al., 2001), the AQ scores of our participants were dependent on their gender, i.e. men scored significantly higher than women (equal variances not assumed): in total AQ ($M = 17.11$, $SD = 8.07$ vs. $M = 12.64$, $SD = 3.71$, $t(34.56) = 2.583$, $p = .014$), AQ-communication subscale ($M = 2.35$, $SD = 2.02$ vs. $M = 1.46$, $SD = 0.92$, $t(34.431) = 2.039$, $p = .049$) and AQ-communication subscale ($M = 2.73$, $SD = 2.51$ vs. $M = 1.43$, $SD = 1.03$, $t(32.763) = 2.461$, $p = .019$). It is also worth to note that the participants’ working memory was positively correlated with their score in the matrix reasoning test ($r(54) = .442$, $p = .001$). Yet, there were no statistically significant differences between “pragmatists” and “logicians” in any of the measured variables, including also gender or age (all tests with $p > .14$). Thus, we cannot conclude that the choice of the pragmatic vs. semantic interpretation of the quantifier some was dependent on any of the measured cognitive or personality traits.

### EEG results

For the statistical analysis of the EEG data, we used the Matlab Fieldtrip package (Oostenveld, Fries, Maris, & Schoffelen, 2011). We performed a non-parametric statistical procedure called cluster-based permutation test (Maris and Oostenveld, 2007). For each subject the ERPs were averaged over trials in each condition, in epochs of 0–1000 ms post-onset and for each channel. Between two compared conditions the data-points of subjects’ averages (time × channel) were compared with a two-tailed dependent t-test. The significantly different ($\alpha = 0.025$) data-points were then clustered according to the time-spatial adjacency. For each observed cluster, the cluster-level statistics were calculated by taking the sum over the t-values for each cluster. Then, the cluster-level p-values were evaluated with a Monte Carlo simulation: For each subject, the ERP averages were randomly swapped between the two conditions. The cluster-level statistics were computed again and the maximum of the cluster-level statistics was taken as the test statistics for this permutation. This procedure was repeated 10,000 times and the p-values of each of the observed cluster-level statistics were estimated as the proportion of permutations that resulted in a higher test statistics than the observed one. This method – primarily designed to determine only whether the compared conditions differ significantly in the chosen epoch – with some

## Table 5. The repeated measures ANOVA of the reaction times (correct responses).

|                | F-stat | p-value | effect size |
|----------------|--------|---------|-------------|
| Quantifier     | $F(1, 52) = 4.791$ | $p = .033$ | $\eta^2 = .084$ |
| Truth          | $F(2, 104) = 7.325$ | $p = .001$ | $\eta^2 = .123$ |
| Quantifier × Truth | $F(1.528, 79.472) = 10.525$ | $p < .001$ | $\eta^2 = .168$ |
| Quantifier × Truth × Group | $F(1.528, 79.472) = 5.290$ | $p = .012$ | $\eta^2 = .092$ |
| Truth × Group  | $F(2, 104) = 2.233$ | $p = .112$ | $\eta^2 = .041$ |
| Quantifier × Group | $F(1, 52) = .043$ | $p = .836$ | $\eta^2 = .001$ |
| Group          | $F(1, 52) = .036$ | $p = .849$ | $\eta^2 = .001$ |

## Table 6. The repeated measures ANOVA results of the reaction times (correct responses) for each quantifier.

### All

|                | F-stat | p-Value | Effect size |
|----------------|--------|---------|-------------|
| Truth          | $F(2, 104) = 9.844$ | $p < .001$ | $\eta^2 = .159$ |
| Truth × Group  | $F(2, 104) = .401$ | $p = .671$ | $\eta^2 = .008$ |
| Group          | $F(1, 52) = .055$ | $p = .816$ | $\eta^2 = .001$ |

### Some

|                | F-stat | p-Value | Effect size |
|----------------|--------|---------|-------------|
| Truth          | $F(2, 104) = 7.508$ | $p = .001$ | $\eta^2 = .126$ |
| Truth × Group  | $F(2, 104) = 6.274$ | $p = .003$ | $\eta^2 = .108$ |
| Group          | $F(1, 52) = .021$ | $p = .885$ | $\eta^2 < .001$ |
reservations has also been used to identify the actual significant clusters of channels and time-points. (For a similar application of the permutation tests see, for instance, Baggio, van Lambalgen, & Hagoort, 2008; Baggio, 2012; Pijnacker et al., 2010). Using a similar approach we interpret the identified clusters as our effects, taking the first data point of the observed cluster to be the effect onset and the last point to be the effect offset.

We used the above-described procedure to pairwise compare averaged ERPs between the S- and A-conditions separately. Thus, for the quantifier some we compared Some-Infelicitous vs. Some-True, and Some-False vs. Some-Infelicitous, which were our critical comparisons, as well as Some-False vs. Some-True, which was the control baseline comparison. For the quantifier all we compared All-False-Primed vs. All-True, All-False-Non-primed vs. All-True and All-False-Non-primed vs. All-False-Primed. This procedure was performed for “pragmatists” and “logicians” separately as well as for all subjects jointly. Foreshadowing the presentation of our results: since the joint-group analysis reflects the averaged effect of the distinctive results of both groups, we do not discuss it here in detail but only report the clusters’ p-values and latencies in Table 9.

### The ERP effects depend on the truth-value judgement: the between-group differences

One of our main hypotheses was that the ERP effects related that the violation of the implicature should be modulated by the subjects’ truth-value evaluation of the underinformative sentences. Thus, we expected differential effects for “pragmatists” and “logicians”. Whereas “pragmatists” were expected to get larger N400 ERPs in the Some-Infelicitous compared to the Some-True condition, for “logicians” we expected a significantly smaller difference in this comparison. The opposite result was predicted for the Some-False vs. Some-Infelicitous comparison, where “logicians” were expected to get a significantly larger N400 effect than “pragmatists”. The visual inspection of the grand averages for both groups (Figure 4) suggested that our hypothesis could be sustained, and the analysis involving

### Table 7. The distribution of gender and the mean values with the standard deviations M(SD) for age and all cognitive tests.

|                | Age     | Gender | AQ   | Working memory | Matrix reasoning | Vocabulary |
|----------------|---------|--------|------|----------------|-----------------|------------|
| **Logicians**  | 24.61(5.01) | 14/14 | 15.07(7.56) | 12.32(2.52) | 11.29(2.32) | 13.29(2.21) |
| **Pragmatists**| 23.77(3.66) | 12/14 | 14.53(5.38) | 11.27(2.75) | 10.73(1.59) | 12.88(2.33) |
| **All subjects**| 24.2 (4.39) | 26/28 | 14.79(6.55) | 11.81(2.66) | 11.02(2.01) | 13.09(2.21) |

### Table 8. The mean values and the standard deviations M(SD) for all AQ-subscales.

|                | Communication | Social skills | Attention to detail | Attention switching | Imagination |
|----------------|---------------|---------------|---------------------|---------------------|-------------|
| **Logicians**  | 1.89(1.85)    | 1.86(2.73)    | 4.79(2.45)          | 4.50(1.95)          | 2.04(1.95)  |
| **Pragmatists**| 1.88(1.31)    | 1.4(2.21)     | 4.42(1.77)          | 4.69(1.67)          | 2.08(2.06)  |
| **All subjects**| 1.89(1.59)    | 1.65(2.48)    | 4.61(2.14)          | 4.50(1.8)           | 2.06(1.98)  |

In the table, we report all clusters with p < .05, whereas empty spaces mean that no clusters with p < .05 were found. Note that after applying Bonferroni correction for multiple comparisons all the reported clusters for “pragmatists” and “logicians” are still significant at the level of p < .05/12 = .0041 (correction for a given group for six comparisons, negative and positive clusters). [ ] marks a marginally (after correction) significant cluster.
Figure 4. Grand averages for all three evaluation-conditions for the quantifier *some* shown at nine centro-parietal electrode locations for (a) "pragmatists" and (b) "logicians".
cluster-based permutation tests that were performed separately for each of the two groups confirmed this hypothesis. In Table 9 we report all the significant negative and positive clusters found by the permutation tests for each of the groups as well as for the joint-group analysis.

The analysis revealed that for “pragmatists” the ERPs for critical words were significantly more negative in condition Some-Infelicitous than in condition Some-True (p < .0005) in the time-window of 264–436 ms post-onset, which overlaps with the time-window of the standard N400 effect. This effect had a global topographical extension and was followed by a significant, over 400 ms-long, wide-ranging positive cluster (482–894 ms, p < .0001), that was maximal at the parietal regions. This observed positivity effect could be identified as the “semantic P600” effect or the late positive component (LPC). Such late positivities have been often observed after N400 effects (Frenzel, Schlesewsky, & Bornkessel-Schlesewsky, 2011; Hagoort & Brown, 2000; Kutas & Hillyard, 1980; Pijnacker et al., 2010; Van Berkum et al., 1999), although their functional sensitivity is still unclear. Following Van Petten and Luka (2012), we adopt a theoretically neutral term Post-N400-Positivity (PNP) to refer to any positivity observed after the N400 effect. In contrast, no significant negative or positive effects were found in the comparison between Some-Infelicitous and Some-True for “logicians”, who evaluated sentences in both conditions as true.

A contrastive result was obtained when condition Some-False was compared with Some-Infelicitous: Whereas “logicians” got a significant N400 effect and a significant PNP effect in this comparison, for “pragmatists” only the N400 effect was significant. Since “pragmatists” gave the same truth-value judgements in both compared conditions, their N400 effect in this comparison could be associated with the priming difference between the compared conditions.

Comparing the baseline Some-True and Some-False conditions gave similar results for both groups: As expected, subjects’ ERPs were more negative in the N400 time-window for the unambiguously false compared to the unambiguously true sentences. This N400 effect was followed by a significant and posteriorly distributed positivity effect starting after 500 ms post-stimulus onset and lasting until over 800 ms post-onset.

The results of the comparisons between the A-conditions were also similar for both groups: Averaged ERPs in both All-False conditions were more negative than in the All-True condition in the time-windows corresponding to the N400 effect. These N400 effects were followed by significant Post-N400-Positivity effects. Comparing the two All-False conditions (All-False-Primed vs. All-False-Non-primed), which differed in the priming of the critical noun, also resulted in a significant N400 effect. This effect was followed by a significant late positivity effect (810–998 ms, p < .0035) only for “logicians”, but there was no significant positivity observed for “pragmatists”. For the comparison of grand averages for the A-conditions see Figure 5.

The sizes of the observed N400 and the PNP effects were then compared between the two groups with independent t-tests. A size of an effect was taken to be an average amplitude difference between the compared conditions in a fixed time-window. The time-windows for computing average sizes of the N400 and PNP effects were selected to be the time-windows of the corresponding significant clusters that were found in a given comparison by the permutation test performed for all subjects, i.e. both groups jointly (see Table 9). The averages were calculated over all the electrodes involved in the effects, i.e. all electrodes that had at least one data point in the significant cluster. This approach allowed us to directly analyse whether the effects observed in the permutation analysis for the whole tested population were significantly different for the two groups.

As expected, the effects in the comparison of conditions Some-Infelicitous and Some-True were significantly larger for “pragmatists” than for “logicians”: the N400 effect (260–436 ms; t(52) = −5.392, p < .001; equal variances assumed in all cases) as well as the PNP effect (500–624 ms, t(52) = 3.506, p = .001). However, the N400 effect in the comparison of conditions Some-False and Some-Infelicitous was significantly larger for “logicians” than for “pragmatists” (254–514 ms; t(52) = −4.333, p < .001), whereas the PNP effect in this comparison was marginally (after correcting for multiple comparisons) larger for “logicians” (544–964 ms; t(52) = 2.238, p = .03). There were no other significant between-group differences regarding the sizes of the N400 effects or PNP obtained in the control comparison Some-False vs. Some-True, as well as in any of the comparisons between the A-conditions (all the p-values ≥ .1; equal variances assumed).

The modulation of the N400 effects by the number of “verifiers” in the scenario

The N400 effect is defined as a relative difference between the test and the baseline condition. Therefore, we were primarily interested in the relative differences between conditions Some-True and Some-Infelicitous and between Some-False and Some-Infelicitous, focusing on the question whether these differences are dependent on the truth-value judgements given by the subjects in the Some-Infelicitous condition. This hypothesis was supported by our data. However, it is also a relevant
Figure 5. Grand averages for all three evaluation-conditions for the quantifier *all* shown at nine centro-parietal electrode locations for (a) "pragmatists" and (b) "logicians".
question how the participants’ truth-value evaluation of the sentences in the Some-Infelicitous condition modulated their ERPs triggered in this condition. A straightforward expectation is that “pragmatists” should have larger N400 ERPs for the Some-Infelicitous sentences compared to “logicians” and that the between-group difference in this condition should be the source of the N400 effects in the Some-Infelicitous vs. Some-True as well as the Some-False vs. Some-Infelicitous comparisons. Yet, a careful inspection of grand averages allows one to notice that the differences between “pragmatists” and “logicians” in these critical N400 effects seem to be due to the difference in the Some-True, respectively Some-False, rather than in the Some-Infelicitous condition. It seems counterintuitive that the two groups should differ in the Some-True or Some-False conditions, rather than in the critical test condition Some-Infelicitous. In order to check whether this observation can be statistically confirmed, we compared the two groups with respect to their mean amplitudes in each of the S-conditions, in the time-window of 250–450 ms and over the whole scalp. As the N400 effects in each comparison and for each group had varying latencies, this time-window as well as the global extension were selected based on the inspection of the grand averages and the results of the permutation tests in all comparisons. It turned out that “pragmatists” and “logicians” indeed differed in conditions Some-True (t(52) = −2.761, p = .008) and Some-False (t(52) = −2.666, p = .01), i.e. in both conditions “logicians” had more negative N400 ERPs than “pragmatists”, but there was no between-group difference in condition Some-Infelicitous (t(52) = −.028, p > .9).

This surprising result can be explained by referring to different strategies that need to be applied by both groups during the verification procedure. One should mention that the S-conditions posed different difficulties for “logicians” and for “pragmatists”. For “logicians” both object categories shown in the S-conditions were potential “verifiers” of the sentence, i.e. could be mentioned at the end of a trial to complete a true sentence, whereas for “pragmatists” only one of the shown object categories was a potential verifier. Thus, whereas the scenario-based expectations regarding the upcoming noun were unique for “pragmatists”, they were non-unique for “logicians”. The reverse is true with regard to potential “falsifiers” in the S-conditions. For “logicians” only a noun denoting neither of the shown object categories would complete a false sentence with the quantifier some. For “pragmatists”, in addition to that, also a noun denoting one of the shown object categories would complete a false some-sentence. These differences between “pragmatists” and “logicians” in the processing demands at the stage of inspecting the pictures might have led to the differences in the N400 ERPs evoked by critical nouns in all S-conditions. A brief comparison with the results for the quantifier all allows us to test this hypothesis. In the case of the A-conditions, always one of the shown object categories was a verifier, whereas the other one was a falsifier, and additionally any noun denoting neither of the shown objects categories would complete a false sentence. Thus, the demands in processing the pictures in the A-conditions by both groups were similar to the demands in processing the pictures in the S-conditions by “pragmatists” only.

In order to statistically explore the hypothesis that the unique vs. non-unique expectations regarding the upcoming word modulated the N400 amplitude, we performed additional comparisons. We compared conditions All-True with Some-True and All-True with Some-Infelicitous for “logicians” and “pragmatists” separately (250–450 ms, scalp average). In all these conditions the test sentences were judged by “logicians” to be true, but in conditions Some-True and Some-Infelicitous “logicians” had non-unique expectations regarding the upcoming noun, whereas in condition All-True they expected a unique noun. In contrast, “pragmatists” had unique expectations in both conditions Some-True and All-True. “Pragmatists” had also unique expectations in condition Some-Infelicitous; however, as these expectations were violated, they consequently judged the sentences to be false. Therefore, the Some-Infelicitous condition was for “pragmatists” similar to the All-False-Primed condition. If our hypothesis is true, “logicians” should have significantly larger N400 ERPs for condition Some-True and Some-Infelicitous relative to condition All-True, but for “pragmatists” no difference is expected between conditions Some-True and All-True. It is, however, expected that “pragmatists” would have larger N400 ERPs for condition Some-Infelicitous than for All-True; yet this expectation is already based on the truth-judgement difference between the compared conditions. In contrast, for “pragmatists” the N400 ERPs should not differ between conditions Some-Infelicitous and All-False-Primed.

Our predictions were borne out by the data. For “logicians” the N400 ERPs (250–450 ms, scalp average) were larger in condition Some-True than in All-True (t(27) = −4.023, p < .001) and larger in condition Some-Infelicitous than in All-True (t(27) = −3.671, p = .001). In contrast, for “pragmatists” the N400 ERPs did not differ significantly between conditions Some-True and All-True t(25) = 1.081 p = .29, whereas – as expected – the “pragmatists” N400 ERPs were larger in condition Some-Infelicitous than in All-True:
t(25) = −7.006, p < .001. Additionally, there was no significant difference in the mean N400 amplitude for “pragmatists” between conditions Some-Infelicitous and All-False-Primed: t(25) = −7.81, p = .42.

The modulation of the N400 through priming

The priming differences between the conditions were also reflected in our ERP results. A significant N400 effect was observed in all cases in which the compared conditions differed with respect to the truth-value judgements given by the subjects. However, comparing two false conditions, i.e. comparing the two All-False conditions for both groups, or comparing conditions Some-False with Some-Infelicitous for “pragmatists”, also resulted in significant N400 effects, even though in these cases the compared conditions did not differ with regard to the truth-value evaluation. The N400 effect in these comparisons can be explained by the priming difference between the compared conditions: The critical words in conditions All-False-Non-primed and Some-False denoted completely new objects which were not depicted in respective scenarios, whereas the critical words in conditions All-False-Primed and Some-Infelicitous referred always to one of the depicted object categories. Thus, only in conditions All-False-Non-primed and Some-False the critical words were not visually primed by the preceding scenarios. This lack of priming gave rise to more negative amplitudes in the N400 time-widows. Interestingly, there was no significant modulation of the Post-N400-Positivity through priming: There was no significant late positivity effect in the comparison Some-Infelicitous vs. Some-False for “pragmatists”, whereas the comparison All-False-Non-primed vs. All-False-Primed resulted in a marginal late positivity effect only after 800 ms post-stimulus onset, which reached significance for the whole group and for “logicians” but not for “pragmatists”.

Post-N400-Positivity

A centro-parietal positivity effect of latency similar to that observed in our study, i.e. from 500 and up to around 800 ms, is generally considered to be typical for syntactic violations (Osterhout and Holcomb, 1992). However, such late positivity effects have also been reported for semantic violations instead of any N400 effects (so-called “semantic P600” effects). Such “semantic P600” effects are often attributed to processes resembling those that are considered to underlie the standard, syntactic P600 effect, namely processes related to re-analysis or prolonged analysis of sentences (Kuperberg, 2013; Van Petten & Luka, 2012). In contrast, in studies related to episodic memory a late positivity effect observed approximately 400–800 ms post-stimulus onset has been described as so-called old/new effect – stimuli recognized as old tend to elicit more positive ERPs than new items correctly rejected (Van Petten and Luka, 2012). This effect, also referred to as late positive component (LPC), is described similarly to the P600 effect as largest over the posterior scalp sites, but unlike the standard P600, it tends to be larger over the left hemisphere (Friedman and Johnson, 2000).

In many studies that primarily focus on semantic violations, late positivity effects have been observed in addition to and after the expected N400 effects. In their review paper, Van Petten and Luka (2012) observe that most of the Post-N400-Positivities tend to have a centro-parietal topography. They suggest that such PNPs are similar to the “semantic P600” effect and reflect processing costs related to an attempted reanalysis. In our study, we observed a modulation of the positivity effect as early as 500 ms post-stimulus onset only when the compared conditions differed with respect to their truth-value evaluation. When the sentences in both compared conditions were evaluated as true (Some-Infelicitous vs. Some-True for “logicians”), there was no late positivity effect. Note that these two conditions did not differ with respect to priming of the critical words. When two false conditions were compared, i.e. Some-Infelicitous vs. Some-False for “pragmatists” or All-False-Non-primed vs. All-False for both groups, which additionally differed with respect to priming of the critical words, we observed a late positivity effect only in the latter case. Moreover, in this case the positivity effect had a different latency than in all those cases in which the compared conditions differed with respect to the truth-value evaluation. Based on these results, we can link the PNP effect in our experiment to the truth-related reprocessing of a sentence: The evaluation of a sentence as false, increased a relative positivity in the amplitude of the recorded signal after around 500 ms post-stimulus onset.

Autistic spectrum quotient and the processing of the scalar implicature

One of our research questions was whether the AQ scores of our subjects modulated the observed ERP effects or the ERP components, especially the N400 in the experimental Some-Infelicitous condition. As mentioned earlier “pragmatists” and “logicians” did not differ significantly with respect to their AQs (total scores, nor any of the subscales). We also did not observe any significant modulation of the N400/PNP in the Some-Infelicitous condition, the critical Some-Infelicitous vs. Some-True, or Some-False vs. Some-Infelicitous
N400/PNP effects, nor of any other effects, by the AQ scores of our subjects. There were no correlation results and the median split analysis (Low vs. High AQ) did not bring significant results either. It is also noteworthy that none of the other measured cognitive traits modulated subjects’ critical ERP effects. As we had a relatively large group of subjects (N = 54), it is unlikely that the lack of a significant effect was due to a low statistical power. Therefore, we can assume that the N400 and PNP effects in the critical comparisons Some-Infelicitous vs. Some-True and Some-False vs. Some-Infelicitous were modulated solely by the differences in the intuitive truth-value evaluation as well as the applied verification strategies, and not by the measured cognitive or personality traits.

Our results seem to counter the results by Nieuwland et al. (2010) and suggest that subjects’ AQ scores do not have any effect on their processing of scalar inferences. However, we should emphasise that the design of our study differs essentially from the design by Nieuwland et al. (2010): They used sentences that referred to world-knowledge, we used a sentence-picture verification paradigm. Although in the experiment by Nieuwland et al. (2010) the high- and low-AQ groups showed different effects in response to the violation of scalar inferences, for the high-AQ group the N400 ERPs were modulated by the LSA values in the tested sentences. In contrast, in our experiment this factor was controlled to stay stable across conditions. It is possible that in the study by Nieuwland et al. (2010) the low- and high-AQ groups were similarly sensitive to the pragmatic infelicity, but people with high AQT were significantly more sensitive to the lexico-semantic relationship in the tested sentences, and in their case the effect of the LSA values overrode the effect of the pragmatic felicity. Moreover, although the high-AQ group had significantly larger N400 ERPs for underinformative compared to informative sentences, there is no data in this study regarding the between-subject differences with respect to the intuitive evaluation of the underinformative sentences. Finally, plausibility judgements that are based on world-knowledge are likely to be associated with different processes than plausibility judgments that are based on short-term memory.

Discussion

Our experiment provides clear evidence that scalar inferences modulate the N400 of healthy individuals independently of their AQTs and independently of the lexico-semantic relationship between the critical words and the main noun phrases in tested sentences. This modulation of the N400 effect by the scalar implicature was only dependent on the person’s truth-value evaluation of the underinformative sentences. In our study, critical words in all unambiguously false sentences elicited larger N400 ERPs than critical words in the compared unambiguously true sentences. This effect was in each case followed by a significant Post-N400-Positivity effect. However, comparing the ERPs elicited by critical words in pragmatically infelicitous sentences with those elicited by critical words in true sentences (Some-Infelicitous vs. Some-True) resulted in an N400 and a late positivity effect only for the pragmatic responders. In contrast, the difference between the N400 elicited by critical nouns in false and infelicitous sentences (Some-False vs. Some-Infelicitous) was significantly larger for the logical responders than for the pragmatic responders. Moreover, the late positivity effect in this comparison was significant only for “logicians”.

The modulation of the sizes of the observed ERP effects by the truth-value evaluation of the underinformative sentences is a particularly important result. It allows to conclude that the observed ERP effects were triggered by the implicature violations only if a person explicitly adopted this implicature as a part of the sentence’s intuitive truth-conditions. It is worth to note that our participants were almost evenly divided into “pragmatists” and “logicians”, which brings up a question of whether this division is based on some inherent personality or cognitive traits of our subjects or whether it is based on their more or less random choice. We cannot fully exclude the possibility that there are in fact two distinct linguistic sub-populations, i.e. those who tend to use some always in the pragmatic meaning, and those who tend to use it in the logical meaning, or perhaps even those that are generally more “pragmatic” or more “logical” in using language. However, this group divide could also be taken to indicate ad hoc semantic decisions of the subjects, who had to choose between two alternative interpretations of the same linguistic expression. This conjecture is reinforced by the fact that there were no other significant differences between the two groups: neither involving their ERP differences in the control comparisons, their age, gender, AQ quotients, nor results of any part of the intelligence test. Thus, one can conclude that the recorded ERP effects in the critical comparisons Some-Infelicitous vs. Some-True and Some-False vs. Some-Infelicitous indicated only the person’s truth-value choice in the evaluation of the underinformative sentences. It should also be emphasized that whereas the N400 effect was additionally modulated by the priming differences between the compared conditions, the late positivity effect was only recorded for the comparisons
where the compared conditions differed with respect to the truth-value evaluation.

One could argue that the form of the task in our experiment, i.e. the need to provide acceptability judgements, forced the participants to adopt some strategies in evaluating the infelicitous sentences. Consequently, the experimental situation would be so far from natural communication that it would be difficult to draw conclusions about the real-life implicature processing. It has been observed that acceptability or plausibility judgements tend to evoke per se some ERP components, e.g. a late positivity on sentence-final words (Friedman, Simson, Ritter, & Rapin, 1975). Acceptability judgements make it also more likely that certain types of semantic violations will evoke the P600 effect (Kolk, Chwilla, van Herten, & Oor, 2003). Therefore, it is sometimes argued that a modulation of the P600 triggered by such acceptability tasks tells us little about language processing in general. However, Kuperberg (2007) points out that to sustain this claim one would need to explain why some types of semantic violations are more sensitive to such judgements than other types. It is also not well justified why acceptability tasks should be far from real-life language use. Real-life communication constantly involves decisions made on the basis of a linguistic input. Knowing sentences’ truth-conditions is essential for successful communication. Thus, understanding how the requirement of a truth-conditional decision modulates the neural dynamics of language comprehension has important relevance for understanding normal language processing in real-life communication. In our experiment, acceptability judgements played a double role. Firstly, it was our aim to investigate to what extent truth-value responses modulate the ERP components triggered by the implicature violation. Secondly, without gathering such judgements it would have been hard to control whether our subjects were focused on the meaning of the presented sentences, especially that in our experiment we were interested in the processing of the sentence meaning when the subject’s attention is actually directed towards the truth-conditions.

It is also an interesting result that the between-group differences in the Some-Infelicitous vs. Some-True as well as Some-False vs. Some-Infelicitous N400 effects were triggered by the differences in the N400 in the Some-True and Some-False conditions, whereas the N400 in the Some-Infelicitous condition was similar for both groups. This result allows to shed light on the verification strategies that can be presumed to underlie the sentence evaluation performed by both groups, as well as on the nature of the N400 component itself. Since the N400 is correlated with the context-based expectancy of the upcoming word, it seems that one can quantitatively measure this expectancy: The more words are expected in a given context, the larger is the N400 for one of the expected words. By contrast, the more unique the expectation regarding the upcoming word, the flatter the N400 for the critical word that is compatible with this expectation. It is clear that “logicians” and “pragmatist” applied different verification procedures for some-sentences. Whereas “logicians” treated both object categories presented in the scenario as potential verifiers of the sentence, “pragmatists” expected only the one that was consistent with the implicature, which made their predictions regarding the upcoming word unique.

Assuming that our ERP results reflect distinct truth-value judgements of the underinformative sentences as well as the verification strategies applied by the two groups, the question arises, what these results tell us with respect to the various theories of scalar implicatures. It should be relatively clear that the strong default view, which presupposes a mandatory and automatic character of scalar implicatures, has to be refuted. First, in support of the existing literature (Bott & Noveck, 2004; Huang & Sneedek, 2009; Tomlinson et al., 2013), our reaction time results provide evidence that the implicature is associated with an increased processing cost. Such results have generally been considered to be incompatible with the default view. Yet, our ERP results also speak against the default account. Unlike the P600 effect, which, as suggested by Hahne and Friederici (1999), can reflect controlled processes, the N400 is commonly understood to be a part of immediate semantic processes and thus taken to mirror automatic processes. Accordingly, the occurrence of the N400 effect associated with the implicature violation may be interpreted as a signature of an incremental integration of the implicature into sentence meaning during the online process of sentence interpretation. However, in our experiment the N400 effect was elicited by the implicature violation only for those subjects who responded pragmatically in the truth-value judgement task. Yet, the default theory is supposed to be a universal theory and it does not differentiate between individual language users. Therefore, this theory predicts that the N400 effect should occur irrespectively of the subject’s final decision to cancel the implicature. Consequently, the default theory is either false, or else it has a non-universal character and is true only with regard to the “pragmatic” linguistic subpopulation.

Yet, if the scalar implicature is not default, does it mean that it is based on a post-propositional inferential processes? In this alternative framework, only the logical meaning of some should be involved in the first-pass semantic processing. Why would then the implicature violation trigger any N400 effect in the first place for any of our subjects?
reflects a degree of preactivation or expectation of the stimulus, that is based on the context proceeding this stimulus, the N400 effect observed for the pragmatic responders in the case of the implicature violation means that the implicature had clearly shaped the expectations of this group of subjects regarding the upcoming word. Hence, the pragmatic responders must have generated the implicature already at the early stage (during the scenario inspection at the latest) and the N400 on the critical word reflected this early integration of the implicature into the sentence meaning. This means that “pragmatists” had generated the implicature at least before the completion of the sentence, which is also consistent with the results by Huang and Snedeker (2009). Thus, our results support those language comprehension models in which the scalar implicature is integrated into the semantic information in a parallel manner, while the sentence unfolds.

Another important question we were trying to answer in our study was to what extent implicatures become part of the intuitive truth-conditions of a sentence. The correlation between the truth-value responses and the recorded ERP effects (especially the N400) supports the view that the implicature can indeed be incorporated into the sentence’s truth-conditional content. What is not explained is why, when no background context is given, the pre-theoretic intuitions of ordinary speakers concerning the truth-conditions of sentences with some are highly divergent. Still, it is an important result that the truth-conditional decisions of our subjects were the sole predictors of their ERP responses. In the light of these results it might be tempting to say that some is simply ambiguous between the pragmatic and the logical meaning. Such a solution would safeguard that the implicature is both optional and involved in the compositional semantics. However, even if there are still some unanswered questions, our results provide clear evidence that any theory of the scalar implicature has to take into account two aspects so far considered to be not compatible: the possibility that the implicature is incrementally integrated into the sentence meaning and a non-default character of this implicature.

**Notes**

1. The frequency value $v$ of a word $w$ is equal to the $\log_2$ of the quotient of the frequency of the word “der” and the frequency of the word $w$ in corpus.
2. In supplementary materials, we provide the list of the noun triples used for the test trials and those used for the fillers as well as the log-frequency values for all the words.
3. As the analysis yields the same significant effects when performed on untransformed data for ease of interpretation, we refer to the untransformed values in Table 4 and Figure 3.

4. The grand averages and the topplots of the effects observed in the whole group analysis are available in supplementary materials. We also present there an example graphical visualization of the cluster-based permutation test results as well as the topographical maps of the effects in the comparisons of the $S$-conditions and $A$-conditions for each of the groups. Furthermore, for a comparison of the cluster-based statistics with a more traditional approach to data analysis, we also provide a repeated measures ANOVA of the EEG data using pre-defined time-windows. This analysis yields similar results as the reported here cluster-based statistics.

5. In each case, we excluded the EOG electrodes, linked mastoids and those anterior electrodes that were discarded from the analysis in the beginning due to technical problems (AF7, AF8, Fp1, Fp2). For the N400 effects in all comparisons, the involved electrodes included basically the whole scalp, except for the electrode F7 in the comparison $S_I$ vs. $S_T$ and F7, FT7 in the comparison $SF$ vs. $SI$. For some of the positivity effects, certain regions were not involved: For the $S_I$ vs. $ST$ PNP the non-involved electrodes were F7, F8, FC6, T8, F5, FT7, FT8; for the $SF$ vs. $SI$ PNP: F3, F8, AF3, AF4, F1; whereas the AFN vs. APF PNP effect had a centro-parietal distribution, i.e. the excluded electrodes were: F7, F3, F4, Fz, F8, FC6, C4, T8, P8, AF3, AF4, F5, F1, F2, F6, FC4, FT8, C6, TP8. The PNP effects in the remaining comparisons involved all electrodes.

6. The EOG electrodes, linked mastoids as well as the anterior electrodes: Fp1, Fp2, AF7, AF8 were excluded.

**Acknowledgments**

We would like to thank Natalia Rak for assistance in collecting the data, Napoleon Katsos for helpful comments on the project, finally the anonymous reviewer for useful remarks that helped us improve this paper.

**Disclosure statement**

No potential conflict of interest was reported by the authors. The icons of the cat and the ball, as featured in Figure 1, were obtained from www.openclipart.org and are free from any copyright restrictions. This site allows unlimited reproduction of the content for non-commercial use and all images are in the public domain.

**Funding**

This research has been supported by the grant “Structure of Memory” from Stiftung Mercator. The authors declare no competing financial interests.

**References**

Baggio, G. (2012). Selective alignment of brain responses by task demands during semantic processing. *Neuropsychologia*, 50 (5), 655–665. doi:10.1016/j.neuropsychologia.2012.01.002.
Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146*, 23–49. doi:10.1016/j.brainsci.2006.12.063

Kuperberg, G. R. (2013). The proactive comprehender: What event related potentials tell us about the dynamics of reading comprehension. In B. Miller, L. Cutting, & P. McCord (Eds.), *Unraveling the behavioral, neurological, and genetic components of reading comprehension* (pp. 176–192). Baltimore, MD: Paul Brookes Publishing.

Kuperberg, G. R., Caplan, D., Sitnikova, T., Eddy, M., & Holcomb, P. J. (2006). Neural correlates of processing syntactic, semantic, and thematic relationships in sentences. *Language and Cognitive Processes, 21*(5), 489–530. doi:10.1080/0169096050094279

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research, 17*(1), 117–129. doi:10.1016/S0926-6410(03)00086-7

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Science, 4*(12), 463–470. doi:10.1016/S1364-6613(00)01560-6

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*, 621–647. doi:10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203–205.

Kutas, M., & Van Petten, C. (1994). Psycholinguistics electrified: Event-related potential investigations. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 83–143). San Diego: Academic Press.

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

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284. doi:10.1080/01638539809545028

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implication*. Cambridge, MA: Massachusetts Institute of Technology.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177–190. doi:10.1016/j.jneumeth.2007.03.024

Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language, 63*(3), 324–346. doi:10.1016/j.jml.2010.06.005

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth isn’t too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science, 19*(12), 1213–1218. doi:10.1111/j.1467-9280.2008.02226.x

Nieuwland, M. S., & Martin, A. E. (2012). If the real world were irrelevant, so to speak: The role of propositional truth-value in counterfactual sentence comprehension. *Cognition, 122*(1), 102–109. doi:10.1016/j.cognition.2011.09.001

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*(7), 1098–1111. doi:10.1162/089892906777563724

Van Berkum, J. J., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience, 11*(6), 657–671. doi:10.1162/089892999563724

Van Berkum, J. J. A., Koornneef, A. W., Otten, M., & Nieuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research, 1146*, 158–171. doi:10.1016/j.brainsci.2006.06.091
Van Herten, M., Kolk, H. J., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research, 22*(2), 241–255. doi:10.1016/j.cogbrainres.2004.09.002

Van Petten, C. (1993). A comparison of lexical and sentence-level context effects in event-related potentials. *Language and Cognitive Processes, 8*(4), 485–531. doi:10.1080/01690969308407586

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*(2), 176–190. doi:10.1016/j.ijpsycho.2011.09.015

Werning, M. (2005). Right and wrong reasons for compositionality. In M. Werning, E. Machery, & G. Schurz (Eds.), *The compositionality of meaning and content, Vol. I: Foundational issues* (pp. 285–309). Frankfurt: Ontos Verlag.

Werning, M., Hinzen, W., & Machery, M. (Eds.) (2012). *The Oxford handbook of compositionality*. Oxford: Oxford University Press.