Gricean expectations in online sentence comprehension: an ERP study on the processing of scalar inferences

Petra Augurzky¹/²*, Michael Franke¹/²/³ & Rolf Ulrich²/⁴

¹Department of Linguistics, University of Tübingen,
²Sonderforschungsbereich 833, University of Tübingen,
³Institute of Cognitive Science, University of Osnabrück,
⁴Department of Psychology, University of Tübingen

*Corresponding author:
Email: petra.augurzky@uni-tuebingen.de
ABSTRACT

There is substantial support for the general idea that a formalization of comprehenders’ expectations about the likely next word in a sentence helps explaining data related to online sentence processing. While much research has focused on syntactic, semantic and discourse expectations, the present ERP study investigates neurolinguistic correlates of pragmatic expectations, which arise when comprehenders expect a sentence to conform to Gricean Maxims of Conversation. For predicting brain responses associated with pragmatic processing, we introduce a formal model of such Gricean pragmatic expectations, using an idealized incremental interpreter. We examine whether pragmatic expectancies derived from this model modulate the amplitude of the N400, a component that has been associated with predictive processing. As part of its parameterization, the model distinguishes genuine pragmatic interpreters, who expect maximally informative true utterances, from literal interpreters, who only expect truthfulness. We explore the model’s non-trivial predictions for an experimental set up which uses a picture-sentence verification paradigm with ERPs recorded at several critical positions in sentences containing the scalar implicature trigger some. We find that Gricean expectations indeed affect the N400, largely in line with the predictions of our model, but also discuss discrepancies between model predictions and observations critically.

Keywords: ERPs; N400; pragmatic processing; predictive processing; probabilistic pragmatic modeling; scalar implicatures
1. Introduction

Interpreting an utterance involves rapidly combining information from linguistic and contextual sources. Comprehenders usually achieve this efficiently, in part by anticipating quite successfully what a speaker is likely to say next in the given context. Such next-word expectancies take into account syntactic structure, lexical material and its semantic meaning, discourse coherence, and the overall message a speaker is likely to convey. Evidence for this prediction-based processing comes from different experimental measures. For instance, self-paced reading times were shown to correlate with the predictability of lexico-syntactic material (Levy, 2008; Smith & Levy, 2013). Moreover, event-related potentials (ERP) studies indicate that the amplitude of the N400 component, a negative deflection with a centro-parietal maximum between 200 and 600 ms post stimulus onset (Kutas & Hillyard, 1984), can be modulated by particularly relevant contextual information, which may give rise to predictions regarding the lexical-semantic properties of upcoming words (see Kutas & Federmeier, 2011, for an overview and alternative interpretations, such as signaling lexical retrieval). Despite its explanatory success, the idea that predictive expectations explain online processing is not uncontroversial, and many crucial details remain unsettled. For instance, there is currently some vivid debate concerning the specific contextual cues that may trigger predictions and about how specific and generalizable such predictions are (DeLong, Urbach & Kutas, 2017; Ito, Martin & Nieuwland, 2017; Yan, Kuperberg & Jaeger, 2017). Moreover, it is astonishing that the vast majority of studies to date focuses on syntactic or lexical-semantic processing with much less emphasis on genuine pragmatic aspects (but see Werning & Cosentino 2017). The present ERP study aims at bridging this gap by investigating the role of pragmatic information as a prediction trigger in on-line sentence comprehension.\(^1\) To this end, we examine predictive effects on the N400 amplitude related to the processing of scalar implicatures.

Scalar implicatures can be triggered by utterances like (1), which contain a scalar item (here: some) for which a semantically stronger, so-called scalar alternative (here: all) exists (e.g. Horn, 1972,

---

\(^1\) The notion prediction has sometimes been distinguished from expectation. For instance, van Petten & Luka (2012) use prediction for a literal anticipation corresponding to a specific lexical item, whereas the term expectation is used for the anticipation of semantic content more generally (the activation of specific lexical features that match the previously processed information). As our studies were not designed to distinguish between these terms, we use these notions interchangeably.
Levinson, 1983). By standard assumptions, sentence (1) is literally true if some or all of the relevant dots are blue. However, comprehenders usually additionally draw the scalar implicature that not all of the relevant dots are blue. The common explanation for this is that listeners expect speakers to adhere to fundamental rules of adequate language use, as described by Grice’s Maxims of Conversation. This entails an expectation that speakers give all the relevant information they are capable of giving. If listeners observe an utterance of (1), they would therefore draw the conclusion that the speaker was not in a position of uttering the stronger sentence (2), arguably because it is false.

(1) Some of the dots are blue.
(2) All of the dots are blue.

Scalar implicatures have been well-researched behaviorally, but considerably less-studied in the neurolinguistic domain. Whereas previous studies largely focused on whether scalar implicatures are generated mandatorily, and whether implicature processing *per se* is associated with processing cost (Bott & Noveck, 2004; Huang & Snedeker, 2009, Noveck & Posada, 2003, Tomlinson & Bott 2013), our study complements this research by focusing on predictive processing. To test whether such Gricean expectations affect sequential utterance processing, we examine neurolinguistic correlates of scalar implicature processing over the course of the sentence by measuring ERPs, which provide a temporal resolution in the millisecond range and also add a qualitative dimension to behavioral tasks. We introduce a model of sequentially updated probabilistic Gricean expectations about utterance continuations and test its predictions regarding the amplitude of the N400.

2. Background: processing of scalar implicatures in the brain - ERPs and expectations

To date, ERP studies on implicature processing are still scarce and do not reveal a consistent pattern of results. One component that has been associated with the processing of quantified sentences in general is the N400 (Augurzky, Bott, Sternefeld & Ulrich, 2016; Freunberger & Nieuwland, 2016; Urbach &
Kutas, 2010; Urbach, DeLong & Kutas, 2015). The N400 has been observed as a processing reflex of so-called underinformative sentences, i.e. sentences with some used in a situation where the speaker could have used all instead (e.g. Some elephants are mammals; see Hunt, Politzer-Ahles, Gibson, Minai & Fiorentino, 2013; Nieuwland, Ditman & Kuperberg, 2010; Zhao, Liu, Chen & Chen, 2015; but see Noveck & Posada, 2003, for an effect going in the opposite direction).

One ERP study which explicitly discusses the role of prediction related to implicature processing is that by Spychalska, Kontinen & Werning (2016). Spychalska and colleagues used a segmented trial structure with the explicit intention of inducing strong next-word expectations in their experimental participants (a design choice we see critical and try to improve on, see below). Each trial first presented German sentence fragments of the form Some/All pictures contain…, which were followed mid-sentence by pictures as in Fig. 1. At this stage, participants were hypothesized to generate expectations about the sentence continuation. The crucial object nouns, for which ERPs were analyzed, were then presented on the screen following the picture. Three types of conditions were compared. For a sentential context Some of the pictures followed by the display in Fig. 1, the noun cats exemplifies the “felicitous” condition (i.e. semantically true and pragmatically felicitous, in which the implicature is drawn and some corresponds to the not all reading), balls exemplifies the “infelicitous” condition (i.e. semantically true and pragmatically infelicitous, in which some corresponds to all), and teeth the “false” condition (i.e. semantically false).

Figure 1: One of the pictures used in Spychalska et al. (2016)

On the noun, the “false” condition elicited the largest N400, but differences in amplitude between the “felicitous” and “infelicitous” condition depended on, so the authors argued, participants’ response types. Based on their binary truth-value judgements, participants were either categorized as pragmatic
responders, if they predominantly judged underinformative sentences with *some* as false, or semantic responders, if they predominantly judged them as true. Pragmatic responders showed a larger N400 in the “infelicitous” condition than in the “felicitous” condition. Semantic responders showed no such difference. According to the authors, these differences were driven by different contextual predictions. Semantic responders expected a true description, so that either *cats* or *balls* was unsurprising. Pragmatic responders expected a true and felicitous utterance, so they expected only *cats* but not *balls*.

Participant-related effects were also found in other studies. For instance, in Nieuwland et al. (2010), consistent N400 effects for sentences like *Some people have lungs* vs. *Some people have pets* were restricted to participants exhibiting low scores in the so-called *Autism-Spectrum Quotient questionnaire* (Baron-Cohen, Wheelwright, Skinner, Martin & Clubley, 2001). According to Nieuwland et al., low autism values are indicative of a pronounced pragmatic sensitivity, and comprehenders with low AQ values are more likely to exhibit a pragmatic N400 effect (p. 329). Finally, in some studies, the N400 was accompanied by a late positive component, which has been suggested to reflect task-related processing strategies associated with the verification process in general (Augurzky et al., 2016; Nieuwland, 2016).

Interestingly, brain responses to underinformative readings seem to be dependent on the sentential position from which ERPs are aggregated. Whereas N400 effects related to implicature processing were observed at positions following the quantifier, a distinct pattern was found when measuring ERPs from the onset of *some*. For instance, Politzer-Ahles, Fiorentino, Jiang & Zhou (2013) tested Mandarin Chinese versions of sentences such as *In this picture, some/all of the girls are sitting on blankets suntanning* preceded by pictures that did or did not match the sentential meaning. Underinformative sentences elicited a sustained negativity as opposed to pragmatically felicitous sentences. This effect was interpreted as reflecting the effortful retrieval of the literal interpretation of *some* after its pragmatic reading had already been activated on the picture, a process comparable to discourse-related revision. Moreover, no ERP effects of implicature processing were observed on the quantifier *some* in a study by Hartshorne, Snedeker, Liem Azar, & Kim (2014), in which conditional and declarative sentences were compared. However, as evidenced by Politzer-Ahles & Gwilliams (2015), implicature-related brain activity on the
quantifier may be difficult to measure using ERPs though it can be detected by source-related measures like fMRI or MEG. By considering ERPs from different sentence positions, we aim to provide further evidence for or against this potential asymmetry.

Finally, though for a study unrelated to scalar implicatures, Werning & Cosentino (2017) propose a concrete model of pragmatic expectations formulated in terms of recently popular probabilistic approaches (Franke & Jäger 2016, Goodman & Frank 2016). Their model assumes that comprehenders reason about the likely utterances of a by-and-large rational speaker who adheres to Grice’s Maxims of Conversation. Concretely, the model derives next-word expectations from the assumption that an initial discourse move by the speaker must have happened for a good reason, in particular as a potential signal that the event to be described later is predictable or not. Werning & Cosentino adopt an interpretation of N400 amplitude as a function of contextual surprisal of the trigger word and demonstrate that their probabilistic model predicts the qualitative differences in relevant N400 amplitudes correctly.

3. Motivation of our design and predictions of a probabilistic model

Building on Spychalska et al. (2016), we investigate whether and how Gricean expectations of informative utterances influence incremental sentence processing. However, we would like to improve on their experimental design in several respects and also add an explicit formal model, which allows us to systematically and clearly derive non-trivial predictions for our more elaborate design. In particular, we take issue with Spychalska et al.’s choice to segment the presentation of the to-be-interpreted sentence with the goal of inducing clear next-word expectations. We rather strive for a more realistic mode of reading sentences, to deflect the worry that the strong next-word expectations are a mere artifact of the segmented presentation of the sentence material. Moreover, we are interested in potential effects of Gricean expectations at different critical positions in the sentences, in order to investigate whether comprehenders indeed hold Gricean expectations from the start and update them sequentially as the utterance unfolds.

To this end, we employed an experimental design comparable to the one from a recent ERP study
on the processing of the quantifier *all* (Augurzky et al., 2016). In that study, different pictorial contexts comparable to those in Fig. 2 below preceded target sentences like *Sind alle Dreiecke blau, die innerhalb des Kreises sind?* (‘Are all triangles blue, that inside-of the circle are?’).

![Figure 2](image)

**Figure 2**: A subset of the experimental context pictures presented in Augurzky et al. (2016)

An N400 on the color adjective *blue* was found for literally false as opposed to true sentences, i.e. following Fig. 2C vs. 2B. When sentences were preceded by complex contexts like A, the amplitude of the N400 on the adjective was intermediate between true and false sentences. Augurzky et al. attributed this finding to a processing strategy in which the truth evaluation had been delayed to a position where a semantic commitment was unambiguously possible. Alternatively, the observed effect could be explained by considering the relative predictability of the adjective, which varies as a function of the pictorial context and the already processed sentence information prior to the adjective. Since Augurzky et al. examined the processing of questions, it is difficult distinguish between these approaches (see Augurzky et al., 2016, for discussion). Therefore, the present study uses declarative clauses and tests whether ERPs at several sentence positions can be explained in terms of comprehenders’ expectations of incoming lexical material. By comparing previous results from the quantifier *all* with the present results from the quantifier *some*, the current study also adds to recent considerations about the on-line processing of semantic and pragmatic aspects of quantificational meaning (Zhan, Jiang, Politzer-Ahles & Zhou, 2017).

To fix a context of utterance, we first presented participants with an abstract picture. Each picture showed geometrical objects (e.g. dots) in two container shapes (e.g. a circle and a square, see Fig. 3). In simple pictures (B and C in Fig. 3), all objects were of identical color, whereas in complex pictures (A and D in Fig. 3), objects within one of the two container forms were of two different colors. Following
each picture, German sentences as in (3) were shown.\textsuperscript{2} ERPs were analyzed from three sentential positions: First, from the onset of the quantifier (einige, ‘some’), second, from the onset of the color adjective (e.g. blau, ‘blue’), and third, from the onset of the geometrical shape noun (coded as left or right, e.g. Quadrat (‘square’) or Kreis (‘circle’)).

Towards a derivation of theoretical predictions for these stimuli, we consider an idealized incremental interpreter who holds probabilistic beliefs about each upcoming word $w_{i+1}$ after hearing an initial sentence fragment $w_1, \ldots, w_i$ in context $C$ (e.g. Levy, 2008). Let $S_{[w_1, \ldots, w_i, w_{i+1}]}$ be the (finite) set of all (relevant, salient) complete sentences the speaker might utter in the concrete utterance context $C$, starting with $w_1, \ldots, w_i, w_{i+1}$ and $M$ be the (finite) set of contextually salient meanings the speaker might wish to communicate. Moreover, let $P(m \mid C)$ be the comprehender’s prior belief about how likely the speaker would want to express meaning $m$. Finally, $P_s(s \mid m, C)$ is (the comprehender’s holistic belief about) the likelihood that the speaker produces a complete sentence $s$ when wishing to communicate meaning $m$. By Bayes rule, the next-word beliefs of an idealized comprehender are derivable as:

$$P(w_{i+1} \mid C, w_1, \ldots, w_i) \propto \sum_{m \in M} \sum_{s \in S_{[w_1, \ldots, w_i, w_{i+1}]}} P(m \mid C) P_s(s \mid m, C)$$

This rational analysis of a comprehender’s next-word beliefs could make very fine-grained

\textsuperscript{2} The postposition of the restrictive relative clause is not grammatically marked in German, unlike in English.
quantitative predictions about the extent to which experimental participants should be surprised by a newly processed word. Unfortunately, quite detailed and possibly controversial additional assumptions are necessary to derive these minute quantitative predictions. For instance, it requires fixing the set $M$ of relevant meanings and their prior probabilities. To sidestep these complications, at least to a certain extent, we start here by exploring a simpler approach. We will compare this simpler model and the full idealized incremental interpreter model in the final discussion.

In the simple model we explore here, comprehenders expect word $w_{i+1}$ to occur next with probability $P(w_{i+1} | C, w_1, \ldots, w_i)$ based on whether it is possible to continue the initial sentence fragment $w_1, \ldots, w_i$ to form a complete sentence compatible with the general pattern of sentences encountered in the current experimental setting such that the resulting full sentence is either (i) a true and pragmatically felicitous description of $C$, (ii) a true but underinformative description of $C$ (e.g., using *some* where *all* would have been true) or (iii) a false description of $C$. We assume that:

$$P(w_{i+1} | C, w_1, \ldots, w_i) \propto \begin{cases} \ a & \text{if } w_1, \ldots, w_i, w_{i+1} \text{ can yield a true and felicitous description of } C \\ \ b & \text{if } w_1, \ldots, w_i, w_{i+1} \text{ can yield at most a true description of } C \\ \ c & \text{if } w_1, \ldots, w_i, w_{i+1} \text{ can not yield a true description of } C \end{cases}$$

where, $a \geq b \gg c$. For participants who, as in previous studies (e.g. Spychalska, Kontinen & Werning, 2016) classify as semantic interpreters by their behavioral response pattern, we will assume that $a = b$, i.e. these participants do not make a difference between continuations that differ with respect to pragmatic felicity. For pragmatic responders we assume that $a > b$.

According to this model, any two continuations which can both yield true and pragmatically felicitous descriptions of the context picture are equally likely; if $a > b$, continuations that can yield true and felicitous sentences are more likely than those that could be at best true; continuations that can only result in false descriptions are very unlikely whenever at least one continuation exists that yields at least a true description of $C$. Linking the expectability of incoming words with the amplitude of the N400 (Frank, Otten, Galli & Vigliocco, 2015; De Long, Urbach & Kutas, 2005), we generally expect that the bigger the difference between $P(w_{i+1} | C, w_1, \ldots, w_i)$ and $P(w'_{i+1} | C, w_1, \ldots, w_i)$ the higher the N400 amplitude after an observation of $w'_{i+1}$ in contrast to an observation of $w_{i+1}$.

This model makes clear and testable categorical predictions for our experimental conditions. On
the quantifier, no matter whether comprehenders expect pragmatic felicity \((a>b)\) or not \((a=b)\), sentence-initial \textit{some} is as likely as \textit{all} in complex conditions A and D. However, for the simple conditions B and C, sentence-initial \textit{some} cannot lead to a true and felicitous sentence given the general pattern of sentences used in this experiment. Therefore, we predict: \(P(\text{some} \mid A, \emptyset) = P(\text{some} \mid D, \emptyset) \geq P(\text{some} \mid B, \emptyset) = P(\text{some} \mid C, \emptyset)\), where the inequality is strict just in case \(a > b\), and \(\emptyset\) is used to denote absence of an initial sentence fragment. An unmet pragmatic expectation at sentence-initial positions is therefore predicted to lead to an increased N400 amplitude for pragmatically infelicitous \((B,C)\) vs. felicitous \((A,D)\) conditions.\(^3\)

On the adjective, we anticipate effects driven by expectations of truth in simple conditions: \(P(\text{blue} \mid B, \text{some} \ldots \text{are})\) should be very high, but \(P(\text{blue} \mid C, \text{some} \ldots \text{are})\) should be very low. Predictions about the color adjective in the complex contexts \((A\text{ and }D)\) are intermediate: \(P(\text{blue} \mid A, \text{some} \ldots \text{are}) = P(\text{blue} \mid D, \text{some} \ldots \text{are}) \approx 0.5.\) Independent of pragmatic sensitivity, we anticipate: \(P(\text{blue} \mid C, \text{some} \ldots \text{are}) < P(\text{blue} \mid A, \text{some} \ldots \text{are}) = P(\text{blue} \mid D, \text{some} \ldots \text{are}) < P(\text{blue} \mid B, \text{some} \ldots \text{are})\). Consequently, we predict the N400 amplitude to be most reduced for the simple true conditions \((B)\), and most increased for the simple false conditions \((C)\), with the complex conditions \((A\text{ and }D)\) in between.

For the shape noun, we do not predict ERP differences for the simple conditions \((B\text{ and }C)\), irrespective of pragmatic sensitivity: \(P(\text{circle} \mid B, \text{some} \ldots \text{blue} \ldots \text{in the}) = P(\text{circle} \mid C, \text{some} \ldots \text{blue} \ldots \text{in the}) \approx 0.5.\) In condition A, \textit{circle} is a true and felicitous continuation, but \textit{square} yields an infelicitous, though true full utterance, so that \(P(\text{circle} \mid A, \text{some} \ldots \text{blue} \ldots \text{in the}) \equiv a/(a+b) \geq b/(a+b) = P(\text{square} \mid A, \text{some} \ldots \text{blue} \ldots \text{in the})\), with a strict inequality whenever \(a > b\). For condition D, \textit{circle} is true and felicitous, but \textit{square} makes the utterance semantically false, so that \(P(\text{circle} \mid D, \text{some} \ldots \text{blue} \ldots \text{in the}) \equiv a/(a+e) \geq e/(a+e) = P(\text{square} \mid D, \text{some} \ldots \text{blue} \ldots \text{in the})\) for pragmatic \((a > b)\) and semantic responders \((a=b)\) alike. We thus expect a more pronounced N400 for \textit{square} than for \textit{circle} in condition D, independent of responder type. We expect a difference in N400 in condition A only for pragmatic responders.

---

\(^3\) Based on the previous heterogeneous results on the quantifier position (see Section 2 and the General Discussion), it is also conceivable that pragmatically infelicitous readings elicit a different brain response at this position (i.e. a sustained negativity, as in Politzer-Ahles, 2013, or no effect at all, as in Hartshorne et al., 2014). In order to deal with these issues, we also included the time windows from Politzer-Ahles et al. into our statistical analyses, in addition to an earlier N400 time window.
4. Experiment

4.1. Methods

4.1.1. Participants
Twenty-five right-handed students (mean age: 24.1 years) from the University of Tübingen took part (9 male), all German native speakers with normal or corrected-to-normal vision paid for participation.

4.1.2. Materials
A 4x2-factorial within-subjects design was used with the factors CONTEXT (A-D) and SHAPE (left, right). Quadruplets of pictures similar to Fig. 3 were generated with different shapes inside two container dorms of different kinds. A total set of twenty quadruplets was created by randomly combining eleven colors, seven shapes and ten container forms. Experimental sentences were as illustrated in (3) above (see Table 1 for a full list).

<table>
<thead>
<tr>
<th>Condition</th>
<th>Functional label</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Context A, left shape complexA1, underinformative</td>
<td>Einige Punkte sind blau, die im Quadrat sind. Some dots are blue, that in-the square are.</td>
</tr>
<tr>
<td>A2</td>
<td>Context A, right shape complexA2, informative</td>
<td>Einige Punkte sind blau, die im Kreis sind. Some dots are blue, that in-the circle are.</td>
</tr>
<tr>
<td>B1</td>
<td>Context B, left shape simpleB1, underinformative</td>
<td>Einige Punkte sind blau, die im Quadrat sind. Some dots are blue, that in-the square are.</td>
</tr>
<tr>
<td>B2</td>
<td>Context B, right shape simpleB2, underinformative</td>
<td>Einige Punkte sind blau, die im Kreis sind. Some dots are blue, that in-the circle are.</td>
</tr>
<tr>
<td>C1</td>
<td>Context C, left shape simpleC1, false</td>
<td>Einige Punkte sind blau, die im Quadrat sind. Some dots are blue, that in-the square are.</td>
</tr>
<tr>
<td>C2</td>
<td>Context C, right shape simpleC2, false</td>
<td>Einige Punkte sind blau, die im Kreis sind. Some dots are blue, that in-the circle are.</td>
</tr>
<tr>
<td>D1</td>
<td>Context D, left shape complexD1, false</td>
<td>Einige Punkte sind blau, die im Quadrat sind. Some dots are blue, that in-the square are.</td>
</tr>
<tr>
<td>D2</td>
<td>Context D, right shape complexD2, informative</td>
<td>Einige Punkte sind blau, die im Kreis sind. Some dots are blue, that in-the circle are.</td>
</tr>
</tbody>
</table>

Per condition, twenty picture-sentence pairs were presented, resulting in 160 experimental sentences. We included 160 filler sentences with the quantifier alle (‘all’). The experiment consisted of 320 sentences. Conditions were evenly spread over ten blocks. To control for positional effects, two
experimental versions were generated. The first block of the first version corresponded to the final block of the second version, the second block of version 1 corresponded to the ninth block of version 2 and so on.

4.1.3. Procedure
Participants were seated in a dimly-lit, soundproof booth in front of a 17” computer screen. Stimuli appeared in a pseudo-randomized order. The experimental session was divided into ten blocks (32 trials per block), with breaks between the blocks. Each trial began with the picture in the center of the screen for 1500 ms. Then the sentence was presented in a word-by-word fashion via RSVP (500 ms per word). After each sentence, participants made a truth evaluation (*Did the preceding sentence truly or falsely reflect the content of the picture?*). After the final word had disappeared, three question marks were shown, signaling that participants now had to answer with “wahr” (*true*) and “falsch” (*false*) by pressing one of two buttons (‘F’ or ‘J’) on the keyboard. The keys for true and false answers were counterbalanced across participants. Participants were asked to make their truth evaluation as quickly as possible. The initial above timeout was 1200 ms and was adapted to the response speed of the participants by using an exponentially weighted moving average (Leonhard, Fernández, Ulrich & Miller, 2011). When participants’ reaction times exceeded the current timeout, they received visual feedback (*Schneller!*, ‘faster’) on the screen. Following the judgment, a blank screen appeared for 500 ms and three exclamation marks in yellow (1200 ms) indicated that participants now could blink until the next picture was presented. A practice session preceded the experimental session. After the experimental session, participants filled in a German version of the *Autism Spectrum Quotient Questionnaire* based on which we could calculate the *AQ-Comm score* (see Baron-Cohen et al., 2001; Nieuwland et al., 2010, for details). Including electrode application, the experimental session lasted between 2 and 2.5 hours.

4.1.4. EEG recording
The EEG was continuously recorded from 32 Ag/AgCl electrodes using a BIOSEMI Active-Two amplifier system: FP1, FP2, AF3, AF4, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5,
Augurzky et al.: “Gricean expectations in online sentence comprehension”

CP1, CP2, CP6, P7, P3, Pz, P4, P8, PO3, PO4, O1, Oz, O2. The electrooculogram (EOG) was recorded by means of electrodes placed at the outer canthus of each eye (horizontal EOG) and above and below the participant’s left eye (vertical EOG). Two electrodes were put on the left and right mastoid for the purpose of off-line referencing. Recordings were sampled at 1024 Hz, and a 100 Hz low-pass filter was used. Off-line, recordings were downsampled to 256 Hz for data analysis, and electrode sites were referenced to linked mastoids. Raw EEG data were filtered using a 0.3-20 Hz bandpass filter. Before entering ERP analyses, individual participant data were automatically and manually screened in order to exclude trials with eye movement and muscular artefacts. Data per participant and per condition were aggregated from the onset of the critical element (quantifier, adjective, and shape noun) to 1000 ms post onset. Afterwards, grand averages were calculated over all participants.

2.1.5. Data analysis

We first determined whether participants could be classified into exhaustive groups of semantic and pragmatic responders according to their sentence-final truth-value judgements (Noveck & Posada, 2001; Spychalska et al., 2016). Comparable to previous studies, we tested whether participants could be classified into semantic and pragmatic responders. To this end, we considered participants’ pragmatic responses to underinformative statements (Noveck & Posada, 2003; Politzer-Ahles et al., 2013). Moreover, we calculated AQ-Comm scores for each participant. If the behavioral data allowed a binary group classification, we determined whether groups differed with respect to the values obtained by the AQ-Comm scores (Spychalska et al., 2016). The between-subjects factor GROUP (2) was added to subsequent analyses if responder-type groups could be determined. In that case, each of the participants was classified as belonging to one of the groups (semantic/pragmatic responders).

ERPs from three sentential positions (quantifier, color adjective, shape noun were analyzed statistically by repeated-measures ANOVAs on mean amplitude values. Four regions of interest (Rois) were introduced (see Augurzky et al., 2016): left anterior (Roi 1: F3, F7, FC1, FC5), right anterior (Roi 2: F4, F8, FC2, FC6), left posterior (Roi 3: CP1, CP5, C3, P3) and right posterior (Roi 4: CP2, CP6, C4, P4). All analyses were calculated within an early N400 time window (300-400ms) and a time window
capturing late positive effects (450-800ms) following Augurzky et al. (2016). As previous results on the quantifier itself were inconclusive, we additionally applied the two time windows Politzer-Ahles et al. (2013) for analyses of the quantifier position, i.e. a 500-1000 ms, as in their Experiment 1, and in a 300-1000 ms time window, as in their Experiment 2.

From quantifier onset, we carried out ANOVAs with the factors COMPLEXITY (simple (B,C) vs. complex (A,D)) x ROI (left anterior, right anterior, left posterior, right posterior). On the color adjective, we carried out ANOVAs with the factors CONTEXT (A,B,C,D) x ROI. From the onset of the shape noun, we carried out ANOVAs with the factors CONTEXT (A,B,C,D), SHAPE (left vs. right) x ROI.

Statistical analyses were carried out in a hierarchical manner. Only significant interactions (p < .05) were resolved. Corrected p-values (Huynh & Feldt, 1970) were chosen when the analysis involved more than one degree of freedom in the numerator. Probability levels were Bonferroni adjusted for the planned comparisons between the four levels of the factor CONTEXT.

4.2. Results

4.2.1. Behavioral Data

Mean error rates and reaction times of clause-final truth value judgments are shown in Table 2. Mean error rates were 8.5%, suggesting overall good task performance. Approximately half of the participants (n=13) consistently gave “pragmatic answers”, i.e. judged underinformative sentences with some to be false in a situation where all would have been true as well. Fig. 4 shows the histogram of the frequency of pragmatic answers by participant. A clear dichotomy is visible, with no participant a borderline case of semantic vs. pragmatic responder. Participants’ AQ-Comm scores ranged from 8 to 30 out of a maximal value of 50, with a median value of 18, but did not allow to split the participants into response groups, as seven of them scored the median value. For between-participants analyses of the reaction times and ERPs, groups were therefore determined by participants’ truth-value judgements (13 pragmatic, 12 semantic responders). The ANOVA on reaction time with the between-factor GROUP and the within-factors, CONTEXT and SHAPE did not reveal any significant group interactions (all p values >
A significant effect of SHAPE ($F(1, 24) = 10.20; p < .01$) was obtained. Descriptively, this effect is caused by an advantage for the right shape in both of the complex conditions. However, the interaction between CONTEXT x SHAPE was only marginally significant ($F(3, 72) = 2.72; p = 0.053$). No main effect of CONTEXT was observed ($p > 0.11$).

Table 2. Behavioral results for the experimental conditions. RTs are provided for semantic and pragmatic comprehenders separately.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Label used for statistical analysis</th>
<th>Functional label</th>
<th>Correct responses (%)</th>
<th>Reaction times (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sem.</td>
<td>Prag.</td>
</tr>
<tr>
<td>A1</td>
<td>Context A, Left shape</td>
<td>complexA1, underinformative</td>
<td>NN</td>
<td>361.5</td>
</tr>
<tr>
<td>A2</td>
<td>Context A, Right shape</td>
<td>complexA2, informative</td>
<td>85.1</td>
<td>322.0</td>
</tr>
<tr>
<td>B1</td>
<td>Context B, Left shape</td>
<td>simpleB1, underinformative</td>
<td>NN</td>
<td>298.6</td>
</tr>
<tr>
<td>B2</td>
<td>Context B, Right shape</td>
<td>simpleB2, underinformative</td>
<td>NN</td>
<td>298.8</td>
</tr>
<tr>
<td>C1</td>
<td>Context C, Left shape</td>
<td>simpleC1, false</td>
<td>93.1</td>
<td>307.2</td>
</tr>
<tr>
<td>C2</td>
<td>Context C, Right shape</td>
<td>simpleC2, false</td>
<td>92.7</td>
<td>332.8</td>
</tr>
<tr>
<td>D1</td>
<td>Context D, Left shape</td>
<td>complexD1, false</td>
<td>74.5</td>
<td>340.4</td>
</tr>
<tr>
<td>D2</td>
<td>Context D, Right shape</td>
<td>complexD2, informative</td>
<td>86.7</td>
<td>289.0</td>
</tr>
</tbody>
</table>

Figure 4: Histogram of participant’s percentage of pragmatic responses

4.2.2. ERP Data

A complete overview of the results is provided in the Appendix.

Quantifier

Fig. 5 shows ERPs from quantifier onset. Simple contexts elicited a positivity compared to complex conditions. The ANOVA with the factors GROUP, ROI und COMPLEXITY revealed effects of
COMPLEXITY in each time window (all p values < .05). Whereas the effect was significant for both groups alike in the 300-400, the 300-1000 and the 450-800 ms time windows (all interactions with Group p > .13), an interaction of GROUP x COMPLEXITY was obtained in the 500-1000 ms time window ($F (1,23) = 5.59; p < .05$). In this window, the effect was significant only for pragmatic responders (semantic responders: $p = .59$; pragmatic responders: $p < .01$). In addition, a ROI x COMPLEXITY interaction was found in the 300-400 ms time window, showing a more frontal distribution of the effect in this window. No further effects reached significance. In sum, we found a positivity for underinformative *some* at the quantifier. At one of the four time windows, this effect became significant only for pragmatic responders.

**Figure 5**: ERPs from the onset of the quantifier

Color Adjective

Fig. 6 shows ERPs from adjective onset. No interactions between the factor GROUP and any of the experimental factors was observed, in line with our hypotheses. In the *early time window* (300-400 ms), a significant effect of CONTEXT ($F (3,72) = 14.74; p < .001$) was obtained. The analyses showed consistent differences between all context types (all p values < .05, except for the comparison between A vs. C ($p = 0.22$), and B vs. D ($p = 1$). Moreover, the interaction of ROI x CONTEXT: $F (9,216) = 5.60; p < .001$) was significant. The analyses showed consistent differences between all context types across Rois (all p values < .05, except for the comparison between A and B (complex$_{A1-A2}$ vs. simple$_{B1-B2}$) in
the left anterior Roi ($p = 0.336$), A and C (all $p$ values > .2), and the comparison between A and D (complexA1+A2 vs. complexD1+D2) in the left anterior ($p = 0.192$) and the left posterior Roi ($p = 0.072$). In the late time window (450-800ms), an effect of CONTEXT was found ($F (3,72) = 14.13; p < .001$), and an interaction of ROI x CONTEXT: ($F (9,216) = 3.78; p < .001$). In each Roi, consistent differences were restricted to the comparison between B and C (simpleB1+B2,underinformative vs. simpleC1+C2,false), as well as between C and D (simpleC1+C2,false vs. complexD1,false+D2,informative) (all $p$ values > .05, except for the comparison between C and D in the right anterior Roi ($p = 0.14$). None of the other comparisons was significant (all $p$ values > .072).

In line with our hypotheses, simple true but underinformative conditions (simpleB1+B2) exhibited the most reduced N400-positivity pattern as opposed to simple false conditions (simpleC1+C2). Descriptively, conditions following the complex context A (complexA1,underinformative+A2,informative) were in between simple true and simple false conditions, though these conditions did not differ reliably from simple false conditions. Interestingly, conditions following the complex context D (complexD1,informative+D2,false) did not significantly differ from simple true conditions. The difference between the complex conditions A and D was not expected under our simple prediction-based model. We will discuss this issue below.

Shape noun

Figures 7 and 8 show ERPs from the onset of the shape noun. The analyses did not show any effects of
SHAPE for the simple conditions, in line with our hypotheses (Figure 7). Moreover, given that we found interactions between GROUP and the experimental factors for the complex conditions, we show ERPs for semantic responders and pragmatic responders separately for these conditions (Figure 8).

![Figure 7: ERPs from the onset of the shape noun for the simple conditions. In the current example, the left shape corresponds to the circle whereas the right shape corresponds to the square.](image)

In the early time window, the ANOVA with the factors GROUP, ROI, CONTEXT und SHAPE revealed an interaction between GROUP x CONTEXT (F (3,69) = 4.21; p < .01), as well as an interaction between GROUP, CONTEXT and SHAPE (F (3,69) = 4.16; p < .01). Whereas semantic responders neither showed a CONTEXT x SHAPE interaction (p > .29) nor a significant ROI x CONTEXT x SHAPE interaction (p = .055), pragmatic responders did (CONTEXT x SHAPE: F (3,36) = 4.48; p < .01; ROI x CONTEXT x SHAPE: F (9,108) = 1.99; p < .05). For semantic responders, an effect of CONTEXT was found (F (3,33) = 9.32; p < .001), with significant differences for B and D (simpleC1+B1,underinformative vs. complexD1,false+D2,informative): F (1,11) = 26.34; p < .01), and C and D (simpleC1+C2,false vs. complexD1,false+D2,informative): F (1,11) = 12.08; p < .05; all other p values > .18). Moreover, a ROI x SHAPE interaction was observed (F (3,33) = 4.08; p < .05), but the levels of SHAPE did not differ in neither ROI (all p values > .27). As the three-way ROI x CONTEXT x SHAPE interaction reached significance for semantic responders, we also looked at the CONTEXT x SHAPE interaction in the single Rois. However, the interaction did not reach significance in neither of the Rois (all p values > .13). For the pragmatic responders, the interaction between CONTEXT
SHAPE was significant in the right posterior Roi \( (F(3,36) = 5.35; p < .05) \), and marginally significant in the right anterior \( (F(3,36) = 3.94; p = 0.064) \) and left posterior Rois \( (F(3,36) = 3.80; p = 0.072) \). In each of these Rois, an effect of SHAPE was obtained for context A: Condition complex\(_A1\), underinformative was significantly more negative than condition complex\(_A2\), informative \( (all \ p values > .05) \). For condition D, a similar effect was marginally significant in the right posterior Roi: Condition complex\(_D1\), false was more negative than condition complex\(_D1\), informative \( (F(1,12) = 4.36; p = 0.059) \). None of the other conditions showed any effects \( (all \ p values > .14) \).

**Figure 8:** ERPs from the onset of the shape noun. In the current example, the left shape corresponds to the circle whereas the right shape corresponds to the square.

In the late time window, no GROUP interactions were found. An effect of CONTEXT was found \( (F(3,72) = 26.46; p < 0.001) \), with significant differences \( (p < .01) \) for each of the single comparisons except for the two complex conditions A vs. D \( (\text{complex}_{A1+A2} \text{ vs. complex}_{D1+D2}; p = .20) \) and the two
simple conditions B vs. C (simpleB1,B2 vs. simpleC+D; \(p = .20\) \(p = .10\)). Moreover, the ANOVA yielded a statistically reliable main effect of SHAPE \((F(1,24) = 19.21; p < .001)\), and an interaction of ROI x CONTEXT \((F(9,216) = 8.69; p < .001)\), showing more consistent effects between the single conditions in the posterior Rois. Finally, the interaction between CONTEXT x SHAPE was significant \((F(3,72) = 7.37; p < .001)\): Significant effects of SHAPE were found for the complex contexts A (complexA1,underinformative vs. complexA2,informative: \(F(1,24) = 9.79; p < .01\)) and D (complexD1,false vs. complexD2,informative: \(F(1,24) = 30.14; p < .001\)), showing that both underinformative and false conditions were more positive in the complex conditions. No effects were found for the simple false and simple true contexts B and C (each \(p > .56\)).

Finally, in order to confirm the descriptive finding that semantic and pragmatic responders exhibited a discrepant behavior with respect to semantically false and pragmatically infelicitous sentences in the early (300-400 ms) time window, we carried out an additional analysis. Specifically, we assessed whether context A (pragmatically infelicitous) in contrast to context D (semantically false) produced differential group effects. These analyses confirmed that the two participant groups behaved differently with respect to the pragmatically felicitous vs. infelicitous conditions (context A: GROUP x SHAPE interaction: \(F(1,23) = 6.23; p < .05\)), but not for the semantically true vs. false conditions (context D: GROUP x SHAPE interaction: \(F(1,23) = 0.01; p = .93\)).

In sum, as expected, no effects of SHAPE were observed for the simple conditions in the early and late time window. With regard to the complex conditions, participant groups exhibited a different behavior: Whereas logically false vs. true sentences differed for semantic and pragmatic responders alike, the pragmatic infelicity was registered exclusively by the pragmatic responders.

5. Discussion

Using a picture-sentence verification paradigm, the current ERP experiment examined the processing of sentences containing the scalar quantifier *some* over the course of the utterance. ERPs were measured at different critical positions in the sentence. We formulated an explicit model of predictive processing,
in which the N400 component is linked to the predictability of incoming words. In that model, comprehenders form semantic expectations about contextual truth alone (semantic responders) or pragmatic expectations about contextual truth and pragmatic felicity combined (pragmatic responders). Pragmatic processing indeed affected the N400. Moreover, we replicated findings of different participant groups based on behavioral responses (Noveck & Posada, 2003; Politzer-Ahles et al., 2013; Hunt et al., 2013; Spychalska et al., 2016) and showed that ERPs differed between groups on late sentence positions. As a whole, the present pattern of results is explained well, though not perfectly, by our probabilistic pragmatic processing model. We will now address the results on the different sentential positions in turn.

5.1. Processes on the quantifier

By measuring ERPs from quantifier onset, we examined brain responses related to pragmatically infelicitous uses of some at the earliest possible position in the sentence. In three out of four time windows, underinformative readings vs. informative readings elicited a positivity for semantic and pragmatic responders alike. Though the model predicts differences between the two readings, the finding of a positivity instead of an N400 is unexpected under the so-far assumed link hypothesis that associates strength of expectations inversely with the N400 amplitude. This raises crucial theoretical questions, not least after the functional interpretation of the positivity in response to pragmatic violations. To our knowledge, only three studies have investigated implicature processing on the quantifier, and each of these studies revealed distinct neurophysiological correlates related to implicature processing.

First, Hartshorne et al. (2014) contrasted declarative clauses like Addison ate some of the cookies before breakfast this morning, and the rest are on the counter with conditional clauses like If Addison ate some of the cookies before breakfast this morning, then the rest are on the counter. According to what they call a higher-level account, implicature calculation depends on the relative informativity of an utterance. Due to this grammatical context-dependency, implicatures should only be calculated in declarative main clauses but not in antecedents of conditionals, and ERPs from the onset of the rest support this idea: the authors found a sustained positivity on the rest in conditional sentences when compared against the declarative sentences. Yet ERPs from quantifier onset did not differ between
declarative clauses and conditionals. While the pragmatic-expectations approach explored here does not explain the sustained positivity on the rest, the absence of an effect on the quantifier in Hartshorne et al.’s study is compatible with the position endorsed here, if we may assume, as seems plausible, that the context prior to the utterance of some provides little evidence for or against a subsequent mentioning of the scalar implicature trigger.

Similar to the present study, Politzer-Ahles et al. (2013) measured ERPs from quantifier onset following pictures supporting an informative or an underinformative reading and found that ERPs differed already at the clause-initial quantifier. However, in contrast to our study, a sustained negativity was observed in response to underinformative sentences. According to the authors, this effect was due to an effortful retrieval of the literal interpretation after inferences had already been calculated on the picture. However, such a pragmatic-revision account cannot explain why we do not observe a similar sustained negativity but an N400 in Spychalska et al. (2016), where the clause-final noun also directly followed the presented picture, and, arguably, inferences had also been calculated before the critical position was reached. Generally, it seems as if underinformative readings at advanced positions stably elicit N400 components across studies, while effects on the quantifier position are more difficult to unify. At present, we can only speculate about potential reasons for the qualitatively distinct results on the quantifier position. Differences between studies include the language under investigation (i.e. Chinese in Politzer-Ahles et al. vs. German in the present study), the number of pictorial contexts (two in Politzer-Ahles, four in the present study), and the level of abstraction of the linguistic and pictorial contexts. It is unclear which of these differences might have elicited the distinct brain response to the pragmatic infelicity.

Finally, Barbet & Thierry (2016) examined the processing of multiple quantifiers on the single-word level in an oddball task. The physical properties of the stimulus (i.e. whether some or all of the letters of a word were printed in green, with the remaining letters printed in white) did or did not correspond to the quantifier meaning. For instance, some presented exclusively in green color corresponded to the underinformative reading of some. Whether the underinformative reading was to be interpreted as a match or as a mismatch was specified by instructions and was varied between blocks. Participant groups were determined via acceptability judgments and other measures such as Empathy scores and
the Autism-Spectrum Quotient (p.4). The P3b response to *some* was reduced relative to an increase in participants’ pragmatism scores. The less the literal reading was accepted, the stronger was the P3b reduction. Interestingly, this effect was restricted to those cases in which the standards were mismatches and deviant targets were matches. However, as quantifiers were presented on the single-word level in that study and occurred in a specific experimental paradigm, a P3b component was clearly expected due to task-inherent demands, and these results do therefore not necessarily generalize to other paradigms.

To sum up, if ERPs on the quantifier position reflect met or unmet expectancies based on the picture information, then it remains difficult to explain why underinformative conditions elicited a positivity instead of an N400 in the present experiment. One possible explanation for the positivity would be that in the current study, the quantifier itself was processed with increased attentional demands, thus leading to an enhanced P3b component (Sassenhagen, Schlesewsky & Bornkessel-Schlesewsky, 2014). While, for instance, the combined felicity and truth-value judgements Politzer-Ahles et al. (2013) only hinged on the quantifier, there are three critical positions in the to-be-processed utterance in our study. Moreover, whereas in Politzer-Ahles et al., informative sentences remained so over the course of the utterance, part of the present sentences involved a shift from locally informative to underinformative. The possibility of such a meaning shift might have involved increased attentional resources compared to previous studies.

An alternative interpretation of the positivity comes into mind when considering another area of pragmatic reasoning, namely the field of presupposition processing. In a recent study, Jouravlev et al. (2016) found a positivity for the processing of presupposition triggers such as *again* in sentences pairs like *Jake had never tipped a maid at the hotel before. Today he tipped a maid at the hotel again, although the hotel paid its maids good wages* when compared to pragmatically unmarked control conditions. The positivity was interpreted as a combined P3b/P600 component that signals a context update of the current mental discourse model due to an incoming pragmatic cue (p. 8). The positivity in the present study might thus be interpreted as reflecting a similar context update based on local inference generation.

---

4 As pointed out by an anonymous reviewer, one possibility to explain the missing N400 pointed out would be that the expectancy of *some* was generally high in the present experiment, in which *some* occurred in 50% of all instances as the first word in the sentence.
If an explanation along these lines is correct, then processes at the position of local inference generation might generally differ from expectancy-related effects that already involve the result of this local inference generation step.

5.2. Processes on the color adjective

Whereas literally true sentences elicited the most reduced N400, literally false sentences elicited the largest effect, with effects of the mixed context A in between these extremes, irrespective of semantic or pragmatic strategies. This is as predicted by our pragmatic expectations account (see also Bott, Augurzky, Sternefeld & Ulrich, 2017; Spychalska et al., 2016), but may be surprising for accounts that expect ERP modulations based on underinformativity alone. In simple contexts B and C, after hearing Some of the dots are…., the assumption that the sentence will be continued by a color adjective that makes the sentence true fixes a strong expectation for blue in context B and red in context C. Moreover, we can explain why the complex context A should result in an intermediate effect. After processing the initial sentence fragment Some of the dots are…., the color adjective blue makes for a true and felicitous description of the picture. Encountering blue is thus neither as surprising as the false blue in context C, nor as predictable as the uniquely true blue in context B. The finding of an N400 as an index of predictions is in line with current studies that interpret the N400 as an index of lexical predictions (DeLong, Urbach & Kutas, 2005; Ito, Corley, Pickering, Martin, & Nieuwland, 2016; Otten, Nieuwland, & van Berkum, 2007). In the present experiment, the N400 was accompanied by a late positivity. Similar effects have been interpreted as reflecting logical-semantic revision (Spychalska et al., 2016) or disconfirmed expectations (Hunt et al., 2013). Recently, it has been suggested that positivities like the current one may be linked to the sentence-verification paradigm itself and reflect task-related effects rather than semantic or pragmatic processing per se. In particular, both Nieuwland (2016) and Augurzky et al. (2016) compared quantifier processing in a sentence verification task and another task that did not require explicit verification. In both studies, the late positivity disappeared in the alternative task. The

5 Though visual inspection suggests that the positivity has a slightly more frontal distribution than the P3b/P600, the interaction with the factor Roi reached significance only in the 300-400 ms time window.
positivity thus could be an instance of the P3b component that reflects increased attentional demands related to information that is highly relevant for task fulfillment (Sassenhagen et al., 2014).

A slightly puzzling observation remains. Our model predicts that blue should be equally expected in contexts A and D, but we observed that the N400 component in condition D was reduced in comparison to that in condition A. If our link hypothesis that the N400 amplitude tracks contextual expectability of a word is correct, the observed finding implies that participants should have considered blue a more likely continuation in context D than in context A. It then remains an open issue what may have caused these expectations. One possibility is that the smaller number of blue dots sticks out perceptually, so that this effect is due simply to something like perceptually induced salience. Another possible explanation is that some is a better description of 3 out of 10 than it is of 7 out of 10. Indeed, recent work on the naturalness or typicality of some for different cardinalities suggests that naturalness ratings of sentences like Some of the dots are blue depend on the number/proportion of blue balls in a presented display (van Tiel, 2014, Degen & Tanenhaus, 2015). It remains an open empirical issue as to whether what explains offline naturalness or typicality ratings also could affect processing-related measures like ERP responses.

It is also interesting to note at this point, a potential difference in qualitative predictions between the simple model we used so far and the more complex, full idealized incremental interpreter model from Section 4. In a nutshell, the simple model predicts no difference between conditions A and D (A = D), but the empirical observation seems to be that blue is more surprising in A than in D (A > D), while the complex model would likely predict the opposite tendency (A < D). Recall that the main difference is that the simple model derives expectations from whether there exist certain categorical types of sentence continuations, the complex model computes a weighted sum over all continuations. This will likely lead to different predictions for the case at hand. In both contexts A and D, there is exactly one true and infelicitous continuation of Some of the dots are blue (in the set of relevant sentences considered so far), namely circle. But while the continuation with square yields a true and infelicitous sentence for context A, it yields a false sentence for context D. This means that, by weighing over all continuations, the full Bayesian model would predict that blue is more expected in context A
than in context D, contrary to what we observe and what the simple model predicts. Of course, much depends on how the ideal incremental interpreter model is set up precisely (e.g., the speaker likelihood function, the set of relevant sentences etc.), and the full Bayesian model too would be able to include additional assumptions about salience of color and/or typicality of quantifiers. We must leave a full exploration of the complex model and an in-depth comparison to future work, and restrict ourselves to the observation that simple and complex models may likely make different predictions which are empirically testable and that, where they do in the present experiment, the complex model is not obviously better than the simple one.

5.3. Processes on the shape noun

On the shape noun, underinformative conditions and semantically false conditions elicited an N400 that was followed by a late positivity. However, the effect was restricted pragmatic responders.

As expected, no effects were found for simple conditions, as each of the two shapes could be equally predicted based on the pictorial information. With regard to the complex conditions, an interesting discrepancy between semantically-driven and pragmatically-driven predictions arises. According to a semantically-driven processing strategy, literally true outcomes were predicted at each sentence position. Therefore, no effects were expected on the shape noun for underinformative conditions, as both shapes are equally compatible with a semantically true outcome. This expectation was confirmed by our data. By contrast, according to a pragmatically-driven processing strategy, semantically and pragmatically unmarked outcomes were predicted, and therefore lexical predictions included the bicolored shape in condition A, resulting in an increased N400 when the unicolored shape was encountered.

5.4. Summary and outlook

In sum, under a prediction-based account, the observed effects can be linked to participants’ expectations given the picture and the current sentence input. Though a straightforward version of a prediction-based account can capture most of the observed effects, it does not capture all the subtle effects found here. Especially the differences on the color adjective between complex conditions are not anticipated
by our model. Also, it remains an open issue as to which ERP components the model predictions should be linked to (late positivity for sentence-initial quantifiers vs. early N400 amplitude at advanced sentential positions).

It has to be noted that the current results are principally also compatible with a lexical-integration account of the N400 (e.g. Hagoort, 2008). According to this view, the N400 signals bottom-up processes associated with postlexical processing stages rather than the pre-activation of lexical features involved in predictive processing (see Szewczyk & Schriefers, 2018, for a recent overview). As a consequence, the lexical properties of an incoming word have to be fully processed before it can be integrated into the already generated sentence representation. By contrast, the predictive approach considers the N400 amplitude as an index of the ease of lexical access given a certain amount of lexical pre-activation, a process that does not necessarily involve a full-fledged semantic representation of the sentence so far.

Though our study was not specifically designed to distinguish between these approaches, the observed pattern of results on the color adjective seems unexpected under a lexical-integration approach. There are at least three possible ways of specifying the lexical-integration approach for our case study, which differ in the way they treat the meaning representation of the initial sentence fragment. First, we could assume that the initial sentence fragment “Some of the dots are…” is interpreted just semantically, without any pragmatic enrichments, at the stage of lexical-integration. In that case, all conditions A, B and D should be equally easy to integrate, for both semantic and pragmatic responders. This, however, is contradicted by the data. Secondly, we could assume that the lexical enrichment of “some” to some-but-not-all is already integrated into the meaning representation of the initial sentence fragment. In that case, the lexical-integration approach would predict that, again, all conditions A, B and D would be equally easy to integrate, contrary to observation. Even if we assumed, as a third possibility, that only in context A and D the initial sentence fragment includes the pragmatically enriched reading of “some” already (because it is supported by the visual context) would the integration approach predict equality between conditions A, B and D. As a consequence, under all three construals of the integration view, the conditions A, B, and D should all elicit comparably reduced N400 amplitudes on the adjective, as they can be locally integrated as true sentences. This pattern is contrary to our observations. Finally, yet
another option would be that the semantic integration on the adjective position requires some local evaluation of the semantic and pragmatic well-formedness, despite the fact that the pragmatic anomaly had already been registered before. Under such a version of the integration view, at least for the pragmatic responders, the true but underinformative condition B would have been expected to be more difficult to integrate than the complex conditions A and D, as it contains an additional pragmatic infelicity. In sum, while the prediction-based account explored in this paper also does not explain all of the facts, the integration-based approach does not appear to fare any better.

Finally, it is an open issue whether our observed effects could also be obtained in different and possibly more realistic settings. In particular, the complexity of the current design did not allow us to include further plausible continuations like Some dots are in the circle, which would allow for an informative sentence continuation even for the simple contexts B and C. As a matter of fact, we consider it plausible that the positivity on the quantifier might be affected by this possibility, or that it might even disappear as soon as a larger number of plausible salient continuations are present in the discourse (see Grodner, Klein, Carbary, & Tanenhaus, 2010, for a study on salient alternatives on the processing of some, and see Augurzky, Schlotterbeck & Ulrich, 2016, for a study on proportional frequency effects on the quantifier all). Thus, more complex models and the comparison of different and more naturalistic experimental settings would offer an interesting test case for future studies.

6. Conclusions

The present ERP study employed picture-sentence verification to examine implicature processing over the course of the utterance. By measuring ERPs at different sentential positions, we tested an explicit model of predictive processing, in which ERP components are linked to the predictability of incoming words derived by comprehenders’ expectations about contextual truth and pragmatic felicity. The current results indicate that pragmatic expectations at different positions may modulate the amplitude of the N400, a component that has been associated with predictive processing. Expectancy-driven

6 We thank an anonymous reviewer for pointing out this issue.
processes affected on-line processing for all participants, but differed subtly depending on the response type, as manifest in truth-value judgements. The current studies are thus compatible with prediction-based pragmatic approaches, in which comprehenders actively predict on-going sentences by considering alternative expressions suitable for pragmatic interpretation. Interestingly, the present results also demonstrate that model predictions may differ as a function of how the ideal incremental interpreter model is set up precisely (e.g., the speaker likelihood function, the set of relevant sentences etc.), and can thus be seen as starting point for future approaches considering different modeling assumptions and including a larger range of observed pragmatic phenomena.

7. Acknowledgements

This research was funded by the German Science Foundation (DFG), in the projects B1 of the SFB833 The Construction of Meaning, and the Institutional Strategy of the University of Tuebingen (ZUK 63). We thank Oliver Bott, Robin Hörnig, Fabian Schlotterbeck and Wolfgang Sternefeld for valuable discussion, and Nadine Balbach, Anna Hänle, Natalie Leschinski and Helena Schütze for assistance during data acquisition. The authors do not have any conflicts of interest.

References


Augurzky et al.: “Gricean expectations in online sentence comprehension”


Franke, M., & Jäger G. (2016). Probabilistic pragmatics, or why Bayes’ rule is probably important for pragmatics. Zeitschrift für Sprachwissenschaft, 35:1, 3-44.


Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). “Some”, and possibly all,


Augurzky et al.: “Gricean expectations in online sentence comprehension”


Augurzky et al.: “Gricean expectations in online sentence comprehension”


Appendix: Detailed statistical results

For the sake of clarity, we present statistical results by referring to the condition labels we used in the ANOVAs. For the convenience of the reader, we provide the functional labels corresponding to the condition labels in Table A.

### Table A: Conditions

<table>
<thead>
<tr>
<th>Context</th>
<th>Shape</th>
<th>Functional label</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>left</td>
<td>complexA1, underinformative</td>
</tr>
<tr>
<td></td>
<td>right</td>
<td>complexA2, informative</td>
</tr>
<tr>
<td>B</td>
<td>left</td>
<td>simpleB1, underinformative</td>
</tr>
<tr>
<td></td>
<td>right</td>
<td>simpleB2, underinformative</td>
</tr>
<tr>
<td>C</td>
<td>left</td>
<td>simpleC1, false</td>
</tr>
<tr>
<td></td>
<td>right</td>
<td>simpleC2, false</td>
</tr>
<tr>
<td>D</td>
<td>left</td>
<td>complexD2, false</td>
</tr>
<tr>
<td></td>
<td>right</td>
<td>complexD2, informative</td>
</tr>
</tbody>
</table>

A1: Results of the ANOVAs from quantifier onset

**300-400 ms**
- **Complexity**: $F (1,24) = 6.59; p < .05$
- **ROI x Complexity**: $F (3,72) = 4.71; p < .01$

<table>
<thead>
<tr>
<th></th>
<th>Left anterior</th>
<th>Right anterior</th>
<th>Left posterior</th>
<th>Right posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Complexity</strong></td>
<td>$F (1,24) = 12.96^{**}$</td>
<td>$F (1,24) = 7.64^{*}$</td>
<td>$F (1,24) = 2.8$</td>
<td>$F (1,24) = 2.9$</td>
</tr>
</tbody>
</table>

- All further effects $p > .27$

**450-800 ms**
- **Complexity**: $F (1,24) = 9.68; p < .01$
- All further effects $p > .31$

**500-1000 ms**
- **Complexity**: $F (1,24) = 7.83; p < .01$
- **GROUP x Complexity**: $F (1,23) = 5.59; p < .05$
  - Semantic responders: **Complexity**: $F (1,11) = 0.30; p = .59$
  - Pragmatic responders: **Complexity**: $F (1,12) = 10.46; p < .01$
- All further effects $p > .11$

---

7 Only significant interactions (p<.05) were resolved.
Augurzky et al.: “Gricean expectations in online sentence comprehension”

300-1000 ms
- **COMPLEXITY**: $F(1,24) = 7.83; p < .01$
- All further effects $p > .13$

A2: Results of the ANOVAs from adjective onset

ANOVA with the factors **GROUP**, **ROI** and **CONTEXT**:
- No interactions between **GROUP** and any other factor.

300-400 ms (N400)
- **CONTEXT**: $F(3,72) = 14.74; p < .001$
  - A vs. B: $F(1,24) = 16.46; p < .001$
  - A vs. C: $F(1,24) = 4.86; p = 0.22$
  - A vs. D: $F(1,24) = 12.72; p < .05$
  - B vs. C: $F(1,24) = 28.01; p < .001$
  - B vs. D: $F(1,24) = 0.16; p = 1$
  - C vs. D: $F(1,24) = 19.6; p < .001$
- **ROI x CONTEXT**: $F(9,216) = 5.60; p < .001$

<table>
<thead>
<tr>
<th>300-450 ms</th>
<th>Left anterior</th>
<th>Right anterior</th>
<th>Left posterior</th>
<th>Right posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CONTEXT</strong></td>
<td>$F(3,72) = 8.9***$</td>
<td>$F(3,72) = 12.08***$</td>
<td>$F(3,72) = 15.78***$</td>
<td>$F(3,72) = 16.66***$</td>
</tr>
<tr>
<td>A vs. B</td>
<td>$F(1,24) = 7.03; p = 0.336$</td>
<td>$F(1,24) = 14.28*$</td>
<td>$F(1,24) = 14.56*$</td>
<td>$F(1,24) = 24.08***$</td>
</tr>
<tr>
<td>A vs. C</td>
<td>$F(1,24) = 2.98; p = 1$</td>
<td>$F(1,24) = 1.65; p = 1$</td>
<td>$F(1,24) = 7.61; p = 0.264$</td>
<td>$F(1,24) = 5.77; p = 0.576$</td>
</tr>
<tr>
<td>A vs. D</td>
<td>$F(1,24) = 8.41; p = 0.192$</td>
<td>$F(1,24) = 15.15*$</td>
<td>$F(1,24) = 10.79; p = 0.072$</td>
<td>$F(1,24) = 12.07*$</td>
</tr>
<tr>
<td>B vs. C</td>
<td>$F(1,24) = 18.38***$</td>
<td>$F(1,24) = 19.99***$</td>
<td>$F(1,24) = 31.26***$</td>
<td>$F(1,24) = 32.12***$</td>
</tr>
<tr>
<td>B vs. D</td>
<td>$F(1,24) = 0.16; p = 1$</td>
<td>$F(1,24) = 0.09; p = 1$</td>
<td>$F(1,24) = 0.09; p = 1$</td>
<td>$F(1,24) = 0.23; p = 1$</td>
</tr>
<tr>
<td>C vs. D</td>
<td>$F(1,24) = 12.97*$</td>
<td>$F(1,24) = 17.83***$</td>
<td>$F(1,24) = 20.1***$</td>
<td>$F(1,24) = 21.68***$</td>
</tr>
</tbody>
</table>

450-800 ms (Positivity)
- **CONTEXT**: $F(3,72) = 14.13; p < .001$
  - A vs. B: $F(1,24) = 10.09; p < .05$
  - A vs. C: $F(1,24) = 4.36; p = 0.288$
  - A vs. D: $F(1,24) = 4.47; p = 0.27$
  - B vs. C: $F(1,24) = 67.89; p < .001$
  - B vs. D: $F(1,24) = 3.72; p = 0.396$
  - C vs. D: $F(1,24) = 22.87; p < .001$
- **ROI x CONTEXT**: $F(9,216) = 3.78; p < .001$
A3: Results of the ANOVAs from shape noun onset

ANOVA with the factors GROUP, ROI, CONTEXT and SHAPE:

- Interactions with GROUP in the early time window

300-400 ms (N400)

GROUP X CONTEXT: $F(3, 69) = 4.21; p < .01$

GROUP, CONTEXT X SHAPE: $F(3, 69) = 4.16; p < .01$

→ separate analyses for semantic and pragmatic responders

Semantic responders:

- CONTEXT: $F(3, 33) = 9.32; p < .001$
  
  A vs. B: $F(1, 11) = 5.92; p = 0.199$
  A vs. C: $F(1, 11) = 3.86; p = 0.452$
  A vs. D: $F(1, 11) = 6.16; p = 0.183$
  B vs. C: $F(1, 11) = 0.03; p = 1$
  B vs. D: $F(1, 11) = 26.34; p < .01$
  C vs. D: $F(1, 11) = 12.08; p < .05$

- ROI X SHAPE: $F(3, 33) = 4.08; p < .05$

300-400 ms

<table>
<thead>
<tr>
<th>SHAPE</th>
<th>Left anterior</th>
<th>Right anterior</th>
<th>Left posterior</th>
<th>Right posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F(1, 11) = 1.378; p = 1$</td>
<td>$F(1, 11) = 0.175; p = 1$</td>
<td>$F(1, 11) = 2.892; p = 0.468$</td>
<td>$F(1, 11) = 4.106; p = 0.272$</td>
</tr>
</tbody>
</table>

Pragmatic responders:

- CONTEXT: $F(3, 36) = 19.64; p < .001$
  
  A vs. B: $F(1, 12) = 15.75; p < .05$
  A vs. C: $F(1, 12) = 23.56; p < .01$
  A vs. D: $F(1, 12) = 0.03; p = 1$
  B vs. C: $F(1, 12) = 2.54; p = 0.821$
  B vs. D: $F(1, 12) = 21.64; p < .01$
  C vs. D: $F(1, 12) = 31.48; p < .001$
Augurzky et al.: “Gricean expectations in online sentence comprehension”

- **Context X Shape**: \( F(3,36) = 4.478; p < .01 \)
  - A: Effect of Shape: \( F(1,12) = 7.17; p < .05 \)
  - B: Effect of Shape: \( F(1,12) = 1.282; p = 0.28 \)
  - C: Effect of Shape: \( F(1,12) = 1.832; p = 0.201 \)
  - D: Effect of Shape: \( F(1,12) = 1.643; p = 0.224 \)

- **ROI X Context**: \( F(9,108) = 6.903; p < .001 \); Interaction **ROI X Context X Shape**: \( F(9,108) = 1.991; p < .05 \)

<table>
<thead>
<tr>
<th>Time Window</th>
<th>Left Anterior</th>
<th>Right Anterior</th>
<th>Left Posterior</th>
<th>Right Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>300-400 ms</td>
<td><strong>F(3,36) = 13.27</strong>***</td>
<td><strong>F(3,36) = 13.15</strong>***</td>
<td><strong>F(3,36) = 23.3</strong>***</td>
<td><strong>F(3,36) = 26.184</strong>***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context</th>
<th>A vs. B</th>
<th>A vs. C</th>
<th>A vs. D</th>
<th>B vs. C</th>
<th>B vs. D</th>
<th>C vs. D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( F(1,11) = 9.97; p = 0.198 )</td>
<td>( F(1,11) = 18.* )</td>
<td>( F(1,11) = 0.00; p = 1 )</td>
<td>( F(1,11) = 3.12; p = 1 )</td>
<td>( F(1,11) = 12.07; p = 0.11 )</td>
<td>( F(1,11) = 20.24* )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Shape</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( F(1,12) = 6.33* )</td>
<td>( F(1,12) = 5.49* )</td>
<td>( F(1,12) = 0.49; p = 0.497 )</td>
<td>( F(1,12) = 9.79; p &lt; .01 )</td>
</tr>
</tbody>
</table>

**450-500 ms (Positivity)**

- No Group interactions
- **Context**: \( F(3,72) = 26.46; p < .001 \)
  - A vs. B: \( F(1,24) = 17.43; p < .01 \)
  - A vs. C: \( F(1,24) = 30.07; p < .001 \)
  - A vs. D: \( F(1,24) = 5.08; p = 0.201 \)
  - B vs. C: \( F(1,24) = 6.74; p = 0.095 \)
  - B vs. D: \( F(1,24) = 31.30; p < .001 \)
  - C vs. D: \( F(1,24) = 41.65; p < .001 \)
- **Shape**: \( F(1,24) = 19.21; p < .001 \)
- **ROI X Context**: \( F(9,216) = 8.69; p < .001 \)
- **Context X Shape**: \( F(3,72) = 7.37; p < .001 \)
  - A: **Shape**: \( F(1,24) = 9.79; p < .01 \)
  - B: **Shape**: \( F(1,24) = 0.03; p = 0.875 \)
  - C: **Shape**: \( F(1,24) = 0.34; p = 0.564 \)
  - D: **Shape**: \( F(1,24) = 30.14; p < .001 \)
Augurzky et al.: “Gricean expectations in online sentence comprehension”

<table>
<thead>
<tr>
<th>300–400 ms</th>
<th>Left anterior</th>
<th>Right anterior</th>
<th>Left posterior</th>
<th>Right posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CONTEXT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>F</em> (3,72) = 11.64***</td>
<td><em>F</em> (3,72) = 11.51***</td>
<td><em>F</em> (3,72) = 39.08***</td>
<td><em>F</em> (3,72) = 38.81***</td>
</tr>
<tr>
<td><strong>SHAPE</strong></td>
<td><em>F</em> (1,24) = 18.37***</td>
<td><em>F</em> (1,24) = 15.34**</td>
<td><em>F</em> (1,24) = 14.10***</td>
<td><em>F</em> (1,24) = 12.61**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CONTEXT</th>
<th>A vs. B</th>
<th>A vs. C</th>
<th>A vs. D</th>
<th>B vs. C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>F</em> (1,24) = 8.173; <em>p</em> = 0.208</td>
<td><em>F</em> (1,24) = 10.99; <em>p</em> = 0.07</td>
<td><em>F</em> (1,24) = 19.65*</td>
<td><em>F</em> (1,24) = 26.20***</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>F</em> (1,24) = 10.69; <em>p</em> = 0.078</td>
<td><em>F</em> (1,24) = 16.79*</td>
<td><em>F</em> (1,24) = 50.93***</td>
</tr>
<tr>
<td></td>
<td><em>F</em> (1,24) = 3.88; <em>p</em> = 1</td>
<td><em>F</em> (1,24) = 0.06; <em>p</em> = 1</td>
<td><em>F</em> (1,24) = 11.70; <em>p</em> = 0.054</td>
<td><em>F</em> (1,24) = 4.31; <em>p</em> = 1</td>
</tr>
<tr>
<td></td>
<td><em>F</em> (1,24) = 2.73; <em>p</em> = 1</td>
<td><em>F</em> (1,24) = 1.96; <em>p</em> = 1</td>
<td><em>F</em> (1,24) = 9.08; <em>p</em> = 0.145</td>
<td><em>F</em> (1,24) = 6.29; <em>p</em> = 0.476</td>
</tr>
<tr>
<td></td>
<td><em>F</em> (1,24) = 15.6*</td>
<td><em>F</em> (1,24) = 10.99; <em>p</em> = 0.07</td>
<td><em>F</em> (1,24) = 45.01***</td>
<td><em>F</em> (1,24) = 43.89***</td>
</tr>
<tr>
<td></td>
<td><em>F</em> (1,24) = 15.30*</td>
<td><em>F</em> (1,24) = 13.11; <em>p</em> &lt; .05</td>
<td><em>F</em> (1,24) = 73.42***</td>
<td><em>F</em> (1,24) = 65.72***</td>
</tr>
</tbody>
</table>