Cognition 166 (2017) 328–343

Contents lists available at ScienceDirect

Cognition

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

Original Article

Incremental generation of answers during the comprehension of questions with quantifiers

Oliver Bott [a,b,*], Petra Augurzky [b], Wolfgang Sternefeld [b], Rolf Ulrich [b]

[a] *Project CiC, XPrag.de, University of Tübingen, Germany*
[b] *Project B1, SFB 833, University of Tübingen, Germany*

ARTICLE INFO

Article history:
Received 14 July 2016
Revised 8 May 2017
Accepted 10 May 2017

Keywords:
Incrementality
Quantifier interpretation
Quantifier restriction
Context integration
Answer preparation
Probe RT task

ABSTRACT

The paper presents a study on the online interpretation of quantified questions involving complex domain restriction, for instance, *are all triangles blue that are in the circle*. Two probe reaction time (RT) task experiments were conducted to study the incremental nature of answer generation while manipulating visual contexts and response hand overlap between tasks. We manipulated the contexts in such a way that the incremental answer to the question changed from *'yes'* to *'no'* or remained the same before and after encountering the extraposed relative clause. The findings of both experiments provide evidence for incremental answer preparation but only if the context did not involve the risk of answer revision. Our results show that preliminary output from incremental semantic interpretation results in response priming that facilitates congruent responses in the probe RT task.

1. Introduction

Language comprehenders have the remarkable ability to automatically restrict the referential domain in such a way that it optimally fits the purposes of a conversation. They do so incrementally, in close contiguity with the linguistic input. In line with the incrementality assumption, a large number of studies have demonstrated that referential expressions are immediately resolved to their referent. This happens automatically in isolated sentences as well as in natural referential communication tasks with large numbers of potential referents (Allopenna, Magnuson, & Tanenhaus, 1998; Brown-Schmidt & Tanenhaus, 2008). Reference resolution is even so fast that it regularly leads to cohort effects well before a referring expression has been fully processed, e.g. hearing the onset of *beaker* activates *beetle*. The restriction of the referential domain is not only essential to establish reference to individual objects but is also required in order to restrict entire sets of objects, i.e. pluralities. Among the linguistic devices whose interpretation strongly depends on this sort of reference are quantificational expressions relating a restrictor set to another set, their nuclear scope (Barwise & Cooper, 1981; Montague, 1973). Usually, the restrictor set is contextually constrained (von Fintel, 1994). For instance, the quantifier *every* in (1) must be restricted to a contextually relevant set of students, most plausibly students taught by the professor, instead of taking the restrictor argument to refer to all students there have ever been.

(1) This professor annoys every student.

The present paper investigates the incremental nature of this update process. We report two experiments investigating whether the interpretation system immediately takes into account the context of utterance of a quantified question and immediately generates an answer to it. We systematically manipulated picture contexts (cf. Fig. 1) and showed them before the presentation of questions with a universal quantifier whose restrictor argument was further restricted postverbally by an extraposed relative clause, cf. (2). Crucially, this construction type involves the integration of a complex, discontinuous restrictor argument with truth-conditionally relevant content coming before and after the nuclear scope argument *blue*.

(2)	Sind	alle	Dreiecke	<u>blau</u>,	die	im	Kreis	sind?
	Are	all	triangles	blue	which	in the	circle	are?
	Are all triangles in the circle blue?							

[*] Corresponding author at: Project CiC, XPrag.de, University of Tübingen, Germany.

http://dx.doi.org/10.1016/j.cognition.2017.05.023
0010-0277/© 2017 Elsevier B.V. All rights reserved.

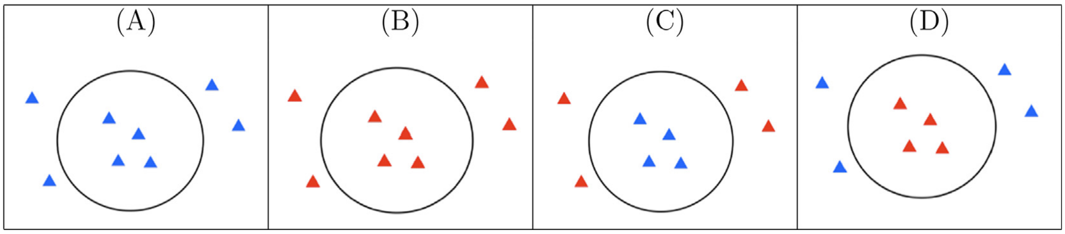
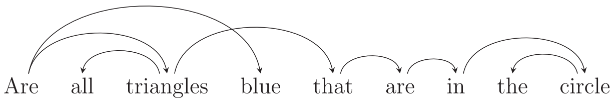


Fig. 1. Picture contexts for the sample item in (2). The answer in condition (A) is "yes, true" at the critical region, i.e. the underlined color adjective, and does not change with further restriction; The answer in (B) starts out "no, false" and does not change with further restriction; incremental interpretation of (C) should start out "no, false", but require a shift to "yes, true" after integrating the relative clause; (D) should start out "no, false" and stay the same until the end of the sentence. Visually, it also involves a complex scenario similar to (C).

Our main interest is the interpretation constructed mid-sentence, i.e. at the color adjective (e.g. *blue*). Crucially, only in some contexts the polarity of the answer can be safely computed already at this point. In other contexts (cf. condition C), the intermediate answer 'no' must later be revised to a 'yes' answer. Since such revision processes may incur processing costs, the present study addresses the question whether the processing system is sensitive to the risk of a potential shift in the polarity of the answer. We used a modifed version of the probe reaction time task (Logan, Zbrodoff, & Fostey, 1983; Miller, Coles, & Chakraborty, 1996; Posner & Boies, 1971) to determine whether the semantic processor incrementally decides on an answer to the unfolding question and whether it immediately takes into account the risk of answer revision depending on the properties of the context.

Sentences such as (2) pose a challenge to incremental semantic interpretation. When encountering the postverbal modification, comprehenders have to add information to the restrictor argument even though the restrictor argument must have already been saturated. Otherwise it would not be possible to incorporate lexical material into the nuclear scope.

To compute the compositional meaning, the processor must thus somehow keep or generate an open slot in order to fit in the upcoming restrictor information. This complexity is also reflected on the structural side. Extraposed relative clauses involve crossing word-word dependencies (see, e.g., Levy, Fedorenko, Breen, & Gibson, 2012). In (2), the auxiliary verb selects for a subject and a verb phrase. The latter dependency crosses another dependency, namely the one between the noun and the relative pronoun. This is illustrated in the following dependency graph:



In terms of phrase structures, crossing dependencies correspond to discontinuous constituents. Establishing these crossing dependencies has been found to create difficulty during comprehension but this difficulty turns out to be mediated by lexically driven syntactic expectations. Levy et al. (2012, Exp. 3) showed that manipulating the expectancy of a post-verbal modification can even neutralize the processing costs generally observed for extraposed post-verbal modifiers. Some determiner phrases (DPs) have determiners that trigger expectations for an upcoming restriction, such as *only those* in *only those executives*. Interestingly, these cases involve a partitive construction that partitions the set denoted by the noun into two subsets, a presupposed set in the denotation of the predicate and a set of alternatives that are incompatible with the predicate. In the Levy et al. study sentences were presented out of the blue, so the presupposed set might have been accommodated (in the sense of Lewis, 1979) upon encountering the DP. Alternatively, the DP could trigger the expectation for upcoming restriction. Semantic considerations along these lines motivate why we consider it highly important to investigate the influence of contextual information on the online interpretation of quantifier restriction.

In general, we may distinguish three alternative hypotheses on how the semantic processor constructs compositional interpretations for sentences such as (2) drawing on contextual information.

Global interpretation: The answer to a question is only generated after the complete question has been processed. Triggers for answer generation are intonational cues (spoken language) or a question mark (written language).

Incremental interpretation, revision-insensitive: An answer is generated at the earliest possible point no matter if there is danger of subsequent reanalysis.

Incremental interpretation, revision-sensitive: An answer is generated at the earliest possible point, but only if the context unambiguously determines its polarity.

At first sight, global interpretation of quantifiers may seem like a Strawman hypothesis. However, with respect to the processing of quantifiers it is actually not clear whether incremental effects should really be expected. A number of ERP studies have addressed incremental interpretation of quantifiers by investigating quantified sentences that do not correspond to the beliefs people hold about the actual world using the N400 as dependent measure. Opinions in the matter are divided ranging from non-incremental over partially incremental to fully incremental views on quantifier interpretation (Freunberger & Nieuwland, 2016; Kounios & Holcomb, 1992; Nieuwland, 2016; Urbach & Kutas, 2010; Urbach, DeLong, & Kutas, 2015). In the first study on quantifier comprehension, Kounios and Holcomb (1992) manipulated quantifier type (*all* vs. *some* vs. *no*) affecting the truth value of sentences such as *all/some/no rubies are gems*. They did not find an effect of quantifier type on the N400 elicited by the last word of the sentence, even though truth-value judgments indicated that participants had interpreted the sentences correctly. The authors concluded that quantifier interpretation is delayed and that the N400 does not reflect late compositional interpretation processes involving quantifiers (see also Fischler, Bloom, Childers, Roucos, & Perry's (1983) investigation of negated sentences like *a robin is (not) a bird* for a lack of truth-value effect). This was taken up by Urbach and Kutas (2010) manipulating quantifier type ('positive' vs. 'negative'; e.g., e.g., *most* vs. *few*) fully crossed with a truth-value manipulation (*most/few farmers grow crops/worms*). While positive quantifiers showed the expected N400 effect, negative quantifiers did not modulate the N400 in the expected direction but showed the opposite effect. End of sentence truth-value judgments were as expected for both quantifier types, though. This led the authors to the conclusion that quantifier interpretation proceeds 'partially incremental' with delayed processing of negative quantifiers. Nieuwland (2016) employed the same factorial design as Urbach and Kutas (2010) but additionally controlled for the expectedness of the critical word. For items with low cloze probability across quantifier conditions effects were similar to Urbach and Kutas (2010)'s study, but for items with a high cloze probability of the critical word, N400 effects were as predicted by the incrementality hypothesis. He concluded that quantifier interpretation can be fully incremental but only if there is sufficiently strong guidance by world-knowledge about typical situations for developing predictions about how the quantified statement will continue. The recent studies by Urbach et al. (2015) and Freunberger and Nieuwland (2016) point in a similar direction. If quantified sentences are embedded in a real-world context (Urbach et al., 2015) or their interpretation is supported by prosodic information in auditory presentation (Freunberger & Nieuwland, 2016), incremental N400 effects are observed for both positive and negative quantifiers. Urbach et al. (2015) furthermore show that the occurrence of a fully crossed interaction in addition depends on the experimental task. The interaction was only observed in a reading-for-comprehension task, but not in a plausibility-rating task.

Another study of interest for our considerations on quantifier processing is an ERP study by Nieuwland, Ditman, and Kuperberg (2010) investigating scalar implicatures. Following a classic violation paradigm they presented sentences such as *some people have eyes* ... that are underinformative due to the scalar implicature of *some* (=*some but not all*) but plausible without the implicature (cf. *some and possibly all people have eyes*). Nieuwland et al. (2010) conducted two ERP experiments investigating an N400 effect for these kind of sentences embedded in larger sentence contexts. In their first experiment, materials were visually presented with a comma after the critical word (*eyes*). In the second, the same materials were presented without a comma. Nieuwland et al. (2010) hypothesized that the critical word would only be in discourse focus in sentences signaling the end of the scope domain, that is in the presence of a comma. However, without a comma the scope of the quantifier extends to the subsequent words allowing for a different discourse focus as exemplified by their fully felicitous example *some people have eyes that are different colors*. Critical words out of focus should be processed more shallowly, and Nieuwland et al. (2010) therefore expected ERP responses not to be modulated by pragmatic factors such as informativeness. In line with these considerations only in the first experiment an N400 effect for underinformative sentences was observed. The obtained findings are highly relevant for the present paper because applying the revision-sensitive incrementality assumption from above leads us to expect basically the same pattern of results without resorting to shallow processing. The absence of a comma can be expected to serve as a cue for additional scope information further downstream the sentence which might result in a revision of the interpretation of the first part of the sentence.

Concerning quantificational restriction, studies using a Stop Making Sense Task (Wijnen & Kaan, 2006), eyetracking during reading (Frazier et al., 2005), and the ERP technique (Kaan, Dallas, & Barkley, 2007) investigated discourses in the following two conditions.

(3) a. Four flowers were put in a vase. Six had a broken stem.
b. Twelve flowers were put in a vase. Six had a broken stem.

The second condition allows for the establishment of an anaphoric relation between the implicit restriction of the bare numeral since eight flowers can stand in a subset relation to the set of twelve flowers introduced in the first sentence. A subset relation is, however, ruled out in the first condition. Therefore, a new set of flowers has to be accommodated. The reported behavioral and eyetracking experiments provide evidence for rather early effects in line with the incremental hypotheses above. In the ERP study, however, the conditions did not differ in their early components but only showed a rather late positivity 900–1500 ms post onset of the numeral. This led Kaan et al. (2007, p. 206) to the conclusion that "the interpretation of quantifiers and the establishment of new discourse referents are rather slow processes". However, since the design crucially involves the accommodation of a restrictor set, it is not clear whether the interpretation of quantificational restriction is also delayed in the normal case where the restrictor set is contextually given.

A recent ERP study conducted in our lab directly addressed the processing of sentences with extraposed relative clauses with respect to the hypotheses stated above (Augurzky, Bott, Sternefeld, & Ulrich, 2016). The sentences tested were highly comparable to the type illustrated in (2), and they were presented after showing visual contexts in the four conditions of Fig. 1. The study comprised of two ERP experiments employing two different tasks: A truth-value judgment task (Exp. 1) and a probe detection task (Exp. 2). ERPs were analyzed relative to the presentation of the color adjective, and of the preposition *inside of/outside of*. Across tasks consistent effects were observed with respect to an early negativity 300–400 ms post onset of the critical stimuli. At the color adjective, the simple 'yes' condition (context A) elicited a larger negativity than the simple 'no' condition (context B) which, in Exp. 1, was followed by a P600 effect. This finding indicates an immediate commitment to an answer in the simple context conditions. The analysis further revealed that, in the same time window, the ERP amplitude in the complex conditions (C) and (D) were in between the two simple conditions. This is consistent with the revision-sensitive incrementality assumption from above. However, the finding that the N400 for complex conditions lied in between simple true and simple false sentences could either be associated with underspecified processing or, alternatively, with increased processing difficulty that was less severe than for the simple false conditions. Indeed, in this study, the full pattern of results could only be interpreted by considering later sentential positions.

Finally, we would like to mention that the first hypothesis from above – the *Global Interpretation Hypothesis* – was explicitly adopted to account for yet another aspect of quantifier interpretation, namely the computation of relative scope in multiply quantified sentences (Bott & Schlotterbeck, 2015). The study reports on an eyetracking during reading and a self-paced reading experiment investigating the online effects of scope computation in sentences with inverse versus linear scope. Their findings indicate scope inversion to happen only at the end of the sentence in line with the global interpretation hypothesis. Thus, finding incremental effects for quantificational restriction would imply that the restrictor and the scope argument of quantifiers are processed in qualitatively different ways. We think that such a finding would be highly relevant for the general issue whether semantic interpretation proceeds fully incrementally. It would set quantifiers apart from other two-place relations such as transitive verbs known to allow for an immediate interpretation of both arguments (cf. Knoeferle, Crocker, & Pickering, 2005).

The three hypotheses are systematically related to each other because they implement different solutions to a goal conflict between three opposing general constraints of sentence interpretation.

(4) a. Minimize working memory costs by not leaving linguistic material semantically unintegrated.
b. Minimize revision costs.
c. Minimize costs related to the anticipation and evaluation of 'good finals'.

The Global Interpretation Hypothesis is in line with the second and third constraint but does so at the cost of violating the first constraint. Revision-insensitive incrementality, on the other hand, realizes the first and the third constraint but violates the second. Finally, the revision sensitive incrementality hypothesis fulfills the first two constraints but does so at the cost of violating the third constraint.

The third constraint is hardly ever explicitly discussed in theories of semantic interpretation. We would therefore like to be more explicit about the revision sensitive incrementality hypothesis. This hypothesis presupposes that the processor is able to access the space of possible continuations of a given sentence and evaluates their effect on possible outcomes of semantic interpretation (Schlenker, 2008). We therefore have to consider semantic anticipation of whole propositions different from lexical or structural expectations prominently discussed in the psycholinguistic literature (see, e.g., Altmann & Kamide, 1999; DeLong, Urbach, & Kutas, 2005). In the following we will briefly sketch how semantic predictions for a question such as (2) might be constructed with respect to the four context types in Fig. 1.

2. Implementing the hypotheses

The revision-insensitive and the revision-sensitive version of the incremental interpretation hypothesis can be thought to resolve the sketched goal conflict in completely different ways. On the one hand, the processing system should minimize the number of dependencies predicted yet to come. On the other hand, syntactic and/or semantic reanalysis should be avoided. While the revision-insensitive incremental interpretation hypothesis weighs the first constraint over the second, the revision-sensitive hypothesis realizes the opposite ranking of constraints. In the following, we will sketch how this decision procedure can be modeled semantically.

2.1. The revision-insensitive incrementality assumption

The revision-insensitive incrementality assumption can be directly implemented using standard representations from compositional semantics. As outlined above, in Generalized Quantifier Theory quantificational determiners denote relations between sets (Barwise & Cooper, 1981). The determiner *all* corresponds to set inclusion. *All triangles are blue* which translates into ALL(TRIANGLE)(BLUE) is true if and only if the set of triangles is a subset of the set of blue things.[1] The standardly assumed lexical entries of the determiners closely mirror syntactic constituency. First, the determiner ALL has to take its restrictor argument TRIANGLE, i.e. the set of triangles, before the scope argument BLUE, i.e. the set of blue entities, can be interpreted (see, e.g., Heim & Kratzer, 1998). The resulting truth conditions can directly be evaluated in the contexts (A)–(D) in Fig. 1. Semantic interpretation will output the value *true* in context (A), where the set of triangles is identical to – and therefore a subset of – the set of blue things, and *false* in the other three contexts due to the existence of non-blue triangles.

Instead of assertive statements the present study concerns the incremental interpretation of questions such as *are all triangles blue?* We adopt a standard analysis according to which a question ?ALL(TRIANGLE)(BLUE) denotes the set of propositions consisting of possible answers to the question, here the set of the polar answers ALL(TRIANGLE)(BLUE), and NOT ALL(TRIANGLE)(BLUE) (Hamblin, 1973; Karttunen, 1977). The contexts (A)–(D) of Fig. 1 are only compatible with one of the answers in this set: The affirmative answer in context A, and the negative answer in the other three contexts. Importantly, the role of contextual information for the revision-insensitive incrementality hypothesis can be characterized as being secondary in nature. The interpretation process is entirely driven by the linguistic combinatorial system, and context comes only into play to filter out incompatible answers. This is fundamentally different in a revision-sensitive version of the hypothesis where context is of utmost importance.

2.2. Accounting for revision sensitivity

For a comprehender to decide whether there is risk of reanalysis implies considering possible continuations of a sentence (for a worked out processing model based on continuation semantics see Bott & Sternefeld, 2017). These continuations can then be employed for taking into account contextual information in a forward-looking fashion. The main theoretical contribution of the present article (going beyond Bott & Sternefeld, 2017) consists in a semantic proposal of how a finite set rather than an unlimited number of continuations is generated. The solution is straightforward. The proposed semantic representations always contain open slots that can be filled later on. At each processing step, contextual information is used in a systematic way to generate the set of possible continuations by filling the open slots with this information. The resulting sets are what we henceforth refer to as *prediction sets*. These prediction sets are incrementally updated over the course of the sentence taking into account previous sentential and contextual information as well as the semantic properties of the actually encountered word.

In the following, we will outline a procedure allowing for a safe decision about whether the polarity of the answer can still change in the light of linguistic input coming in further downstream the sentence. The computations are summarized in Fig. 2.

For the purpose of exposition let us first consider the assumed processing steps in the *simple yes* condition. First, the picture context is shown (step 0) before the presentation of the sentence up to the adjective (step 1). Encoding this context is equivalent to generating a list of properties present in the picture, that is the properties listed in the context representation. As commonly assumed in semantics, these properties correspond to sets of objects, e.g. BLUE as well as TRIANGLE correspond to the set {triangle 1, ..., triangle 8}. The context representation is minimal and only includes properties that have positive instances, thus, the property RED is not among the items in the list of properties.

The incrementally developing question is evaluated against this context. The sentence-initial auxiliary serves as a processing instruction to build up a question (indicated by '?'). This question involves the universal quantifier (ALL), which in turn takes two semantic arguments, a restrictor argument (TRIANGLE & P) and a scope argument (BLUE). The variable P in the restrictor argument

[1] Here, and throughout the rest of the paper, we will use notation with capital letters when referring to semantic representations with model-theoretic interpretations. The corresponding natural language expressions are always in italics using English translations.

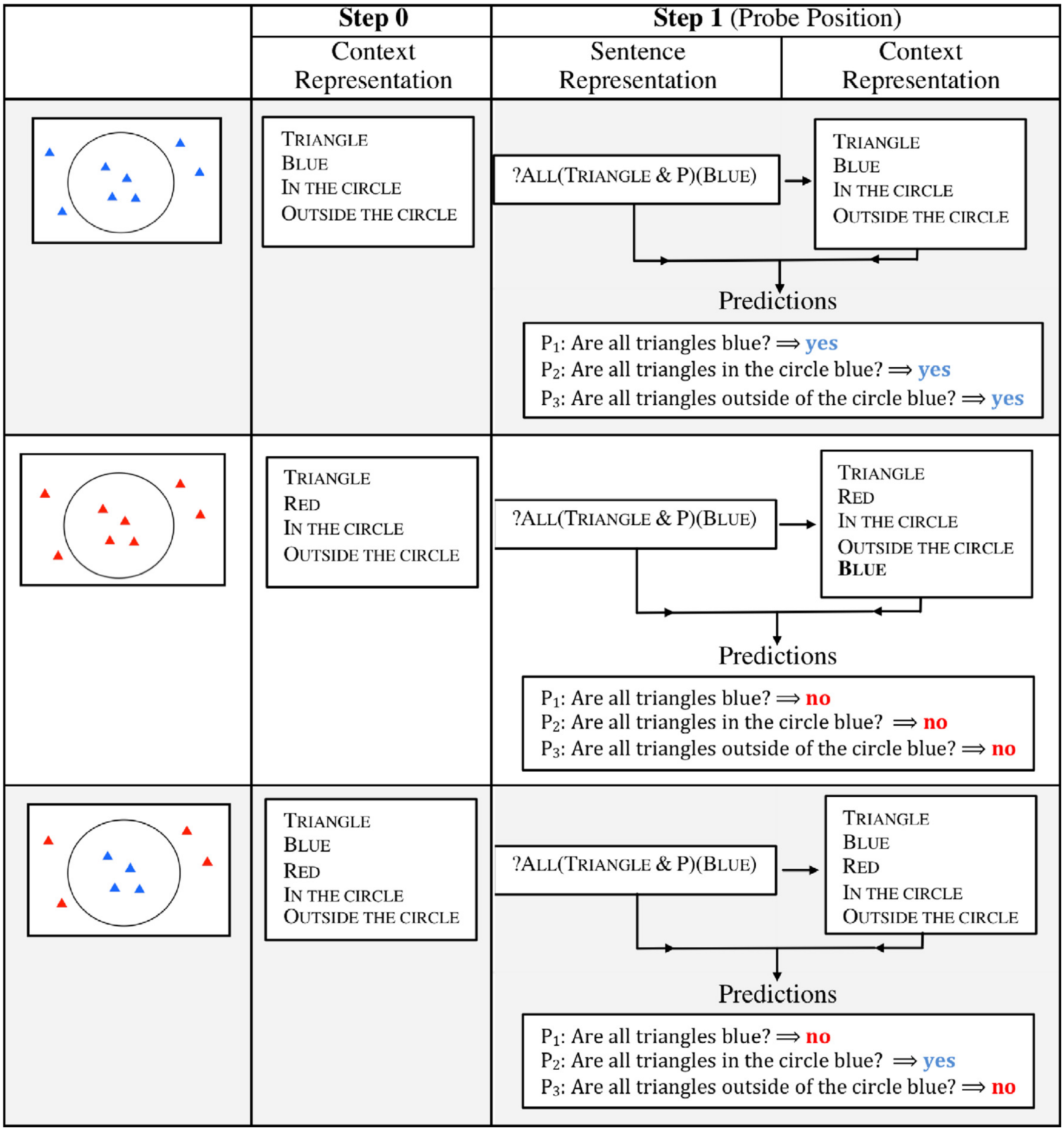


Fig. 2. Summary of the assumed processing steps for the revision-sensitive incrementality hypothesis for *simple yes* contexts, *simple no* contexts, and *complex* contexts. Steps labeled 0 and 1 correspond to the presentation of the context at the beginning of the trial and the interpretation after having processed the color adjective, respectively. The context representation is open to further updates if required by the sentence. A sentence-driven update happens in the *simple no* contexts when interpreting the adjective *blue*. A polar answer is only generated in case all possible continuations summarized in the set of predictions are of the same polarity. Thus, only *simple yes* and *simple no*, but not *complex* contexts lead to the immediate generation of an answer.

is crucial. P is a continuation variable, that is a variable which corresponds to an open slot where further information can be filled in at a later processing step (Barker, 2002; Bott & Sternefeld, 2017).[2] In general, we assume that each piece of linguistic information comes with one or more continuation variables. Quantificational determiners, for instance, introduce two open slots, one for the restrictor and another one for the scope argument. Since we are only concerned with incremental updates of the restrictor, we simply drop the second continuation variable. However, in order to incrementally interpret more complex scope arguments such as *have eyes that are different colors* (Nieuwland et al., 2010) a continuation variable in the scope argument is needed, too (HAVE EYES & Q). An elegant feature of this type of semantics is that it is essentially equivalent to a semantics without continuations. This is because open slots can always be filled with the trivial continuation, the tautology TRUE. ?ALL(TRIANGLE & TRUE)(BLUE) is equivalent to ?ALL (TRIANGLE)(BLUE).

We propose that all properties from the context representation can, in principle, be used to compute the set of possible continua-

[2] Our main motivation for resorting to continuation semantics is that in order to spell out left-to-right derivations for the developing sentence we have to be able to systematically deal with non-constituents. In the present case this becomes particularly clear if we consider the incremental interpretation of the first two words of the question (*are all* ...). The question operator ? selects for a proposition but the quantifier ALL is essentially of the wrong type because it is not a proposition (yet) but ultimately denotes a relation between properties/sets. As a consequence, in standard frameworks composition becomes only possible *after* having interpreted the restrictor and the scope argument. If we think about ALL from the perspective of continuation semantics, however, where all kinds of expressions are open propositions (here, ALL (P)(Q)), we can build up semantic representations from left to right and continuation semantics provides us with a mechanism to administer the remaining open slots. The interested reader is referred to Bott and Sternefeld (2017) for a comprehensive discussion on the consequences of the incrementality assumption for theories of semantic interpretation.

tions. An important constraint is that the resulting set has to be restricted to questions about informative and consistent propositions. Thus, the prediction set does neither include questions about tautologies (e.g., *are all blue triangles blue?*) nor about contradictions (e.g., *are all red triangles blue?*). Given the limited number of properties in the context representation the resulting prediction set contains exactly the three questions listed in P_1 to P_3. P_1 results from plugging in the trivial continuation (setting P to TRUE), and P_2 and P_3 use the properties IN THE CIRCLE and OUTSIDE THE CIRCLE, respectively. In the given context, all of the questions in the prediction set receive a *yes* response. To put it differently, no matter how the sentence proceeds, the polarity of the answer is unambiguously *yes*. Therefore, an answer can be safely prepared without the risk of later revision.

Similarly for the *simple no* condition. Here, the evaluation of the prediction set reveals that the answer is unambiguously *no*. However, processing this condition involves a linguistically triggered update of the context representation with severe consequences for the incremental construction of the prediction set. By assumption, the initial context representation does not contain the property BLUE because there are no blue objects. However, linguistic processing of the adjective *blue* makes it necessary to check this property. The update consists in adding the property BLUE to the context representation: a property that denotes the empty set.

Note that this additional step provides us with a semantic notion of surprisal. The adjective *blue* is unexpected in the *simple no* context. This semantic surprisal can be related to the degree of overlap between the prediction sets computed at a processing step n as compared to an earlier processing step $n-1$. Let us therefore consider the prediction sets computed for earlier parts of the question before the adjective and compare them to the prediction set P_1 to P_3 computed at step 1. The partial question without the adjective *are all triangles* translates into ?ALL(TRIANGLE & P)(Q) with two continuation variables. Generating predictions about possible values for P and Q results in a set of predictions that essentially does not make use of BLUE because this property is not among the properties in the context representation. Thus, neither of the predictions P_1 to P_3, which all involve the property BLUE, are part of the prediction sets for earlier parts of the sentence. Consequently, at step 1 a completely new prediction set has to be build. This is different in the *simple yes* context where the prediction set for step 1 is a monotonic extension of the prediction sets computed earlier on. Note that non-monotonicity only concerns the update of the respective prediction sets before and after the adjective. The update of the context representation simply consists in a conservative extension by adding the property BLUE.

Finally, let us turn to the complex condition. Evaluation of the set of predictions shows that the polarity of the answer cannot be unambiguously determined. Therefore, further input is needed, and no answer will be generated. In Fig. 2 we have only included one of the complex conditions, but as the reader can easily check the other context (context (D) in Fig. 1) works exactly the same. Again, one of the predicted question continuations evaluates into *yes*, and the other two into *no* ruling out a safe answer, yet.

We are now in the position to derive the revision-sensitive incremental interpretation hypothesis.

> **Revision-Sensitive Incrementality:** If the evaluation of the set of possible continuations of a question given a context results in a prediction set for which all correct answers have the same polarity, an answer is immediately generated. If the answers to the questions in the prediction set differ in polarity, however, no answer is generated.

It is important to note that the latter option does not mean that the question has not been interpreted. Rather to the contrary, it means that the context does not allow to generate a definite answer yet. This hypothesis can be closely related to the generation of probabilistic expectations. In case the semantic value of a quantified sentence depends on further restriction, the processor can be expected to anticipate further restriction, for instance, in the form of an extraposed relative clause. On the other hand, in simple contexts where the simple propositions logically entail propositions including additional restrictor information, Gricean reasoning can be employed to generate the expectation that no further restriction will follow. Under such circumstances, encountering an extraposed relative clause can be expected to lead to processing difficulty.

The system just outlined generates and evaluates all possible continuations of incomplete sentences taking into account all the relations that are contextually given. In more naturalistic settings than the idealized contexts from Fig. 1 with only a very small number of properties the sketched procedure will very soon become computationally intractable. Instead of being able to generate the complete prediction set allowing for safe decisions the processor will then be forced to compute a more constrained prediction set that contains only some of all combinatorily possible continuations. We are well aware of this fact and would therefore like to emphasize that our proposal is an idealization. We will come back to this issue below in the general discussion.

2.3. The present study

In order to assess answer generation mid-sentence without disrupting the interpretation process we employed a variant of the probe RT task paradigm (e.g., Miller et al., 1996). The primary task was a linguistic truth evaluation task involving question answering and the probe RT task was a tone judgment task. The linguistic task required participants to first inspect a visual context as illustrated in Fig. 1, then they received a quantified question sentence with extraposed relative clauses, which was presented with rapid serial visual presentation (500 ms per word). Immediately after the last word of the question participants provided a positive or negative response by pressing a button with the right or left index finger, respectively.

Embedded in this linguistic task tone probes were presented 400 ms after the onset of the color adjective and participants had to judge whether the tone was a high or a low tone (Exp. 1) or whether the tone was presented to the left or the right ear (Exp. 2). Crucially, the response fingers in this secondary task were the same as in the primary linguistic task. Thus responses in the probe RT task were either congruent or incongruent with the correct responses to the partial question sentence including the color adjective. If answer generation including response preparation happens incrementally, we expected to find compatibility effects of the linguistic task on the secondary tone judgment task.

The probe RT task has been shown to be highly sensitive in probing partial output from the primary task. More specifically, this task has been employed in previous two-choice RT studies to assess whether preliminary information in the primary task can activate responses before processing of the primary task has been completed (Band & Miller, 1997; Ilan & Miller, 1998; Miller, 1985; Miller et al., 1996). The basic idea of the probe task is that if such preliminary information is available, responses to the probe stimulus should be facilitated when it requires the same response alternative as the primary task. For example, in one study the primary task was a mental rotation task and in which auditory tones were embedded to probe preliminary response activation Ilan and Miller (1998, Exp. 3). Participants in this study saw either a normal or the mirror image of a character that was either upright or rotated. The character's identity determined the response hand and participants were to respond only when the character was presented in its normal form. The major question in this study was whether information about the character's identity would already prime the response before mental rotation has been completed. Choice-RT responses to auditory probe stimuli that appeared either in the left or the right ear were faster when the preliminary information from the primary task was congruent with the probe response than when it was incongruent. The result of this study shows that preliminary output from the primary task can produce response preparation before the task has been completed. Analogous effects of preliminary output have been observed with the probe RT task for other primary tasks (Ilan & Miller, 1998; Miller, 1985; Miller et al., 1996).

The present study used the probe RT task for the first time to investigate whether incremental linguistic information is available before the sentence is complete. In Experiment 1 (Exp. 1), we used an arbitrary mapping between hands and tones to probe partial linguistic output from sentence processing whereas in Experiment 2 (Exp. 2) the mapping between tones and responses was spatially compatible. Above response priming, it seems that the probe RT task is also sensitive to the extent of cognitive load that is associated with the processing of the primary task (Logan et al., 1983; Posner & Boies, 1971).

We would like to point out that even though both experiments reported in the present paper crucially involved a secondary task in addition to the linguistic task, these secondary tasks themselves were intended to induce as little cognitive load as possible and therefore guarantee language interpretation to proceed as smoothly as it can be given our research questions. This should be especially true for the congruent mapping of the tone detection task in Exp. 2.

In the next section we will lay out the predictions by the three hypotheses combining our semantic considerations from above with the just outlined assumptions about response facilitation for congruent responses.

2.3.1. Predictions

Both experiments employ a 2 × 2 × 2 (COMPLEXITY × POLARITY × RESPONSE COMPATIBILITY) within design. The context manipulations regarding COMPLEXITY and POLARITY are illustrated in Fig. 1 (repeated for convenience in Fig. 3 with condition labels relevant for the probe RT task), and RESPONSE COMPATIBILITY was manipulated by either mapping the correct response in the tone judgment task (task 2) to the response for the locally correct answer or by mapping it to the other hand. We considered the locally correct answer at the color adjective, which is a yes response in the *simple yes* context (A) and a no response in the other three contexts (B), (C) and (D).

The global interpretation hypothesis predicts no task interference. Processing in the linguistic task is expected to only involve the stimulus encoding of the color adjective and hence task 1 locally lacks a stage of response selection. Consequently, task 2 should not be affected by task 1, neither with respect to a general dual-task effect nor with respect to response preparation of the locally congruent answer. However, the linguistic task at the end of the sentence should clearly show that comprehenders eventually select the correct answer. Conditions (A) and (C) should elicit yes responses and conditions (B) and (D) should yield no responses.

For both versions of the incrementality hypothesis, in its revision insensitive as well as in its revision sensitive version, general task interference as well as response facilitation are expected. We start with the revision insensitive version because in this case the predictions turn out to be simpler.

The revision-insensitive hypothesis states that for all four visual contexts the semantic processor generates an answer immediately when the color adjective is encountered. Only for the *simple yes* context (A) the local answer is yes. Since computing a no answer should impose more processing load than a yes answer an interaction between COMPLEXITY and POLARITY is predicted with the *simple yes* context (A) leading to shorter tone judgment RTs than the other three visual contexts (B), (C) and (D). Furthermore, equally strong compatibility effects are expected for all four context types. This should result in a main effect of RESPONSE COMPATIBILITY but no interactions with the other two factors.

Turning to the revision sensitive hypothesis, let us first consider the monochrome, simple contexts *simple yes* (A) and *simple no* (B). For those, the predictions are the same as for the revision-insensitive hypothesis. One difference between the two hypotheses, however, is that for the revision-insensitive version the general difference between the *simple yes* context (A) and the *simple no* context (B) was just stipulated, but given the semantic model from above, this difference can actually be derived. Please recall that in the case of the *simple yes* context (A) the context representation already contains the property BLUE and the set of possible questions can directly be generated from this context set. For the *simple no* context (B) the situation is different. Here, the encoding of the property BLUE is only triggered by the color adjective and a predicate denoting the empty set has to be added to the context set. Only after this additional step, evaluation of the possible question continuations becomes possible. We therefore directly gain an explanation why the *simple no* context (B) should enhance cognitive load in the linguistic task relative to the *simple yes* context (A).

For the two-color, *complex* contexts (C) and (D), the predictions of the revision-sensitive hypothesis differ from the revision-insensitive alternative. In our model above, we outlined that in these contexts the semantic processor does not arrive at a decision because the polarity of the answers of the different question continuations differed from each other. This evaluation process clearly involves a response selection stage but this time the semantic processor does not output an answer. The evaluation of the more complex context representation than in the case of the *simple yes* context (A) should lead to generally longer judgment times in task 2 but in contrast to the two monochrome contexts to no response compatibility effects. Thus, the revision-sensitive version of the incrementality hypothesis predicts interactions between the two context factors and RESPONSE COMPATIBILITY.

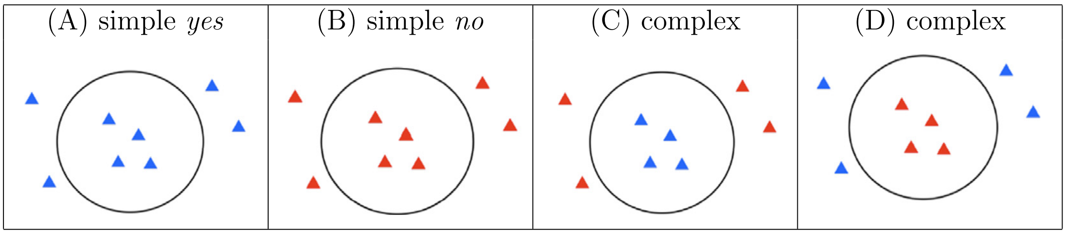


Fig. 3. Contexts for the question *are all triangles blue* ... with condition labels relevant for the probe RT task. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Experiment 1

Experiment 1 tested these predictions in a probe RT task experiment. At the beginning of each trial, one of four visual contexts (cf. Fig. 1) was presented, followed by a question with the quantifier *alle* (*all*). The question was presented word-by-word with rapid serial visual presentation. (5) is an example.

(5)	Sind	alle	Dreiecke	blau,	die	im	Kreis	sind?
	Are	all	triangles	blue	that	in the	circle	are?
	‘Are all triangles blue that are inside?’							

Contexts either consisted of simple (A,B), or complex (C,D) pictures. In each picture, geometrical objects were presented within and outside of a container form. In the simple pictures, all of the small objects were of identical color, whereas in the complex pictures the objects within and outside of the container form were of different colors. For examining incremental processes on the color adjective, we carried out a probe RT task experiment involving an arbitrary mapping between tones in the tone task and their respective response hands. In this experiment we used low and high tones who had to be discriminated by a reaction with the left or right index finger. By arbitrary mapping we refer to the fact that there does not exist any natural association between the dimension low/high and the dimension left/right. The arbitrariness of the mapping makes response selection in the secondary task a nontrivial task and we therefore expected to see clear congruency effects.

3.1. Methods

3.1.1. Materials

The experimental materials were constructed according to a $2 \times 2 \times 2$ factorial within-design with the factors context (simple (contexts A/B), complex (contexts C/D)), polarity (*‘yes’ answer* (contexts A/C), *‘no’ answer* (contexts B/D)), *TONE* (high tone, low tone). In addition, the within-factor *PREPOSITION* (*in*, *outside of*) was included as a counterbalancing factor. Picture contexts were generated as quadruplets via MS PowerPoint. In each picture, at least three geometrical objects of the same shape (eight shapes: triangles, squares, rectangles, etc.) appeared inside and at least three outside a container form (same eight shapes, e.g. a circle). For each quadruplet, the color of the objects inside vs. outside the container was manipulated as illustrated in the sample contexts in Fig. 1. A total set of 64 quadruplets was generated. Each picture context was presented twice: once, together with a question involving the preposition *in* (*inside-of*), and once, together with a question involving the preposition *außerhalb* (*outside-of*). (6) is a sample question.[3]

(6)	a.	Sind	alle	Dreiecke	blau	die	im		Kreis	sind?
		Are	all	triangles	blue	that	in the		circle	are?
		Are all triangles blue that are inside the circle?								
	b.	Sind	alle	Dreiecke	blau,	die	außerhalb	des	Kreises	sind?
		Are	all	triangles	blue	that	out of	the	circle	are?
		‘Are all triangles blue that are outside the circle?’								

The counterbalancing factor *PREPOSITION* thus was included into the experimental design as an additional within factor. Note that for the complex conditions a context that instantiated condition C with *in* appeared as condition D with *outside of* and vice versa. Four lists of question-picture pairs were generated in a latin square design. Each included a total of 128 picture-sentence pairs, i.e. experimental items. Neither the items nor the distractor items included any punctuation marks.

The factor *TONE* was manipulated by assigning half of the items in each contexts a low tone and the other half a high tone. Low tones had a frequency of 440 Hz, whereas high tones had a frequency of 880 Hz, ie. they were one octave above the low tones. In summary, each of the resulting eight experimental conditions in the contexts A–D was tested 16 times per list and participants answered a total of 32 questions for each of the contexts A–D.

Because the experiment involved a tone decision and a linguistic task with two answer alternatives both, each of the four lists was realized in four versions. This way, the answer hands with which the single, binary choices had to be made in each task could be fully counterbalanced (version 1: *high tone* – left hand; *low tone* – right hand; *true* – left; *false* – right; version 2: *high tone* – left hand; *low tone* – right hand; *false* – left; *true* – right; and so on). Finally, for each of the 128 items in each list and version of the experiment it was labeled whether the response hand in the tone task was congruent with the response hand for the locally correct answer to the question at the color adjective (ignoring the relative clause).

Within each list, the 128 experimental trials were distributed to four experimental blocks with 32 experimental items per block. This was done using a latin square design that ensured that each context appeared only once within the same block and that each condition appeared equally often (four times) in each block.

To minimize strategic effects, a set of 64 fillers (32 *true*) was added to each list. The fillers were question-picture pairs that involved different quantifiers like *weniger als n* (’less than n’; 25% of all fillers), coordination structures (25%), universally quantified questions ending on the color adjective (25%), as well as further sentences containing *all* (25%; e.g., *Are all circles red that are in the picture?*). The fillers also involved a tone judgment on the fourth word of the sentence. Per block, 16 filler items were included (8 true), resulting in a total of 48 trials per block. The total experiment thus consisted of 192 trials.

3.1.2. Procedure

Participants were seated in a dimly-lit, sound-shielded booth in front of a 19” computer screen. Tones were presented via headphones, and loudness was adapted beforehand individually according to participants’ demands. Stimuli appeared in a pseudo-randomized fashion, in which maximally two items of the same condition appeared in succession. The experimental session was divided into four blocks with short breaks in between.

At the beginning of each trial, the picture appeared in the center of the screen for 1500 ms. Then the picture disappeared, and the single words of the sentence were presented word-by-word via RSVP (500 ms per word). 400 ms after the onset of the adjective, a high or low tone was bilaterally presented, and participants

[3] In order to ensure that a local semantic commitment is principally possible on the adjective, we controlled our materials for potential silent-prosodic confounds during reading (Fodor, 2002). Therefore, the sentence materials involve an implicit prosody that is both compatible with a (potentially following) restrictive relative clause and a sentence-final prosodic contour. This is one of the reasons why we presented yes–no questions instead of declarative sentences. As questions involve a clause-final rise, their prosody is highly comparable to relative clause continuations (see Augurzky, 2006; Poschmann & Wagner, 2016 who report rising F0 contours preceding restrictive relative clauses). Controlling for silent prosody should thus ensure that sentences were ambiguous with respect to whether a restrictive relative clause will follow.

had to judge whether the tone was high or low by pressing one of two buttons ('F' or 'J') on the keyboard. In the experimental sentences, the words of the restrictive relative followed via RSVP, undisrupted by the tone task. Given that the following words were presented 500 ms each, the tone task was thus completed before the end of the question. In the filler items, tones were randomly presented at different sentential positions, and the tone task was also completed before the end of the question. After the final word had disappeared, three question marks were shown, signaling that participants now had to make a truth evaluation (*Did the preceding sentence truly or falsely reflect the content of the picture?*). Participants answered with 'wahr' (true) and 'falsch' (false) by pressing one of two buttons ('F' or 'J') on the keyboard. Crucially, the keys for high and low tones, as well as for *true* and *false* answers involved the same response keys. A paper sign with the response labels (true/false, high/low) was presented at the bottom of the monitor in a centered position.

As noted above, the tone-key and the answer-key mappings were counterbalanced across participants. Participants were asked to make both their tone decision and their truth evaluation as quickly as possible. The initial timeout for the tone task was 800 ms, and for the truth evaluation, it was 1200 ms. Both timeouts were 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. After they had provided an answer, a blank screen appeared for 500 ms, and the next trial began. The experimental session was preceded by a short practice with 12 trials. A session took between 35 and 50 min.

3.1.3. Participants

Sixty-four right-handed students (mean age: 23.2 years) from the University of Tübingen took part in the study (27 male). They were native speakers of German with normal or corrected-to-normal vision and were paid for their participation. Participants were randomly assigned to the four versions of the four lists in such a way that each was tested four times. Three of the participants were excluded from the statistical analyses due to a large proportion of errors or timeouts (more than 50% of all trials).

3.1.4. Statistical analysis

Statistical analyses were computed on the data from the tone task and the linguistic task, respectively. For each task we performed analyses on the RT data as well as on the proportions of errors. RT data were log-transformed before computing inference statistics. The random effect for item was only included in the analysis of the linguistic task. All Generalized Linear Mixed Effects analyses were computed in the statistical software environment R using the package *lme4* (Bates, Mächler, Bolker, & Walker, 2015).

For the analyses of the proportions of correct answers in tone task, we first excluded all trials in which no judgment had been provided within a 1000 ms time interval after tone onset (this affected 6.6% of all trials).[4] The data were pooled over the complex contexts C and D because at the color adjective these two context types instantiate the same condition. Consequently, statistical analyses were performed on a 3 × 2 within design including the fixed factors CONTEXT (*simple 'yes'* (context A), *simple 'no'* (context B), *complex* (contexts C/D)), CONGRUENCY (*same response hand*, *different response hand*), and their interaction. The proportions of errors were subjected to logit mixed-effects model analyses including the fixed effects and the random intercepts as well as slopes for CONTEXT and CONGRUENCY for participants.[5] Significant effects including the three-level factor CONTEXT and/or significant interactions were further broken down by analyzing subsets of the data. For the analysis of the RT data, only correct trials with RTs of at most 1000 ms were included into the linear mixed-effects models analyses (excluding 15.0% of all trials). These models included the fixed effects and had maximal random effects structure, that is random intercepts plus random slopes for CONTEXT, CONGRUENCY, and their interaction for participants. Error probabilities were determined by conducting likelihood-ratio tests between mixed-effects models differing only in the presence or absence of the fixed effect under consideration. Significant effects including the three-level factor CONTEXT and/or significant interactions were further broken down by analyzing subsets of the data.

For the analysis of the interpretation data, we first eliminated nine trials for which the system had failed to log responses (0.1% of the data). Error rates as well as reaction times were then subjected to generalized linear mixed-effects model analyses with the fixed effects of CONTEXT (*simple*, *complex*), POLARITY (*'yes' answer*, *'no' answer*), and their interaction. The errors were analyzed in logit mixed-effects model analyses including the fixed effects as well as random intercepts and random slopes for CONTEXT and CONGRUENCY for participants and items. For the analysis of the RT data only correct trials and RTs below 1000 ms were included into a linear mixed-effects models analysis (excluding 22.0% of the data; 3.8% due to their latencies, and 18.2% erroneous responses). The analysis included the full fixed-effects structure and also had maximal random effects structures, i.e. random intercepts plus random slopes for CONTEXT, POLARITY, and their interaction for participants and items. Error probabilities were determined by conducting likelihood-ratio tests between mixed-effects models differing only in the presence or absence of the fixed effect under consideration.

In order to examine whether any potential congruency effects had to do with participants getting used to the task and/or the high proportion of extraposed relative clauses in the experiment, we carried out additional mixed-effects model analyses including the fixed effect and the random slopes of HALF OF THE EXPERIMENT (*first* vs. *second half*) in the analyses. These additional analyses were only computed for the proportion of errors and RTs of correct answers in the tone task.

3.2. Results

3.2.1. Tone task

Fig. 4 shows the mean proportions of errors and the judgment RTs of correct judgments in each of the six conditions. All conditions had at most 10% errors except for the incongruous *simple 'no'* condition, which led to errors 16.0% of all trials. The logit mixed-effects model analysis revealed that the interaction between CONTEXT and CONGRUENCY was highly significant ($\mathcal{X}^2(2) = 21.27; p < .01$). In follow-up logit mixed-effects analyses the congruent and the incongruent conditions were compared in each context type. The pairwise comparison revealed that the error rates in the *simple 'yes'* context did not differ between congruent and incongruent tone judgments ($\mathcal{X}^2(1) = .77; p = .38$). However, in the *simple 'no'* context the 9.8% difference in error rates between the incongruent condition and the congruent conditions was

[4] Here, and in all analyses to follow an a priori defined cut-off value of 1000 ms was applied. A constant cut-off is standardly employed in RT research (e.g., Ulrich & Miller, 1994). Note that an a priori defined cut-off value avoids the problem of post hoc flexibility in data analysis (cf. Simmons, Nelson, & Simonsohn, 2011). Crucially, truncating the RT distribution on the right tail should result in smaller effects than they really are, or, in other words, would lead to a more conservative estimate of the observed effect.

[5] None of the reported logit mixed effects model analyses did include random slopes for the interaction because the models failed to converge.

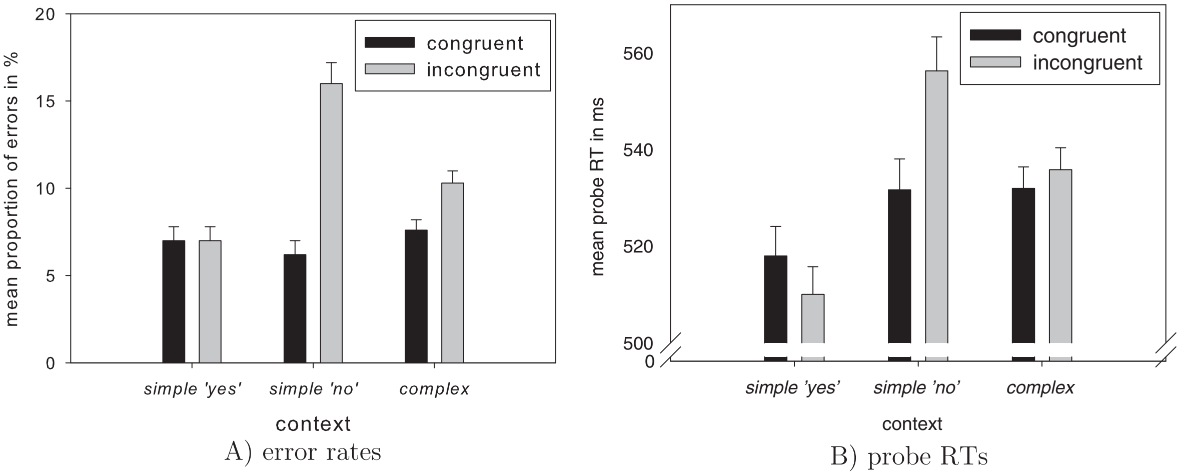


Fig. 4. Mean error rates (panel A) and probe RTs of correct judgments (panel B) in the tone task of Exp. 1. The error bars correspond to one standard error of the mean.

highly significant ($\mathcal{X}^2(1) = 10.61; p < .01$). Similarly, for the complex contexts the 2.7% difference turned out to be reliable ($\mathcal{X}^2(1) = 4.20; p < .05$).

A similar pattern of results was observed in the RT analysis on correct responses. The two *simple 'yes'* conditions received on average the fastest judgments with RTs of 518 ms in the congruent and 510 ms in the incongruent condition. Both the *simple 'no'* and the *complex* conditions had RTs well above 530 ms. CONGRUENCY led to RT differences in the *simple 'no'* conditions. In this context type, the incongruent condition was judged on average 25 ms slower than the congruent condition. The respective mean judgment RTs were 557 ms and 532 ms. Following complex contexts mean RTs of congruent vs. incongruent tone judgments were approximately equally fast with 532 ms in the congruent and 536 ms in the incongruent condition. The global mixed-effects model analysis on log-transformed RTs revealed a marginally significant interaction between CONTEXT and CONGRUENCY ($\mathcal{X}^2(2) = 5.69; p = .06$). The follow-up analyses revealed that only the *simple 'no'* context led to a reliable CONGRUENCY effect ($\mathcal{X}^2(1) = 5.81; p < .05$) but neither the RTs in the *simple 'yes'* context ($\mathcal{X}^2(1) = 0.65; p = .42$) nor in the *complex* context ($\mathcal{X}^2(1) = .32; p = .57$) differed significantly between congruent and incongruent responses. The same pattern of effects was found in linear mixed-effects model analyses on the untransformed RT data with the only difference that this time the interaction between CONTEXT and CONGRUENCY turned out to be fully significant ($\mathcal{X}^2(2) = 6.52; p < .05$).

Furthermore, we conducted an analysis testing the predicted main effect of CONTEXT. Above, we derived the prediction that the judgment RTs in the *simple 'yes'* conditions would be shorter than the overall RTs in the other two context types. In order to test this prediction, we converted CONGRUENCY to a sum-coding numeric representation and conducted a likelihood-ratio test between mixed-effects models differing only in the presence or absence of a main effect of CONTEXT (Levy, 2014). The likelihood-ratio test provided evidence for a main effect of CONTEXT on the log-transformed judgment RTs ($\mathcal{X}^2(2) = 27.45; p < .01$). The *simple 'yes'* context led to reliably shorter judgment RTs relative to the other two context types (*simple 'yes'* vs. *simple 'no'/complex*: estimate = $-.035, t = -4.03$). The same pattern of results was obtained in analyses on untransformed data (main effect of CONTEXT: $\mathcal{X}^2(2) = 26.21; p < .01$; *simple 'yes'* vs. *simple 'no'/complex*: estimate = $-25.81, t = -3.37$).

Finally, we conducted analyses including HALF OF EXPERIMENT to test whether the observed effects were consistently present during the entire course of the experiment. Specifically, we investigated the possibility that the lack of effect in the complex conditions may have resulted from participants adaptating to the high probability of extraposed relative clauses in the experimental materials. Table 1 presents the mean error rates and RTs for the two halves of the experiment. Overall, participants made less errors in the second than in the first half of the experiment (*first half* 10% vs. *second half* 8%), and got faster during the course of the experiment (*first half* 571.5 ms vs. *second half* 491.5 ms). Apart from these general differences the relative differences between the experimental conditions followed the same pattern for both halves. The congruency effect in the RTs of the *simple 'no'* conditions was even numerically stronger in the first half than in the second, as was the difference between the *simple 'yes'* context relative to the other contexts. The simplest model able to account for the RT data was a mixed effects model including the fixed effect of HALF as well as the interaction and the main effects of CONTEXT and CONGRUENCY and no other fixed effects (model comparison between this model and the saturated model, both models with the same random-effect structure (intercept plus random slopes for HALF and the interaction term CONTEXT × CONGRUENCY): $\mathcal{X}^2(5) = 1.63; p = .90$). This shows that the only effect of HALF on RT was the general speed-up effect. The analysis of error rates also provided evidence for a main effect of HALF but no interaction with the other two factors. The simplest model able to account for the data at hand was one with a simple fixed effect of HALF, and the main effects of CONTEXT and CONGRUENCY. In summary it can be stated therefore that participants made fewer errors in the second half than the first half of the experiment, but in both halfs they were better in the *simple 'yes'* context than in the other context types, and they made fewer errors in congruent trials than in incongruent ones.

3.2.2. Linguistic task

Fig. 5 presents the mean proportions of errors and the answer RTs of correct responses in the linguistic task. Error rates varied as a function of COMPLEXITY and POLARITY. Participants made more errors for complex contexts, and they made more errors if the answer was 'no' than when it was 'yes'. The logit mixed-effects model analysis revealed no significant interaction between the two factors ($\mathcal{X}^2(1) = .20; p = .65$). Model comparisons between the maximal model including both fixed effects with simplified models from which one of the fixed effects was removed revealed significant effects of COMPLEXITY ($\mathcal{X}^2(1) = 17.86; p < .01$) and of POLARITY ($\mathcal{X}^2(1) = 33.39; p < .01$).

Table 1
Mean error rates and mean RTs for the first and second half of Exp. 1.

	Errors	RTs	
	M (in %)	*M* (in ms)	*SD*
First half			
Simple 'yes'			
Congruent	7.6	560	181
Incongruent	7.6	550	181
Simple 'no'			
Congruent	7.9	568	188
Incongruent	17.4	604	194
Complex			
Congruent	9.2	573	185
Incongruent	10.6	574	188
Second half			
Simple 'yes'			
Congruent	6.4	478	166
Incongruent	6.3	473	148
Simple 'no'			
Congruent	4.7	494	177
Incongruent	14.7	508	177
Complex			
Congruent	6.1	493	173
Incongruent	10.0	497	171

M = mean and SD = standard deviation.

The analysis of reaction times revealed a similar pattern albeit no significant effects of POLARITY. RTs for complex contexts took on average 30 ms longer than answers to the simple contexts. In the first step of the statistical analysis on log-transformed answer RTs the saturated model was compared with a model from which the interaction term was removed from both the fixed effects as well as the random slopes for participants and items. The model comparison revealed no significant interaction between COMPLEXITY and POLARITY ($\mathcal{X}^2(1) = 1.28; p = .26$). We then performed model comparisons between the resulting simplified model including both main effects with models from which one of the fixed effects was removed while keeping the random effects structure maximal. While the effect of COMPLEXITY turned out to be reliable ($\mathcal{X}^2(1) = 17.03; p < .01$) the effect of POLARITY was not significant ($\mathcal{X}^2(1) = .02; p = .88$).

Overall, the reaction times were very fast with a mean answer RT of only 295 ms. RTs as short as these are very untypical for psycholinguistic experiments on quantifier interpretation. The very fast responses indicate that participants may have computed their answer well before the question marks appeared on the screen.

3.3. Discussion

The RT effects in the tone judgment task show clear influences of the linguistic task on the tone judgment task. The observed interactions in error rates as well as judgment RTs can only be explained in terms of the incrementally ongoing linguistic interpretation. Also, the very quick answers in the linguistic task indicates that participants had computed an answer before they reached the end of the sentence. Thus, our findings are only consistent with some form of the incrementality hypothesis. Extending what is known from a large body of work on incremental interpretation the reported findings show that the incremental nature of online interpretation even includes the dynamic generation of answers to quantified questions. The reported findings rule out the globality hypothesis stated above.

Coming back to the revision-insensitive versus revision-sensitive version of the incrementality hypothesis from the introduction the results are not fully consistent with either version. Let us first consider the revision-insensitive incrementality hypothesis. A crucial prediction was that the linguistic task should lead to congruency effects in all three context types. In particular, *simple 'no'* contexts were expected to lead to exactly the same effects as *complex* contexts do. This is because these two context types should give rise to a local *'no'* response interfering with the response selection of the *'yes'* hand. However, consistent congruency effects in error rates and judgment RTs were observed for simple contexts that required a *no* response (and to a certain extent, also for the complex contexts).[6] Also, the results disconfirm the predictions for the *simple 'yes'* conditions. Although this context type imposed significantly less load on the secondary task – in line with the predictions of both versions of the incrementality hypothesis – we did not observe any congruency effect for this context type. The latter result is also unexpected under the revision-sensitive version of the incrementality hypothesis, even though the effects regarding the other two context types are fully consistent with this hypothesis. Except for this discrepency, however, the results fit the predictions of our revision-sensitive account.

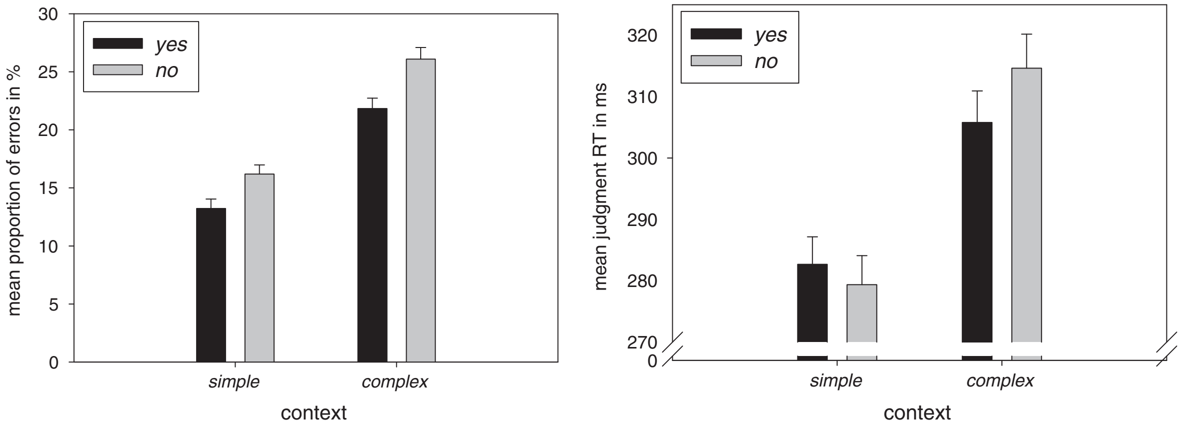
So, how can the lack of a congruency effect be explained in the *simple 'yes'* contexts? The RT analysis shows that the *simple 'yes'* contexts were in fact the contexts that imposed the smallest cognitive load. One possible explanation would be that answer preparation in the linguistic task did not interfere with response selection in the tone task because the answer preparation of a *'yes'* judgment in the simple context may happen so fast that potential response facilitation effects had already decayed when response selection was at issue in the tone task. We will follow up on this line of thinking in the General Discussion where we will employ a slightly modified version of the outlined model from the introduction.

Before drawing any premature conclusions from this somewhat surprising result we would like to know whether the obtained pattern of results is reliable and allows for replication. Furthermore, we would like to address an issue that may be problematic about the present results. The proportion of errors in the linguistic task was relatively high across all kinds of contexts. On average participants gave about 20% wrong answers. This may be interpreted as indication that the probe RT task employed in the present experiment was in fact overtaxing for the participants. Consequently, it is somewhat unclear whether the reported experiment was able to tap into incremental interpretation processes in all conditions. It is possible that the lack of a congruency effect in the complex conditions is just an artefact of the testing conditions and the stimuli in the complex conditions were actually too complex to be evaluated online. Therefore, we would like to see the same pattern of results in a probe-detection task with a secondary task that is considerably easier.

4. Experiment 2

We conducted a second experiment with a much simpler secondary tone task. Instead of an arbitrary mapping of tones to response hands, stimuli and responses targeted the same dimension in a non-arbitrary fashion. Stimuli were tone probes presented to the left or the right ear and responses had to be provided by the ipsilateral response hand. The aims of the present experiment were thus twofold. firstly, we wanted to see whether the results from Experiment 1 could be replicated. Secondly, we expected consider-

[6] It is noteworthy that there was an error rate effect of congruency in the complex condition. Since this effect, which is unexpected under our account, was not reliable across experiments (cf. Exp. 2), we ignore it in the discussion.



A) error rates

B) judgment RTs

Fig. 5. Mean error rates (panel A) and RTs of correct answers (panel B) in the primary task of Exp. 1. The error bars correspond to one standard error of the mean.

ably fewer errors in the present experiment than in the previous one due to reduced cognitive load.

4.1. Methods

4.1.1. Materials

The same picture and sentence materials as in Experiment 1 were used. Items were spread over lists in an identical manner as in Experiment 1. In the present Experiment, we again created different versions for each list in order to fully counterbalance the keys with which the single decisions of the two tasks were made. In order to create these versions, we took the versions of the previous experiment and replaced high tones by tones to the left ear and low tones by tones to the right ear, so that half of the items in each condition occurred with a tone to the left and the other half with a tone to the right. All tones had a frequency of 880 Hz.

4.1.2. Procedure

In contrast to Experiment 1, the present experiment employed a different tone task. This time, participants had to decide whether the tone had been presented to their left or to their right ear. Crucially, tone judgments were always non-arbitrary. Participants always had to react to tones presented to their left ear with their left hand and to tones presented to their right ear with their right hand. Everything else was identical to the procedure of the previous study. An experimental session took about 40 min.

4.1.3. Participants

Sixty-four right-handed students (mean age: 23.7 years) from the University of Tübingen took part in the study (27 male). They were native speakers of German with normal or corrected-to-normal vision and were paid for their participation. None of them had participated in the previous experiment. Four participants were randomly assigned to each version of each list.

4.1.4. Statistical analysis

We employed the same kinds of statistical analyses as in the previous experiment. Errors in the tone task include trials in which participants gave a wrong response (3.6% of all trials), or failed to provide a judgment within a time limit of 1000 ms (2.0% of the trials). The analyses of RTs in the linguistic task only included answers provided within 1000 ms after the presentation of the question mark (excluding 2.6% of the trials from the analyses).

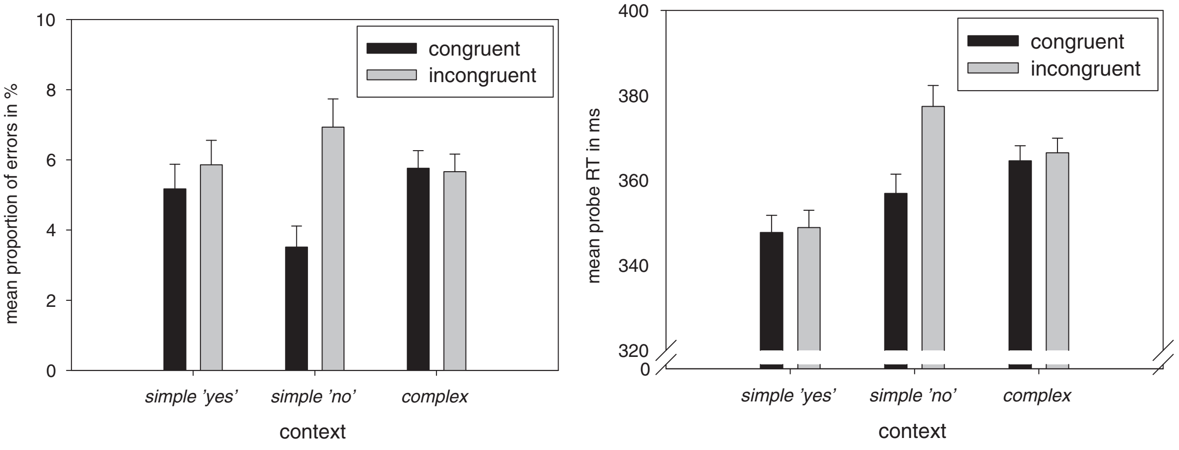
4.2. Results and discussion

4.2.1. Tone task

Fig. 6 presents the proportions of errors and the mean judgment RTs of the tone judgments. Even though participants made relatively few errors (overall error rate 5.5%) there were slight differences between conditions. While the congruent and the incongruent conditions in the *simple 'yes'* and the *complex* contexts had rather similar error rates, the error rates again differed between congruent and incongruent tone judgments in the *simple 'no'* context. Incongruent judgments led to more errors (6.9%) than did congruent judgments (3.5%). A logit mixed-effects analysis on the proportion of errors revealed a significant interaction between CONTEXT and CONGRUENCY ($\mathcal{X}^2(2) = 9.85; p < .01$).[7] Breaking down this interaction, the follow-up analyses revealed a significant CONGRUENCY effect in the *simple 'no'* context ($\mathcal{X}^2(1) = 9.82; p < .01$), but no reliable differences in the *simple 'yes'* ($\mathcal{X}^2(1) = 0.37; p = .54$) and the *complex* contexts ($\mathcal{X}^2(1) = .35; p = .56$). Thus, even though there were on average much fewer trials in which participants failed to provide a correct tone judgment in the present experiment than in Experiment 1, the pattern of effects was highly similar in the two experiments.

The same applies to the analysis of the RTs of correct judgments, too. The only context type that gave rise to a CONGRUENCY effect was again the *simple 'no'* context, while congruent and incongruent responses were equally fast in the other two context types. Furthermore, the *simple 'yes'* context imposed the smallest cognitive load as evidenced by the on average shortest judgment RTs. In the linear mixed-effects analysis of log-transformed RTs the CONTEXT×CONGRUENCY interaction did not reach significance ($\mathcal{X}^2(2) = 4.05; p = .13$. Instead of an interaction, the analysis provided evidence for a reliable main effect of CONGRUENCY ($\mathcal{X}^2(1) = 5.03; p < .05$). On the basis of our prior findings from Exp. 1 and guided by the theoretical considerations in the introduction we nevertheless computed follow-up analyses for each of the three context types. These analyses revealed that the 15 ms difference between the congruent and the incongruent *simple 'no'* conditions was significant ($\mathcal{X}^2(1) = 6.32; p < .05$), but that neither the *simple 'yes'* context ($\mathcal{X}^2(1) = 0.15; p = .70$) nor the *complex* con-

[7] The global model only included the random intercept for participants and no random slopes because models with more complex random-effects structures did not converge. The pairwise comparisons of the congruent vs incongruent *simple 'yes'* and *simple 'no'* contexts included the random intercept as well as the slopes for CONGRUENCY.



A) error rates

B) probe RTs

Fig. 6. Mean error rates (panel A) and probe RTs of correct judgments (panel B) in the tone task of Exp. 2. The error bars correspond to one standard error of the mean.

text ($\mathcal{X}^2(1) = 1.00; p = .32$) showed reliable effects of congruency. In addition to this congruency effect, the main effect of context was again highly significant ($\mathcal{X}^2(2) = 20.21; p < .01$). Follow-up analyses on subsets of the data showed that the *simple 'yes'* context conditions gave rise to generally faster tone judgments than the other two context types, the *simple 'no'* context conditions ($\mathcal{X}^2(1) = 14.36; p < .01$), and the *complex* context conditions ($\mathcal{X}^2(1) = 18.06; p < .01$).

As in Exp. 1 additional analyses were computed involving the factor half of experiment. The descriptive statistics are presented in Table 2. Mean RTs were on average 64 ms faster in the second half than they were in the first half of the experiment. Again, the pattern of results was remarkably consistent across the first and the second half. The simplest model able to account for this pattern of data was a mixed model including the main effects of half, context, and congruency but no interactions between the fixed effects (model comparison between this model and the saturated model: $\mathcal{X}^2(5) = 5.89; p = .32$).

The present experiment replicated the findings from Experiment 1. Taken together, the two experiments provide firm evidence that only the *simple 'no'* context gives rise to a congruence effect. Moreover, the lack of congruency effect in the complex condition could be replicated in a considerably simpler dual-task experiment employing a non-arbitrary stimulus–response mapping from tones to tone judgments. That the modified tone task with a non-arbitrary mapping was in fact simpler than the one employed in Experiment 1 becomes evident by directly comparing the magnitude of the error rates and the judgment RTs between the two experiments. The present experiment gave rise to only about half as many errors relative to Experiment 1 and the mean judgment RT was more than 150 ms faster than in Experiment 1. The modified secondary task employed here should thus be easy enough to allow for an as natural interpretation of the questions as one can get in a probe RT paradigm. Nonetheless, congruency effects were again completely absent in the complex conditions. From this, we can conclude that in contrast to the *simple 'no'* context really no answer is generated in the complex contexts. This speaks against the revision-insensitive incrementality hypothesis and in favor of some sort of revision-sensitive version. However, the lack of congruency effect in the *simple 'yes'* context calls for a modification of the account sketched in the introduction. We will come back to this issue in the General Discussion below.

Table 2
Mean error rates and mean RTs for the first and second half of Exp. 2.

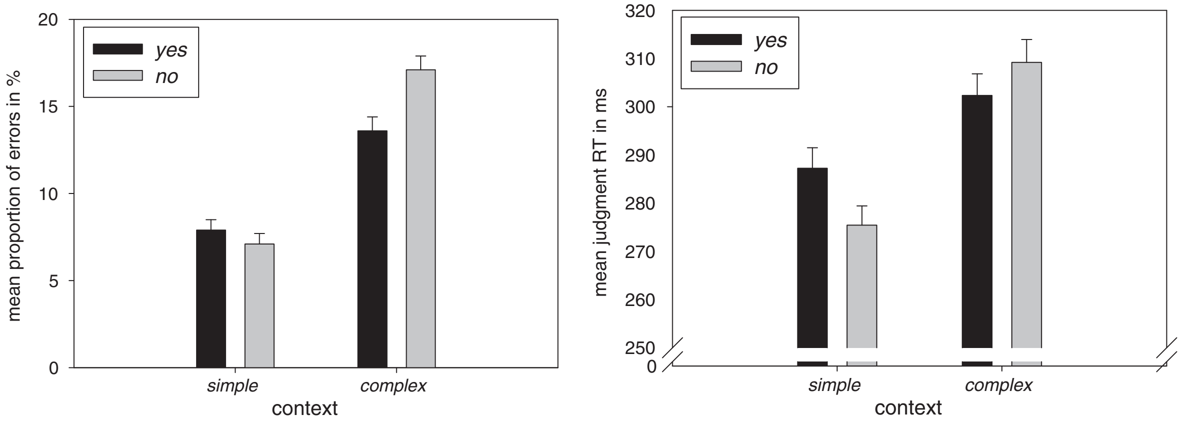
	Errors	RTs	
	M (in %)	*M* (in ms)	*SD*
First half			
Simple 'yes'			
Congruent	4.4	378	132
Incongruent	5.0	379	126
Simple 'no'			
Congruent	3.5	398	152
Incongruent	7.0	413	167
Complex			
Congruent	6.0	400	154
Incongruent	6.5	399	150
Second half			
Simple 'yes'			
Congruent	5.9	323	104
Incongruent	6.7	327	117
Simple 'no'			
Congruent	3.5	322	109
Incongruent	6.8	337	114
Complex			
Congruent	5.6	333	123
Incongruent	4.9	340	128

M = mean and *SD* = standard deviation.

4.2.2. Linguistic task

The performance in the linguistic task also largely fit the results of the previous experiment. Fig. 7 presents the mean proportions of errors and the RTs of correct responses in the linguistic task. Error rates again varied as a function of complexity. Participants made more errors for the complex contexts. We computed a logit mixed-effects model analysis with the fixed effect of complexity as well as random intercepts and random slopes for complexity for participants and items. The log-likelihood ratio test comparing this model with a model from which only the fixed effect was removed revealed that the main effect of complexity was significant ($\mathcal{X}^2(1) = 36.53; p < .01$). Overall, the proportion of incorrect responses to the questions was lower in the present (10.0% errors overall) than in the previous experiment (19.4%). This provides further evidence for far less task interference in the present than in Exp. 1.

The analysis of log-transformed RTs of correct responses to the questions revealed only a main effect of complexity. Models including the interaction complexity×polarity or the main effect of



A) error rates

B) judgment RTs

Fig. 7. Mean error rates (panel A) and RTs of correct answers (panel B) in the linguistic task of Exp. 2. The error bars correspond to one standard error of the mean.

POLARITY did not fit the data better than the model without ($\mathcal{X}^2(2) = .85; p = .65$). The reduced model was compared to a model from which the fixed effect of *COMPLEXITY* was removed. The log-likelihood ratio test provided evidence for an effect of COMPLEXITY ($\mathcal{X}^2(1) = 31.28; p < .01$). To summarize, the findings from the linguistic task showed that questions about complex contexts were more difficult than questions about simple contexts. However, the performance was generally good for all four context types. As in Exp. 1, answers to questions were provided very quickly with a mean RT of only 293 ms suggesting that the computation of the answer must have actually happened before the question mark appeared on the screen.

5. General discussion

We started out with three hypotheses about the time course of semantic interpretation varying in their degree of incrementality. The reported findings from two probe RT task experiments examining the incremental answer preparation during the online interpretation of quantified questions were not fully consistent with any of the hypotheses, albeit the effects turned out to be most consistent with the proposed revision-sensitive incrementality account. Our study thus provides further evidence for the incremental interpretation of quantifiers (Freunberger & Nieuwland, 2016; Urbach et al., 2015). Focusing on the processing of quantificational restriction, our study investigated another aspect of quantifier processing than most other studies investigating the incrementality of quantifier interpretation. The results of the present experiments complement and corroborate the findings and their interpretation by Augurzky et al. (2016) who presented ERP evidence for the revision-sensitive incrementality assumption. Relating the findings of the present study to their results it seems indeed to be the case that quantifier restriction is interpreted incrementally but that the processor takes into account the risk of reanalysis further downstream the sentence in incrementally generating answers to an unfolding question. The unique contribution of the present study consists in showing that this incremental interpretation process can even result in motor programming of a response to an incomplete question, that is it can even affect the final outcome of language interpretation conceived of as an input-output system.

Finding consistent patterns of effects across the first vs. the second half in the two experiments suggests that the observed effects cannot be explained in terms of the high probability of extraposed relative clauses in the tested experimental materials. In the first half of the experiment participants should have weaker expectations about questions ending in extraposed relative clauses. Nevertheless the cross-task effects in the first half of the experiment were even numerically stronger than in the second half. We therefore feel confident in claiming that the lack of answer facilitation in the complex context cannot receive an explanation solely based on the likelihood of upcoming restriction in the context of the present experiments. Instead it should generalize to incremental answer generation in general.

The finding that goes against the predictions of the revision-sensitive incrementality account was the complete lack of facilitation effects in the *simple 'yes'* contexts, ie. the context type for which immediate answer preparation should be most likely to occur. Why did we not find any congruency effect in this condition? Here is what we consider a likely explanation. The very short RTs in the linguistic task and in particular in the simple context conditions suggest that participants had already generated their answer before the end of the sentence. Moreover, this context type imposed the lowest cognitive load of all context types on the secondary tone judgment task. We therefore consider it highly plausible that participants had already finished generating the 'yes' answer **before** they got to the tone task. The answer preparation must in fact have occurred so early that any traces related to answer generation in the linguistic task had already fully decayed in the central answer selection stage. Since reactions in the tone task occurred about 900 ms post adjective onset in Exp. 1 and 700 ms post adjective onset in Exp. 2 answer generation in the linguistic task thus most likely occurred even before the color adjective was encountered. Otherwise, congruency effects would probably still be expected.[8]

However, as the proposed revision-sensitive model stands, the earliest point at which an answer can be generated is the color adjective. This becomes clear if we consider the *simple 'yes'* contexts once more. But this time we focus on the point when the quantifier is encountered, ie. before having accessed the scope information (e.g., *are all triangles...*). In this case, the semantic representation of the yet unfinished question (7) will contain two open slots, the first P reserved for further restrictor information and the second Q for the upcoming scope information.

[8] It is usually assumed that congruency effects emerge at the level of decision making (Kornblum, Hasbroucq, & Osman, 1990). Therefore, once a decision has been made no task interference is expected to occur.

(7) ?ALL(TRIANGLE & P)(Q)

When this partial question is evaluated in the *simple 'yes'* context, we get – among other alternatives – the following two possible continuations of the sentence. Again, without loss of generality, we will only list the propositions corresponding to the 'yes' answers and ignore 'no' answers.

(8) a. ?ALL(TRIANGLE)(BLUE)
b. ?ALL(TRIANGLE)(IN THE CIRCLE)

In the given context, the first possible continuation, corresponding to *are all triangles blue*, is evaluated as 'yes, true' while the second, corresponding to *are all triangles in the circle*, is false. Hence, no safe answer can be generated and the processing system should have to wait for the scope argument before an answer can be computed. Accordingly, an answer facilitation effect should definitely be observed on the color adjective.

We think that the derivation shows that our proposal of how revision-sensitivity could be modeled takes into account way too many possibilities. The way our proposal was formulated in the introduction it was intended as a maximally safe decision procedure allowing the processor to incrementally evaluate which point the polarity of an answer is uniquely determined. As it stands, this procedure outputs a decision only in case the risk of answer revision corresponds to a value of zero. Intuitively but also as indicated by our data, this idealizing assumption must actually be considered unrealistic. For instance, in the linguistic task employed here it was always the dimension of color that was the relevant property for the scope argument. There were no questions asking about the spatial localization of the objects. Therefore, (8-b) seems intuitively a very unlikely continuation for the initial part of the question and we would therefore only expect (8-a) to be included into the set of possible continuations. If we rule out (8-b) from the prediction set, we can immediately account for the anticipation of the color adjective and therefore prediction-based answer generation in the *simple 'yes'* contexts. Going one step further, the answer polarity can even be determined for the string *are all*... because the only entities in the model that constitute a real plurality are the triangles. If this is correct, answer generation in the *simple 'yes'* contexts could have occurred two words before the color adjective explaining why we did not observe any facilitation effects for congruent answers any more.

This explanation has interesting consequences for the *simple no* condition as well. Note that the generation of an answer before the color adjective should happen in this condition too. Crucially, however, the predictive system would generate a yes answer to a set of completely different questions than the one actually asked! For our example at hand the interpretation system would generate a yes answer to a set of questions all involving the property RED. Upon encountering the color adjective *blue*, this prediction set has to undergo a complete revision resulting in a set of new predictions involving BLUE instead of RED. In our model this corresponds to a non-monotonic update of the respective prediction sets where none of the predictions of the set computed before the adjective survives its integration. Furthermore, in terms of the prepared answer this implies that the predicted yes answer has to be suppressed and a no response has to be programmed instead. Both aspects are likely to result in increased processing demands in line with the observed processing load in this condition.

The just outlined explanation for the lack of congruency effect in the *simple 'yes'* context does not mean that comprehenders do not have to parse the rest of the question, though. Although the continuation of the question does not change the polarity of the answer, the presence of further restriction leads to a semantically different interpretation. That is, under the present account, it makes a clear difference whether the processor entertains the inter pretation ?ALL(TRIANGLE)(BLUE) or rather the interpretation ?ALL (TRIANGLE & IN THE CIRCLE)(BLUE). We would therefore like to emphasize that even though the polarity of the answer is clear right from the outset, our account still predicts that participants have to engage in question monitoring in order to properly understand the meaning conveyed.

The just proposed modification of our original account can be conceived of as a pragmatic filter on the combinatorially possible continuations principally provided by the incremental semantics outlined in the introduction. Probably, a number of factors contribute to this pragmatic filter. We have already referred to the internal probability distribution within the experiment where the scope argument was always a color adjective, but other factors such as typicality or informativity will probably also play a huge role in determining what should included into the set of relevant continuations. The current proposal should therefore ultimately be linked to the growing body of work on anticipation conducted within psycho- and neurolinguistics but also work on the mentioned pragmatic factors within experimental pragmatics. In particular, the probability of upcoming restriction should have clear processing consequences. If the overall context makes further restriction highly unlikely because extraposed relative clauses are practically never used in the context of an experiment or by a particular speaker, the prediction set should be constrained accordingly. For the complex condition in the present study this means that – given a different experimental context with no extraposed relative clauses – the prediction set might be constrained to the pair of questions ?ALL(TRIANGLE,BLUE) and ?ALL(TRIANGLE,RED). To be able to deal with such pragmatic effects the present proposal has to be turned into a probabilistic theory, though. We have to leave this important issue for future research.

Finally, the reported research adds an important aspect to work on extraposed relative clauses. By investigating the role of contextual information, our experiments identified a factor that may bear an influence on how easy it is to integrate extraposed relative clauses. In those cases where additional information can change the semantic value of a given sentence, extraposed relative clauses may actually be expected. This could be directly relevant for the finding with respect to *only those* in Levy et al.'s (2012) study where *only those executives* can only be fully interpreted if the relevant set is further restricted. If, however, the context is such that the semantic value of the sentence without the relative clause logically determines the semantic value of all possible sentence continuations, extraposition should be unexpected.

To summarize, our study has contributed data from a novel paradigm to the question whether quantificational restriction is updated incrementally during online interpretation. To our knowledge, this is the first application of a probe RT task to incremental semantic interpretation. The data suggest that semantic processing immediately takes into account the context of utterance in determining the quantificational domain. The processing system thereby seems to be set up in a way to optimize a solution to a number of opposing constraints: most importantly the working memory load due to withholding an answer contra the revision costs of answer revision. The highly incremental effects suggest that late effects in other studies on the online processing of quantificational restriction such as Kaan et al. (2007) are not primarily due to delayed processing of restrictor information but could be related to other aspects of discourse processing. Finally, our results indicate that restrictor information may be processed in a fundamentally different way than the relative scope of quantifiers in multiply quantified sentences. For the latter, Bott and

Schlotterbeck (2015) argued explicitly in favor of the *Global Interpretation Hypothesis*, a hypothesis incompatible with the data reported in the present paper. If it turns out that the time course of interpretation is in fact qualitatively different for these two aspects of quantification, a cognitively realistic quantification theory will have to consider different processing mechanisms for them.

Acknowledgments

This research has been funded by the German Science Foundation (DFG) in the Projects B1 *Incrementality of Semantic Interpretation* in the Collaborative Research Centre SFB 833 *The Construction of Meaning* and the project CiC *Composition in Context* of the Priority Program *XPrag.de New Pragmatic Theories based on Experimental Evidence*. We would like to thank Fabian Schlotterbeck for suggestions on an earlier version of this paper. Finally, we would like to thank Jochen Saile for invaluable programming advice.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247–264.

Augurzky, P., 2006. *Attaching relative clauses in German – The role of implicit and explicit prosody in sentence processing*. Leipzig: MPI Series in Human Cognitive and Brain Sciences.

Augurzky, P., Bott, O., Sternefeld, W., & Ulrich, R. (2016). Are all triangles blue? – ERP evidence for the incremental processing of german quantifier restriction. *Language and Cognition*, 1–34.

Band, G. P. H., & Miller, J. (1997). Mental rotation interferes with response preparation. *Journal of Experimental Psychology: Human Perception and Performance, 23*(2), 319–338.

Barker, C. (2002). Continuations and the nature of quantification. *Natural Language Semantics, 10*(3), 211–242.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy, 4*, 159–219.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bott, O., & Schlotterbeck, F. (2015). The processing domain of scope interaction. *Journal of Semantics, 32*(1), 39–92.

Bott, O., & Sternefeld, W. (2017). An event semantics with continuations for incremental interpretation. *Journal of Semantics, 34*(2), 201–236.

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Realtime investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive Science, 32*(4), 643–684.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117–1121.

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

Fodor, J. (2002). Prosodic disambiguation in silent reading. In M. Hirotani (Ed.). *Proceedings of the 32nd annual meeting of the north east linguistic society (nels)* (Vol. 32, pp. 112–132). Amherst, MA: GSLA.

Frazier, L., Clifton, C., Rayner, K., Deevy, P., Koh, S., & Bader, M. (2005). Interface problems: Structural constraints on interpretation? *Journal of Psycholinguistic Research, 34*, 201–231.

Freunberger, D., & Nieuwland, M. S. (2016). Incremental comprehension of spoken quantifier sentences: Evidence from brain potentials. *Brain Research, 1646*, 475–481.

Hamblin, C. (1973). Questions in Montague English. *Foundations of Language, 10*, 41–53.

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar*. Oxford: Wiley-Blackwell.

Ilan, A. B., & Miller, J. (1998). On the temporal relations between memory scanning and response preparation. *Journal of Experimental Psychology: Human Perception & Performance, 24*(5), 1501–1520.

Kaan, E., Dallas, A. C., & Barkley, C. M. (2007). Processing bare quantifiers in discourse. *Brain Research, 1146*, 199–209.

Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and Philosophy, 1*, 3–44.

Knoeferle, P., Crocker, M. W., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition, 95*(1), 95–127.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility – A model and taxonomy. *Psychological Review, 97*, 253–270.

Kounios, J., & Holcomb, P. J. (1992). Structure and process in semantic memory: Evidence from event-related potentials and reaction times. *Journal of Experimental Psychology: General, 121*(4), 459–479.

Leonhard, T., Fernández, S., Ulrich, R., & Miller, J. (2011). Dual-task processing when task 1 is hard and task 2 is easy: Reversed central processing order? *Journal of Eperimental Psychology: Human Perception & Performance, 37*(1), 115–136.

Levy, R. (2014). Using R formulae to test for main effects in the presence of higher-order interactions. (unpublished manuscript).

Levy, R., Fedorenko, E., Breen, M., & Gibson, E. (2012). The processing of extraposed structures in English. *Cognition, 122*, 12–36.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic, 8*, 339–359.

Logan, G. D., Zbrodoff, N., & Fostey, A. (1983). Costs and benefits of strategy construction in a speeded discrimination task. *Memory & Cognition, 11*(5), 485–493.

Miller, J. (1985). Discrete and continuous models of divided attention. In M. I. Posner & O. S. M. Marin (Eds.). *Attention and performance* (Vol. XI, pp. 251–281). Hillsdale, NJ: Erlbaum.

Miller, J., Coles, M. G., & Chakraborty, S. (1996). Dissociation between behavioral and psychophysiological measures of response preparation. *Acta Psychologica, 94*, 189–208.

Montague, R. (1973). The proper treatment of quantification in ordinary English. In K. Hintikka, J. Moravcsik, & P. Suppes (Eds.), *Approaches to natural language* (pp. 247--270). Dordrecht: D. Reidel.

Nieuwland, M. S. (2016). Quantification, prediction, and the online impact of sentence truth-value: Evidence from event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(2), 316–334.

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

Poschmann, C., & Wagner, M. (2016). Relative clause extraposition and prosody in german. *Natural Language and Linguistic Theory, 34*(3), 1021–1066.

Posner, M. I., & Boies, S. J. (1971). Components of attention. *Psychological Review, 78*(5), 391–408.

Schlenker, P. (2008). *Be articulate*: A pragmatic theory of presupposition projection. *Theoretical Linguistics, 34*(3), 157–212.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science, 22*, 34–80.

Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General, 123*, 34–80.

Urbach, T. P., DeLong, K. A., & Kutas, M. (2015). Quantifiers are incrementally interpreted in context, more than less. *Journal of Memory and Language, 83*, 79–96.

Urbach, T. P., & Kutas, M. (2010). Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language, 63*(2), 158–179.

von Fintel, K. (1994). *Restrictions on quantifier domains* Ph.D. dissertation. Amherst, MA: University of Massachusetts.

Wijnen, F., & Kaan, E. (2006). Dynamics of semantic processing: The interpretation of bare quantifiers. *Language and Cognitive Processes, 21*, 684–720.