Embedded disjunctions and the best response paradigm[1]

Anton BENZ — Leibniz–Centre General Linguistics, ZAS, Berlin
Nicole GOTZNER — Leibniz–Centre General Linguistics, ZAS, Berlin

Abstract. The current study investigates implicature of embedded disjunctions. We employ a paradigm in which implicatures are inferred indirectly from action choices of test subjects. This avoids meta–linguistic judgements on which previous studies relied. The focus is on four different types of implicature that may be triggered by embedded disjunctions in a situation with a competent speaker. We distinguish between local and global scalar implicatures, exhaustive and existence implicatures. The results provide evidence that varieties all four types of implicature have been inferred by a majority of subjects.

Keywords: experimental pragmatics, embedded implicature, disjunctions.

1. Introduction

The connective '*or*' is one of the core examples illustrating different types of quantity implicature. In (1), the disjunction gives rise to three quantity implicatures, the *scalar* implicature which says that the sentence with '*or*' replaced by '*and*' is false, the so–called *ignorance* or *clausal* implicatures which says that, for all that the speaker knows, it is possible that any of the disjuncts may be true or false, and the exhaustive implicature which says that Kate did not find, for example, the green marble, if she has a green marble:

(1) Kate found her blue or her red marble.
   - $\rightsquigarrow$ Scalar: Kate did not find her blue and her red marble.
   - $\rightsquigarrow$ Ignorance: $\Diamond$ /$\Diamond\neg$ Kate did find her blue marble;
     $\Diamond$ /$\Diamond\neg$ Kate did find her red marble;
   - $\rightsquigarrow$ Exhaustive: Kate did not find any other marble except the blue or the red one.

In this example, the connector is not embedded. The ignorance implicature is inconsistent with a situation in which the speaker knows the actual state of the world. Hence, un-embedded disjunction is generally not licensed in such situations. This marks a difference to embedded '*or*'. For example, '*All of the girls found their red or their blue marble*' can be uttered by a competent speaker.

The current study investigates the complex sentences with embedded disjunctions in (2).

(2) a. All of the girls found their red or their blue marble.
   b. Some of the girls found their red or their blue marble.
   c. All of the girls found their red, their blue, or their green marble.
   d. Some of the girls found their red, their blue, or their green marble.

[1]This work was supported by the Bundesministerium für Bildung und Forschung (BMBF) (Grant Nr. 01UG1411), and the Deutsche Forschungsgemeinschaft (DFG) (Grant Nr. BE 4348/4-1). We are grateful to Felicitas Enders, Mady Thonicke, and Danny Belitz for preparing the experimental stimuli.

In particular, we test the status of the following types of implicature:

(3) (A) the embedded scalar implicature of the disjunction, e.g. for (2a) the implicature '*all either r or b*' leading to an exclusive reading of '*or*';
(B1) the global implicature from (2b) to none $r \wedge b$, and from (2d) to none $r \wedge b$, none $r \wedge g$, and none $b \wedge g$;
(B2) the global implicature from (2b) to not all $r \vee b$, and from (2d) to not all $r \vee b \vee g$;
(C) the exhaustive implicature from (2a) and (2b) to *none found their green marble*;
(D) the strong existence implicature of the embedded disjunctions, for example from (2a) all $(r \vee b)$ to some $r \wedge$ some b.
(D') the weak existence implicature of the embedded disjunctions, for example from (2a) all $(r \vee b)$ to *it is possible that some r* and *it is possible that some b*.

The weak existence implicature (D') is implied by the strong existence implicature (D). We treat them as two variants of the same type of implicature. Likewise, (B1) and (B2) are treated here as two variants of the same type of implicature. The labels have to be understood descriptively. By using them, we do not mean to commit to any specific theoretical framework.

Throughout, we only consider situations in which the speaker is commonly known to be competent, i.e. knowledgeable of the true state of the world. A further assumption is that the objects that can be connected by disjunction are also commonly known. For example, in a scenario in which (2a) is uttered, it will be assumed that each of the girls owns a commonly known finite set of marbles, each with a distinct colour, that could be found by the girl.

Theories make different predictions concerning the implicatures in (3). For example, Chierchia (2004) predicts (A) and a weaker version of (B1), e.g. (2b) $\rightsquigarrow$ (some $r \dot{\vee} b$ and $\neg$all $r \vee b$) but does not address (D). Franke (2009) predicts (A) and (B1), and Sauerland (2004) (B1) and for (2a) a weaker version of (A) $(\neg \text{all}\, r \wedge b)$. None of the theories predicts (C), and (D) is only explained by Sauerland (2004) and Crnič et al. (2015) for the sentences with universal quantifier. All theories predict (B2). This is only a sample of the theories about embedded implicature that could be considered (e.g. Asher, 2013; Chierchia et al., 2012; Benz, 2012; Potts et al., 2016). None of the theories addresses all types of implicature, and not all of them are specific enough to be testable.

We present clear experimental evidence that the implicature of types (A), (B1), (C), and (D') can be drawn reliably in a scenario that is based on a game theoretical design (*best response paradigm*, Gotzner and Benz, 2018). In contrast to previous experimental studies (e.g. Geurts and Pouscoulous, 2009; Chemla and Spector, 2011), our paradigm provides an organic setting that avoids meta-linguistic judgement. In the first section, we discuss the general methodological motivation for the best response paradigm. The second section presents the experiments and their results, and the third discusses their evidence for the different types of implicature listed in (3). For example, we will see that there is no evidence for the type (B2) implicature being inferred reliably, which is surprising as they are predicted by all theories. In the fourth section we compare our results to the more recent study of Crnič et al. (2015), who found evidence for strong existence implicatures (D) in absence of implicatures of type (A) and (C).

2. Embedded disjunctions: some theory and some experimental issues

A large body of experimental research on implicatures has emerged over the past decade. The majority of experiments concentrate on questions of acquisition (Noveck, 2001; Papafragou and Musolino, 2003; Katsos and Bishop, 2011), the time course of implicature processing (Noveck and Posada, 2003; Huang and Snedeker, 2009; Grodner et al., 2010; Tomlinson et al., 2013), or the question whether they are generated by default or triggered in context (Breheny et al., 2006). There have been considerably fewer studies on embedded implicatures (Chemla 2009; Geurts and Pouscoulous 2009; Clifton Jr and Dube 2010; Chemla and Spector 2011; Potts et al. 2016; Crnič et al. 2015; Gotzner and Romoli 2017). These studies have employed various paradigms, picture verification tasks (e.g. Geurts and Pouscoulous, 2009), inferencing tasks (e.g. Chemla, 2009), graded acceptability tasks (e.g. Chemla and Spector, 2011), and picture selection tasks (Clifton Jr and Dube, 2010). In these studies, sentences as (2a) are considered ambiguous between a semantic and one or more pragmatic interpretations. The task is to determine whether there is a sub-population that interprets the test sentence in accordance with the critical pragmatic interpretation. There has been a sharp controversy about methodological issues. For example, on the one side it has been argued that inferencing tasks inflate the proportion of pragmatic interpretations (Geurts and Pouscoulous, 2009), and that graded acceptability judgements and picture selection tasks are susceptible to typicality effects (Geurts and van Tiel, 2013; van Tiel, 2014), so that the evidence for embedded implicatures provided by experiments based on these designs may be doubted. On the other side, it has been argued that picture verification tasks induce subjects to interpret semantically, and therefore to underestimate the real proportion of subjects adopting the critical pragmatic interpretation (Clifton Jr and Dube, 2010; Benz and Gotzner, 2014). With the exception of Crnič et al. (2015), the studies were only concerned with embedded scalar implicatures, and generally showed only low proportions of subjects choosing the critical interpretation.

The low proportions may seem unproblematic if the goal is to show that subjects *can* arrive at certain interpretations. In the standard neo-Gricean theory of conversational implicature (Levinson, 1983), implicatures are not alternative readings of a sentence but supplements to semantic content and part of communicated meaning. To show that an interpretation is implicated in this stronger sense, it has to be shown that it is understood by all addressees on a par with semantic content. We are therefore interested in the question which potential implicatures are reliably inferred such that they can count as part of communicated meaning, and which are not. Experiments show a certain degree of random behaviour. We, therefore, can only try to determine, however, which propositions are inferred with high probability, where the term '*high probability*' introduces a certain amount of vagueness.

Pragmatics is about language in use. We, therefore, devised a scenario in which critical sentences are used for communicating facts that the addressee needs to know for subsequent decision making. Our initial hypothesis was that all four types of implicature (A) to (D) are drawn reliably. We will see that the observed response pattern indicates that implicatures of type (A), (B1), (C), and (D') are drawn reliably.

3. The Best Response paradigm

The following experiment takes advantage of the fact that interpretations can be indirectly inferred from action choices of interpreters. In this way, meta-linguistic judgements can be circumvented. As we have argued elsewhere, meta-linguistic judgements, in particular, picture verification tasks, bias subjects towards literal interpretation (Benz and Gotzner, 2014). The presence of a substantial group of literally interpreting subjects dooms any attempt at showing that certain implicatures are communicated reliably. We therefore developed a scenario in which utterance selection and interpretation are embedded in a cooperative action selection task.

3.1. Methods

3.1.1. Participants

Participants with US IP addresses were recruited on Amazon's Mturk platform and were screened for their native language. In total, 20 native English speakers (mean age: 32.7, 13 female, 7 male) took part in the experiment.

3.1.2. Scenario and task

In Gotzner and Benz (2018), we presented the basic version of the best response paradigm. For the current experiment, we use the same basic scenario with minor modifications. Participants in our experiment were presented with a scenario involving four girls who each own a set of three special edition marbles, consisting of a blue, a green and a red marble (a scenario introduced by Degen and Goodman 2014, which we extended with an action-based task). While the girls are playing, the marbles get lost and they have to find them again. Participants in our experiment were told that the mother of the girls wants to reward them depending on how many marbles they find. In particular, participants were presented with the following reward system in the instructions.

(4) Reward system:
- chocolate: all 3 marbles
- candy: 2 marbles
- gummy bear: 1 marble
  - green gummy bear: green marble
  - red gummy bear: red marble
  - blue gummy bear: blue marble
- pretzel stick: 0 marbles

The participants's task in the experiment is to buy sweets as rewards for the four girls depending on the statement the mother utters. After participants had read the instructions, they were asked control questions about the number of marbles each girl owns and which reward type a girl gets depending on how many marbles she found. Then, participants were given an example

item involving the statement '*None of the girls found any of their marbles*', in which case they should buy pretzels and nothing else.

3.1.3. Experimental items

In the main part of the experiment, participants were asked to give binary responses (YES/NO) for each of the six types of sweets: chocolate, candy, gummy bears (red, green or blue) and pretzel sticks. There were two types of critical test sentences: either simple or complex disjunction was embedded under *all* and *some*. The test sentences are shown in (5). Our four critical sentences were repeated twice in the experiment.

(5) a. All of the girls found their red or their blue marble.
b. Some of the girls found their red or their blue marble.
c. All of the girls found their red, their blue, or their green marble.
d. Some of the girls found their red, their blue, or their green marble.

An example experimental trial with All $(r \vee b)$ and a possible response choice is presented in (6). Participants had to check one of two radio buttons ('YES' or 'NO') for each type of sweets.

(6)

The mother says:	'All of the girls found their red or blue marble'
chocolate	○ YES ⊙ NO
candy	○ YES ⊙ NO
green gummy bear	○ YES ⊙ NO
red gummy bear	⊙ YES ○ NO
blue gummy bear	⊙ YES ○ NO
pretzel stick	○ YES ⊙ NO

In addition, participants saw 15 filler items such as the statement 'Sue and Kate found some of their marbles' as well as the original test sentences used in Gotzner and Benz (2018) involving the quantifier *some* embedded under *all* and *some* itself. Hence, each participant saw 23 items in total.

3.2. Results

Figure 1 shows the mean percentage of YES responses across reward type for simple disjunctions and Figure 2 shows the data for complex disjunctions (Table 4 in the appendix details the percentages of each combination of action choices per condition). We computed two separate logit mixed models for simple and complex disjunctions. Our dependent variable was binary (choice of sweet: 1 or 0) and we included the fixed factors quantifier condition (All $(r \vee b)$ vs. Some $(r \vee b)$), reward type (candy, chocolate, red, green, blue gummy bear, pretzel stick), their interaction as well as random factors for participants and items. As reference level we chose the condition (All $(r \vee b)$, candy) and (All $(r \vee b \vee g)$, candy) respectively.

Figure 1: % YES responses for sentence All $(r \vee b)$ (left column) and Some $(r \vee b)$ (right column) across reward type. Error bars represent SEM.

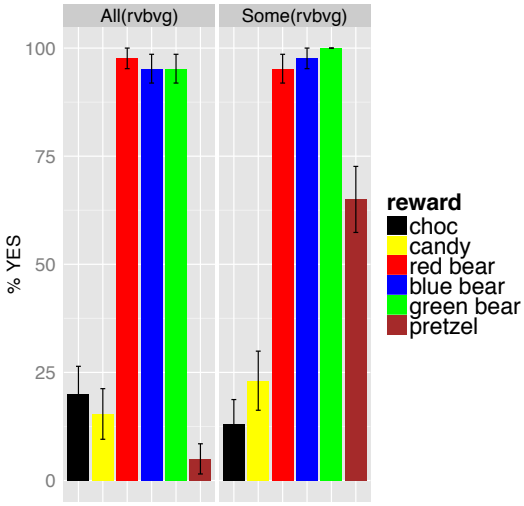

Figure 2: % YES responses for sentence All $(r \vee b \vee g)$ (left column) and Some $(r \vee b \vee g)$ (right column) across reward type. Error bars represent SEM.

Mixed model for simple disjunctions. The model showed a significant difference across reward type: Participants chose candy less often than red bears (p <.001) and blue bears (p <.001). There was no difference between the two quantifier conditions at the baseline level or between candy and the other two reward types. However, the model revealed an interaction between quantifier condition and reward type (pretzel : Some $(r \vee b)$: p <.0001). Table 5 in the appendix shows a summary of the mixed model.

Mixed model for complex disjunctions. The model showed a significant difference across reward type: Participants chose candy less often than red bears, green and blue bears (all p's <.0001). There was also a marginal difference compared to the condition with pretzels (p = .09) as well as an interaction across condition and reward type (pretzel : Some $(r \vee b \vee g)$ p <.001). There were no further main effects nor any interactions. Table 6 in the appendix shows a summary of the mixed model.

4. Evaluation of results

We tested the four sentences presented in (2). The task of the participants was to buy sweets for the four girls. They were asked to give binary responses (yes/no) for each of the six types of sweets: chocolate, candy, red, blue, and green gummy bears, and pretzel sticks. If subjects draw the implicatures (A)–(D), then their expected response pattern is that shown in (7).

(7)

condition		choc.	candy	red b.	blue b.	green b.	pretzel
(2a)	All $(r \vee b)$	no	no	yes	yes	no	no
(2b)	Some $(r \vee b)$	no	no	yes	yes	no	yes
(2c)	All $(r \vee b \vee g)$	no	no	yes	yes	yes	no
(2d)	Some $(r \vee b \vee g)$	no	no	yes	yes	yes	yes

As we have seen, the observed choices conform to this prediction. The only exception are the choices with respect to pretzels in the *some*-conditions. We observe a larger proportion of subjects buying no pretzels (about 30%),[2] which is inconsistent with there being a group of girls that found none of their marbles. This implicature follows from an utterance of (2b) by the following reasoning: First, with (B2) Some $(r \vee b)$ implicates ¬all $(r \vee b)$; second, (C) implies that none found the green marble. Together, this implies that some found none. If subjects would not adhere to (C), then they should have bought either chocolate, candy or green gummy bears in the '*some* $(r \vee b)$' condition (2b). As they did not buy them, the problem must be the implicature of type (B2). Hence, the experiments indicate that there is a significant proportion of subjects that do not draw the (B2) implicature, i.e. global scalar implicature from embedding '*some*'. However, we can conclude that the hypothesis that subjects draw implicatures (A), (B1), (C), and (D) is *consistent* with the experimental results.

We next consider the more difficult question of how far the results shown in Table 4 in the appendix *imply* that subjects infer (A), (B1), (C), and (D). As the percentage values follow the

[2] A mixed model analysis with disjunction embedded under *some* and the reward type pretzels as reference level revealed that this condition differed significantly from all other reward types (all p's $< .001$, both for simple and complex disjunctions).

response pattern shown in (7), with the exception of the pretzel results for (2b) and (2d), we abstract from the precise numerical values and concentrate on the evidence provided by the categorical pattern in (7). What we have to show is that for all possible belief states that are consistent with the subjects's choices of sweets, the implicatures generated by (A), (B1), (C), and (D) hold true.

In our scenario, each observable world can be identified with a 4-tuple (m_1, m_2, m_3, m_4) of sets of marbles $m_i \subseteq \{b, g, r\}$. The set m_i represents the set of marbles found by the i^{th} girl. There are $(2^3)^4 = 4096$ different observable worlds. The reward system with different kinds of sweets distinguishes $2^6 - 1 = 63$ relevant states, of which $\sum_{i=1}^{4} \binom{6}{i} = 56$ can be instantiated by four girls. In the following, we mean by '*possible world*' always one of the 63 possible worlds defined by the reward system, and not one of the observable worlds.

We consider the four conditions (2a) to (2d) one after the other. First, let us consider (2a):

(2a) All of the girls found their red or their blue marble.

This sentence is semantically consistent with 24 possible worlds, 8 of which are shown in Table 1. The other 16 worlds can be found by making one copy of the table and replacing the 1s in the '*blue*' column by 0s, and then by making another copy and replacing the 1s in the '*red*' column by 0s. Of the 24 worlds only one world is consistent with the pragmatic choice as indicated by the ✓ in the last column.

Possible worlds (8 of 24)

pretzl	blue	green	red	candy	choc	cons
0	1	1	1	1	1	–
0	1	1	1	1	0	–
0	1	1	1	0	1	–
0	1	1	1	0	0	–
0	1	0	1	1	1	–
0	1	0	1	1	0	–
0	1	0	1	0	1	–
0	1	0	1	0	0	✓

Information states consistent with choice

inf. state	pretzl	blue	green	red	candy	choc
I	0	1	0	1	0	0
II	0	1	0	1	0	0
	0	1	0	0	0	0
III	0	1	0	1	0	0
	0	0	0	1	0	0
IV	0	1	0	0	0	0
	0	0	0	1	0	0

Table 1: Target sentence (2a) *All of the girls found their red or their blue marble.*

If the addressee infers an exact state of the world, then we can conclude that she believes to be in the world indicated by ✓. However, there are three other belief states that are consistent with the choice of only red and blue gummy bears. They are also shown in in Table 1. The experimental set up does not allow us to distinguish between the four states. We now have to check whether the implicatures predicted by (A), (B), (C), and (D) are supported by all four of them. (A) is the embedded implicature that none found both the red and the blue marble; if the addressee believes that this implicature might be false, then there must exist a belief world in which it is appropriate to buy hard candy; as in neither of the four belief states there exists such a world, we can conclude that the addressee does not believe it possible that the implicature is

false. Hence, the embedded implicature (A) is true. (B1) and (B2) only apply to the '*some*' sentences. The exhaustive implicature (C) that none found the green marble is also true in all belief states. We can also immediately see that the weak existence implicature (D') holds: each belief state contains a world with a 1 in the *red* column, and a world with a 1 in the *blue* column; hence, in each belief state it is possible that a girl found a red marble and it is possible that a girl found a blue marble. The stronger existence implicature (D) is violated in information states II, III, and IV: in each of these states there is a world in which either the *red* column contains a 0 or the *blue* column.

We next consider condition (2b):

(2b) Some of the girls found their red or their blue marble.

This sentence is semantically consistent with 48 possible worlds, 9 of which are shown in Table 2. The 48 worlds consist of the 24 worlds consistent with (2a) together with a copy of these worlds where the 0s in the '*pretzel*' column are replaced by 1s. Of the 48 worlds only two worlds are consistent with the experimental results: one world consistent with the choice of the subjects who bought pretzels, and one world consistent with the choice of the subjects who did not by them. The two worlds are indicated by ✓ in the last column of Table 2.

Possible worlds (9 of 48)

pretzl	blue	green	red	candy	choc	cons
1	1	1	1	1	1	–
1	1	1	1	1	0	–
1	1	1	1	0	1	–
1	1	1	1	0	0	–
1	1	0	1	1	1	–
1	1	0	1	1	0	–
1	1	0	1	0	1	–
1	1	0	1	0	0	✓
0	1	0	1	0	0	✓

Information state consistent with choice (1 of 57)

pretzl	blue	green	red	candy	choc
1	1	0	1	0	0
1	1	0	0	0	0
1	0	0	1	0	0
0	1	0	1	0	0
0	1	0	0	0	0
0	0	0	1	0	0

Table 2: Target sentence (2b) *Some of the girls found their red or their blue marble.*

For (2b), there are many more belief states than for condition (2a). One belief state is shown in Table 2. We can generate 57 belief states by taking all subsets of this state which contain a world with a 1 in the '*blue*' column and a world with a 1 in the '*red*' column.

If the addressee believes to be in one of the worlds indicated by ✓, then she believes in the implicature generated by (A), (B1), (C), and (D). If it is the world with a 1 in the '*pretzel*' column, then she also draws the global scalar implicature for embedding '*some*' (B2). Our results showed that this is true only for about two-third of the subjects. If no belief state contains a world where candy or chocolate has to be bought, then the addressee must believe in the (A) implicature. As all belief states are subsets of the one shown in Table 2, this holds true. The global implicature of embedded '*or*' (B1) says that none of the girls found the red and the blue marble. This follows from the fact that no belief state contains a world in which candy or chocolate has to be bought. The global implicature (B2) stating that not all girls found the red

A	B	D	E	F
A	B			F
A		D	E	F
A		D		F
A	B			F

Table 3: A critical item probing for existence implicature from Crnič et al. (2015: p. 15).

or the blue marble is violated if a belief state contains a world with a 0 in the '*pretzel*' column. We have seen that such belief states are not ruled out by our data. There is a larger proportion of subjects whose choice of sweets is consistent only if they are not drawing the (B2) implicature, as we have argued before. The exhaustive implicature (C) that none found the green marble is true in all belief states. We can again immediately see that the weak existence implicature (D') holds. The stronger existence implicature (D) is violated in all but three belief states.

For the remaining conditions (2c) and (2d), the claims follow by analogous arguments.

(2c) All of the girls found their red, their blue, or their green marble.
(2d) Some of the girls found their red, their blue, or their green marble.

The motivation for testing these sentences in addition to (2a) and (2b) was that deriving the embedded scalar implicatures (A) for double disjunction in contrast to single disjunction is more difficult in localist frameworks (Chierchia et al., 2012). However, our results do not indicate any significant difference between them.

5. Comparison with Crnič et al.

Our aim was to test whether the different types of implicature associated with embedded '*or*' are part of communicated meaning. Within common error ranges, subjects followed the response pattern shown in (7), except for '*some*' sentences, for which a substantial proportion of subjects failed to infer that some of the girls found none of the marbles. This pattern suggests that indeed implicatures from (A), (B1), (C), and (D') are reliably drawn in our scenario. As mentioned earlier, the goal of our study contrasts with that of many other studies of embedded implicature. A relevant example is that of Crnič et al. (2015) who tested for strong existence implicature in a picture verification task. They only considered sentences of type (2a), i.e. no sentences with embedding '*some*':

(8) Every box contains an A or a C.

A critical picture is shown in Table 3. For this picture, a significant proportion of subjects rejected sentence (8) while almost all accepted the following sentences:

(9) a. Every box contains a B or a D.
b. Every box contains an A or a B.
c. Every box contains an A or an F.

As the strong existence implicature (D) from sentence (8) is false in the given situation, it can be inferred that some subjects draw this implicature. As subjects accepted (9b) and (9c), they must have ignored the embedded scalar implicature (A), as well as the weaker global scalar implicature from '*all* $p \vee q$' to '$\neg$*all* $p \wedge q$'. None of the sentences in (9) is consistent with the exhaustive implicature (C). Hence, the results differ significantly from our study. We assume that this difference is due to the different tasks the subjects had to perform. As mentioned before, picture verification tasks push subjects to purely semantic interpretations. In contrast, we provided them with an action selection task which does not directly involve reasoning about truth conditions. This does not necessarily entail that one design is deficient, and the other does it right. They simply test interpretation in different contexts. The question is what we want to find out about language use. Crnič et al. (2015) wanted to show that there exist a non-empty group of subjects that draw existence implicatures without drawing other implicatures. They are not concerned with showing that they are drawn by almost all subjects. The critical sentence was accepted in 78.4% of all occasions, in contrast to 97% and 93.1% for sentences (9b) and (9c). Hence, the existence implicature has been inferred in at most 22% of all cases in which a critical sentence was presented. In terms of individuals, they found that only 6 out of 51 subjects drew the existence implicature consistently. It is clear that these percentages could not count as evidence that they are drawn reliably.

A drawback of the design we used is that it is not sensitive to the distinction between weak and strong existence implicature. We would have to rule out uncertainty about interpretations on the hearer side. If their belief states consist of only one world, for example of the worlds marked by ✓ in Tables 1 and 2, it would be obvious that the strong existence implicature is inferred. In order to show that uncertainty can be ruled out, we have to develop our paradigm further.[3]

6. Conclusions

The results of our study provide clear evidence that participants compute local scalar implicatures (A), global implicature for strengthened '*or*' (B1), exhaustive implicaures (D), and weak existence implicature (D'). If uncertainty about interpretation on the hearer side can be ruled out, the strong existence implicature (D) can be assumed. We were concerned with implicature as part of communicated meaning. Hence, we had to show that a very high proportion of subjects draw these implicatures. As shown in the introduction, none of the existing theories can account for all these implicatures. In our previous discussions, we have not addressed the issue of how the observed implicature can be accounted for theoretically. The contrast between picture verification tasks, for example (Geurts and Pouscoulous, 2009; Crnič et al., 2015), and the best response paradigm (Gotzner and Benz, 2018) show that context specific parameters, as for example shared assumptions about the speaker's knowledge about the exact state of the world, and general domain tasks that define relevant meaning differences, play a crucial role in implicature generation.

[3]To simply add an answer option '*I don't know*' to the options '*yes*' and '*no*' would not be enough as this option introduces new uncertainties about its interpretation.

References

Asher, N. (2013). Implicatures and discourse structure. *Lingua 132*, 13–28.

Benz, A. (2012). Implicatures of complex sentences in error models. In A. Schalley (Ed.), *Practical theories and empirical practice*, pp. 273–306. Amsterdam: John Benjamins.

Benz, A. and N. Gotzner (2014). Embedded implicatures revisited: Issues with the truth-value judgment paradigm. In J. Degen, M. Franke, and N. D. Goodman (Eds.), *Proceedings of the Formal & Experimental Pragmatics Workshop*, Tübingen, pp. 1–6.

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

Chemla, E. (2009, May). Universal implicatures and free choice effects: Experimental data. *Semantics and Pragmatics 2*(2), 1–33.

Chemla, E. and B. Spector (2011). Experimental evidence for embedded scalar implicatures. *Journal of Semantics 28*(3), 359–400.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax / pragmatics interface. In A. Belletti (Ed.), *Structures and Beyond*, pp. 39–103. Oxford: Oxford University Press.

Chierchia, G., D. Fox, and B. Spector (2012). Scalar implicature as a grammatical phenomenon. In C. Maienborn, K. von Heusinger, and P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning*, Volume 3, pp. 2297–2331. Berlin: De Gruyter Mouton.

Clifton Jr, C. and C. Dube (2010, July). Embedded implicatures observed: A comment on Geurts and Pouscoulous (2009). *Semantics and Pragmatics 3*(7), 1–13.

Crnič, L., E. Chemla, and D. Fox (2015). Scalar implicatures of embedded disjunction. *Natural Language Semantics 23*, 271–305.

Degen, J. and N. D. Goodman (2014). Lost your marbles? The puzzle of dependent measures in experimental pragmatics. In P. Bello, M. Guarini, M. McShane, and B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pp. 397–402.

Franke, M. (2009). *Signal to Act: Game Theory in Pragmatics*. Ph. D. thesis, Universiteit van Amsterdam. ILLC Dissertation Series DS-2009-11.

Geurts, B. and N. Pouscoulous (2009, July). Embedded implicatures?!? *Semantics and Pragmatics 2*(4), 1–34.

Geurts, B. and B. van Tiel (2013). Embedded scalars. *Semantics and Pragmatics 6*(9), 1–37.

Gotzner, N. and A. Benz (2018). The best response paradigm: A new approach to test implicatures of complex sentences. *Frontiers in Communication 2*(21).

Gotzner, N. and J. Romoli (2017). The scalar inferences of strong scalar terms under negative quantifiers and constraints on the theory of alternatives. *Journal of Semantics 35*, 95–126.

Grodner, D. J., N. M. Klein, K. M. Carbary, and M. K. Tanenhaus (2010). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition 116*, 42–55.

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

Katsos, N. and D. V. Bishop (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition 120*(1), 67–81.

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

Noveck, I. (2001). When children are more logical than adults: Investigations of scalar implicature. *Cognition 78/2*, 165–188.

Noveck, I. and A. Posada (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language 85*, 203–210.

Papafragou, A. and J. Musolino (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition 86*, 253–282.

Potts, C., D. Lassiter, R. Levy, and M. C. Frank (2016). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics 33*, 755–802.

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy 27*, 367–391.

Tomlinson, J., T. M. Bailey, and L. Bott (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language 69*, 18–35.

van Tiel, B. (2014). *Quantity Matters: Implicatures, Typicality and Truth*. Ph. D. thesis, Radboud Universiteit Nijmegen.

7. Appendix

7.1. Instructions

Ann, Sue, Mary and Kate are sisters. They are really into collecting marbles. Each of the sisters has a set of 3 special edition marbles. One set consists of a green, a blue, and a red marble. Each set is special, so the girls know which marble belongs to whom. Their mother is playing a game with them. She hides their marbles in the house. Then the girls search for their sets. In the end, they receive a reward.

- If a girl finds
  - call 3 of her marbles, she will win a chocolate
  - 2 of her marbles, she will win a candy
  - 1 of her marbles, she will win
    a red gummy bear, if she found her red marble
    a green gummy bear, if she found her green marble
    a blue gummy bear, if she found her blue marble
  - 0 of her marbles, she will win a pretzel stick as consolation prize.

The mother is really good at hiding marbles. So, it is really difficult for the girls to find them. The girls never help each other, and if one of them spots the marble of another one, then she ignores it. In this experiment you will read sentences that were uttered by the mother after she checked the marble bags.

Before you begin the experiment, please answer the following question: How many special edition marbles does each sister own? ___
What reward will a girl get when she has found...
all 3 of her marbles? ___
only 2 of her marbles? ___
only 1 of her marbles, if she found the red one? ___
only 1 of her marbles, if she found the green one? ___
only 1 of her marbles, if she found the blue one? ___
none of her 3 marbles, if she found the red one? ___

The Main Task:

Please decide which sweets you would buy as a reward for the girls depending on the mother's statement.

7.2. Detailed percentages of action choice across conditions

Reward/condition	All $(r \vee b)$	Some $(r \vee b)$	All $(r \vee b \vee g)$	Some $(r \vee b \vee g)$
candy	18%	15%	20%	13%
chocolate	8%	10%	15%	23%
green bear	8%	18%	95%	100%
red bear	93%	95%	98%	95%
blue bear	93%	93%	95%	98%
pretzel stick	11%	69%	5%	65%

Table 4: Different reward choices for simple and complex disjunctions embedded under *all* and *some*

	Estimate	SD	z-value	p-value
(Intercept)	-1.7518	0.5292	-3.310	0.000932
choc	-1.2033	0.8227	-1.463	0.143590
red bear	4.6708	0.7963	5.866	0.000
blue bear	4.6600	0.7943	5.867	0.000
green bear	-1.2074	0.8235	-1.466	0.142592
pretzel	-0.8132	0.7560	-1.076	0.282065
Some $(r \vee b)$	-0.2592	0.6805	-0.381	0.703291
choc : Some $(r \vee b)$	0.6246	1.1228	0.556	0.577997
red bear : Some $(r \vee b)$	0.7245	1.1758	0.616	0.537774
blue bear : Some $(r \vee b)$	0.2909	1.1019	0.264	0.791762
green bear : Some $(r \vee b)$	1.4368	1.0685	1.345	0.178716
pretzel : Some $(r \vee b)$	3.8622	0.9969	3.874	0.000107

Table 5: Summary of logit mixed effects model for simple disjunction (n = 474, log-likelihood = -150.9)

	Estimate	SD	z-value	p-value
(Intercept)	-2.44296	0.69573	-3.511	0.000446
choc	0.52841	0.73897	0.715	0.474567
red bear	7.20754	1.27769	5.641	0.000
blue bear	6.40951	1.04230	6.149	0.000
green bear	6.41877	1.04435	6.146	0.000
pretzel	-1.71050	1.01189	-1.690	0.090951
Some $(r \vee b \vee g)$	0.72414	0.73288	0.988	0.323119
choc : Some $(r \vee b \vee g)$	-1.54500	1.07413	-1.438	0.150329
red bear : Some $(r \vee b \vee g)$	-1.53248	1.48651	-1.031	0.302576
blue bear : Some $(r \vee b \vee g)$	0.04538	1.47837	0.031	0.975511
green bear : Some $(r \vee b \vee g)$	20.60431	273.67643	0.075	0.939986
pretzel : Some $(r \vee b \vee g)$	4.45654	1.23688	3.603	0.000314

Table 6: Summary of logit mixed effects model for complex disjunction (n = 488, log-likelihood = -117.7)