Embedded implicature in a new interactive paradigm[1]

Anton BENZ — *Leibniz Centre General Linguistics*
Nicole GOTZNER — *Leibniz Centre General Linguistics*
Lisa RAITHEL — *Potsdam University*

Abstract. Previous research on scalar implicature has primarily relied on metalinguistic judgment tasks and found varying rates of such inferences depending on the nature of the task and contextual manipulations. This paper introduces a novel interactive paradigm involving both a production and a comprehension component, thereby fixing a precise conversational context. The main research question is what is reliably communicated by *some* in this communicative setting, when the quantifier occurs in unembedded positions as well as embedded positions. Our new paradigm involves an action-based task from which participants' interpretation of utterances can be inferred. It incorporates a game–theoretic design, including a precise model to predict participants' behaviour in the experimental context. Our study shows that embedded and unembedded implicatures are reliably communicated by *some*. We propose two cognitive principles which describe what can be left unsaid. In our experimental context, a production strategy based on these principles is more efficient (with equal communicative success and shorter utterances) than a strategy based on literal descriptions.

Keywords: scalar implicature, embedded implicature, experimental pragmatics, game–theoretic pragmatics.

1. Introduction

In the current paper, we introduce a new experimental paradigm to test implicatures in an interactive scenario. We provide comprehension data on a variety of utterance combinations involving one or multiple scalar terms.

Implicatures of complex sentences have been a controversial topic of discussion. A variety of theoretical approaches have been developed (e.g. Chierchia et al. 2012, Sauerland 2004, Franke 2009, Benz 2012, Pavan 2013, Potts et al. 2016), and conflicting experimental evidence has been produced (e.g. Geurts and Pouscoulous 2009, Chemla and Spector 2011, van Tiel 2014). The relevant complex sentences are those in which an implicature trigger like '*some*' is embedded under a quantifier, which may itself be an implicature trigger. For example, the sentence (A-E) '*Each girl found some of her marbles*' potentially gives rise to the inference that each girl found some but not all of her marbles. In the course of this debate, a view took hold according to which sentence meaning is highly ambiguous, and different implicatures are just different readings that language speakers may entertain (in particular Chierchia et al. 2012). In this paper, we instead are guided by the standard neo–Gricean view (Levinson 1983) that considers implicature as part of communicated meaning. Therefore, our main research question

[1] We would like to thank Danny Fox, Uli Sauerland, Jack Tomlinson, Bob van Tiel and the audience of Sinn und Bedeutung and the XPRAG Conference in Cologne for helpful comments on this work. This research was supported by the Bundesministerium für Bildung und Forschung (BMBF) (Grant Nr. 01UG1411), and the Deutsche Forschungsgemeinschaft (DFG) (Grant Nr. BE 4348/4-1). Author names appear in alphabetical order. AB and NG contributed equally to writing this paper and LR programmed the system for the experiments.

© 2018 Anton Benz, Nicole Gotzner and Lisa Raithel. In: Uli Sauerland and Stephanie Solt (eds.),
Proceedings of Sinn und Bedeutung 22, vol. 1, ZASPiL 60, pp. 205–221. ZAS, Berlin.

is: What can be reliably communicated by sentences containing embedded or un-embedded '*some*'? In the following, we operationalise this research question and develop a new interactive experimental paradigm that involves the production and interpretation of embedded '*some*'. We started out with the following basic idea: A speaker who wants to communicate a certain proposition can express all he wants to express literally, or he may take advantage of implicature, and leave certain aspects unsaid. This will lead to a shortening of utterances. Hence, our main research question can be reformulated as follows: To what extent can a description be shortened without jeopardizing communicative success? The shortest descriptions will then reveal all the implicatures that can be communicated reliably. To turn this idea into a testable theory, we formulated two cognitive principles that guide the elimination of linguistic material related to embedded '*some*': (ENA-Elim) the simplification of '*some but not all*' to '*some*', and (N-X-Elim) the elimination of '*none found X*'. For example, together they allow the simplification of literal '*Some found all, some some but not all, and none none*' to '*some all and some some*'. Our assumption was that utterance simplifications based on (ENA-Elim) and (N-X-Elim) communicate the intended message as reliably as the corresponding literal description, and all further simplification leads to unreliable communication.

With utterances composed of sentences of the form (X-Y) '*X of the girls found Y of the marbles*' with *X* and *Y* chosen from quantifier phrases '*none*', '*some*', '*any*', '*some but not all*', '*some and possibly all*', and '*all*', seven different worlds can be semantically distinguished depending on whether there are some who found none, some who found some but not all , or some who found all. As a next step towards a testable hypothesis, we defined a critical production strategy for the seven possible worlds applying the two elimination rules to a literal production strategy.

The main hypotheses we tested in our experiments were the following: (I) The critical strategy is as successful at communicating the state of the world as the corresponding literal strategy; (II) any further reduction of utterance length leads to a considerable decrease in communicative success. In the following, we present an experimental study that tests the efficiency of the critical strategy for all seven worlds. Specifically, we tested whether this strategy is communicatively successful, and how it compares to strategies pursued by naive participants, in particular whether they produce shorter utterances, and if so, whether these utterances are still successful. Our experiments indicate that the critical strategy is among the shortest strategies with almost maximal communicative success.

This paper is organized as follows. First, we review theoretical and experimental research on embedded implicature. Second, we provide the background assumptions for our new interactive paradigm. Third, we present two experiments implementing the paradigm. Finally, we compare our new experimental paradigm to previously-used paradigms and discuss the implications of the findings for theories of implicature.

2. Embedded implicature: theory and experiments

Consider the following scenario: there are four girls that have to clean up their rooms and find their marbles with which they played before. If one parent says '*Some of the girls found their marbles*', then the other parent can infer that not all of the girls found them. Grice (1975) explained this inference from the assumption that the speaker is truthful and follows the so-called *maxim of quantity*, which requires utterances to be as informative as required. In a situation in which all girls found their marbles, a truthful parent could have said both '*all found them*' and '*some found them*'. The first alternative is more informative and, presumably, the additional information is also relevant, hence, the maxim of quantity would compel the parent to say '*all*'. As s/he said '*some*', the probable reason is that not all found their marbles. It follows that some but not all must have found them.

This reasoning was systematised by Horn (1972, 1989), Gazdar (1979), and others. Their model is known as the neo–Gricean model of scalar implicature.[2] In this model, the two alternatives '*some*' and '*all*' form a *scale*, which means that they are equally complex, and that sentences with '*some*' are logically weaker than the corresponding sentences with '*all*'. If the speaker chooses the weaker alternative, then *normally* the addressee is entitled to infer that the stronger alternative is false. This inference is called an *implicature*, and since it is triggered by the scale ⟨*all,some*⟩, it is called a *scalar* implicature.

The problem of implicatures of complex sentences can be formulated as follows: How does the neo–Gricean model have to be modified if '*some*' occurs in the scope of another quantifier, or other logical operator? Two critical examples are shown in (1).

(1) **a)** All of the girls found some of their marbles.

b) Some of the girls found some of their marbles.

In both sentences, '*some*' occurs in upward entailing contexts. The rule that these sentences implicate the negation of all sentences resulting from a replacement of '*some*' by '*all*' predicts that (1a) implicates (2a), and (1b) all three sentences in (2).

(2) **a)** It is not the case that all of the girls found all of their marbles.

b) It is not the case that all of the girls found some of their marbles.

c) It is not the case that some of the girls found all of their marbles.

Here, the whole sentences resulting from replacing '*some*' with '*all*' are negated, therefore these implicatures are called *global* implicatures. There is also the possibility of applying negation locally. This means that the negation of '*all*' is embedded where '*some*' occurs in the sentence. This rule predicts the following additional implicatures:

(3) **a)** All of the girls found some but not all of their marbles.

b) Some of the girls found some but not all of their marbles.

c) Some but not all of the girls found some of their marbles.

[2] See (Levinson 1983: Ch. 3) and (Levinson 2000: Ch. 2) for a summary.

d) Some but not all of the girls found some but not all of their marbles.

Sentence (3a) is the local implicature of (1a), and (3b), (3c), and (3d) are local implicatures of (1b).[3]

There exist a variety of theoretical accounts of implicature in complex sentences which make different predictions. In particular, there has been a controversial debate about locally embedded implicatures (see Sauerland 2010, Geurts and van Tiel 2013 for an overview of the debate). Approaches can be divided into structural accounts that predict local implicatures by integrating them into compositional semantics (Chierchia 2004, Fox 2007, Chierchia et al. 2012), or by generalising the neo–Gricean approach so that expected implicature can be derived as global implicature (Sauerland 2004, Geurts 2010). Other approaches derive them from requirements on discourse relations (Asher 2013), or pragmatically from the interaction between speaker and hearer in game–theoretic and probabilistic models (Franke 2009, Benz 2012, Pavan 2013, Potts et al. 2016).

The approaches also make different predictions about the context dependence and strength of implicatures. For example, Chierchia (2004) assumed that the local implicatures predicted by his theory are default inferences, whereas newer grammatical accounts consider them alternative readings which may or may not be preferred (Chierchia et al. 2012). In such an approach, (1b) is considered ambiguous between its standard semantic meaning, and (3b), (3c), and (3d). In probabilistic accounts, there may be a dominant interpretation, but, in general, all semantically possible interpretations receive some positive probability (Potts et al. 2016). Other approaches predict a unique interpretation, which is, however, in some specified manner dependent on context. In the standard neo–Gricean theory, conversational implicatures are part of communicated meaning (Levinson 1983: Ch. 3, p. 131). This suggests that they are communicated as reliably as semantic meaning. Such a strong claim was, however, until now, not supported by the experimental literature. In the case of un–embedded '*some*', proportions of subjects inferring the implicature can be high but for embedded '*some*' they tend to be rather low. Reported numbers for embedded scalars range from 0% (Geurts and Pouscoulous 2009) to 40% (Chemla 2009).[4] In the same study, Geurts and Pouscoulous report values between 34% and 93% for un–embedded '*some*', depending on the test paradigm. If implicatures are communicated as reliably as literal content, the proportion of subjects inferring implicatures should be close to ceiling. With a few exceptions in the case of un–embedded '*some*', this has generally not been observed. Hence, experimental evidence seems to lend support to grammatical and probabilistic accounts that are consistent with high degrees of uncertainty in utterance interpretation.

In the following experiment, we show that embedded implicatures can be communicated as reliably as literal meaning. As the experimental literature demonstrates, we can only hope

[3] However, (3b)–(3d) are already implied by the global implicatures of (1b) in (2) such that (3a) is the only local implicature that is not implicated globally.

[4] Other studies report values that lie between these extremes, see (Chemla and Spector 2011, Benz and Gotzner 2014, Potts et al. 2016, Franke et al. 2017). Clifton Jr and Dube (2010) used a picture selection task and reported 71% of subjects arriving at local implicature in one of their experiments (Exp. 1, p. 7). However, their study may be affected by typicallity effects as van Tiel (2014) argued.

to show this for certain contexts. So, the question arises, for which contexts can we expect implicature to be inferred reliably?

For Grice (1975), an implicature is an inference towards the speaker's intended meaning. The inference is based on the assumption that the speaker adheres to the conversational maxims, which include the maxim of quantity, and the over–arching *cooperative principle*, which states that the speaker contributes to an '*accepted purpose or direction of the talk exchange*' in which s/he and the hearer are engaged (Grice 1989: p. 26). Grice's maxim of quantity requires the speaker to provide enough information and not more than this. In sum, an implicature must be the speaker's intended meaning providing neither more nor less information than is required by a recognisable purpose of the talk exchange. A requirement that is not explicitly listed by Grice is the competence assumption: it must be shared knowledge between speaker and hearer that the speaker is competent enough to contribute the required information. If Grice's was right about the role of the cooperative principle and speaker's intentions, then a sentence produced non–conversationally should generate no conversational implicature.

Given this background, we may consider the picture verification task by Geurts and Pouscoulous (2009), which yielded particularly low proportions of subjects answering in accordance with embedded implicature. In this experiment, the test sentence was not produced by a recognisable speaker, it is not an utterance, there is no addressee, there is no recognisable *purpose of the talk exchange*, and, hence, there is no intended message that could be sought out behind its literal meaning. The situation is detached from purposeful conversation, and, hence, lacks a central precondition in Grice's theory.[5] To different degrees, all picture verification, graded acceptability and inferencing tasks are affected by this problem.[6]

For this reason, Gotzner and Benz (2018) designed an experimental paradigm which avoided metalinguistic judgments and aimed at implementing Grice's conversational requirements for generating implicature. They used a game–theoretic design in which interpretations are read off from test subjects' choice of action. Grice's purpose or direction of the talk exchange is provided by an explicit decision problem, choosing a set of rewards based on the interpretation of an utterance. In the experimental scenario, each of four girls owns a set of four special edition marbles (extending the scenario by Degen and Goodman 2014). The marbles get lost during play, and in the end they have to find them again. Their mother motivates them by promising rewards which depend on how many of their marbles they find. A girl gets (i) chocolate if she finds all 4 of her marbles, (ii) candy if she finds fewer than 4 of her marbles and (iii) a gummy bear when she finds none of her 4 marbles (as a consolation prize). The task of the participants is to buy sweets for the four girls depending on the statements the mother utters. For example, if the mother says (N-Any) '*None of the girls found any of her marbles*' participants should only buy gummy bears. Participants were asked to give binary responses (yes/no) for each of the three types of sweets: chocolate, candy and gummy bears. Subjects were instructed to

[5] In fairness, it has to be pointed out that Geurts and Pouscoulous (2009) intended to disprove Chierchia's (2004) assumption that embedded implicature are default inferences triggered by the logical form of sentences. Their results may pose problems for this particular semantic theory.

[6] Of the aforementioned studies by (Chemla 2009, Chemla and Spector 2011, Benz and Gotzner 2014, Potts et al. 2016, Franke et al. 2017), that by Chemla (2009) is arguably the least affected. He also reports the highest percentages of pragmatically answering subjects.

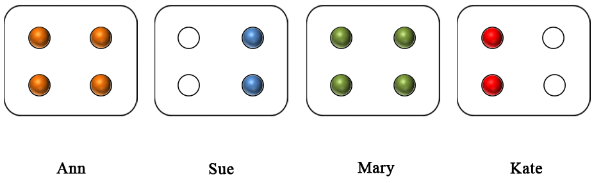

Figure 1: Picture showing boxes of four girls with the marbles they have found.

buy all the sweets that are needed but not more than that. If the mother says '*all found some marbles*', then for subjects drawing the local implicature '*all found some but not all*' the best response is to buy hard candy only. If they only draw the weaker implicature '*not all found all*', then it is better to buy both hard candy and chocolate. If the mother says '*some found some marbles*', then subjects inferring the global implicatures listed in (2) should buy gummy bears and hard candy but no chocolate. Gotzner and Benz (2018) implemented this scenario in an MTurk experiment. Subjects saw sentences produced by the mother and had to decide by ticking off yes/no buttons which of the sweets they had to buy.

The results indicated that subjects draw the strong local implicature (97%) for test sentence '*All of the girls found some of their marbles*', and the strong global implicature (87%) for test sentence '*Some of the girls found some of their marbles*'.[7] Hence, this experiment showed that, in a context that satisfies Grice's conversational requirements, controversially discussed embedded implicatures can be reliably drawn.

One limitation of the study by Gotzner and Benz (2018) is that it only tested the comprehension of certain embedded implicatures in two possible worlds. In the current study, we develop an interactive version of the best response paradigm, which provides both comprehension and production data for a variety of utterance combinations in seven possible worlds. The main research question we address in this collaborative scenario is: To what extent can speakers shorten their description of a state of affairs without jeopardizing communicative success? The shortest descriptions will then reveal all the implicatures that can be communicated reliably in a given communicative context.

3. The interactive best response paradigm: Background

In the following, we describe the background assumptions for our interactive best response paradigm. Let us again consider the marble scenario from Gotzner and Benz (2018). A situation in which two girls found all of their marbles and two found some of them is shown in Figure 1. The mother can describe this situation by saying, for example, '*Ann found all of her marbles,*

[7] There was a surprisingly high percentage of subjects not buying gummy bears for the '*some some*' sentence (24%), indicating that subjects had problems inferring implicature E-N from E-E. The study compared predictions of four theories: a localist (Chierchia 2004), a globalist (Sauerland 2004), and two game–theoretical (Franke 2009, Benz 2012). All theories agreed that subjects should buy hard candy for sentences A-E and E-E, gummy bears for the E-E sentence, but not for the A-E sentence. Hence, only the values for chocolate were critical to the comparison.

Mary found all, Sue found some, and Kate found some.' As it does not matter how the individual girls performed in the marble scenario, only whether there are girls that found none, some, or all of the marbles, the mother could also say (E-A & E-E) '*Some of the girls found all of their marbles, and some found some.*' Intuitively, this should communicate enough information for the addressee to buy the appropriate sweets. However, it is not a literal description of the situation. The second use of '*some*' leaves open whether or not all found all marbles. Hence, the mother could have said more precisely (E-A & E-ENA) '*Some of the girls found all of their marbles, and some found some but not all.*' This is not a literal description either as it leaves open the possibility of some finding nothing. To rule out this possibility, the mother should have said (E-A & E-ENA & N-N) '*Some of the girls found all of their marbles, some found some but not all, and none found none.*' If we start with the full literal description of the scene, then the short description E-N & E-E can be derived by first eliminating the '*not all*' part of '*some but not all*', and then by elimination of '*none found all*', as shown in (4).

	description	
(4)	E-N & E-ENA & N-A	*literal*
	E-N & E-E & N-A	*elimination*: ENA → E
	E-N & E-E	*elimination*: N-A → –

Our hypothesis is that all that can be eliminated by these two rules can be left unsaid without reducing chances of communicative success. If more is left unsaid, i.e. if the utterance is shorter than E-N & E-E in the situation of Figure 1, then communication becomes unreliable. The two rules can then be used to derive the shortest reliable descriptions of each possible world. To do this, we first have to define what the possible worlds and their possible descriptions are. We begin with the latter.

We consider sentences of the form (Q-Q′) '*Q of the girls found Q′ of their marbles,*' where Q and Q′ were one of the quantifiers '*some*' or '*all*'. To describe the situation in Figure 1, the mother may also want to use '*none*' and '*some but not all*'. She may also want to use '*some and possibly all*', and '*any*' in a negative context. To produce literal descriptions of situations it is also sometimes necessary to build conjunctions of Q-Q′ sentences. We use abbreviations for referring to these sentences. If Q and Q′ are the quantifiers '*all*', '*some*', or '*none*', then the following abbreviations are used:

	A-A	all found all	E-A	some found all	N-A	none found all
(5)	A-E	all found some	E-E	some found some	N-E	none found some
	A-N	all found none	E-N	some found none	N-N	none found none

For the more complex construction '*some but not all*' we write ENA. For '*any*' we write '*Any*'.We abbreviate conjunctions by combining sentences with ' & '.

With these sentences, it is possible to distinguish seven possible worlds that are definable by whether or not the sentences E-A, E-ENA, and E-N are made true by them. We use pictograms for referring to these worlds. They are shown and defined in the next table:

(6)

E-N	E-ENA	E-A	world
1	0	0	□
0	1	0	▨
0	0	1	■
1	1	0	□▨
1	0	1	□■
0	1	1	▨■
1	1	1	□▨■

In the marble scenario each situation is represented by one of the possible worlds, for example, Figure 1 represents world ▨■.

Next, we define a literal description of each of the possible worlds by conjoining their defining basic sentences in (6), except for the first three worlds for which universally quantified or negated basic descriptions exist. Then we simplify these descriptions by application of the two elimination rules.

We derive two production strategies, the critical strategy defined by elimination rules and the corresponding literal strategy. They are shown in (7).

(7)

world	critical strategy	literal strategy
□	N-Any	N-Any
▨	A-E	A-ENA
■	A-A	A-A
□▨	E-E & E-N	E-ENA & E-N & N-A
□■	E-A & E-N	E-A & E-N & N-ENA
▨■	E-A & E-E	E-A & E-ENA & N-N
□▨■	E-A & E-E & E-N	E-A & E-ENA & E-N

As we stated before, our assumption is that the application of the two elimination rules will not change communicative success. This means that the critical production strategy has the same degree of communicative success as the literal strategy. Communication is successful if the hearer interprets an utterance as intended by the speaker. *Degree* of communicative success can then be measured by the proportion of utterances that are correctly understood. We further assume that any additional eliminations will lead to utterances that are too short to communicate successfully.

4. Experiments

4.1. Goals and rationale

The goal of the first experiment is to implement an interactive version of the best response paradigm involving a comprehension and a production side. This experiment is set up as a game involving groups of up to 4 participants in the lab. The system always pairs two participants, a speaker and hearer. The speaker is shown a picture and his task is to describe the state of affairs with up to five sentences. Then, this utterance is sent to another participant, the comprehender. The comprehender's task is to choose a set of rewards, reflecting his interpretation of the speaker's utterance. Communication between the two individuals is successful, if the hearer has chosen the appropriate set of rewards for the state of affairs the speaker described. In our analysis, we measure the relative success rate and utterance length of different production strategies based on the comprehension data. In Experiment 1a, we test the critical strategy defined in Section 3 and compare it to a strategy based on literal descriptions. The main research question of the second experiment is whether the critical strategy can be further shortened without jeopardising communicative success. We will first present the methodology of both experiments together and then describe the results.

4.2. Methods

4.2.1. Apparatus

For our experiments, we programmed a system in Python using the GUI toolkit wxPython[8], which allowed us to implement a game with four participants. Participants were seated in a lab with four computers separated by a booth. The computers (DELL Optiplex 3020, 4GB RAM, Windows 8.1 Enterprise) each had an LG monitor with a resolution of 1920 × 1080 and a refresh rate of 64 Hz (15.62 ms). The system controlled stimulus presentations and pairings of participants. The system itself is based on a server-client architecture, where each client corresponds to a participant, while the server connects those clients, sends messages back and forth, pre- and post-processes the data and saves the results.

In general, the system allows to run experiments with either two or four subjects. Furthermore, it is possible to use only one (or three) computers, while the second (the fourth) PC/participant is replaced by the system itself (as was done in Experiment 1b), acting according to a predefined plan to investigate production strategies in a controlled manner.

4.2.2. Experiment 1a

Participants Participants were recruited via a subject pool of the Psychology Department from Humboldt University. In total, 38 German participants (21 female, 17 male, mean age: 29.3) took part in the experiment. Participants took the experiment in groups of varying sizes: there were groups with 4 players, groups with 2 players, and groups with 3 players in addition to the experimenter, who played the critical strategy (see Section 3). 8 participants took part in the version with 4 players (2 groups), 10 participants in the version with two players (5 groups) and 18 participants in the version with 3 players (6 groups). Finally, 2 participants played a version with 1 player in addition to the experimenter (2 groups). These two participants were not included in the analysis reported below.

[8] `https://www.wxpython.org/`

The mother says:	'Each girl found all of her marbles'	
chocolate	⊙ YES	○ NO
candy	○ YES	⊙ NO
gummy bear	○ YES	⊙ NO

Table 1: Example item *each-all* with example response choice, participants were asked to check a radio button for each type of sweets.

Scenario Participants in our experiment were presented with a scenario involving six girls who each own a set of four special edition marbles (extending the basic best response paradigm by Gotzner and Benz 2018).[9] While the girls are playing the marbles get lost and they have to find them again. Participants in our experiment were told that the nursery school teacher of the girls wants to reward them depending on how many marbles the girls find. In particular, participants were presented with the following reward system in the instructions:

A girl gets:

- chocolate if she finds all 4 of her marbles
- candy if she finds fewer than 4 of her marbles
- a gummy bear when she finds none of her 4 marbles (as a consolation prize).

Experimental tasks Participants were randomly assigned to two different roles in the experiment: a speaker or a comprehender. The speaker saw a picture showing the marbles each girl had found, representing all seven possible worlds. The seven worlds we distinguished corresponded to the model presented in Section 3.

The task of the speaker was to describe the picture so that the comprehender can buy the appropriate sweets for the girls. Participants were presented with a sentence frame and they were required to fill in two blanks. They were allowed to type in one of the following words or phrases: *all, some, none, some but not all, some and possibly all* and *any* (in German). Participants were allowed to produce up to five sentences to describe a given picture. Participants' responses were checked for spelling by the system. If they used a word which was not allowed, the corresponding box was highlighted and they had to correct their response.

When the speaker was done describing the picture, the comprehender received his message. The comprehender's task was to select the appropriate kind of sweets for the six girls depending on the message he received. An example trial with the utterance '*Each girl found all of her marbles*' and the appropriate response choice is presented in (1). Participants gave their response by checking one of two radio buttons for each type of sweets.

[9] In this experiment we introduced six girls rather than four in order to avoid referring to a single entity with *some*. Even though the basic semantics of *some* is existential, the quantifier most naturally denotes a set of at least two items (see for example Degen and Tanenhaus 2015 and van Tiel 2014).

In Experiment 1a, we used a confederate, the experimenter, who produced the critical utterances outlined in Section 3.

Procedure At the start of a session, participants were presented with instructions describing the basic setup of the experiment. We told them about the scenario and the different roles they have to take during the experiment. After participants had read the instructions, they performed seven practice trials to learn the reward system used in the comprehender's task. During practice trials, participants saw a picture representing the state of the world and had to chose the appropriate sweets (while during test trials, participants chose the approriate sweets based on an utterance produced by the speaker). The system checked the responses and reported an error if participants chose the wrong sweets.

In the main part of the experiment, participants were assigned to the two different roles in succession. That is, in a given experimental block, a participant either described a picture or interpreted an utterance he received. In these critical trials, no feedback was given by the system so that participants were not biased to pursue a certain interpretation. Each participant took every role 3 times during the course of the experiment. Hence, there were 6 experimental blocks in total. The system always paired two participants for a given world-message pair. For example, the first participant produced a description of the picture and then the second participant received this description and had to chose the reward depending on the statement(s). The pairing of the subjects varied from round to round to make sure each participant plays with every other participant and adopts both roles.

One experimental block consisted of 7 trials representing the different worlds (randomized across the different blocks). The system waited until all participants made their responses and then the next trial was initiated. While the producer typed in a description of the current picture, the comprehender had to wait and vice versa. In the 4 and 3 participant versions, we obtained a total of 82 observations (production/comprehension pairs). In the 2 participant versions, there were 41 observations in total.

4.2.3. Experiment 1b: Shortening strategy

Participants In total, 20 German participants (13 female, 7 male, mean age: 31.0) took part in the second experiment. In Experiment 1b, there were four groups with 3 players and the critical production strategy was fed in by the system. In two sessions 4 participants took part and the production data of these participants were saved and replaced by the computer strategy.

Materials Participants were presented with the same instructions and scenario as in Experiment 1a.

In Experiment 1b, we tested whether the critical strategy can be further shortened and therefore included the following three simple utterances:

1. *Some of the girls found some of their marbles* (E-E) □▨
2. *Some of the girls found all of their marbles* (E-A) □■
3. *Some of the girls found none of their marbles* (E-N) □▨■

In worlds □ N-E, ▨ A-E and ■ A-A we used the same critical utterances as in Experiment 1a. And for world ▨■ we tested the utterance N-N, which is not relevant for the shortening of utterances.

Procedure The procedure was the same as in Experiment 1a except that there were no groups in which the experimenter took part. Instead of using the experimenter as a confederate, the shortening strategy was fed in by the system. That is, if only 3 participants played the game, the critical messages were sent by the computer. In 2 groups, 4 participants came and we saved the production data of the fourth participant and fed in the critical strategy instead. The comprehension data were used from all participants.

4.3. Results

4.3.1. Experiment 1a

We analysed participants' success rate (expected utility) as a function of whether the hearer selected the appropriate sweets depending on the picture the speaker saw. Only if the hearer selected all required sweets correctly was the choice considered a success. Overall, the success rate was quite high (89.7 %), showing that participants understood the task. We, then evaluated how successful different production strategies were, also taking into account utterance length. A t-test showed that the critical strategy was significantly more successful than the average participant strategy (t = -3.85, p-values <.001) and it was also significantly shorter in terms of mean utterance length (t = 6.13, p-values <.001). Table 4.3.1 compares the success rate of the critical and literal strategy in each individual world. Interestingly, when participants produced exact descriptions such as *Each girl found some but not all of her marbles* the communicative success was not better compared to utterances where the short form was used. Hence, for each world the critical strategy was at least as successful as the literal strategy and shorter in terms of utterance length.

4.3.2. Experiment 1b

To show that the critical strategy is the most efficient one, we need to establish that shortening utterances any further lowers communicative success. In Experiment 1b, we replicated the findings concerning the success rate of the utterances also used in Experiment 1a (detailed results are shown in (8) in the Appendix). In the following, we focus on the results of the critical

world	critical strategy	# int.	% success	literal strategy	# int.	% success
□	N-Any	49	100%	N-Any	49	100%
▨	A-E	36	94%	A-ENA	54	93%
■	A-A	114	99%	A-A	114	99%
□▨	E-E & E-N	37	95%	E-ENA & E-N & N-A	12	100%
□■	E-A & E-N	52	96%	E-A & E-N & N-ENA	16	88%
▨■	E-A & E-E	41	98%	E-A & E-ENA & N-N	13	100%
□▨■	E-A & E-E & E-N	48	100%	E-A & E-ENA & E-N	29	97%

Table 2: Results Exp. 1a: Success rate of critical and literal strategy per world (# int: absolut number of items interpreted by subjects).

short utterances	□	▨	■	□▨	□■	▨■	□▨■
E-E	-	32%	-	21%	-	5%	42%
E-A	-	-	11%	-	-	16%	74%
E-N	11%	6%	-	17%	6%	-	61%

Table 3: Results Exp. 1b: Success rate of shortening strategy per world

shortening strategy. Table 4.3.2 details the interpretation data for the critical short utterances.

The success rate of the short utterances was lower than that of the critical strategy. We computed a one sample t-test with the lowest success rate of the critical strategy as expected value (94 %), which found the differences to be significant (t = 6.25, p <.05).

Finally, in Table 4, we present an overview of the average success rate and utterance length of the critical strategy, the literal strategy and participants' average strategy (taking into account the data from both experiments for all seven worlds).

strategy	mean utterance length	%success
average	2.09	89%
critical	1.71	97%
literal	2.5	93%

Table 4: Comparison of mean utterance length and success rate of different production strategies (average of Experiments 1a and 1b)

In sum, these data demonstrate that the critical strategy is maximally efficient in the sense that it is equally successful as the corresponding literal strategy and cannot be shortened without introducing interpretative uncertainty.

5. Discussion

In two experiments we tested our new interactive paradigm. We showed that participants reliably communicate embedded and unembedded implicatures in our interactive setting. This confirms Grice's central requirement for implicature: contextual relevance. Our data confirmed our main hypotheses: The critical strategy is as successful as the corresponding literal strategy, and shortening it further significantly reduces communicative success. The results, thereby, support the hypothesis that the two proposed elimination principles (ENA-Elim and N-X-Elim) characterize what can be left unsaid.

Whereas previous experimental studies focused on the comprehension of a few test sentences in isolation, we have gathered data on a variety of utterance combinations in a precise communicative context. Some previous studies had already indicated that embedded implicatures exist (Chemla 2009, Clifton Jr and Dube 2010, Chemla and Spector 2011, Benz and Gotzner 2014, Potts et al. 2016, Franke et al. 2017, Gotzner and Romoli 2017). However, the experimental paradigms used by these studies have been critized for being unnatural or being prone to typicality effects (see especially Geurts and van Tiel 2013, van Tiel 2014). What is more, our goal was to show that, in a context that makes certain implicatures relevant, they should be reliably communicated, that is as successfuly as corresponding literal descriptions. In our new interactive best response paradigm, we have implemented contextual relevance as an explicit decision problem, chosing a set of rewards. We believe that our action-based task, which distinguishes between relevant readings, is the crucial reason why implicatures are communicated successfully (see Gotzner and Benz 2018). In turn, the meta-linguistic tasks used in previous studies (inferential and truth value judgments) seem to highlight the ambiguity between implicature-based responses and literal interpretations of an utterance.

We now turn to theoretical implications of the current results. The model we based our critical strategy on was developed as a refinement of the game–theoretic model of (Benz 2012), but, for the purposes of this paper, we can keep a relatively theory–neutral position. However, there are two sentences that are partcularly problematic for globalist theories (e.g. Sauerland 2004). They are E-E '*Some of the girls found some of their marbles*', and E-E & E-A '*Some of the girls found some, and some found all*'. Our model predicts that E-E will fail to reliably communicate the state of the world, and that E-E & E-A communicates that the actual world is ▨■. Gricean globalism predicts that E-E implicates that not A-E '*all some*' and not E-A '*some all*', and, hence, that E-E implicates □▨. For E-E & E-A we find the stronger alternative A-E & E-A, hence, Gricean globalism predicts the negation of A-E & E-A, and, therefore, that the speaker meant □■ or □▨■. However, we have seen that it is reliably interpreted as ▨■. We find here a clear conflict between our experimental results and the globalist principle by which sentences implicate the negation of their stronger alternatives. Other theories, in general, do not make predictions that are specific enough to decide whether they are in conflict with our model or not. This does not mean, however, that there are no problems. For example, there is no simple explanation in the standard localist model of (Chierchia et al. 2012) for why E-E & E-A implicates that none found none.

6. Conclusions

Our experiments demonstrated that, in an interactive context involving a speaker and a hearer, embedded implicatures are reliably communicated. We also presented a critical production strategy that was defined by two rules that allow simplifications of literal descriptions. These rules were i) the rule that '*some but not all*' can be simplified to '*some*', and ii) the rule that conjuncts stating that '*none found X*' can be eliminated. In our experiments, the critical strategy was maximally efficient in the sense that it a) communicated the state of the world as reliably as the literal strategy from which it was derived, and b) could not be shortened further without loosing communicative success.

Our new paradigm opens up the possiblity to investigate a variety of sentences of particular theoretical interest in a controlled manner. The advantage is that the sentences are embedded in a natural communicative situation in which subjects are more strongly immersed in the experimental setting. The software that we developed can be used to test speaker-related and other contextual factors, for example by using a confederate. This is done in such a way that subjects do not notice that sentences have not been produced by an actual dialogue partner. On request, we will make the system available to researchers. We hope that our new paradigm will spark further research on implicatures in interactive settings with controlled dialogue.

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.

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.

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.

Degen, J. and M. K. Tanenhaus (2015). Processing scalar implicatures: A constraint-based approach. *Cognitive Science 39*, 667–710.

Fox, D. (2007). Free choice and the theory of scalar implicatures. In U. Sauerland and P. Stateva (Eds.), *Presupposition and Implicature in Compositionsl Semantics*, pp. 71–120. Basingstoke: Palgrave Mcmillan.

Franke, M. (2009). *Signal to Act: Game Theory in Pragmatics*. Ph. D. thesis, Universiteit van

Amsterdam. ILLC Dissertation Series DS-2009-11.

Franke, M., F. Schlotterbeck, and P. Augurzky (2017). Embedded scalars, preferred readings and prosody: An experimental revisit. *Journal of Semantics 34*, 153–199.

Gazdar, G. (1979). *Pragmatics: Implicature, Presupposition, and Logical Form*. New York: Academic Press.

Geurts, B. (2010). *Quantity Implicatures*. Cambridge: Cambridge University Press.

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*.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and Semantics*, Volume 3, pp. 41–58. New York: Academic Press.

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge MA: Harvard University Press.

Horn, L. R. (1972). *On the Semantic Properties of the Logical Operators in English*. Ph. D. thesis, Indiana University.

Horn, L. R. (1989). *A Natural History of Negation*. Chicago: University of Chicago Press.

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

Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicatures*. Cambridge, MA: MIT Press.

Pavan, S. (2013). Quantity implicatures and iterated admissibility. *Linguistics and Philosophy 36*, 261–290.

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.

Sauerland, U. (2010, January). Embedded implicatures and experimental constraints: A reply to Geurts & Pouscoulous and Chemla. *Semantics and Pragmatics 3*(2), 1–13.

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

A. Summary results

Results for critical and literal strategy in Experiment 1b (# int: number of items that had been presented to subjects for interpretation)[10]:

[10] The absolute numbers of literal utterances were lower in Exp. 1b than in Exp. 1a due to the lower number of participants.

(8)

world	critical strategy	# int	% success	literal strategy	# int	% success
□	N-Any	37	100%	N-Any	37	100%
▨	A-E	24	92%	A-ENA	25	92%
■	A-A	60	95%	A-A	60	95%
□▨	E-E & E-N	5	100%	E-ENA & E-N & N-A	3	67%
□■	E-A & E-N	22	86%	E-A & E-N & N-ENA	6	67%
▨■	E-A & E-E	5	100%	E-A & E-ENA & N-N	1	0%
□▨■	E-A & E-E & E-N	8	100%	E-A & E-ENA & E-N	10	90%

Results for critical and literal strategy for the accumulated data of both experiments (# int: number of items that had been presented to subjects for interpretation):

(9)

world	critical strategy	# int	% success	literal strategy	# int	% success
□	N-Any	86	100%	N-Any	86	100%
▨	A-E	60	93%	A-ENA	79	92%
■	A-A	174	98%	A-A	174	98%
□▨	E-E & E-N	42	95%	E-ENA & E-N & N-A	15	93%
□■	E-A & E-N	74	93%	E-A & E-N & N-ENA	22	82%
▨■	E-A & E-E	46	98%	E-A & E-ENA & N-N	14	93%
□▨■	E-A & E-E & E-N	56	100%	E-A & E-ENA & E-N	39	95%

Note that the number of items presented to subjects include those that had been produced by a confederate (experimenter in Exp. 1a, system in Exp. 1b).

Notation Quantifiers within one utterance are separated by ‘-’ and ‘&’ represents conjunction of multiple utterances; A = *all*, E = *some*, N = *none*, ENA = *some but not all*.