Linguistics and Philosophy (2021) 44:1099–1130
https://doi.org/10.1007/s10988-020-09310-x

ORIGINAL RESEARCH

Embedded implicature: what can be left unsaid?

Anton Benz[1] · Nicole Gotzner[1]

Published online: 19 November 2020

Abstract
Previous research on scalar implicature has primarily relied on meta-linguistic 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 side and a precise conversational goal. The main research question is what is reliably communicated by *some* in this communicative setting, both when the quantifier occurs in unembedded and 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, thereby 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 that describe what can be left unsaid. In our experimental context, a production strategy based on these principles is more efficient (with equal communicative success but shorter utterances) than a strategy based on literal descriptions.

Keywords Scalar implicature · Embedded implicature · Experimental pragmatics · Game theoretic pragmatics

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 and 4-2, as part of the priority program *New Pragmatic Theories based on Experimental Evidence* (SPP 1727). We are grateful to Lisa Raithel for programming the system for the experiments.

✉ Anton Benz
benz@leibniz-zas.de

Nicole Gotzner
gotzner@leibniz-zas.de

[1] Leibniz-Centre General Linguistics (ZAS), Schützenstrasse 18, 10117 Berlin, Germany

1 Introduction

Grice (1975) distinguished between two components of communicated meaning: *what is said* and what is *implicated*. The first component is the content that is semantically expressed by an utterance, and the second component is the content that is pragmatically inferred from the assumption that the speaker is cooperative and follows certain maxims. For example, if Kate is a girl who has to clean up her room and find all of her marbles with which she had played before, then, if one parent is asked '*How many of her marbles did Kate find?*', the answer '*Kate found some of her marbles*' would clearly communicate that she did not find all of them. Literally, the answer only states that Kate found some and possibly all of them. The proposition that Kate did not find all has to be inferred from shared contextual and pragmatic knowledge. For Grice, a necessary condition for this proposition to be a *conversational implicature* is that '*the speaker thinks (and would expect the hearer to think that the speaker thinks) that it is within the competence of the hearer to work out, or grasp intuitively,*' that it is implied by context and pragmatic principles of conversation (1975, p. 50). If the speaker can rely on the hearer's ability to infer the implicature, and if this ability is shared knowledge between speaker and hearer, then this allows the speaker to simplify utterances and to encode only part of the intended message in semantics and leave the rest to pragmatics. Levinson (2000, Ch. 1) argued that it is a key function of implicature to improve the efficiency of information exchange by allowing the speaker to produce less linguistic material without increasing the risk of miscommunication. For (1), this means that the implicature from '*some*' to '*some but not all*' allows the speaker to simplify (1a) to (1b) by eliminating the '*but not all*' part while communicating the same meaning.

(1) a. Kate found some *but not all* of her marbles.
b. Kate found some of her marbles.

Guided by this communicative function of implicature, we investigate the following research question: How much can a speaker simplify (complex) sentences containing '*some*' and replace the eliminated semantic meaning by implicated meaning? An answer to this question will tell us which implicatures are reliably understood by the hearer, and, hence, can be left unsaid by the speaker.

Our main interest lies in the interpretation of complex sentences in which '*some*' is embedded under a quantifier, for example, the sentences (2a) and (2c), which may implicate (2b) and (2d), respectively. Sentences of this form have been of particular theoretical interest since competing accounts make different predictions about the implicatures arising from them (for an overview, see e.g. Geurts 2009; Sauerland 2012; Geurts and van Tiel 2013; Gotzner and Benz 2018). To understand which implicatures are reliably communicated and can thus be left unsaid by the speaker, we consider all state of affairs distinguished by the double-quantified sentences of the form (2), as well as complex sentences conjoining several sub-sentences, as in (3). We investigate the interpretation of such sentences in a new interactive experimental paradigm involving a speaker and hearer task.

(2) a. All of the girls found some of their marbles.
b. ⇝ None of the girls found all of their marbles.
c. Some of the girls found some of their marbles.
d. ⇝ Some of the girls found some but not all of their marbles.

(3) a. Some of the girls found some of their marbles, and some found all.
b. Some of the girls found some of their marbles, and none found all.

To preview the answer to our research question: The speaker can simplify complex sentences containing '*some*' up to the limit marked by a pre-defined critical speaker strategy. For each state of affairs, the critical speaker strategy chooses a maximally efficient utterance in the sense that a) this utterance communicates the state of affairs with the same rate of success as a literal description, and that b) there are no shorter utterances that are equally successful. The strategy is defined by shortening literal descriptions according to two *elimination rules*: One rule that allows for the simplification of *some but not all* to *some*, and another rule that allows for the elimination of sub-sentences that have '*none*' as an outer quantifier, for example, the elimination of '*none found all*'.

This paper is organized as follows. In Sect. 2, we motivate our research questions and the design of our new experimental paradigm, which is an interactive version of the best response paradigm (Gotzner and Benz 2018). Then, we introduce elimination rules defining our critical strategy and we formulate our hypotheses in Sect. 3. Section 4.3 presents two experiments that test our hypotheses. We will discuss the relevance of our experimental results to theories of embedded implicatures in Sect. 5. In general, the results indicate that basic generalized neo-Gricean accounts (e.g. Sauerland 2004) lead to over- and undergeneration problems.

2 Motivation and research questions

The theoretical controversy about embedded implicatures motivated several experimental studies on implicatures of complex sentences, most of them focusing on sentences like (2a) (Geurts and Pouscoulous 2009; Clifton and Dube 2010; Chemla and Spector 2011; Geurts and van Tiel 2013; Potts et al. 2016; Franke et al. 2017; Gotzner and Romoli 2017). These studies have, in part, provided evidence that sentences like (2a) give rise to embedded implicatures (e.g. Chemla and Spector 2011). However, reported rates of embedded implicature in previous studies were generally low (between 0% and 40%; see van Tiel et al. 2018 for a comparison of different tasks), hence it is an open issue whether sentences of type (2a) can be reliably understood as (2b), and how sentences of greater complexity are interpreted. We argue that previous research leaves an important conversational task of implicature unexplored: making conversation more efficient. Moreover, many experimental paradigms used in previous research face the methodological issue that they did not implement an actual speaker or a precise conversational goal.

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 overarching *cooperative principle*, which states that the speaker contributes to an '*accepted purpose or direction of the talk exchange*' in which they and the hearer are engaged (Grice 1989, p. 26) 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.

Gotzner and Benz (2018) designed a scenario that aimed at implementing Grice's conversational requirements, the so-called *best response paradigm*. In particular, Grice's purpose or direction of the talk exchange was provided by an explicit decision problem. In the experimental scenario, each of four girls owns a set of four special edition marbles (an extension of the marbles scenario by Degen and Goodman 2014). The marbles get lost during play, and in the end, the girls have to find them again. Their mother motivates them by promising rewards which depend on how many of their marbles they find. The task of the participants is to buy sweets for the four girls depending on the statements the mother utters. The rewards distinguish all possible readings of critical sentences, hence the participant's interpretation can be read-off from his or her choice of rewards.

The results of the Gotzner and Benz study indicated that participants draw an embedded implicature (all found some but not all, 97%) for test sentence (2a) (henceforth A-E for *all-some*), and the global implicature (none found all, 87%) for test sentence (2c) (henceforth E-E for *some-some*). This experiment showed that, in a context that satisfies Grice's conversational requirements, controversial embedded implicatures are reliably drawn. However, the data revealed an unexpected problem with the potential implicature from '*Some found some*' (E-E) to '*Not all found some*', i.e. '*Some found none*' (E-N). This implicature is predicted by all theories. Yet a substantial percentage of participants did not derive it (24%). Even if we interpreted this finding as a sign of uncertainty, it differs remarkably from the results for other implicatures. We conjecture that the uncontroversial implicature from E-E to E-N is not an implicature that can be reliably communicated, even if Grice's conditions are satisfied. This assumption will be investigated further in the current experiment. If our conjecture is correct, then the dividing line between communicated implicature and non-communicated ones must run differently from what previous theories expected. To test our conjecture, we address the following research question: '*What are the shortest sentences that can still reliably communicate the state of the world?*' A collection of such sentences will define what we call *the corner of efficiency*, outlined in Sect. 3.

The idea underlying our research question is based on the following assumptions: A speaker who wants to communicate a certain proposition can express all they want to express literally, or they may take advantage of implicature, and leave certain aspects unsaid. This will lead to a shortening of utterances. There will be a limit to the extent that an utterance can be shortened. When the speaker has taken advantage of all the implicatures that can be communicated reliably and starts shortening them further, then miscommunication will set in. This allows us to transform the research question '*Which implicatures can be communicated reliably?*' into the question '*How much can a speaker's description be shortened without jeopardizing communicative success?*'. An answer to the latter question will imply an answer to the former. To answer our research

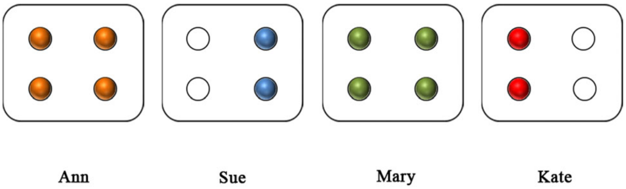


Fig. 1 Picture showing boxes of four girls with the marbles they have found

questions, we develop an interactive version of the best response paradigm, that will allow us to gather interpretation and production data on sentences with embedded and unembedded *some*.

Our experiments will show that subjects can understand implicatures of complex sentences as reliably as literal content. Earlier experiments failed to show this. Our experimental set up is motivated by the desire to create a natural communicative situation that satisfies strong Gricean conditions for drawing an implicature: the speaker has to be cooperative, competent, and the implicated meaning must be contextually relevant. We assume that creating such a context is the main reason why subjects drew implicatures in our experiments much more reliably than in previous experiments.

3 Elimination rules and the corner of efficiency

In this section, we introduce two rules that tell us how exhaustive literal descriptions can be simplified without losing communicative success. We call them *elimination rules*. They will define the critical speaker strategy, which is predicted to be the shortest production strategy that reliably communicates a state of affairs. To understand the rationale behind this, let us consider a concrete example from the marble scenario in the best response paradigm. A situation in which two girls found all of their marbles and two found some of them is shown in Fig. 1.

A speaker who has to describe this situation could say that '*Ann found all of her marbles, 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 speaker 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. However, it is not a literal description of the situation. The embedded '*some*' leaves open whether or not all found all. Hence, the speaker 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 speaker 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-A & E-E can be derived by first eliminating the '*not all*' part of '*some but not all*', and then by elimination of '*none found none*', as shown in (4).

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

Our hypothesis is that all that can be eliminated by these two rules can be left unsaid without reducing the chance of communicative success. If more is left unsaid, i.e. if the utterance is shorter than E-A & 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 state of affairs.

Elimination rules are speaker oriented, telling him/her what can be left unsaid. In contrast to that, other theories approach pragmatics from the interpretation perspective. In such theories, the effect of both elimination rules would be modelled by some sort of *exhaustification* and *only*-operators, i.e. by the assumption that everything that was not mentioned but would have been relevant is not the case (e.g. van Rooij and Schulz 2004; Fox 2007). Grammatical accounts, for example, explain the effects of ENA-elimination (ENA-Elim rule) by the presence of optional, hidden *only*-operators in the logical form of sentences (Chierchia et al. 2012). For N-N-elimination (N-X-Elim rule, X some quantifier phrase) the connection to exhaustification can be seen if the speaker's utterance is considered an answer to the implicit question (under discussion) '*How many of the girls found all, some, or none of their marbles?*' Her answer implies that there is a group of girls that found some of their marbles, and that there is a group of girls that found all of them. She did not mention that there are girls that found none. Hence, exhaustification could lead to the conclusion that there are *only* girls that found all or *only* some of their marbles. Hence, our speaker centred approach partly draws on the same motivation that underlies current semantic approaches. However, this does not mean that both approaches arrive at the same conclusions. We will discuss some critical examples in Sect. 5.

The idea that implicature can be explained by shared rules that allow for systematic syntactic simplifications of utterances goes back to Benz (2009, 2012). The elimination rules (ENA-Elim) and (N-X-Elim) in the present form were introduced in Gotzner and Benz (2018) together with a third rule that allowed the elimination of '*some none*' from utterances (E-N-Elim). The latter rule predicted that E-E (*some found some*) and E-A (*some found all*) implicate that *some found none*. As explained in the previous section, the experimental results reported by Gotzner and Benz (2018) are in conflict with the (E-N-Elim) rule. This led to the hypothesis that only (ENA-Elim) and (N-X-Elim) are rules that allow for simplifications of utterances from which addressees can still reliably reconstruct the originally intended full sentences.

Next, we show how to derive short utterances for each state of affairs with the help of elimination rules. To do this, we first have to define what the states of affairs are, and what the sentences are that might be used for describing them. We begin with the latter issue.

We consider sentences of the form (Q an Q') '*Q of the girls found Q*' *of their marbles*', where Q and Q' are one of the quantifiers '*some*' or '*all*'. To describe the situation in Fig. 1, the mother may also want to use '*none*', '*some and possibly all*' and ‘*some but not all*’. In a negative context, she may want to use ‘*any*’. 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:

(5)

A-A	all found all	E-A	some found all	N-A	none found all
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*’.[1] Hence, in addition to the abbreviations in (5), we consider N-any (*none found any*), and the sentences in (5) with E replaced by ENA, e.g., A-ENA (*all found some but not all*), E-ENA (*some found some but not all*), ENA-ENA (*some but not all found some but not all*), etc. We abbreviate conjunctions by combining sentences with & .

With these sentences, it is possible to distinguish seven state of affairs, or *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 to refer to these worlds. They are shown and defined in the following 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	□▨■

Hence, □ is the world in which none of the girls found any of her marbles, ▨ the one in which all found some but not all, and ■ the one in which all found all. The other worlds have mixed proportions of girls finding none, some, or all of their marbles. For example, Fig. 1 shows a situation that represents world ▨■.

Next, we produce a literal description of each of the 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 applying the two elimination rules, see Table 1.

From Table 1, we can derive two production strategies: the critical strategy defined by elimination rules and the corresponding literal strategy, as shown in (7).

[1] The experiments were conducted in German. N-any, therefore, stands for ‘*Keines-irgendeine*’, and N-E for ‘*Keines-einige*’.

Table 1 All state of affairs with literal descriptions and their simplified descriptions. First, if possible, elimination rule 'ENA → E', simplifying '*some but not all*' to '*some*', is applied, then rule 'N-X → -', which removes all conjuncts starting with '*none of the girls*'. . . .

□	▨	■
N-any	A-ENA A-E	A-A
□▨	□■	▨■
E-N & E-ENA & N-A E-N & E-E & N-A E-N & E-E	E-A & E-N & N-ENA E-A & E-N & N-E E-A & E-N	E-A & E-ENA & N-N E-A & E-E & N-N E-A & E-E
□▨■		
E-A & E-ENA & E-N E-A & E-E & E-N		

(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, we assume that applications of the two elimination rules will not change communicative success. Hence, our main hypotheses are:

(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 lengths makes the resulting strategy communicatively less successful than the literal strategy.

Before we introduce our experiments in the next section, we consider an expected consequence of (I) and (II) that further clarifies the relation between critical strategy and the strategies of individual speakers. Let us assume that we collect production data of several individuals, and we also know how sentences are interpreted on average. The behaviour of each individual can be represented by a probabilistic production strategy, i.e. by the probabilities with which they chose utterances for different worlds. For each production strategy we can calculate the average length of utterances,[2] and the average error rates (based on the interpretation data). Hence, for each individual strategy our data provides a pair (l, e) of average utterance length l and average error rate e. If we now draw a diagram that shows the average length of strategies on the x-axis, and the average error rates on the y-axis, then a pattern as that shown in Fig. 2 should emerge.

[2] A simple measure that counts the conjuncts and adds a small amount for occurrences of '*some but not all*' and '*some, and possibly all*' will suffice here.

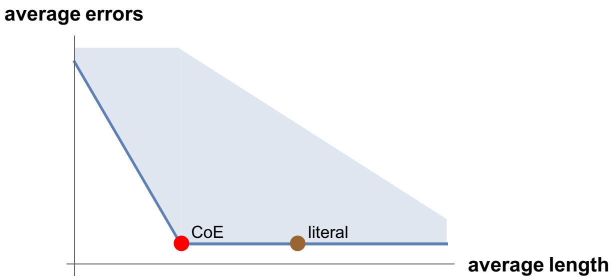


Fig. 2 The corner of efficiency (red dot). Expected pattern for diagramme with average error rate of production strategies over average length of utterances

Human strategies define points in the shaded area. The most successful ones should be close to the x-axis. They will not be exactly on the x-axis due to natural random errors in the interpretation data. Literal descriptions are longer. As more and more information is eliminated from them, they become shorter. So our assumptions predict that we will see a line closely parallel to the x-axis which stretches from the right with strategies that produce utterances that are very long on average to the left until it reaches some point where utterances will become too short to be successful. From this point onward to the left the average error rates will increase. This is the corner where the most efficient production strategies will be located, i.e. the production strategies with the shortest average length of utterances and still full communicative success. We call this the *corner of efficiency*. It is a logical consequence of hypotheses (I) and (II) that the corner of efficiency exists and that the critical production strategy is located there.[3]

If we say that the critical strategy produces the shortest utterances on average with full communicative success, then this needs further qualification. It may well be that interlocutors converge on more efficient strategies, for example, after repeated interaction.[4] If one interlocutor learns that the other one always means □▨ when they say E-E, then this would allow them to use a strategy that is more efficient than the critical one. In the end, a learning process may lead to a strategy that produces only short utterances. A hypothetical candidate is shown in (8).

(8)

world	short production strategy
□	N-any
▨	A-E
■	A-A
□▨	E-E
□■	E-A
▨■	N-N
□▨■	E-N

[3] It is, however, not the only possible strategy located there.

[4] For experiments investigating the emergence of communication strategies in multi-player experiments see for example Fay et al. (2010) and Yoon and Brown-Schmidt (2017).

Our claim is not that interlocutors can never learn to use such a short strategy, only that randomly drawn hearers will interpret this strategy less successfully on average than the literal strategy or the critical strategy.

4 Experiments

4.1 Goals and rationale

As stated before, our main hypotheses are: (I) The critical strategy is as successful at communicating the state of the world as the corresponding literal strategy; (II) a short speaker strategy is communicatively less successful than the literal strategy. To test these hypotheses, we implemented an interactive version of the best response paradigm (Gotzner and Benz 2018) involving a comprehension and a production side. The experiment divided into two parts, Experiment 1a and 1b. In the first part, we gathered interpretation data about the critical strategy, and in the second part interpretation data about a short strategy. To gather these data, we used a design with a confederate to produce the targeted sentences.

Our experiments were set up as a game involving groups of up to 4 participants in the lab. Participants in the experiment take turns in two roles, the speaker and the comprehender. 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. Her task is to choose a set of rewards, reflecting her interpretation of the speaker's utterance. Communication between the two individuals is successful, if the comprehender 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 each experimental trial, 2 participants were paired for a given production and comprehension trial. The pairings were controlled by the computer system and they changed in each experimental block consisting of seven trials. In Experiment 1a, a confederate took part in the experiment if a group of participants only consisted of three people. The confederate always played the critical strategy defined in (7). In Experiment 1b, the role of the confederate was taken over by the system itself. That is, in experimental groups consisting of three participants the pre-defined strategy was played by the system, so that the critical utterances were transmitted to the participant when one role was vacant. If a group consisted of four participants, the utterances produced by one participant were not sent to the comprehender but instead the system sent the pre-defined critical strategy. Since pairings changed in each experimental block, participants had no way of finding out who had produced the utterances. For this reason, we analyse comprehension data from utterances produced by the confederate/system or participants together. We will first present the methodology of both experiments together and then describe their results together. Finally, we present an overall evaluation of the critical strategy for both experiments and compare it to other simpler strategies.

4.2 Methods

4.2.1 Apparatus

For our experiments, we programmed a system in Python using the GUI toolkit wxPython[5], 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 postprocesses the data and saves the results.

In general, the system allows to run experiments with either two or four players since in each round, two players are paired for a production-interpretation trial. In Experiment 1a, if only one or three participants showed up, the vacant role was filled by a confederate who played the pre-defined critical strategy. In Experiment 1b, the role of a confederate was taken over by the system itself. It was possible to fill a vacant role by the system, or to replace a participant's utterances by the system if no role was vacant. The system always produced sentences according to a predefined plan, that is the critical strategy.

4.2.2 Experiment 1a: critical versus literal strategy

Participants Participants were recruited via a subject pool of the Psychology Department from Humboldt University. In total, 38 native German participants (21 female, 17 male, mean age: 29.3) took part in the experiment. They 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 as defined in (7). 8 participants took part in the version with 4 players (2 groups), 10 participants in the version with 2 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.

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) which was based on the marbles scenario by Degen and Goodman (2014)).[6] 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

[5] https://www.wxpython.org/.

[6] 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 2 items (see for example Degen and Tanenhaus 2015 and van Tiel 2014).

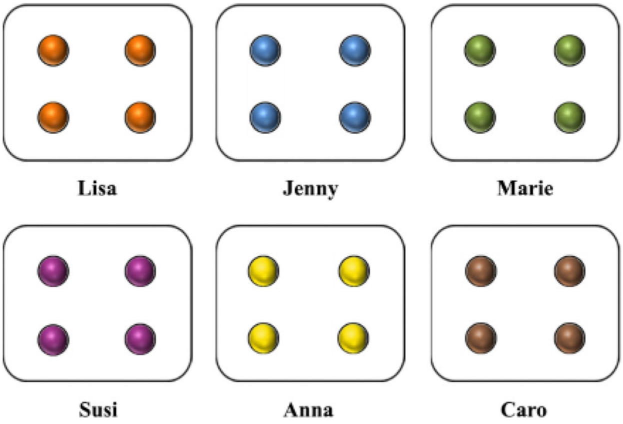


Fig. 3 Example picture

marbles the girls find. In particular, participants were presented with the following reward system in the instructions (all instructions are found in Appendix 1):

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).

Participants' task Participants were randomly assigned to the two different roles in the experiment: a speaker and a comprehender. The speaker saw a picture representing one of seven different states of the world. For example, the speaker saw a picture of the girls' marble boxes in which each girl found all 4 of her marbles (see Fig. 3). The seven worlds corresponded to the seven worlds presented in (6) in Sect. 3. Each world was instantiated by six items in total.

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. Figure 5, Appendix 1, shows an example screenshot. Participants' responses were checked for spelling and appropriateness of words by the system. If participants 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 describing the state of the world. His task was to select the appropriate 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 (9). A screenshot is provided in Fig. 6, Appendix 1. Participants gave their response by checking one of two radio buttons for each type of sweets. In Experiment 1a, the critical utterances in (7) were given by a confederate. The confederate participated in the 6 groups with 3 participants. In each group, the confederate produced each sentence of the critical strategy 3 times such that he contributed $6 * 3 = 18$ tokens of them in total.

(9)

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

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 appropriate sweets based on an utterance produced by the speaker). The system checked the responses and reported an error if participants chose the wrong sweets. The participants then had to try again until they had selected the correct combination of sweets. In a second practice round, participants were trained to use the interface. They saw a picture with three girls who each had four marbles of the same color with colors different for different girls. The participants then had to produce a sentence of the form '*Lisa found the orange marbles, Jenny found the blue marbles, and Marie found the green marbles.*' In this way they practiced how to produce descriptions with several conjuncts with the user interface. Participants were informed that they can produce at most 5 conjuncts.

In the main part of the experiment, participants changed between the roles of producer and comprehender. That is, a participant either described a picture or interpreted an utterance. No feedback was given by the system to avoid any biases about interpretation. The main part started with a trial in which first a picture was shown to each producer. The producers described the picture using the interface and then pressed a send button. The comprehenders waited and did nothing during this time. When the send button was pressed, each description was sent to a comprehender who then had to decide which sweets to buy. During this time, the producers waited. When the comprehenders had made their choice and pressed a send button, the results were stored and the next trial started.

The experiments were either run with 2-producer-comprehender pairs (4 participants or 3 participants and a confederate) or 1-pair (2 participants or 1 participant and a confederate). In the 2-pairs version, one trial consisted of two participants that produced and two participants that interpreted. In the 1-pair version, one participant produced and one interpreted. It was not known to the participants which roles the other participants were assigned to, in particular, they did not know with whom they were paired as producer or comprehender. This means, they did not know who will interpret the sentences they produced, if they are in the producer role, nor who produced the sentences they have to interpret, if they are in the comprehender role. Each participant took each role for each of the 7 worlds 3 times during the course of the experiment. The pairing of subjects into producers and comprehenders varied in such a way that for each world every participant played with every other participant in both roles. In the 2-pairs version, every participant was paired with every other participant in both producer-comprehender configurations only once for each world. This means, that the experiments consisted of $3 * 2 * 7 = 42$ trials. In the 1-pair version, they were paired thrice, which means that the experiment again consisted of $3 * 2 * 7 = 42$ trials. In the 2-pairs version, we obtained a total of 84 observations (2 per trial), and in the 1-pair version 42 (1 per trial).

Trials were organized into 6 experimental blocks. One experimental block consisted of 7 trials representing the different worlds (randomized across the different blocks). The producer and comprehender role was fixed for each block. This means, if a participant started in the first trial of a block as a producer, then he remained producer in all remaining trials of the block. The roles changed when the next block started.

4.3 Experiment 1b: short strategy

Participants In total, 20 German participants (13 female, 7 male, mean age: 31.0) took part in the experiment. In Experiment 1b, there were 4 groups with 3 players and 2 groups with 4 players. The critical production strategy was fed in by the computer. In the sessions in which 4 participants took part, 1 participant was replaced by the system in the speaker role. The production data of the 2 replaced participants were saved. Due to a technical problem, the data of one of them were lost.

4.3.1 Materials

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

In contrast to Experiment 1a, a confederate played by the system itself was administered in all experiments. In 2 sessions with 4 participants, the production data of one human participant were saved in a separate file and replaced by the confederate's strategy. From the participants perspectives, experiments with 3 or 4 participants were indistinguishable. The confederate followed the short production strategy shown in (10). In each session, each sentence was produced 3 times by the confederate, hence, for each world the system contributed $6 * 3 = 18$ utterance tokens of each sentence.

(10)

world	short production strategy
□	N-any
▩	A-E
■	A-A
□▩	E-E

world	short production strategy
□■	E-A
▩■	N-N
□▩■	E-N

The goal of Experiment 1b was to show that a short strategy leads to a significant reduction of communicative success. If the length of critical utterances is further reduced, three utterances can result: E-E (*some found some*), E-A (*some found all*), and E-N (*some found none*). We, therefore, defined a strategy that included these sentences. Of the other utterances shown in (10), three are repeated from the critical strategy of Experiment 1a (N-any, A-E, A-A). In addition, we tested the utterance N-N for exploratory purposes.

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

4.4 Descriptive results

Overall, our two experiments yielded interpretation data for 86 different utterance combinations (86 in Experiment 1a and 62 in Experiment 1b). In this section, we provide descriptive statistics on the data of Experiments 1a and 1b. For hypothesis testing and further analysis, we combine them into one data set. This will be done in Sect. 4.5. To justify this move, we also show that the core statistical measures for the two data sets are not significantly different.

The interpretation data of both experiments are hosted on the server of the Open Science Framework (OSF). The results of Experiment 1 are available at https://osf.io/675pr/?view_only=e5a502aaba82416e8c3eb2d3eb1719a2 and those of Experiment 2 are found at https://osf.io/gdy5c/?view_only=e5a502aaba82416e8c3eb2d3eb1719a2. The tables do not include results on sentences containing ‘*some and possibly all*’ and ‘*any*’. The latter results have been addressed in Benz and Gotzner (2019).

4.4.1 Experiment 1a

We measure the speaker’s success rate (expected utility) by the probability with which a comprehender selects the appropriate sweets for the picture that the speaker sees.[7] Only

[7] For each interaction we recorded the world w, the utterance u that the producer chose, and the choice of sweets by the comprehender. Each choice of sweets could be identified with the unique world v in which the choice was appropriate. Hence, every single interaction provided a triple (w, u, v) of world, utterance, and interpretation. For each utterance u the average probability $A(v|u)$ with which an occurrence of u is interpreted as v can be calculated as follows:

$$A(v|u) = \frac{N(u, v)}{\sum_{v'} N(u, v')}, \tag{11}$$

where $N(u, v)$ is the number of utterance-interpretation pairs (u, v) occurring in interactions (w, u, v). $A(\,.\,|u)$ specifies the probability with which human subjects interpreted utterance u on the average. It can, therefore, be identified with the *average human interpretation* strategy. A *production strategy* S specifies for each world w and utterance u a probability $S(u|w)$ with which utterance u is produced in world w when following S. The *expected utility* of strategy S against the average interpretation strategy A is then defined as follows (note that there are 7 state of affairs):

$$EU(S) = \sum_{w} \frac{1}{7} \sum_{u} S(u|w)\, A(w|u). \tag{12}$$

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

world	critical strategy	N	success	literal strategy	N	success
□	N-any	43	100%	N-any	43	100%
▒	A-E	30	93%	A-ENA	54	93%
■	A-A	108	99%	A-A	108	99%
□▒	E-E & E-N	31	94%	E-ENA & E-N & N-A	12	100%
□■	E-A & E-N	46	96%	E-A & E-N & N-ENA	16	88%
▒■	E-A & E-E	35	97%	E-A & E-ENA & N-N	13	100%
□▒■	E-A & E-E & E-N	42	100%	E-A & E-ENA & E-N	29	97%

if the comprehender selected all required sweets correctly was the choice considered a case of successful communication. Overall, the average human success rate was quite high (88 %), showing that there were few cases of miscommunication overall.

In Experiment 1a, we wanted to gather enough data on the critical strategy defined in (7) to test whether it is as successful as the corresponding literal strategy, also defined in (7). The utterances A-A in world ■, and N-any (German: *keines irgendeine*) in world □ are shared by both strategies. Table 2 summarizes the data for world-utterance pairs defined by the two strategies. The descriptive statistics indicate that the success rates of critical and literal utterances were comparable. The average success rates of both strategies turns out to be equal (97%). Among individual utterances, there were two utterances in the critical strategy (E-E & E-N and E-A & E-E) with a slightly lower success rate than the corresponding literal utterances. We analyse these sentences individually in Sect. 4.5.3.

4.4.2 Experiment 1b

To show that the critical strategy is maximally efficient, we need to establish that shortening utterances any further lowers communicative success. In Experiment 1b, we wanted to collect more data on sentences that can result from shortening the critical strategy. The average human success rate was again quite high (83%). The mean success rate of the critical strategy was almost identical to that of Exp. 1a (1b: 96%, 1a: 97%). However, there were differences for the success rate of the literal strategy that can be explained by the low number of observed occurrences. A table with detailed results on the critical and literal strategy can be found in the Appendix in Table 7. In the following, we focus on the results for utterances that result from shortening critical utterances. Shortening critical utterances can result in three sentences: E-E (*some found some*), E-A (*some found all*), and E-N (*some found none*). These are the critical short utterances that are relevant to us. Table 3 details their interpretation data.

The success rates of critical short utterances are lower than those of the full critical utterances. All short utterances were most frequently interpreted as referring to □■.

Footnote 7 continued
The expected utility $EU(S)$ is equal to the probability with which strategy S is successful against the interpretation strategy A. It, therefore, provides the average success rate of S given A. The definition implies that only *exact* matches of (speaker) intended world and interpretation count as successful.

Table 3 Results Exp. 1b: Success rates of critical short utterances per world (N: absolute number of items interpreted by subjects)

Short utterances	N	□	▨	■	□▨	□■	▨■	□▨■
E-A	19	-	-	11%	-	-	16%	74%
E-E	19	-	32%	-	21%	-	5%	42%
E-N	18	11%	6%	-	17%	6%	-	61%

4.5 Statistical evaluation

We first show that, for core measures, there are no statistical differences between Experiments 1a and 1b. Next, we compare the critical, literal, and human average strategies for the combined data set. We then turn to individual sentences and show that the critical utterances can not be shortened without losing communicative success. In the final explorative section, we explore the success of other shorter strategies that are not derived by shortening the critical strategy.

4.5.1 Testing for differences between Experiments 1a and 1b

First, we verified that Experiments 1a and 1b did not differ significantly with respect to two core measures: the means of critical success rates and the success rates of the average production strategy of participants. For the critical strategy, the observed mean in Exp. 1a was 0.97, and in Exp. 1b 0.96. For the average human production strategy the observed means were 0.88 and 0.83 in 1a and 1b, respectively. We calculated Pearson's correlation coefficient for the two vectors defined by the success rates of the critical strategy in 1a and 1b as $\rho = 0.97$. We also calculated Pearson's correlation coefficient for the success rates of the shared utterances of the average human production strategies of 1a and 1b.[8] The correlation turned out to be very high $\rho = 0.96$.

For testing the significance of the absolute differences between the success rates of the average human production strategies, we used bootstrap methods and resampled production and interpretation data 10000 times, respectively. We paired them and calculated for each paired set success rates of the critical and the average human production strategy (the latter calculated from resampled production data). This was done for both the data of Experiment 1a and 1b, and for the combined data, so that we had three sets of resampled data with 10,000 pairs of production and interpretation data each. We then compared the respective means of the resampled experiments. For the critical strategy, the mean difference between success rates in 1a and 1b was 0.008 (two-sided p= 0.69, $[-0.03, 0.04]$).[9] For the average human production strategy, the mean differences between success rates of Exp. 1a and Exp. 1b was 0.04 (two-sided

[8] As there are more utterances in Exp. 1a than in 1b, we had to restrict the comparison to the shared utterances; i.e. for $USet = \{u \mid \exists w, v : (w, u) \in Exp_{1a} \wedge (v, u) \in Exp_{1b}\}$ we compared $\{(w, u, succ_{1a}(u|w)) \mid u \in USet \wedge (w, u) \in Exp_{1a} \cup Exp_{1b}\}$ and $\{(w, u, succ_{1b}(u|w)) \mid u \in USet \wedge (w, u) \in Exp_{1a} \cup Exp_{1b}\}$. This way, the comparison is based on 61 sentences and 86 world-utterance pairs.

[9] Values > 0 indicate that Exp. 1a had higher success rates, while values < 0 indicate that Exp. 1b had higher success rates. The corresponding differences for the joint data and 1a and 1b are, of course, even smaller: Exp. 1a−joint = 0.005 and Exp. 1b−joint = -0.002.

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

Strategy	Mean utterance length	%success
literal	2.5	93%
average	2.09	87%
critical	1.71	97%
short	1	66%

p= 0.08, [−0.004, 0.08]). The corresponding differences for the joint data and Exp. 1a was −0.01, and that for the difference between joint data and Exp. 1b was 0.02.

In sum, we found no significant differences of success rates for the critical strategy and the average human production strategy. This justifies the conclusion that the methodological differences between Experiments 1a and 1b did not affect participant's behaviour, so that both data sets can be joined for the statistical analysis of production strategies and individual utterances.

4.5.2 Evaluation of strategies

In this section we test one of our core hypotheses: that the critical strategy as a whole has a success rate that is not lower than that of the literal strategy. We also show that it is significantly more successful than the average human production strategy. 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 for the combined data of both experiments for all seven worlds.[10] We also added the average success rate and utterance length for the full short strategy (8) for orientation (without theoretical significance to our overall argument).

For testing the significance of the differences between success rates, we used bootstrap methods, resampling interpretation data 10000 times. In parallel, we also resampled the raw production data that define the average human strategy 10000 times. We paired them and calculated for each paired set success rates of the average human production strategy. We calculated the expected success rate of critical and literal strategy for all 10000 sets of resampled interpretation data. As before, we calculated the differences for these values for the resampled means and determined confidence intervals. The mean difference between critical and literal strategy was 0.032 with a confidence interval [−0.06, 0.074] (p= 0.058 for the one-sided test *literal*<*critical*). The mean difference between critical and human average was 0.086, with a confidence interval of [0.063, 0.109] (p<0.001 for the one-sided test *average*<*critical*). Finally, the mean difference between literal and human average was 0.054, with a confidence interval [0.017, 0.087] (p= 0.004 for the one-sided test *average*<*literal*).

In sum, these data demonstrate that the critical strategy is at least as successful as the literal strategy and more successful than the average human production strategy. At

[10] For the comparison between strategies a simple measure of utterance length was chosen: It was defined as the number of conjuncts plus 1/2 for each occurrence of one of the complex phrases ENA (*some but not all*) and EPA (*some and possibly all*). For example, E-Any had length 1, E-A & E-E length 2, ENA-A & ENA-ENA & ENA-N length 5, etc. The average utterance length of a strategy S was then calculated as $\sum_w 1/7 \sum_u S(u|w)\ \mathit{length}(u)$.

the same time, the data in Table 4 show that it produces shorter utterances on average than both the literal and the average production strategy.

4.5.3 Analysis of individual sentences

For the combined data, each sentence of the critical strategy has a success rate that is equal or higher than the success rate of the corresponding literal strategy, see Table 7 in the Appendix. However, in Experiment 1a, two sentences (E-E & E-N and E-A & E-E) had lower rates than their literal counterparts. We, therefore, used Barnard's exact test for each pair of critical and corresponding literal utterances to test whether the difference is significant. Barnard's exact test is an alternative to Fischer's exact test that does not presuppose a hypergeometric distribution. It generally leads to much sharper p-values than non-parametric tests. However, the difference between success rates of the critical sentences and their literal counterparts failed to be significant (p= 0.27 and 0.45, respectively, for the one-sided test).[11] Hence, the lower rates in Exp. 1a are not statistically significant.

The main hypothesis that we have to verify in this section concerns the sentences that can result from shortening utterances of the critical strategy. These sentences are E-A, E-E, and E-N. The accumulated results are shown in (15) in the Appendix. For each sentence, it was most likely interpreted as ◨. We therefore compared their success rate with respect to ◨ with the success rate of E-A & E-E & E-N, which is the corresponding critical sentence that communicates ◨. We used a Mann-Whitney U test for each of the sentences. Compared to Barnard's exact test, the p-values are much more conservative, hence, it is more difficult to reject the null hypothesis that the short utterances and the critical utterance have the same success rates.[12] In all three cases, the Mann-Whitney U test showed that the differences between success rates of short utterances and the critical utterance are significant (p<0.001 for all three tests). We can, therefore, conclude that the second experimental hypothesis, that the critical strategy cannot be shortened without compromising communicative success, is borne out by the data.

4.5.4 Further analyses of individual sentences in the experimental corpus

In the following we present further results that are based on post-experimental analyses of the combined data of Experiments 1a and 1b. As mentioned before, the experiments

[11] Calculations were done with R-package 'Barnard' downloaded from CRAN-repository, see https://github.com/kerguler/Barnard, with parameter for Z statistic set to *unpooled Wald variance*. As a reminder, the interpretation data consist of pairs (u, w) representing instances in which utterances u was interpreted as w. To compare how successful utterances are in communicating some world v, we consider the binary random variable S_v with values $S_v(u, w) = 1$ if $v = w$ (u successfully interpreted as v), and $S_v(u, w) = 0$ otherwise. The test was applied to the 2×2-matrix defined by number of failures and successes of critical compared to number of failures and successes of literal; e.g. for E-E & E-N vs. E-ENA & E-N & N-A, we calculated the test for the 2×2-matrix (2, 0; 29, 12).

[12] We consider the binary random variable S_v with values $S_v(u, w) = 1$ if $v = w$ (u successfully interpreted as v), and $S_v(u, w) = 0$ otherwise. Hence, for each $u \in$ {E-A, E-E, EN} the Mann-Whitney U test tested whether the binary sets $\{S_v(u, w) \mid (u, w)$ a data item$\}$ and $\{S_v(u_v, w) \mid (u_v, w)$ a data item$\}$ of success values have the same median, with $v =$ ◨ and $u_v =$ E-A & E-E & E-N.

Table 5 All utterances consisting of two conjuncts that have been interpreted as referring to the world in which some of the girls found none, some, or all of their marbles. (N: absolute number of items interpreted by subjects)

Utterance	N	□▨■
E-A & E-E	40	2%
E-A & E-ENA	16	6%
E-A & E-N	68	6%
E-E & E-N	36	3%
E-N & ENA-A	19	26%
E-N & ENA-E	4	25%

Table 6 Short utterances that have been interpreted at least 5 times and are neither part of the critical or literal strategy, nor the result of shortening critical utterances (N: absolute number of items interpreted by subjects).

Utterance	N	□	▨	■	□▨	□■	▨■	□▨■
ENA-ENA	6	-	17%	-	67%	-	-	17%
N-A	8	12%	38%	-	50%	-	-	-
N-N	25	16%	4%	-	-	-	80%	-

provided interpretation data on 86 different sentences. We surveyed them with the aim of finding other shorter utterances that could replace critical utterances without reducing the strategies' communicative success. If such sentences exist, the replacement would lead to a more efficient strategy. If they do not exist, it follows that the critical strategy is indeed a maximally efficient strategy. It is not, however, the *unique* maximally efficient strategy, as there are equally short and successful alternatives for some of the critical utterances. For example, N-any could be replaced by A-N, and E-A & E-E by A-E & E-A.

As the critical strategy produces utterances consisting of 2 conjuncts for worlds □▨, □■, and ▨■, and an utterance consisting of 3 conjuncts for □▨■, we searched for utterances of length 1 that could communicate one of the four worlds, and utterances consisting of 2 conjuncts that could communicate □▨■.

We first consider utterances consisting of two conjuncts. There are 6 of them that were interpreted as □▨■ with some positive probability. As world □▨■ contains groups of girls that found none, some, and all of their marbles, only conjuncts with an outer existential quantifier are expected. If such a sentence has only two conjuncts, then the sentence must be a variant of one of the critical sentences with two conjuncts, of which we know that they reliably refer to other worlds than □▨■. From this, it can already be expected that no sentence with two conjuncts can reliably communicate □▨■. The 6 sentences are shown in Table 5. For none of the sentences, the probability with which they refer to □▨■ does exceed 26%. This confirms that, indeed, none of them can reliably communicate □▨■.

We next consider sentences consisting of one conjunct. The question is whether any of them could replace a longer sentence of the critical strategy without reducing the strategy's expected success. We only considered sentences that had been interpreted at least 5 times. There are 3 such sentences that do not already belong to the critical, the literal, nor the 3 short utterances considered in Table 3. The three sentences are shown in Table 6.

In order to show that none of the short utterances in Table 6 could improve the critical strategy, we compared their success rates for each of the worlds □▒, □■, ▒■, and □▒■ with the success rate of the corresponding critical utterance. In each case, a Mann-Whitney U test was calculated that showed that the short utterances are significantly less successful than the corresponding critical utterances (highest p-values: 0.04 for ENA-ENA and world □▒, 0.02 for N-N and world ▒■, and < 0.001 for N-A and world □▒).[13]

Summing up, none of the critical utterances can be replaced by shorter utterances consisting of one or two conjuncts without losing communicative success. Hence, our data show that the critical strategy is a maximally efficient strategy, i.e. all strategies with a success rate at least as high as that of literal production strategies must produce equally long or longer utterances.

5 Discussion

The model we developed defining a critical speaker strategy and our experimental results pose a challenge to existing accounts of implicature. We motivated our search for a maximally efficient production strategy by the following heuristic: a speaker who wants to communicate the identity of a certain world can describe it literally, or they may take advantage of implicature and leave certain aspects unsaid. This will lead to shorter utterances. There is a limit to the extent that utterances can be successfully shortened, such that shortening them beyond this limit will lead to miscommunication. Our two experiments showed that participants reliably communicate embedded and unembedded implicatures if shortening is done in accordance with the two elimination rules (ENA-Elim and N-X-Elim) introduced in Sect. 3. The critical strategy is as successful as the corresponding literal strategy, and shortening it further significantly reduces communicative success.

Each participant provides a probabilistic production strategy which is defined by the probability with which they chose an utterance u to describe a world w. As shown before, we can calculate the average length l of utterances and the average error rates e for this strategy (based on the average human interpretation strategy). The graph on the left side of Figure 4 shows the length-error pairs (l, e) of all participants.[14] In Sect. 3, we argued that a pattern as that shown on the right side of Fig. 4 (repeated from Fig. 2) should emerge, with the shaded area showing the distribution of human production strategies, and the *corner of efficiency* the area where the most efficient strategies are located. Our hypothesis was that the critical production strategy is located there. As can be seen from the graph in Fig. 4, the data follow the hypothesized pattern and the critical strategy is located in the corner of efficiency.

[13] We again considered the binary random variable S_v with values $S_v(u, w) = 1$ if $v = w$ (u *successfully interpreted as* v), and $S_v(u, w) = 0$ otherwise. For each world $v \in \{$□▒, □■, ▒■, □▒■$\}$, each short utterance u, and corresponding critical utterance u_v, a Mann-Whitney U test was calculated to determine whether there is a significant difference between the medians of the corresponding sets $\{S_v(u, w) \mid (u, w) \text{ a data item}\}$ and $\{S_v(u_v, w) \mid (u_v, w) \text{ a data item}\}$.

[14] 55 blue small points: 36 from Exp. 1a, and 19 from Exp 1b; 1 participant from Exp. 1b was completely replaced by the system; 2 participants from 1a were excluded because they took part in a 1-person group.

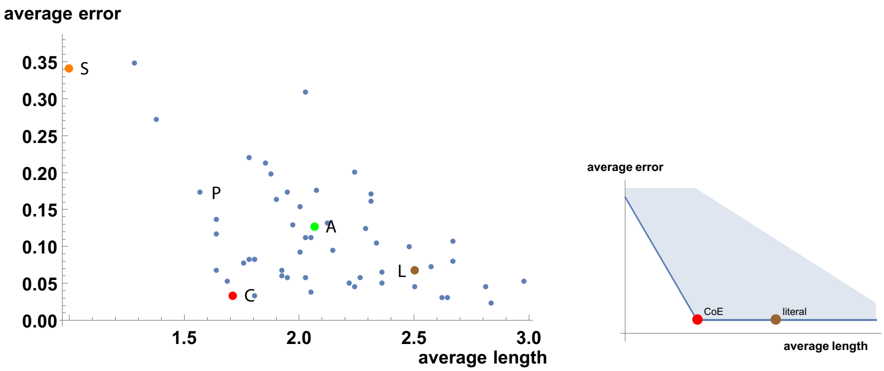


Fig. 4 Average length and error rates of individual strategies. **Left:** Small blue points represent strategies of individual participants. P: an individual participant; A: average human production strategy; C: critical strategy (7); L: literal strategy (7); S: short strategy (8); x-axis: average utterance length of strategy; y-axis: average error rate of strategy (1− success rate). **Right:** expected pattern (2); CoE: corner of efficiency; literal: literal strategy

As we have already emphasized before, we consider, for the purposes of this study, the elimination rules ENA-Elim and N-X-Elim as heuristic principles only.[15] We are not going to provide a model of our results, and likewise cannot enter a full discussion of the theoretical implications of our results for the various frameworks of implicature. However, we want to mention some obvious problems that they raise for *generalized neo-Gricean*, i.e. *basic globalist accounts*.[16] Central to this account is the idea that a sentence implicates the negation of all logically stronger structural alternatives. For example, basic globalism predicts that E-E implicates that not A-E '*all some*' and not E-A '*some all*', and, hence, that E-E implicates □▒. In the case of E-E & E-A there are the stronger alternatives A-E & E-A and E-E & A-A. Hence, basic globalism predicts the negation of A-E & E-A and A-A, and, therefore, that the actual world must be an element of {□■, □▒■}. The following table compares the predictions of generalized neo-Griceanism to our results (see accumulated data in the Appendix, Table 7 and (15)):

		short	predicted	observed
(13)	some some	E-E	{□▒}	□▒ (22%)
	some some & some all	E-A & E-E	{□■, □▒■}	▒■ (98%)

[15] The main obstacle for a theory based on elimination rules is the *generalisability* to the full set of possible utterances built up from *some, all, none, any,* and *some but not all*.

[16] The best worked out account is Sauerland (2004). However, only a small fragment of this theory is relevant to our paper. The extensions necessary, for example, for handling disjunctions are irrelevant to the logically much simpler sentences with upward entailing contexts considered in our study. The following discussion only applies to the basic version. It should also be mentioned that the distinction between global and embedded implicatures is less clear-cut when considering extensions of the basic models. For example, Russell (2006) observed that the embedded implicature N-A of utterance A-E can be accounted for globally under the assumption that an utterance of A-E leads to the negation of the structural alternative E-A.

Furthermore, generalized neo-Griceanism predicts for both sentences that they implicate *not all some*, i.e. *some none*. However, only 57% make this inference in the case of E-E, and only 2% in the case of E-A & E-E. In comparison, if the sentences are interpreted by randomly choosing one of the worlds consistent with semantic meaning, then the implicature would hold 50% of the time. Hence, observed rates are at chance level for E-E, and below chance for E-A & E-E.

In other cases, the predictions of generalized neo-Griceanism are too weak. In particular, embedded implicatures, which should not exist according to traditional globalist approaches, are clearly attested in our data. The following table shows some sentences for which basic globalism makes predictions that are too weak:

		short	predicted	observed
(14)	all some	A-E	{▨, ▨■}	▨ (93%)
	some none & some some & some all	E-A & E-E & E-N	{□■, □▨■}	□▨■ (100%)
	some none & some all	E-A & E-N	{□■, □▨■}	□■ (93%)

We mentioned before in Sect. 3 that elimination rules are speaker oriented principles that theories that approach pragmatics from interpretation would reconstruct in terms of exhaustification rules. One could conjecture that it is possible to solve the problems posed by our examples within a globalist framework if elimination rules are translated into alternative based exhaustification rules: whenever (N-X-Elim) is applied, add E-X to the alternative set and negate it. For example, that E-A & E-E implicates that N-N (*none none*) could be accounted for by assuming that E-N (*some found none*), (or any one of E-A & E-N, E-E & E-N, or E-A & E-E & E-N) is an alternative. By negating the alternative, the observed implicature *none none* follows.[17] Such a globalist account works as long as it is only applied to sentences that were derived by the (N-X-Elim) rule. As a general principle it fails. For example, if it is applied to E-A, then adding E-N as alternative would produce the implicature *none none*, and, hence, mean that E-A implicates ▨■ by negation of alternatives E-N and A-A. As the data on short utterances in Table 3 and in (15) show, only a minority of subjects chose this interpretation. In Sect. 1 of the Appendix, we discuss further examples of our data that are problematic for grammaticalism (Chierchia et al. 2012).

Some previous experimental studies have already provided evidence that embedded implicatures exist (Chemla 2009; Clifton and Dube 2010; Chemla and Spector 2011; Benz and Gotzner 2014; Potts et al. 2016; Franke et al. 2017; Gotzner and Romoli 2017). These studies employed experimental designs involving meta-linguistic judgements, as e.g. *picture verification tasks*, *inferential tasks*, and *graded acceptability tasks*, and focused on the comprehension of a few test sentences in isolation. Our data show that the proportion of participants drawing these implicatures successfully can be as high as the proportion of participants interpreting literal descriptions successfully.

[17] A solution along these lines was proposed by one of the reviewers.

We believe that our action-based task, which distinguishes between relevant readings and avoids judgements about truth and logical entailments, is the crucial reason why the proportion of participants deriving implicatures is much higher in our paradigm (see Gotzner and Benz 2018 for further discussion). Specifically, the rewards participants need to choose are contingent on their interpretation of the speaker's utterance. Note that in our interactive paradigm, the speaker has full knowledge and they also know about the decision problem the comprehender is facing. The comprehender, in turn, knows that the speaker has full knowledge of the world they are describing. We assume that these aspects are crucial to the high communicative success we observed in the current experiments and that changing them would increase the rate of communication errors.

As a final point of discussion, we would like to highlight the methodological implications of our study. In the majority of previous experiment on implicature, test sentences were not produced by a recognisable speaker, there is no addressee nor a recognisable *purpose of the talk exchange*, and, hence, there is no intended message that could be sought out behind the sentence's literal meaning. The experimental situation is often detached from purposeful conversation, and, hence, lacks a central precondition for implicatures in Grice's theory. To different degrees, picture verification, graded acceptability as well as inferential tasks are affected by this problem. Our study shows that, by implementing a precise conversational goal, a hearer's dominant interpretation of relevant sentences can be determined. Thereby, the study highlights the Gricean preconditions for deriving implicatures.

6 Conclusions

Grice defined an implicature as an inference towards the speaker's intended meaning that arises in cooperative conversation satisfying certain requirements concerning rational communication. We tried to implement these requirements in an interactive setting with an action selection task that guarantees that the meaning differences we are interested in are contextually relevant. Our experiments demonstrate that in a communicative context that satisfies Grice's requirements contested embedded implicatures can be communicated as reliably as literal meaning. We also presented a critical production strategy that was defined by two rules that allow simplifications of literal descriptions. These rules were the rule that '*some but not all*' can be simplified to '*some*', and 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 losing communicative success. We also predicted that our critical strategy will be located in what we called the *corner of efficiency* that emerges when considering the average length and error rates of all human production strategies. This prediction was borne out in our data. Concerning the theoretical consequences of our findings, it turned out that the results are particularly problematic for generalized neo-Gricean accounts.

More generally, our new paradigm opens up the possibility 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 a way such 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 dialogues.

Funding Open Access funding provided by Projekt DEAL.

Appendix

A Summary results

Notation Quantifiers within one sentence are separated by '-', and '&' represents the conjunction of sentences; A = *all*, E = *some*, N = *none*, ENA = *some but not all*.

Table 7 shows the results for the critical and literal strategy in Experiments 1a and 1b, as well as the accumulated results of both experiments. (N: number of items that had been presented to subjects for interpretation. The absolute numbers of literal utterances were lower in Exp. 1b than in Exp. 1a due to the lower number of participants.)

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

The table in (15) shows the accumulated data of both experiments for sentences that result from shortening the critical strategy even further (N: number of items that had been presented to subjects for interpretation):

(15)

Short utterances	N	□	▨	■	□▨	□■	▨■	□▨■
E-A	25	–	–	8%	–	–	20%	72%
E-E	23	–	39%	–	22%	–	4%	35%
E-N	22	9%	5%	–	14%	5%	–	68%

Other utterances produced by participants with their respective frequency and success rate per world (accumulated from Exp. 1a and 1b):

(16)

Utterance	N	□	▨	■	□▨	□■	▨■	□▨■
A-N	75	100%	–	–	–	–	–	–
E-A & E-ENA	16	–	–	–	–	–	94%	6%
ENA-A & ENA-ENA & ENA-N	12	–	–	–	–	–	–	100%
ENA-A & N-N	11	–	–	–	–	9%	91%	–
ENA-A & ENA-N	9	–	–	–	11%	89%	–	–
ENA-ENA & ENA-N	9	–	–	–	100%	–	–	–
E-A & E-E & N-N	8	–	–	–	–	–	100%	–
E-E & E-N & N-A	8	–	–	–	88%	–	–	12%
N-A	8	12%	38%	–	50%	–	–	–
ENA-A & ENA-ENA	7	–	–	–	–	–	100%	–
ENA-ENA	6	–	17%	–	67%	–	–	17%
ENA-A & ENA-N & N-ENA	5	–	–	20%	–	80%	–	–
ENA-ENA & ENA-N & N-A	5	–	–	–	100%	–	–	–
ENA-N & N-A	5	–	–	–	80%	20%	–	–

The full list of interpreted sentences is available at https://osf.io/pq8ca/?view_only=e5a502aaba82416e8c3eb2d3eb1719a2.

B Instructions

Lisa, Jenny, Marie, Susi, Anna und Caro sind in derselben Kitagruppe. Sie lieben es, Murmeln zu sammeln. Jede der Schwestern besitzt einen Satz mit 4 einzigartigen Murmeln. Während die Mädchen spielen, gehen die Murmeln oft verloren. Da

Table 7 Result: Success rate (%) of critical and literal strategy per world (N: absolute number of items interpreted by subjects)

(a) Overview results for critical strategy

		Exp. 1a		Exp. 1b		Exp. 1a+b	
world	critical strategy	N	success	N	success	N	success
□	N-any	43	100%	37	100%	80	100%
▨	A-E	30	93%	24	92%	54	93%
■	A-A	108	99%	60	95%	168	98%
□▨	E-E & E-N	31	94%	5	100%	36	95%
□■	E-A & E-N	46	96%	22	86%	68	93%
▨■	E-A & E-E	35	97%	5	100%	40	98%
□▨■	E-A & E-E & E-N	42	100%	8	100%	50	100%

(b) Overview results for literal strategy

		Exp. 1a		Exp. 1b		Exp. 1a+b	
world	literal strategy	N	success	N	success	N	success
□	N-any	43	100%	37	100%	80	100%
▨	A-ENA	54	93%	25	92%	79	92%
■	A-A	108	99%	60	95%	168	98%
□▨	E-ENA & E-N & N-A	12	100%	3	67%	15	93%
□■	E-A & E-N & N-ENA	16	88%	6	67%	22	82%
▨■	E-A & E-ENA & N-N	13	100%	1	0%	14	93%
□▨■	E-A & E-ENA & E-N	29	97%	10	90%	39	95%

ihre Erzieherin nun möchte, dass die Mädchen aufräumen, hat sie beschlossen sie zu belohnen, wenn sie ihre Murmeln finden.

Ein Mädchen bekommt

- ein Stück Schokolade, wenn es alle 4 ihrer Murmeln findet
- ein Bonbon, wenn es weniger als 4 ihrer Murmeln findet
- ein Gummibärchen, wenn es keine ihrer 4 Murmeln findet.

Nachdem die Mädchen aufgeräumt haben, legen sie ihre Murmeln auf den Tisch. Die Erzieherin muss dann in den Supermarkt gehen, um alle Süßigkeiten zu kaufen, die sie als Belohnung für die Mädchen braucht. Ein Mädchen kann jeweils nur eine Belohnung erhalten. Damit die Mädchen nicht enttäuscht sind, ist es wichtig die richtigen Süßigkeiten zu kaufen.

Im folgenden Experiment werden Sie abwechselnd eine von zwei verschiedenen Rollen einnehmen. Das System weist Ihnen zufällig eine Rolle zu, entweder:

(1) Den Auftraggeber: Sie werden Bilder von den Murmeln, die die Mädchen gefunden haben, sehen. Stellen Sie sich vor, dass die Erzieherin der Mädchen im Supermarkt ist und Sie anruft, weil sie vergessen hat, die Murmeln zu überprüfen. Ihre Aufgabe ist es nun, einen Satz zu bilden, so dass die Erzieherin der Mädchen weiß, welche Süßigkeiten sie kaufen muss.

ODER

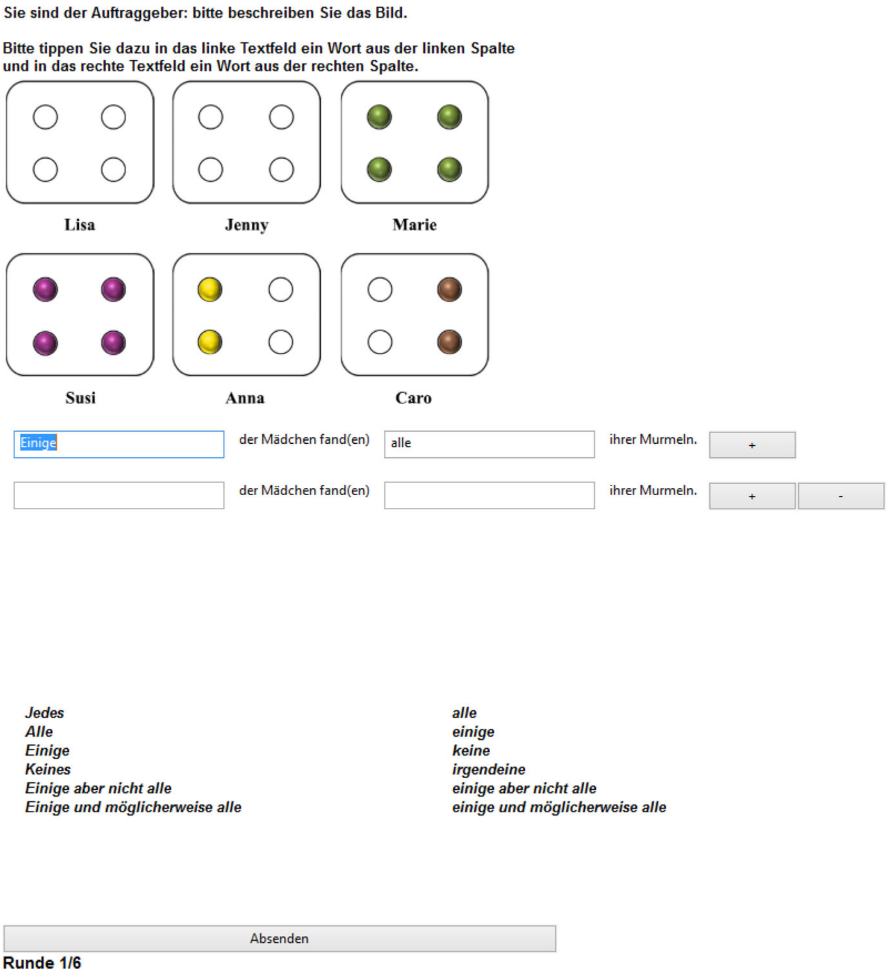
(2) Den Einkäufer: Sie sind die Erzieherin. Stellen Sie sich vor, Sie werden von jemandem angerufen, der bei den Mädchen zu Hause ist und weiß, welche Murmeln sie gefunden haben. Nun müssen Sie entscheiden, welche Süßigkeiten gebraucht werden. Bitte bedenken Sie, dass es wichtig ist alle Süßigkeiten zu kaufen, die benötigt werden. Sie kriegen dafür eine Anweisung von einem anderen Mitspieler.

Bevor Sie mit dem Experiment beginnen, werden einige Proberunden durchgeführt, bei denen Sie üben sollen, welche Süßigkeiten gebraucht werden, je nachdem wie viele Murmeln die Kinder gefunden haben.

C Screenshots

Figure 5 shows an example screen participants saw in the experiment when playing the role of producer.

Figure 6 shows an example screen participants saw in the experiment when playing the role of interpreter.



Sie sind der Auftraggeber: bitte beschreiben Sie das Bild.

Bitte tippen Sie dazu in das linke Textfeld ein Wort aus der linken Spalte
und in das rechte Textfeld ein Wort aus der rechten Spalte.

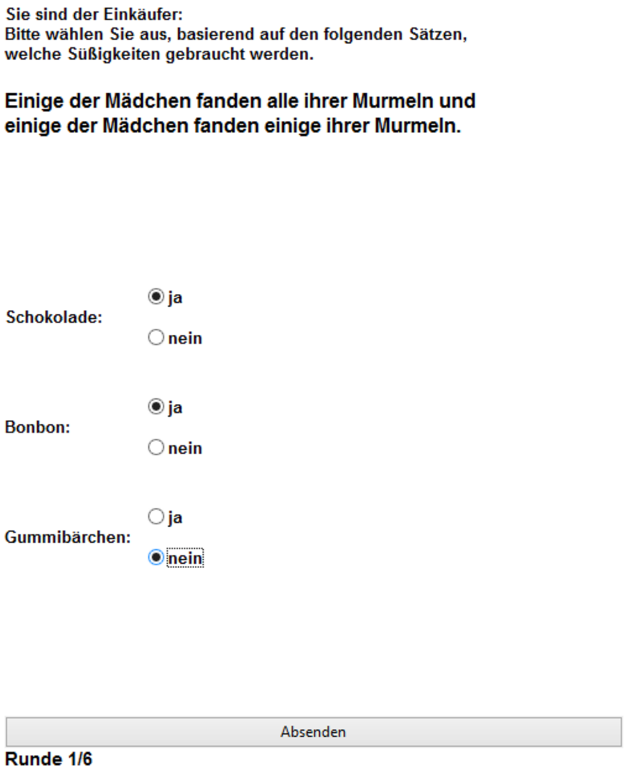
Einige der Mädchen fand(en) alle ihrer Murmeln. +

der Mädchen fand(en) ihrer Murmeln. + -

Jedes	*alle*
Alle	*einige*
Einige	*keine*
Keines	*irgendeine*
Einige aber nicht alle	*einige aber nicht alle*
Einige und möglicherweise alle	*einige und möglicherweise alle*

Runde 1/6

Fig. 5 Screenshot: Example of production task



Sie sind der Einkäufer:
Bitte wählen Sie aus, basierend auf den folgenden Sätzen, welche Süßigkeiten gebraucht werden.

Einige der Mädchen fanden alle ihrer Murmeln und einige der Mädchen fanden einige ihrer Murmeln.

Schokolade: ● ja ○ nein

Bonbon: ● ja ○ nein

Gummibärchen: ○ ja ● nein

Runde 1/6

Fig. 6 Screenshot: Example of comprehension task

D Evaluation of the grammatical account

Local and global accounts are structural accounts in so far as implicatures are derived from purely structural properties of sentences. Global accounts provide a unique pragmatic interpretation. In the more recent grammatical versions of localism, embedded implicatures are derived by inserting optional *only* operators in the logical form of sentences (Chierchia et al. 2012). If an operator is inserted, it negates all alternatives in its scope if its negation is consistent with standard semantics. Depending on the author, only logically stronger alternatives are negated, or non-weaker alternatives may also be negated. In general, this leads to a wide range of possible readings. We consider here only the basic cases, i.e. no recursive application of *only*-operators, no modifications of alternative sets, and negation only of stronger alternatives. The table in (17) shows the predicted readings for the A-E sentence. The first reading is the *literal*, the second the *embedded*, and the third the *global* reading of A-E. Note that the global reading is here just a special case of the structurally predicted readings.

	All of the girls found some of their marbles. (A-E)			
(17)	(a)	all	some	▨, ■, ▣
	(b)	all	O [some]	▨
	(c)	O [all	some]	▨, ▣

The second table shows the readings predicted for utterance E-A & E-E. As in the case of A-E in (17), the first line (a) shows the *literal* reading of E-A & E-E, and the last line (k) the *global* reading. The other readings are embedded readings predicted by inserting silent O below sentence level. All of the readings are too weak such that additional principles are needed for deriving the observed interpretation (▣).

	Some of the girls found all and some found some of their marbles.					
(18)	(a)	some	all	some	some	■, ◨, ▣, ▥
	(b)	some	all	some	O [some]	▣, ▥
	(c)	some	all	O [some]	some	◨, ▥
	(d)	some	all	O [some]	O [some]	▣, ▥
	(e)	some	all	O [some	some]	⊥
	(f)	O [some]	all	some	some	◨, ▣, ▥
	(g)	O [some]	all	some	O [some]	▣, ▥
	(h)	O [some]	all	O [some]	some	◨, ▥
	(i)	O [some]	all	O [some]	O [some]	▣, ▥
	(j)	O [some]	all	O [some	some]	⊥
	(k)	O [some	all	some	some]	◨, ▥

As this type of account derives implicatures purely on the basis of structural properties, it does not straightforwardly predict a dominant interpretation in specific conversational contexts. Sometimes the *strongest meaning hypothesis* (SMH) (Dalrymple et al. 1994; Chierchia 2013) is invoked for selecting between readings. SMH predicts that the strongest interpretation will be chosen when subjects have to choose between alternative readings. This correctly predicts that A-E implicates ▨. In the case of E-E, however, it selects the same faulty interpretation as the global account, namely ◫. In the case of E-A & E-E, it fails to select a unique reading. It has also been observed by Franke et al. (2017) that subjects do not always follow SMH when choosing between global and embedded readings. At least for the sentences considered in our study, it is not a reliable criterion for predicting utterance interpretations.

References

Benz, A. (2009). Implicatures of irrelevant answers and the principle of optimal completion. In P. Bosch, D. Gabelaia, & J. Lang (Eds.), *7th international Tbilisi symposium on logic, language, and computation, TbiLLC 2007 Tbilisi, Georgia, October 2007, revised selected papers* (pp. 95–109). Berlin: Springer.

Benz, A. (2012). Errors in pragmatics. *Journal of Logic, Language, and Information*, *21*, 97–116.

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

Benz, A., & Gotzner, N. (2019). Quantifier *irgendein* and local implicatures. *Snippets*, *37*, https://doi.org/10.7358/snip-2019-037-bego.

Chemla, E. (2009). Universal implicatures and free choice effects: Experimental data. *Semantics and Pragmatics*, *2*(2), 1–33. https://doi.org/10.3765/sp.2.2.

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

Chierchia, G. (2013). *Logic in grammar: Polarity, free choice, and intervention*. Oxford: Oxford University Press.

Chierchia, G., Fox, D., & Spector, B. (2012). Scalar implicature as a grammatical phenomenon. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An international handbook of natural language meaning* (Vol. 3, pp. 2297–2331). Berlin: De Gruyter Mouton.

Clifton, C., & Dube, C. (2010). Embedded implicatures observed: A comment on Geurts and Pouscoulous (2009). *Semantics and Pragmatics*, *3*(7), 1–13. https://doi.org/10.3765/sp.3.7.

Dalrymple, M., Kanazawa, M., Mchombo, S., & Peters, S. (1994). What do reciprocals mean? *Semantics and Linguistic Theory*, *4*, 61–78.

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

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

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, *34*(3), 351–386.

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

Franke, M., Schlotterbeck, F., & Augurzky, P. (2017). Embedded scalars, preferred readings and prosody: An experimental revisit. *Journal of Semantics*, *34*, 153–199. https://doi.org/10.1093/jos/ffw007.

Geurts, B. (2009). Scalar implicatures and local pragmatics. *Mind and Language*, *24*(1), 51–79.

Geurts, B., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics and Pragmatics*, *2*(4), 1–34. https://doi.org/10.3765/sp.2.4.

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

Gotzer, N., & Benz, A. (2018). The best response paradigm: A new approach to test implicatures of complex sentences. *Frontiers in Communication, 2*(21). https://doi.org/10.3389/fcomm.2017.00021.

Gotzner, N., & Romoli, J. (2017). The scalar inferences of strong scalar terms under negative quantifiers and constraints on the theory of alternatives. *Journal of Semantics*, *35*(1), 95–126. https://doi.org/10.1093/jos/ffx016.

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

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

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicatures*. Cambridge, MA: MIT Press.

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

van Rooij, R., & Schulz, K. (2004). Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information*, *13*, 491–519.

Russell, B. (2006). Against grammatical computation of scalar implicatures. *Journal of Semantics*, *23*, 361–382.

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

Sauerland, U. (2012). The computation of scalar implicatures: Pragmatic, lexical or grammatical? *Language and Linguistics Compass*, 36–49.

van Tiel, B. (2014). *Quantity matters: Implicatures, typicality and truth*: Radboud Universiteit Nijmegen dissertation.

van Tiel, B., Noveck, I., & Kissine, M. (2018). Reasoning with ‘some’. *Journal of Semantics*, *35*(4), 757–797.

Yoon, S. O., & Brown-Schmidt, S. (2017). Aim low: Mechanisms of audience design in multiparty conversation. *Discourse Processes*, *55*, 566–592.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.