Implicatures in (non-)monotonic environments[1]

Nicole GOTZNER — *University of Osnabrück*
Anton BENZ — *Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)*

Abstract. It has been shown that *some* embedded under universal and existential quantifiers can locally be interpreted as *some but not all*. This implicature can be reliable if supported by pragmatic context. We present experimental evidence showing that this local implicature can also be understood reliably if *some* is embedded under non–monotonic operators like *only* and *exactly*. Moreover, these operators reliably lead to a *others none* interpretation. We present a decomposition analysis which splits non-monotonic operators into a positive upward entailing and a negative downward-entailing component, where the positive component is responsible for the local *but not all* implicature, and the negative one for the *others none* interpretation.

Keywords: scalar implicature, embedded implicature, quantifiers, downward entailment, non-monotonicity

1. Introduction

A key function of implicature is to improve the efficiency of communication (Levinson, 2000: Ch. 1). If an implicature is *reliably* communicated, it allows the speaker to produce less linguistic material while obtaining the same degree of communicative success as with more complex literal descriptions. Consider the utterance *Kate found some of the marbles* in (1). If hearers infer from this that Kate found some but not all of the marbles and that no other person found any marbles, a speaker can use the much shorter sentence (1a) to convey the same meaning as the longer sentence (1c).

(1) a. Kate found some of the marbles.
b. Kate found some but not all of the marbles.
c. Kate found some but not all of the marbles, and the others found none.

It is uncontroversial that a simple sentence like (1a) carries a *some but not all* and an *others-none* reading in the right context. But sentences with multiple scalar items may trigger a variety of different implicatures and have posed several challenges to theories of implicature (see Gotzner and Romoli 2022 for a recent overview). Here, we focus on cases in which *some* is embedded under a non-monotonic operator such as *exactly n*, see (2). Sentences of this type are of particular theoretical relevance since potential alternatives are logically independent from the assertion. Thus, potential implicatures cannot be derived in a standard (neo-)-Gricean manner (Levinson, 1983; Horn, 1989) and are not even predicted to arise. As we will see, grammatical approaches (Chierchia et al., 2012; Chemla and Spector, 2011) also face a challenge when we consider different numerals that associate with *exactly*.

[1]This work was supported by the Bundesministerium für Bildung und Forschung (BMBF) (Grant Nr. 01UG1411), and the Deutsche Forschungsgemeinschaft (DFG), Grant Nr. BE 4348/4-1 and 4-2, as part of the priority program *New Pragmatic Theories based on Experimental Evidence* (SPP 1727) and the Emmy Noether grant awarded to Nicole Gotzner (GO 3378/1-1). We are grateful to Lisa Raithel for programming the system for the experiments.

 In: Daniel Gutzmann and Sophie Repp (eds.): Proceedings of Sinn und Bedeutung 26, pp. 340–358. University of Cologne.

(2) a. Exactly *n* persons found some of the marbles.
b. Exactly *n* persons found some but not all of the marbles.
c. *n* persons found some but not all of the marbles, and the others found none.

Our first goal is to test which implicatures are communicated in different non-monotonic contexts. To this end, we present an experiment that compares the non-monotonic contexts created by *only P* (where P is a proper name), *only one person*, *exactly one person*, and *exactly two*, see (3), and show that all of these non–monotonic operators *reliably* trigger the *some but not all* reading in (3a), in addition to the *others-none* reading in (3b).

(3) Only Kate / only / exactly *n* of the girls found some of their marbles.
a. $\rightsquigarrow$ Kate / *n* found some but not all of their marbles. (x-sbna)
b. $\rightsquigarrow$ None of the other girls found any marbles. (others-none)

Our second goal is to provide an account of the theoretical challenge posed by our results. We will argue that non-monotonic operators can be decomposed into an upward and a downward-entailing component (following e.g., Alxatib 2014; Bar-Lev 2018; Elliott and Marty 2019). This way, the attested readings in (3a) and (3b) can be derived in a uniform manner. We sketch two ways in which a decomposed analysis of non-monotonic operators can be incorporated into a grammatical theory of implicature and algorithmic model by Benz and Gotzner (2021), which provides an account of how a variety of different utterance combinations are interpreted in context (and how they are produced). Overall, our findings highlight the complex interaction between entailment relations, sentence structure and contextual considerations.

This paper is organized as follows. First, we present the theoretical and experimental background on implicatures of sentences with multiple scalars. Then, we introduce our experimental paradigm that tests which implicatures are reliably communicated (rather than testing the mere existence of an implicature). Finally, we present our theoretical proposal and discuss the relevance of contextual considerations in the derivation of embedded implicatures.

2. Theoretical and experimental background

Grice and Horn considered implicatures to be inferences about speaker meaning (Grice, 1975; Horn, 1972). A special case are *scalar* implicatures. As Horn (1972) observed, they follow a simple pattern: there is a set of alternatives $\mathscr{A}$ of comparatively complex items that are ordered by logical strength. This set is called a *scale*. If $A(\_\_)$ is a sentence frame into which an element of $\mathscr{A}$ is inserted, then this implicates the falsity of all sentences resulting from inserting stronger alternatives.[2] Looking at implicatures in this way and considering them more abstractly, it was an obvious thought to regard them as a special kind of default logical inference triggered by the logical form of sentences (e.g. Levinson 2000; Chierchia 2004).

For Grice, implicatures are part of communicated meaning. This means that they are part of the speaker's intended meaning of an utterance, and that the hearer is able to work out that they are (Grice, 1975: p. 50). In this respect, there are a number of empirical questions that we may ask about implicatures:

[2]Sometimes it is assumed that the primary implicature says only that the speaker does not believe in the truth of the alternatives, and that the primary implicature is then strengthened in a second step saying that the speaker believes that the alternative is false (e.g. Sauerland 2004, see also Soames 1982; Horn 1989).

(4) Questions about implicatures:
   a. *Existence*: Does the reading exist?
   b. *Reliability*: Can the reading be communicated reliably?
   c. *Regularity*: Is the reading regularly derived by default?

Existence refers to the potential availability of an implicature. Evidence for this may come from different sources, for example that language users, in some contexts, endorse a given reading or that they indicate a sentence fits best to a situation in which the reading is satisfied. However, existence is not enough to make an implicature part of communicated meaning. For this, it has to be reliably understood by addressees, and to reliably accompany the production of a sentence. Being reliable may, however, still require the help of context, for example contextual relevance. Hence, one may also ask whether a reading is regularly associated with the utterance of a sentence, independent of contextual relevance.

As mentioned above, the most controversial are cases in which a scalar term is embedded under a logical operator, so-called *embedded implicatures*. The operator that embeds a scalar expression may impose different monotonicity patterns. *Upward entailing* (UE) operators create a context A(.), in which the stronger A(*all*) implies A(*some*). In contrast, *downward entailing* (DE) operators reverse entailment relations such that A(*some*) implies A(*all*). A classic example is negation. Operators that are neither upward- nor downward-entailing are called non-monotonic (NM), for example *only* and *exactly*. Since A(*all*) and A(*some*) are logically independent in NM contexts such cases are of critical interest to theories of implicature. In the following, we present the existing experimental evidence for embedded implicatures in UE, DE and NM environments and discuss its relevance to the theoretical debate.

The first experiment testing embedded implicatures was carried out by Geurts and Pouscoulous (2009). In this experiment, participants were asked to judge whether sentences involving a potential embedded implicature are true or false given a situation represented by a picture. In this sentence-picture verification task, none of the participants seemed to have derived an embedded implicatures in UE contexts, as in (5) (though there was evidence for such a reading in an inference task, see van Tiel 2014; Benz and Gotzner 2014 for a methodological discussion).

(5) All squares are connected to some of the circles

Geurts and Pouscoulous (2009) disproved Chierchia's (2004) assumption that embedded implicature are default inferences triggered by the logical form of sentences (see also Magri 2009). Yet this experiment does not provide counter-evidence against the existence of embedded implicatures. In modified versions of the paradigm, Chemla and Spector (2011) found evidence that the embedded *some but not all* implicature is a potential reading of sentence (5) (see also Clifton Jr and Dube 2010). The embedded implicature in UE contexts (*all-some*) can be accounted for, for example, in grammatical approaches (e.g. Chierchia et al. 2012). Different approaches also agree that implicatures in DE contexts should be dispreferred since entailment relations are reversed, which is consistent with the experimental data (Geurts and Pouscoulous 2009; Chemla and Spector 2011; Potts et al. 2016, but see Chierchia 2013 for cases in which focal stress induces implicatures in DE contexts).

Chemla and Spector (2011) added a crucial test case involving non-monotonic quantifiers (6), in which theories make diverging predictions. For this case, alternatives do not stand in an entailment relation. Thus, it has been argued that the embedded implicature can only be derived if one allows for local enrichment of *some* to 'some but not all'. Chemla and Spector (2011) found evidence for the embedded/local reading in (6b) interpreted and thus interpreted their results as lending prima facie support for a grammatical account of implicature. But the authors also sparked a debate about whether embedded implicature can be derived globally by assuming non-standard alternatives. For the case of *exactly 1*, when we globally negate the alternative *exactly 1 square is connected to all of the circles*, we get a *some but not all* (*sbna*), in addition to the *others-none* reading due to the semantics of *exactly 1*, see (5). When we locally enrich *some*, the *others-none* is not derived, as in (6b).

(6) Exactly one square is connected to some of the circles
   a. **Global**: One square is connected with sbna of the circles and no other square is connected with any of the circles
   b. **Local**: One square is connected with some of the circles and the other squares may be connected with no or all circles

Interestingly, when we consider numerals higher than one such as *exactly two*, the *others-none* reading can neither be derived via standard global nor local mechanisms (see also Section 6 for a more detailed discussion). For this reason, we decided to test a variety of different non-monotonic operators to explore whether Chemla and Spector (2011)'s results generalize.

A variety of further studies have provided evidence that embedded implicatures for different kinds of doubly-quantified sentences *exist* (Potts et al., 2016; Gotzner and Romoli, 2017; van Tiel et al., 2018; Franke and Bergen, 2020). But none of the studies mentioned so far address question (4b) about which implicatures are *reliably communicated*. This was the focus of Gotzner and Benz (2018); Benz and Gotzner (2021), who developed an experimental scenario that strictly controlled for Grice's conversational requirements. In these experiments, embedded implicatures were reliably communicated, with a success rate equal to literal descriptions.[3] In the following section, we introduce the so-called best response paradigm in more detail. We return to a detailed discussion of the theoretical implications after presenting the results of our current experiments on non-monotonic operators.

3. Best response paradigm and critical production strategies

The present study uses a version of the best response paradigm by (Gotzner and Benz, 2018; Benz and Gotzner, 2021). In the BR paradigm, the purpose of the talk exchange is given by an action selection problem. For the purpose of the current study, we adapted the scenario of Benz and Gotzner (2021) so that we can read off numerical interpretations from the participants' responses. In one of our experimental scenarios, there are 6 children, each owning a set of 4 special edition marbles with which they play. After playing, they have to find them and put them into their boxes again. The experiments were done in pairs, one participant saw a picture showing how many marbles each of the 6 children has found, and the other participant had to buy rewards: a gold medal for each child that finds all 4 of the marbles, a silver medal for each

[3]Interestingly though, some global implicatures predicted by all theories were not reliably communicated.

child that finds fewer than 4, and a bronze medal for each one that finds none, as a consolation prize. Hence, in sum, 6 medals had to be bought. (7) shows the critical sentences. We decided to test *only* because this operator has been assumed to be non-monotonic but has some downward-entailing properties (for example, it licenses NPIs, see von Fintel 1999; Chierchia 2013).

(7) a. Sasha found some of the marbles.
b. Only Sasha found some of the marbles.
c. Only one child found some of the marbles.
d. Exactly one child found some of the marbles.
e. Exactly two children found some of the marbles.

The critical interpretations are those in (8). We test whether (7a) and (7b) are reliably interpreted as (8a), whether (7c) and (7d) are reliably interpreted as (8b), and whether (7e) is reliably interpreted as (8c).

(8) a. Sasha found sbna of the marbles, the rest found none.
b. One child found sbna of the marbles, the rest found none.
c. Two children found sbna of the marbles, the rest found none.

There are 28 possible combinations of medals that sum up to 6. We call a combination of 6 medals a *possible world*. This is motivated by the fact that the potential interpretations of critical sentences can be identified with a combination of rewards. For example, if a participant interprets (7e) *exactly 2 children found some of their marbles* as meaning that 2 found some but not all, and 4 found none, then this corresponds to a combination of 2 silver medals and 4 bronze medals.

We selected 16 possible worlds and prepared appropriate pictures. Each possible world was represented by 3 different picture items. The worlds are shown in Table 1. The 16 worlds are relevant to production only. When interpreting a sentence, then any combination of medals, hence, any world, can be chosen by the comprehender. The worlds were chosen such that speakers have an opportunity to produce critical sentences without limiting their expectations about the state of affairs.

selected worlds:			
006	060	204	240
015	105	213	501
024	123	222	510
051	150	231	600

Table 1: Selected worlds with required bronze, silver, and gold medals (colour code: orange-grey-yellow, left to right).

As explained before, the interpretation of sentences can be read off from the response of the participant who buys the medals. If the participant interprets (7a) or (7b) as (8a), then they choose the medals that are appropriate for 510. If they interpret (7c) or (7d) as (8b), then, again, they choose medals appropriate for 510. Finally, It can be seen that they interpret (7e) as (8c), if they choose medals appropriate for 420.

Testing whether sentences are *reliably* interpreted in a certain way poses a methodological problem. How can *reliability* be measured? We use two baselines with which to compare the frequencies with which critical interpretations are chosen: the *literal baseline* defined by literal descriptions of states of affairs, and the *human baseline* defined by the participants' choice of utterances.

Literal baseline criterion. The frequency with which the critical interpretation is chosen can be compared with the frequency with which participants choose the critical interpretation for a sentence that expresses it literally.

Human baseline criterion. The frequency with which the critical interpretation is chosen can be compared to the average success rate of utterances produced by human participants.

The average success rate of humans should be high. As they are prone to errors, however, their average success rate should be below the success rate of a strategy that only produces sentences that are reliably interpreted. Hence, the human baseline criterion provides a lower limit of what can be called *reliable*. Hence, the goal of our experiments is to find out whether the success rate of critical sentences is *higher* than the average human success rate, and possibly as high as that of literal sentences. To set the frequencies into a meaningful context, we also compared them to the frequencies with which alternative sentences are interpreted that result from replacing *some* by the stronger *all*. The alternative sentences are shown in (9).

(9) a. Sasha found all of the marbles.
b. Only Sasha found all of the marbles.
c. Only one child found all of the marbles.
d. Exactly one child found all of the marbles.
e. Exactly two children found all of the marbles.

4. Experiments

4.1. Goals and rationale

We test whether the critical sentences in (7) are as reliably interpreted as corresponding literal descriptions, and more reliably as the average human descriptions. If we call *literal* the production strategy defined by literal descriptions produced by human participants, and *human* the average strategy of humans, then the hypotheses are:

(10) (I) $literal \approx critical$, (II) $critical > human$.

We further expect the success rates of the alternative sentences in (9) to be lower than those of average human utterances.

To test these hypotheses, we implemented an interactive version of the best response paradigm (Benz and Gotzner, 2021) with a comprehension and a production side. To have more interpretation data about critical sentences, we used a design with a confederate to produce them.

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. Their 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 measured how reliably an utterance was interpreted by comprehenders.

sentence	**critical meaning**
Critical sentences and their critical meanings	
P-E	5 1 0
Exact1-E	5 1 0
Exact2-E	4 2 0
Only1-E	5 1 0
OnlyP-E	5 1 0
control & explorative test sentences	
Exact1-E & Rest-N	5 1 0
OnlyP-E & Rest-N	5 1 0
OnlyP-E & Rest-A	0 1 5
OnlyP-E & P-A	4 1 1
Exact2-E & 2-A	2 2 2
comparison with other experiments	
A-E	0 6 0

Table 2: Sentences produced by system and their meanings.

4.2. Methods

4.2.1. Apparatus

For our experiments, we programmed a system in Python using the GUI toolkit wxPython[4], which allowed us to implement a game with four participants. Participants were seated in a lab with four computers separated by booths. 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 running experiments with either two or four players who had to be present at the same time. In both variants, one participant could be replaced by the confederate system when in producer role. If one participant did not show up, the system could take over the vacant role. When in the vacant receiver role, the system stored the sentence that was sent to it, but did not interpret it. Human participants could not notice it, if one of the roles was filled by the system.

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

English	German	short
All children – some	Alle Kinder – einige	A-E
Exactly one of the children – some	Genau eines der Kinder – einige	Exact1-E
Exactly one of the children – all	Genau eines der Kinder – alle	Exact1-A
Exactly two of the children – some	Genau zwei der Kinder – einige	Exact2-E
Exactly two of the children – all	Genau zwei der Kinder – alle	Exact2-A
Exactly one of the children – some & the rest – none	Genau eines der Kinder – einige & der Rest – keine	Exact1-E & Rest-N
Exactly two of the children – some & two children – all	Genau zwei der Kinder – einige & zwei Kinder – alle	Exact2-E & 2-A
Only one of the children – some	Nur eines der Kinder – einige	Only1-E
Only one of the children – all	Nur eines der Kinder – alle	Only1-A
Only Sascha – some	Nur Sascha – einige	OnlyP-E
Only Kim – all	Nur Kim – alle	OnlyP-A
Only Leo – some & the rest – none	Nur Leo – einige & der Rest – keine	OnlyP-E & Rest-N
Only Leo – some & the rest – all	Nur Leo – einige & der Rest – alle	OnlyP-E & Rest-A
Only Toni – some & Nicki – all	Nur Toni – einige & Nicki – alle	OnlyP-E & P-A
Nicki – some	Nicki – einige	P-E
Toni – all	Toni – alle	P-A

Table 3: Sentences produced by confederate system.

4.2.2. Participants

For the experiments, participants were invited in groups of 4. Participants were recruited via the online recruitment system LingEx of ZAS Berlin and Humboldt University for linguistic experiments. In total, 61 native German participants (48 female, 11 male, 1 unspecified; mean age: 29.2) took part in the experiment. Due to no-show participants, the experiments were run in groups of varying sizes: there were groups with 3 or 4 players, groups with 2 players, and 1 groups with 1 player. In all experiments, the confederate system played a production strategy that produced sentences for which we needed more interpretation data, see Table 3. 28 participants took part in the version with 4 players (7 groups), 24 participants took part in the version with 3 players (8 groups), 8 participants in the version with 2 players (4 groups), and 1 participant in the version with 1 player. To guarantee uniformity of the experimental situation, we excluded the versions with 1 and 2 participants from analysis. Furthermore, we excluded all groups in which at least one participant produced the same sentence in 3 or more different worlds. One group was excluded as, by mistake, no confederate took part. We set as target 3 valid groups per scenario. We ran experiments until this goal was reached. As we prepared three scenarios, we gathered valid results for 9 groups that entered evaluation (5 groups with 3, and 4 groups with 4 participants), 3 for each scenario, *football*, *pumpkin*, and *marble*.

4.2.3. Scenario

We developed 3 scenarios in each of which participants had to do the same tasks: the *marble*, *pumpkin*, and *football* scenario. As guiding example, we present the *marble* scenario, which has been successfully used in (Gotzner and Benz, 2018; Benz and Gotzner, 2021), and modified from there for the present purposes. This scenario involved six children who each own a set of four special edition marbles. While the children 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 children wants to reward them depending on how many marbles the children find. In particular, participants were presented with the following reward system in the instructions (all instructions are found in the supplementary material appended at the end of the paper):

A child gets:
- a gold medal if she finds all 4 of her marbles
- a silver medal if she finds fewer than 4 of her marbles
- bronze medal when she finds none of her 4 marbles (as a consolation prize).

In the *pumpkin* scenario, the children sold pumpkins at a school fair. Each had 4 pumpkins on sale, and they receive medals depending on how many of their four pumpkins they managed to sell. In the *football* scenario, the children had to shoot soccer balls into holes in a football wall. They received medals depending on how many of the 4 holes they hit.

4.2.4. 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 16 different states of the world. For example, the speaker saw a picture of the girls' marble boxes in which 5 children found all 4 of their marbles, and 1 found none (see Figure 1). Each world was instantiated by three items in total.

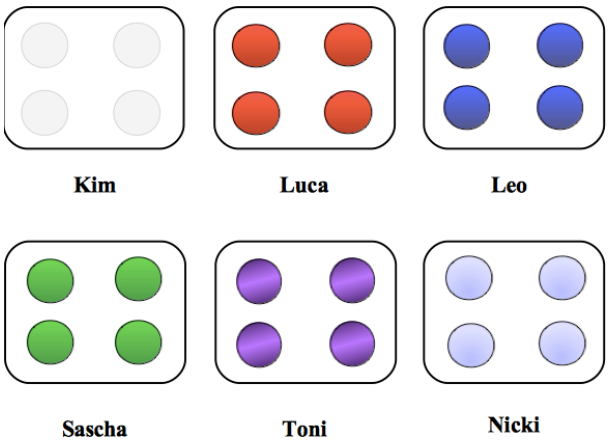

Figure 1: Example picture: Item for world 1 0 5.

The task of the speaker was to describe the picture so that the comprehender can buy the appropriate medals for the children. For the situation in Figure 1, this means that the comprehender has to buy 1 bronze medal and 5 gold medals. Participants in speaker role were presented with a sentence frame, placed underneath the picture, where they were required to fill in two blanks. They were allowed to type in one of the following words or phrases shown in Table 3 in the supplementary material. Participants were allowed to produce up to five sentences to describe a given picture. Figure 1 in the supplementary material, 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.[5]

[5]DOI 10.17605/OSF.IO/WY5S9

When the speaker was done describing the picture, the comprehender received the message describing the state of the world. The comprehender's task was to select the appropriate medals for the six children depending on the speaker's message. The comprehender only saw the sentence produced by the speaker, but had neither information about the picture for which it was produced, nor about the identity of the speaker. An example trial with the utterance *All children found all of their marbles* and the appropriate response choice is presented in (11). Participants gave their response by choosing a number between 0 and 6 from a dropdown menu for each of the medals. If the sum of the three number did not equal 6, participants had to choose again.

(11) The teacher says: **All children found all of their marbles**

6	0	0

4.2.5. 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. All instructions can be found in the supplementary material. After reading the instructions, participants went through three practice rounds where they were trained in using the reward system, the interface, and the linguistic options they have for describing the pictures with the children's results.

The interface consisted of sentences with two slots. An example is shown in (12).

(12) [ ] found [ ] of their marbles [+] [−]

Participants could type in the two slots. If they wanted to produce more sentences, they could press a + button whereupon a second line of the same format appeared. Participants could open up to five lines of sentences.

The main part of the experiment started with all participants in the role of producers, i.e. a picture was shown to each participant for which they had to produce a description using the interface. This was repeated two times more, so that each participant described 3 pictures. After the end of this production block, all participants switched to the comprehender role. Here a sentence was shown to them (no picture) that had been produced by another participant in the production block, and they had to choose the medals that they considered appropriate. This again was repeated three times whereupon roles changed again. No feedback was given by the system to avoid any biases about interpretation, and no communication between participants outside the system was allowed. All in all, the experiment consisted of 16 blocks of production and comprehension trials. Each block consisted of 3 trials, which in turn consisted of 4 speaker–comprehender pairs.[6] Participants could not know with whom they are paired in the trials. We

[6]Participants were numbered 1 to 4. One trial consisted, for example, of the pairs $(1,4),(2,3),(4,1),(3,2)$. The first number is the producer, the second the comprehender. A picture was associated with each pair. When the system reached the trial for the first time, then the associated picture was shown to the participant in the first

selected 16 worlds which were instantiated by 3 picture items each. Each participant described each world once for each other participant. Hence, participants produced 48 descriptions each (16×3). As the system replaced one participant in the role of producer, it also produced 48 times. The production of sentences was triggered by pictures of worlds. Table 4 in the supplementary material shows which sentence was triggered by which world. Picture of worlds were only used as triggers and were invisible to participants. There is no relevant sentence–meaning relation between trigger worlds and sentences produced by the confederate system.

4.3. Results

As explained before, we gathered valid results for 9 groups of participants (5 groups with 3, and 4 groups with 4 participants). This means that the participants interpreted 1488 sentences (31 participants 48 sentences each). Our aim is to test whether the critical utterances have a success rate comparable to literal descriptions (*literal*) and a higher success rate than average human utterances (*human*), see (10) repeated as (13).

(13) (I) *literal* $\approx$ *critical*, (II) *critical* $>$ *human*.

The success rate is measured as follows. For each trial, experimental results consist of a triple (v,u,w) consisting of a world v in which an utterance u was produced, which in turn was interpreted as world w. The interpretation is successful if $v = w$. If $N(v,u)$ is the number of times that humans produced u in v, and $N(u,w)$ the number of times they interpreted u as w, then the average human production strategy is given by the conditional probabilities $P(u|v) = N(v,u)/\sum_{u'} N(v,u')$ and the average human interpretation strategy by the conditional probabilities $P(w|u) = N(u,w)/\sum_{w'} N(u,w')$. The average human success rate is then given by $\sum_v p(v) \sum_u P(u|v)P(v|u)$. If w_u is the critical interpretation of an utterance u, the success rate of producing u in w_u is given by $P(w_u|u)$.

As expected, the success rates of literal strategies and the human average strategy turned out to be high. For literal utterances (excluding utterances with *rest*, *only*, and *exactly*), the average success rate was 99.6%. For average human utterances, it was $88,2\%$.

Table 4 summarizes the results on critical sentences. It shows the number of times that the different sentences have been interpreted and their success rates with respect to the average human interpretation strategy. Their critical meaning is the sum of their literal meaning and two implicatures: the strengthening of *some* to *some but not all* (*sbna*), and the inference that *others-none*. Table 4 also shows how often participants drew these implicatures.

To test whether the differences between success rates are significant, we calculated one-sided p-values with non–parametric bootstrapping (10000 iterations). Of the 5 critical utterances, the differences between OnlyP-E and Exact2-E to the human average are significant ($p = 0.01$ and $p = 0.03$, respectively). The average success rate of the four utterances with *only* and *exactly* is also significantly greater *human*.[7]

position, and, when the system reached the trial again after switching roles at the end of the block, each participant in the second position had to interpret the sentence produced by the one in the first position.

[7] *4 critical*: $\mu = 94.7\%$, CI $[90,0,98.5]$;*human*: $\mu = 88,2\%$, CI $[86.4,90.1]$, *4 critical* $>$ *human*: $p < 0.01$)

As can be seen in Table 4, participants always strengthened *some* to *some but not all*. This means that the success rate with which a critical utterance was interpreted by its critical meaning is identical to the rate with which the *others-none* implicature occurred. Hence, the results fully support the hypothesis (13) with respect to local strengthening of *some*, and partly for the *others-none* implicature.

sentence	**meaning**	#Int	%Int	E $\rightsquigarrow$ ENA	Rest N
P-E	5 1 0	28	78,6	100	78,6
Exact1-E	5 1 0	28	89,3	100	89,3
Exact2-E	4 2 0	27	96,3	100	96,3
Only1-E	5 1 0	30	93,3	100	93,3
OnlyP-E	5 1 0	34	97,1	100	97,1

Table 4: Results for critical sentences and their implicatures (#Int: number of times utterance has been interpreted; %Int: percentage of interpretations as critical meaning (success rate of utterance); E $\rightsquigarrow$ ENA: local strengthening of *some* to *sbna*; Rest N: the *others-none* implicature.

To set the results of critical sentences into perspective, we provide an overview of alternatives with which their success rates may be compared. Table 5 shows results for sentences with embedded *some* and Rest-E or Rest-A. This means that for these sentences only strengthening of *some* to *sbna* is required. The results show that this implicature is reliably drawn. There is no difference between the sentences containing Rest-E and those containing Rest-A.

not included in confederate strategy				included in confederate strategy			
sentence	**meaning**	#Int	%Int	**sentence**	**meaning**	#Int	%Int
P-E & Rest-N	5 1 0	2	100	Exact1-E & Rest-N	5 1 0	31	96,8
Only1-E & Rest-N	5 1 0	6	100	OnlyP-E & Rest-N	5 1 0	43	97,7
Only1-E & Rest-A	0 1 5	6	100	OnlyP-E & Rest-A	0 1 5	37	100

Table 5: Sentences with *Rest-A* or *Rest-N* (others-none/all)

Table 6 shows results for critical sentences in which *some* has been replaced by *all*. As expected, none of these sentences has a reliable interpretation, although certain interpretations are much more frequent than others.

sentence	#Int	0 5 1	1 4 1	2 3 1	3 2 1	4 1 1	5 0 1
P-A	27	22,7	3,7	14,8	–	–	59,3
Exact1-A	28	28,6	3,6	–	7,1	–	60,7
Only1-A	27	37,9	–	6,9	–	–	55,2
OnlyP-A	30	43,3	3,3	3,3	–	–	50,0
sentence	#Int	0 4 2	1 3 2	2 2 2	3 1 2	4 0 2	
Exact2-A	27	33,3	–	7,4	–	59,3	

Table 6: Results for critical sentences with *some* (E) replaced by *all* (A).

Further results are shown in the supplementary material.[8] We should also mention additional results about Rest-E. *Rest* is an operator which creates an upward entailing context. Hence, the example is comparable to the frequently studied *all–some* sentence. If embedded under *the rest*, participants strengthened *some* to *some but not all* for 99% of all occurrences. Further, the *all–some* sentence (A-E) itself was reliably interpreted as referring to 0|6|0 for 100% of all occurrences (67 out of 67).

5. Theoretical challenge and decomposition analysis

We presented an experiment that compared the NM contexts created by *only Kate*, *only one person*, *exactly one person*, and *exactly two*, see (3) repeated below as (14). The cases *exactly two* and *only-some* have not been tested previously. Our results go beyond previous studies by showing that in all of these examples embedded *some* is reliably strengthened to *some but not all (sbna)*. The inference that the others found nothing can be used by the speaker with a success rate that is at least as high as that of average human utterances.

(14) Only Kate / only/exactly *n* of the girls found some of their marbles.
   a. $\rightsquigarrow$ Kate / *n* found some but not all of their marbles. (x-sbna)
   b. $\rightsquigarrow$ None of the other girls found any marbles. (others-none)

The condition *exactly two* is of particular theoretical interest. Following, Chemla and Spector (2011) the implicatures from *Exactly 1 found some* to *1 found sbna and the others found none* can be derived by globally negating the non–stronger alternative *exactly 1 found all*. However, this explanation does not generalize to *exactly n* with $n > 1$ ("exactly 1 some $\wedge \neg$ exactly 1 all $\Rightarrow$ none all", whereas "exactly 2 some $\wedge \neg$ exactly 2 all" is consistent with some all). Hence, the global derivation would leave it open whether there is a girl that found all marbles. The implicature does also not follow by locally embedding the scalar implicature, as this would make *exactly n found some* equivalent to *exactly n found sbna*, which is weaker than the attested readings. A further challenge is that *only* is Strawson downward entailing as it licences NPIs (e.g., *Only one person lifted a finger*, von Fintel 1999). On the other hand, we do not expect scalar implicatures in DE environments because they lead to weakening (Chierchia, 2013). We summarize the challenges posed by our data below.

(15) **Theoretical challenges**:
   Local mechanism fails to derive others-none reading
   Global mechanism fails to derive others-none reading for exactly 2
   Only is Strawson DE but implicatures in DE environments lead to weakening

To meet these challenges, we propose an analysis that decomposes non-monotonic quantifiers into two components: one positive UE component (the prejacent) and a negative DE component (negated alternatives) (Elliott and Marty 2019, based on Alxatib 2014; Bar-Lev 2018; see also Denić and Sudo 2022 for a decomposed analysis of *exactly* to account for donkey anaphora). In the UE component *Kate found some of her marbles*, the standard *some but not all* implicature is derived. In the DE component, the literal *some* is negated (i.e., alternatives of the form *Mary found some of her marbles*). For *only*, it is standardly assumed that this operator introduces

[8] DOI 10.17605/OSF.IO/WY5S9

two asymmetric meaning components, following Horn 1969. A decomposition of *exactly* can be motivated based on the focus-sensitivity of this operator (Sauerland 2013; Elliott and Marty 2019). (16) derives the attested readings for *exactly*, and (19) provides the derivation for *only*.

(16) Exactly two children found some of the marbles
Decompose: $\text{Exactly}_n\,\Phi \Rightarrow n\,\Phi \wedge \neg\, n+1\,\Phi \wedge \neg\, n+2\,\Phi \ldots$
UE: n some $\rightsquigarrow$n sbna
DE: $\neg n+1$ some $\wedge \neg n+2$ some $\wedge \ldots$
UE: Two children found sbna of the marbles
DE: Not (three children found some of the marbles)
In sum: Exactly 2 children found some but not all of the marbles and the others found none

Only one/two can be treated along similar lines, such that *Only n* is decomposed into:

(17) $\text{Only}_n\,\Phi \Rightarrow n\,\Phi \wedge \neg\, n+1\,\Phi \wedge \neg\, n+2\,\Phi \ldots$

In the general case, there is a set of alternative expressions $\mathscr{A}$ that can replace each other in a certain position of a formula Φ. If a is an element of $\mathscr{A}$, and Φ_a a formula in which a occurs, then it holds:

(18) $(\text{Only}_{\mathscr{A}}\, a)\,\Phi_a$ iff Φ_a and for all $b \in \mathscr{A}$ for which $\Phi_a \not\Rightarrow \Phi_b$ it follows that Φ_b is false.

For *Only Kate*, the alternative set has to consist of all plural objects that can be formed by the children in our scenario.

(19) Only Kate found some of the marbles
UE: Kate some $\rightsquigarrow$ Kate sbna
DE: $\neg$ Mary found some $\wedge \neg$ Sue found some, ...
In sum: Kate found some but not all of the marbles and the others found none

Via the decomposition assumption, the implicatures by all these non-monotonic operators can be derived in a uniform manner. This is consistent with the experimental data indicating that the *sbna* implicature is attested for all these operators. At the same time, the analysis is keeping with the standard assumption that scalar implicatures are dispreferred in DE environments. Since the negative component of *only* and *exactly* is a DE environment, *some* is interpreted as an existential here. That is, the literal semantics of *some* is negated rather than the *all* alternative. This leads to the *others-none* reading.

This decomposition analysis can be integrated in existing theories of embedded implicatures. Here, we sketch two possibilities: how it can be integrated in grammaticalist accounts Chierchia (2004); Chierchia et al. (2012), and how it can be integrated in the framework of elimination rules proposed by Benz and Gotzner (2021).

We begin with Chierchia's recursive version of grammaticalism. Following example (95) of Chierchia (2004: p. 67f), one arrives at the readings (20a) and (20b) for *Exactly 2*:

(20) Exactly two children found some of their marbles.
a. Exactly two children found sbna of their marbles.
b. Exactly two children found some or all of their marbles.

Chierchia assumes that a sentence like (20) is ambiguous between the strong interpretation (20a) and the plain literal interpretation (20b). In our analysis, (20a) is the result of strengthening *some* in the upward entailing component of (20), and *the others–none* part of the literal (20b) is the downward entailing component that blocks strengthening. The critical interpretation can be reached by conjoining (20a) and (20b). In upward and downward entailing environments, conjoining the meaning of alternatives does not produce new meanings: If S and W are two elements of a scale $\mathscr{A}$, and $A(\,.\,)$ an upward or downward monotonic sentence frame, then $A(S) \wedge A(W)$ is equivalent to either either $A(S)$ or $A(W)$. If $A(\,.\,)$ is nonmonotonic, then $A(S) \wedge A(W)$ is generally stronger than either $A(S)$ or $A(W)$. Hence, Chierchia's old recursive version of grammaticalism could explain the existence of the critical reading if it is assumed that a sentence can implicate the conjunctions of all recursively calculated readings.

This, then, also provides us with a strategy for integrating our account in newer versions of grammaticalism that employ invisible, optional only–operators to explain implicatures (Fox, 2007; Chierchia et al., 2012). The reading in (20a) emerges if it is assumed that *some* is in the scope of an invisible only–operator O as in (21). The literal reading (20b) emerges if one assumes that the sentence (20) contains no invisible only–operators.

(21) Exactly two children found O some of their marbles.

As in the case of Chierchia's recursive theory, the critical reading can be derived if one assumes that all possible grammaticalist readings are conjoined. For upward and downward entailing contexts such a conjunction rule would just be equivalent to the *strongest meaning hypothesis* (Dalrymple et al., 1994), which was invoked by proponents of grammatical accounts to resolve ambiguities resulting from different insertions of only–operators (e.g. Chierchia, 2013).

Next, we sketch how our solution can be combined with the proposal made by Benz and Gotzner (2021) (see also Benz 2012; Gotzner and Benz 2018). They provided heuristic principles that explain how speakers can simplify sentences and generate utterances with embedded *some* starting from literal descriptions of a situation. In our experimental scenario, a picture representing world 4|2|0 could be described literally by the conjunction of the following sentences:

(22) Niki found sbna of the marbles,
Toni found sbna of the marbles,
Sasha found none of the marbles,
...

In Benz and Gotzner (2021) two elimination rules were introduced, one that allowed replacement of *sbna* by *some*, and one that allowed elimination of sub-sentences starting with *none of the*. Here, only the first rule is relevant.

We use (18) to define an additional elimination rule for *only*-introduction. Let Φ_a and Ψ be two sub-sentences of (22); assume that $(\text{Only}_{\mathscr{A}}\, a)\Phi_a$ is true and that $b \in \mathscr{A}$ is such that $\Phi_a \not\Rightarrow \Phi_b$. Then:

(23) Ψ can be eliminated if Φ_a entails Ψ or $\neg\Phi_b$ entails Ψ.

Note that this rule eliminates only semantically redundant material.

The sentence *Only 2 found some of their marbles* can be generated from (22) by first replacing *sbna* by *some*, and then apply rule (23) to eliminate all sub–sentences of the form *P found none of the marbles*.[9]

6. Discussion

6.1. Decomposition and asymmetry

The proposed decomposition analysis has its roots in the asymmetric analysis of *only* by Horn (1969). Horn (1996) presents a number of cases, which strongly suggest that the two meaning components introduced by *only* do not have the same status (contra a symmetric analysis, e.g., Atlas (1991)). For example, the negation of (24) in (25) also implies the prejacent that Wilma guess the secret word, which Horn (1996) takes as an argument that the prejacent is presupposed.

(24) Only Wilma guessed the secret word
a. prejacent: Wilma guessed the secret word
b. negated alternatives: Sue did not guess the secret word

(25) Not only [Wilma]F guessed the secret word

Another key argument in favor of the asymmetric analysis is that *only* licenses NPIs. Together with other recent work by Elliott and Marty (2019), we proposed to treat *only* and *exactly* alike. There is evidence that *exactly* licenses the NPI *any* (Alexandropoulou et al., 2020). Another argument in favour of a uniform treatment is that *exactly* and *only* share the property if focus-sensitivity with Sauerland (2013). Interestingly, Horn also briefly mentions that *exactly* can be split into an UE and a DE component: *exactly n — at least n + at most n* . Note, however, that this analysis would actually make the wrong prediction for the cases we investigated. Specifically, it would predict that *Exactly n found some* is interpreted *Exactly n found some and possibly all*.

While different in their implementation and assumptions, the idea of decomposition is in line with recent approaches, suggesting that implicatures can be triggered in the presupposed component of meaning (e.g., (Spector and Sudo, 2017)). While the data we presented here do not allow for an argument about the exact status of the prejacent (e.g. whether the prejacent is presupposed or implicated), they do provide further evidence in favor of an asymmetric analysis of *only* and a corresponding treatment of *only*.

6.2. Role of context and reliability

We showed that a decomposition analysis can be integrated in grammatical theories of implicature and the framework of elimination rules, which does not postulate silent grammatical operators. While our data do not decide between competing theories of implicature, we may

[9]Where the rule has to applied to each sub-sentences with *P found none of the marbles* = Ψ, and Φ_X = *X found some of the marbles*.

note that the grammatical framework does not provide a principled account of how listeners decide between competing parses. For that purpose, the strongest meaning hypothesis has been postulated (see (Gotzner and Benz, 2018) for cases in which this principle does not make the right predictions). In turn, rational choice approaches such as the Rational Speech Act (RSA) model (Potts et al., 2016) or our model of elimination rules (Benz and Gotzner, 2021) already incorporate a concrete account of how listeners reason in a given context. Thus, in this case one integrated account can derive potential implicatures as well as predict which ones are reliably understood in a given context.

Yet it might still be necessary to incorporate information about the grammatical structure of sentences into rational choice approaches. A productive line of recent research by Franke and Bergen (2020) explores the interplay of grammatical structure with contextual prior expec–tations. They provide experimental data in favor of an RSA model that integrates grammatical approaches with Gricean reasoning. This model predicts which of the conventionally associated readings for sentences is most likely to be chosen for a given sentence. It may thus be necessary to integrate components of different frameworks to capture the intricate inference patterns speakers and listeners arrive at in different contexts. Our move towards addressing the question about reliability rather than the existence of an implicature sheds new light on existing debates: If the experimental setup verifies conversational requirements, do the same kinds of inference patterns pertain? We have made our production and comprehension data publicly available via the OSF repository and hope that researchers which find a number of exciting test cases, which may inspire novel debates.

7. Conclusions

We presented an experiment demonstrating that different non-monotonic operators reliably trigger a *some but not all* and an *others none* implicature. Our experimental paradigm verifies Gricean requirements and allowed us to show that these implicatures are part of communicated meaning. Our data pose several challenges to existing accounts of implicature. We presented a decomposition analysis, which splits non-monotonic operators into a positive upward entailing and a negative downward-entailing component. This analysis is in line with standard assump-tions about monotonicity and implicature and explains why scalar implicatures occur more often in non-monotonic than usual downward-entailing contexts such as negation. Finally, our findings shed light on the 'asymmetry wars', where they favour an asymmetric analysis of *only* (e.g. Horn, 1969, 1996 vs. Atlas, 1991, 1993).

8. Supplementary material

Further material, including all experimental data, can be found at DOI 10.17605/OSF.IO/WY5S9

References

Alexandropoulou, S., L. Bylinina, and R. Nouwen (2020). Is there any licensing in non-DE contexts? An experimental study. In M. Franke, N. Kompa, M. Liu, J. L. Mueller, and J. Schwab (Eds.), *Proceedings of Sinn und Bedeutung 24*, Volume 1, pp. 35–47.

Alxatib, S. (2014). Free choice disjunctions under *only*. In J. I. und Leland Kusmer (Ed.), *Proceedings or NELS 44*, pp. 15–28.

Atlas, J. D. (1991). Topic/comment, presupposition, logical form and focus stress implicatures: The case of focal particles *only* and *also*. *Journal of Semantics 8*(1-2), 127–147.

Atlas, J. D. (1993). The importance of being *only*: Testing the neo-Gricean versus neo-entailment paradigms. *Journal of Semantics 10*(4), 301–318.

Bar-Lev, M. (2018). *Free Choice, Homogeneity, and Innocent Inclusion*. Ph. D. thesis, Hebrew University of Jerusalem.

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

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.

Benz, A. and N. Gotzner (2021). Embedded implicature: What can be left unsaid? *Linguistics and Philosophy* (44), 1099–1130.

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. (2013). *Logic in Grammar: Polarity, Free choice, and Intervention*. 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.

Dalrymple, M., M. Kanazawa, S. Mchombo, and S. Peters (1994). What do reciprocals mean? In M. Harvey and L. Santelmann (Eds.), *Proceedings of SALT 4*, pp. 61–78.

Denić, M. and Y. Sudo (2022). Donkey anaphora in non-monotonic contexts. *Journal of Semantics 39*(3), 443–474.

Elliott, P. D. and P. Marty (2019). Exactly one theory of multiplicity inferences. *Snippets 37*, 24–25.

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

Franke, M. and L. Bergen (2020). Theory-driven statistical modeling for semantics and pragmatics: A case study on grammatically generated implicature readings. *Language 96*(2), e77–e96.

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

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

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

Gotzner, N. and J. Romoli (2022). Meaning and alternatives. *Annual Review in Linguistics 8*, 213–234.

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.

Horn, L. R. (1969). A presuppositional analysis of only and even. In R. I. Binnick, A. Davison, G. M. Green, and J. L. Morgan (Eds.), *Chicago Linguistic Society 5*, pp. 97–108.

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.

Horn, L. R. (1996). Exclusive company: *Only* and the dynamics of vertical inference. *Journal of Semantics 13*(1), 1–40.

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.

Magri, G. (2009). A theory of individual-level predicates based on blind mandatory scalar implicatures. *Natural Language Semantics 17*, 245–297.

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. (2013). Presuppositions and the alternative tier. In T. Snider (Ed.), *Proceedings of Semantics and Linguistic Theory SALT 23*, pp. 156–173.

Soames, S. (1982). How presuppositions are inherited: A solution to the projection problem. *Linguistic Inquiry 13*(3), 483–545.

Spector, B. and Y. Sudo (2017). Presupposed ignorance and exhaustification: How scalar implicatures and presuppositions interact. *Linguistics and Philosophy 40*(5), 473–517.

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

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

von Fintel, K. (1999). NPI licensing, Strawson entailment, and context dependency. *Journal of Semantics 16*(2), 97–148.