Exclusive disjunction

Michael Franke
Bob van Tiel

Abstract

If someone says 'Donald ate a pretzel or a donut' the hearer may infer that Donald did not eat both a pretzel and a donut. This exclusive reading of 'or' is often explained as a scalar implicature. We tested this explanation by investigating how the robustness of the exclusive reading of 'or' is influenced by three contextual factors: relevance, competence, and prior probability. We found that only prior probability has a significant effect on the robustness of the exclusive reading, thus disconfirming the scalar implicature account. Instead, we propose that the exclusive reading of 'or' is a probabilistic inference based on world knowledge.

1 Introduction

Introductions to logic usually distinguish between two readings of 'or': an inclusive and an exclusive reading (e.g., McCawley, 1981; Copi & Cohen, 2005). The inclusive reading corresponds to the meaning of logical disjunction. According to this reading, '*A* or *B*' is true if at least one and possibly both of *A* and *B* are true. The exclusive reading is more strict in that it excludes the possibility that both *A* and *B* are true. So on its exclusive reading, '*A* or *B*' is true if exactly one of *A* and *B* is true. To illustrate, consider:

(1) Joe supports Donald or Hillary.

On its inclusive reading, this sentence is true whenever Joe supports Donald, Hillary, or both. The exclusive reading, which is arguably more prominent in this example, excludes the latter possibility. That is, on its exclusive reading, (1) is true whenever Joe supports Donald or Hillary, but not both.

The presence of these two readings might suggest that 'or' is associated with two different lexical entries (e.g., Basson & O'Connor, 1960; Baum, 1996; Rescher,

1964). There are, however, good reasons to reject such a lexicalist approach. Perhaps the most compelling reason is that, in some contexts, 'or' can only receive one interpretation. To illustrate, consider the following sentence, in which 'or' occurs in the scope of negation:

(2) Joe does not support Donald or Hillary.

This sentence has only one interpretation, namely that Joe supports neither Donald nor Hillary (cf. Crain, 2008). This interpretation corresponds to the inclusive reading of 'or'. It does not have an interpretation corresponding to the exclusive reading; in other words, it does not have a reading according to which Joe either supports both Donald and Hillary, or neither of them. Since 'or' is systematically monosemous in certain contexts—negation being one of them—it follows that the two readings of 'or' cannot simply be due to a lexical ambiguity.

An attractive alternative to the lexicalist approach stems from Grice's (1975) theory of conversational implicature. According to Grice, natural conversation is governed by the assumption that speakers are cooperative; that is, they attempt to further the purpose of the discourse by means of their utterances. Speakers are cooperative by adhering to four maxims that enjoin their utterances to be truthful, informative, relevant, and clear. Sometimes hearers have to make ancillary assumptions in order to align a speaker's utterance with the assumption of cooperativity. Such ancillary assumptions are called *conversational implicatures*. An example, from Grice's own work, is given in the following conversation:

(3) A: I am out of petrol.
B: There is a garage around the corner.

B's utterance would be irrelevant, and hence uncooperative, if he knew that the garage was closed or did not sell petrol. Since A assumes that her interlocutor is cooperative, she thus concludes that the garage is open and sells petrol.

Horn (1972) was the first to argue that the exclusive reading of 'or' can be explained as a conversational implicature, too, based on the assumption that the primary meaning of 'or' is inclusive. To illustrate, consider (1) again. Assuming that the primary meaning of 'or' is inclusive, the speaker of (1) could have been more informative, and hence cooperative, by saying 'Joe supports Donald and Hillary.' Why didn't she? Presumably because she does not believe that this alternative is true. This is a weak inference that is compatible with a situation in which the speaker is unsure about whether or not Joe supports both Donald and Hillary. This weak inference can be strengthened if there is reason to believe

that the speaker knows whether or not Joe supports both Donald and Hillary. This assumption is often called the *competence assumption*. If the competence assumption is sufficiently plausible, it follows that, according to the speaker, Joe does not support both Donald and Hillary.

This specific kind of conversational implicature is often called a *scalar implicature* because it is assumed that 'or' forms a lexical scale with 'and', and that alternatives are generated by substituting the scalemates 'or' and 'and'. Other lexical scales are ⟨some, all⟩, ⟨warm, hot⟩, and ⟨intelligent, brilliant⟩ (cf. e.g., van Tiel, van Miltenburg, Zevakhina, & Geurts, 2016).

The implicature account straightforwardly explains the absence of exclusive readings under negation. Consider (2) again. Here, unlike in the case of (1), the alternative with 'and' is less informative than the utterance itself. After all, (2) is only compatible with a situation in which Joe supports neither Donald nor Hillary, whereas the sentence with 'and' is also consistent with situations in which Joe supports exactly one of Donald and Hillary. So the speaker was already maximally informative and hence no implicature is derived.

Although the implicature account has since become the standard in the literature (e.g., Chevallier, Noveck, Nazir, Bott, Lanzetti, & Sperber, 2008; Chierchia, Fox, & Spector, 2012; Fox, 2007; Sauerland, 2004; Geurts, 2010), it is not without its problems, as we will see in the next section. Therefore, we set out to experimentally test the adequacy of the implicature account. To that end, we tested the effect of three factors on the robustness of exclusive readings: (i) the *relevance* of the sentence with 'and' to the hearer, (ii) the *competence* of the speaker about the truth of the sentence with 'and', and (iii) the *prior probability* that the sentence with 'and' is true. These factors are usually taken to influence the robustness of scalar implicatures. Hence, if the implicature account is correct, we expect these factors to influence the robustness of the exclusive reading of 'or', too.

For comparison, we also investigated how relevance, competence, and prior probability influence the robustness of a bona fide scalar implicature, namely the inference from 'some' to 'not all'. To illustrate, consider:

(4) Some of my friends support Donald.

An utterance of this sentence may convey that not all of the speaker's friends support Donald. The derivation of this upper-bounding construal runs analogous to that of the exclusive reading of 'or': the speaker could have been more informative by saying 'All of my friends support Donald'. Why didn't she? Presumably because she does not believe that this alternative is true. If, moreover, the speaker

knows whether or not all of her friends support Donald, it follows that she believes not all of her friends support Donald.

As we will see, it turns out that there are marked differences in how relevance, competence, and prior probability influence the robustness of the exclusive reading of 'or' compared to how they influence the robustness of the inference from 'some' to 'not all'. These findings will force us to reconsider the implicature account of the exclusive reading of 'or'.

In the next section, we explain our three factors of interest in more detail. Afterwards, we outline the details of our experiments and discuss the results.

2 Relevance, competence, and prior probability

Relevance

Almost all theorists agree that the robustness of scalar implicatures is modulated by various extralinguistic factors (e.g., Chierchia et al. 2012; Geurts 2010; Horn 1972; Franke 2009, but see Chierchia 2004; Levinson 2000; Storto & Tanenhaus 2005). Some of these factors are specific to certain theories, while others are shared across competing theories. Relevance is of the latter kind in that it features in almost all theories of scalar inferences. However, relevance is a multivocal notion, and there are at least two construals that should be distinguished: relevance for the discourse purpose and relevance for the hearer's interest (Geurts, 2010).

The notion of relevance for the discourse purpose can be illustrated with the following example (from van Kuppevelt 1996):

(5) A: How many of the boys were at the party?
B: Some of the boys were at the party.

Discourse purposes are usually equated to questions, which can be implicit or, as in this case, explicit (Roberts, 2012). A piece of information is relevant if it contributes to answering this question. In (5), the information that not all of the boys went to the party is relevant to the discourse purpose because it narrows down the space of possible answers. Compare this to a situation in which B's answer addresses the question 'Were some of the boys at the party?' In this situation, the upper-bounding construal is irrelevant, since the question is already resolved when 'some' is interpreted as 'at least some and possibly all'. Hence, it is hypothesised, no scalar implicature will be derived in this situation.

In this paper, however, we will mainly be concerned with the second notion of relevance: relevance to the hearer's interest. This construal is intimately connected to *relevance theory* (Sperber & Wilson, 1995).

According to relevance theory, the relevance of an inference is determined by two factors: (i) the positive cognitive effects it has on the hearer—i.e., to what extent it makes "a worthwhile difference to the individual's representation of the world" (Wilson & Sperber, 2002, p. 251)—and (ii) the effort needed to process that inference. In what follows, we focus on the first factor, assuming that processing effort remains invariant across the scenarios that we tested (cf. Chevallier et al. 2008 for a series of experiments on 'or' that manipulate processing effort).

The central tenet of relevance theory is that communication is aimed at maximising relevance, so that hearers only derive inferences whose positive cognitive effects outweigh their processing cost. Hence, relevance theory predicts that the robustness of a scalar implicature is an increasing function of its importance to the hearer. To illustrate, consider:

(6) Donald won the primaries in Nebraska or Indiana.

According to relevance theory, the robustness of the exclusive reading of this sentence depends on how important it is to the hearer whether Donald won the primaries in both Nebraska and Indiana. Suppose, for example, that the hearer made a large bet that Donald would not win both primaries. In that case, she would be more likely to arrive at an exclusive reading than if, for example, she made a large bet that Donald would win at least one of the primaries.

Hence, if relevance theory is correct, we expect the robustness of the inference from 'some' to 'not all' to be positively affected by the relevance of the upper-bounded reading to the hearer. If, in addition, the implicature account is correct, we expect the same effect on the exclusive reading of 'or'.

Competence

In the previous section, we have seen that competence plays an important role in the derivation of scalar implicatures. Initially, pragmatic reasoning yields a weak inference according to which the speaker does not believe that the more informative alternative is true. This weak inference can be strengthened if the competence assumption holds; that is, if it is assumed that the speaker knows whether or not the more informative statement is true. If so, it follows that, according to the speaker, the more informative statement is false.

Although the workings of the competence assumption tend to be unproblematic for most scalar expressions, they are rather precarious in the case of 'or'. Indeed, the need to invoke competence to arrive at an exclusive reading of 'or' gives rise to what Zondervan (2010) calls the *speaker expertise paradox* (cf. Geurts, 2006). To explain this paradox, consider (1) once again:

(7) Joe supports Donald or Hillary.

We have already seen that someone who utters this sentence implies that Joe does not support both Donald and Hillary. Presumably, however, the utterance also implies that the speaker does not know whether Joe supports Donald and that she does not know whether Joe supports Hillary. After all, if the speaker knew that Joe supports Donald, she should have said 'Joe supports Donald', and mutatis mutandis if she knew that Joe supports Hillary. However, in order to arrive at the exclusive reading through pragmatic reasoning, it has to be assumed that the speaker knows whether Joe supports both Donald and Hillary.

More abstractly, in order to derive the exclusive reading of a disjunction '*A* or *B*', it has to be assumed that the speaker is ignorant about the truth of *A* and *B* individually, but knowledgeable about the truth of '*A* and *B*'. Such an epistemic state is possible but intuitively improbable, which contradicts the observation that exclusive interpretations are far from uncommon.

In order to arrive at a more decisive verdict, however, we tested the effect of competence on the robustness of the exclusive reading of 'or' and the upper-bounded construal of 'some'. Previous experimental research has shown that the robustness of the upper-bounded reading of 'some' increases with the competence of the speaker (Goodman & Stuhlmüller, 2013). We expect to replicate this effect in our study. If, moreover, the exclusive reading is due to a scalar implicature, we expect a similar effect in the case of 'or'.

Prior probability

Compared to the previous two factors, the role of prior probabilities is more controversial. Some theorists have marginalised the importance of prior probabilities (Geurts, 2010), while others have assigned them an important role in the derivation of scalar implicatures (e.g., Russell, 2012; Frank & Goodman, 2012; Franke & Jäger, 2016). According to at least some these accounts, the robustness of a scalar implicature is an increasing function of the prior probability that it is true. To illustrate, consider the following sentence once again:

(8) Donald won the primaries in Nebraska or Indiana.

Theories that are sensitive to prior probabilities predict that, if it is deemed extremely unlikely that Donald wins the primaries in both states, someone who hears an utterance of (8) will be more likely to arrive at an exclusive reading of 'or', compared to when the hearer is expecting Donald to win both primaries.

Interestingly, a number of theorists have argued that the robustness of exclusive readings is exhaustively determined by prior probabilities (e.g., Rubin & Young, 1989; Yanal, 1988). These authors hold that 'or' is monosemous and always interpreted inclusively. The appearance of an exclusive reading is caused by probabilistic reasoning about possible states of the world. An especially clear example that illustrates this point is:

(9) Joe voted for Donald or Hillary.

Here, it seems that the speaker's utterance rules out the possibility that Joe voted for both Donald and Hillary. According to the aforementioned theorists, this inference is not due to an exclusive reading of 'or' but rather to an inclusive reading along with the commonsense information that one can only vote once in a democratic election. In this way, then, world knowledge is responsible for seemingly exclusive interpretations of 'or'. Importantly, this position is distinct from the implicature account because it holds that the derivation of exclusive readings does not involve any reasoning about the intentions of the speaker.

Previous experimental research has confirmed that the robustness of the upper-bounded reading of 'some' is positively affected by the prior probability that it is true (Degen, Tessler, & Goodman, 2015). We expect to replicate that result in our study. In addition, if exclusive readings are due to scalar implicatures, we expect a similar effect in the case of 'or'.

Predictions

In summary, we have discussed three factors that have variously been taken to increase the robustness of scalar implicatures: (i) the relevance of the more informative statement to the hearer, (ii) the competence of the speaker about the truth of the more informative statement, and (iii) the prior probability that the more informative statement is false.

In the next section, we discuss two experiments in which we measured the effect of relevance, competence, and prior probability on the robustness of, on the one hand, the upper-bounded construal of 'some' (Exp. 1) and, on the other

hand, the exclusive reading of 'or' (Exp. 2). We hypothesise that all three factors influence the robustness of the upper-bounded construal of 'some'. More concretely, we hypothesise that the upper-bounding construal of 'some' should be more prominent under higher contextual relevance, higher speaker competence, and lower a priori probability of the corresponding sentence with 'all'. If the implicature account is correct, we expect that the exclusive reading of 'or' patterns with the upper-bounded construal of 'some' in these respects.

Since Exp. 1 and 2 share the same methodology, we discuss them in tandem. Afterwards, in Exp. 3, we discuss and test an alternative version of the implicature account, according to which statements of the form '*A* or *B*' are read, in effect, as '*A* and nothing else or *B* and nothing else'.

3 Experiments 1 and 2

Design

The goal of Exps. 1 and 2 was to determine the effect of relevance, competence, and prior probability on the strength of the upper-bounded construal of 'some' (Exp. 1) and the exclusive reading of 'or' (Exp. 2). To measure these three factors, we designed different stories that systematically varied along the three relevant dimensions. We used a slider-rating task to assess participants' intuitive judgement of all four notions of interest: strength of the inference, relevance, competence, and prior probability.

Intuitive judgements for the four notions of interests were measured between-participants: each participant only answered one of the four relevant test questions for a given story. This was to prevent cross-contamination of answers, e.g., asking first about relevance might influence subsequent answers about the strength of the inference.

Participants

203 (Exp. 1) and 200 (Exp. 2) participants were drafted on Amazon's Mechanical Turk and paid 80 US$ cent.[1] Payment was not contingent on any of their

[1]Mechanical Turk is a website where workers perform so-called 'Human Intelligence Tasks' (HITs) for financial compensation. It has been shown that the quality of data gathered through Mechanical Turk equals that of laboratory data (e.g., Buhrmester, Kwang & Gosling 2011, Schnoebelen & Kuperman 2010; Sprouse 2011).

responses. Only workers with an IP address from the United States and with a rate of accepted HITs of at least 90% were eligible for participation.

Materials

The materials in both Exp. 1 and 2 consisted of 16 vignettes (see Appendix A for the full material). Each story came with some background information and an utterance of a statement containing 'some' (Exp. 1) or 'or' (Exp. 2) by one of the characters in the story. For example:

> **Example story from Exp. 1**
> Lucy has to give a talk in front of a big audience of psychologists. She is going to criticize one of the dominant theories of schizophrenia. Afterwards, Jacob, who was in the audience, chatted with his neighbors.
>
> He tells Lucy: 'Some of the people enjoyed your talk.'
>
> **Example story from Exp. 2**
> Danny and Alex reserved a squash court but Alex still has to buy a racket and a pair of shoes. Danny is talking to Alex's girlfriend Jill who just went to the sports store with him.
>
> Jill says to Danny: 'Alex bought a racket or a pair of shoes.'

Each story was associated with three control statements which were either certainly true, certainly false, or of uncertain truth value, given the background information. Moreover, each story was associated with four target statements gauging (i) the strength of inference, (ii) the relevance of the truth of the more informative statement for the hearer, (iii) the competence of the speaker, and (iv) the prior probability of the stronger statement. The four target statements associated with the examples above were:

> **Example target statements from Exp. 1**
>
> *Inference*: From what Jacob said we may conclude that not all of the people enjoyed Lucy's talk.
>
> *Relevance*: It is important for Lucy to know whether all of the people enjoyed her talk.
>
> *Competence*: Jacob knows whether all of the people enjoyed Lucy's talk.
>
> *Prior*: All of the people enjoyed Lucy's talk.

Example target statements from Exp. 2

Inference: From what Alex's girlfriend said we may conclude that Alex did not buy both a racket and a pair of shoes.

Relevance: It is important for Danny to know whether Alex bought both a racket and a pair of shoes.

Competence: Alex's girlfriend knows whether he bought both a racket and a pair of shoes.

Prior: If Alex bought a racket, it is likely that he also bought a pair of shoes. / If Alex bought a pair of shoes, it is likely that he also bought a racket.

Statements *Inference*, *Relevance*, *Competence* (Exps. 1 and 2), and *Prior* (Exp. 1) were single statements. Statements *Prior* (Exp. 2) were pairs of symmetric conditional statements, where each targeted the intuitive probability that, given one disjunct, the other would be true as well. We reasoned that this makes for more natural statements and that it may give us more reliable measures than having participants rate a single statement containing 'and'.[2]

The stories were created to ensure sufficient variability across the three dimensions of interest (i.e., relevance, competence, and prior probability). We classified each story according to whether we felt it to be high or low on each dimension, thus making for eight types of stories. In both experiments, we had two vignettes for each type. For example, we expected the example story from Exp. 1 to score high on relevance, but low on competence and prior probability, and the example story from Exp. 2 to score high on all three dimensions. A full list of the 32 stories, together with our intuitive type-classification, can be found in Appendix A. [mf: insert appendices with materials]

[2]In order to ascertain that nothing hinges on this decision, we conducted a follow-up experiment in which we paid 70 participants on Amazon's Mechanical Turk 50 US$ cent to judge the stories from Exp. 2 followed by a control statement and the following target statement:

(i) It is likely that Alex bought both a racket and a pair of shoes.

Each participant saw six randomly sampled stories, which were presented without the statement containing 'or', just as in Exp. 2.

We excluded one participant for not identifying as a native speaker of English and another one for bad performance on the control questions, using the same criterion as for Exp. 2. Per-vignette means of the given ratings for both types of statements gauging prior probability were highly correlated ($r \approx 0.89, p < 0.0001$). For all of the analyses reported in the main text, nothing of substance changes if we include the ratings for (i) instead.

Procedure

The experiment started with instructions:

> In the following, you will be presented with 8 short background stories. Please read them very carefully. We ask you to rate 2 or 3 statements for each background story. Please indicate, using an adjustable slider, how likely you think a statement is true based on the background story.

Next, we presented a simple background story which was not used in the main experiment, followed by three annotated examples to illustrate the use of the slider bar. One example was clearly true, another clearly false, and the last uncertain.

In the main part of the experiment, every participant saw eight randomly sampled stories, one of each story type, in random order. Each story was followed first by one random control statement and then the statement(s) associated with one of the four factors of interest (inference strength, relevance, competence, or prior probability). Each participant rated each of the four statement types exactly twice, but never in direct succession. When the prior statements were presented, only the background story was provided, but not the statement with 'or', so as to make sure that answers are based on expectations about worldly events alone, unmodulated by information based on pragmatic inferences from utterances. All other question types had the background story and the disjunctive statement with 'or' appear on the screen. The two prior statements in Exp. 2 were presented individually, one after the other, in random order.

Ratings of statements were elicited by asking "How likely do you think it is that the statement is true, given the information in the background story?" together with a continuous slider ranging from "certainly false" to "certainly true." An example of a trial from Exp. 2 is given in Figure 1.

3.1 Data preparation

We coded the slider-ratings as real numbers ranging from 0 ("certainly false") to 1 ("certainly true"). Ratings for the two conditional statements in the prior condition of Exp. 2 were averaged.

Seven participants were excluded from the analysis because they were not self-reported native speakers of English. We also removed another eleven participants for obviously deviant answers (e.g., blindly alternating between maximal

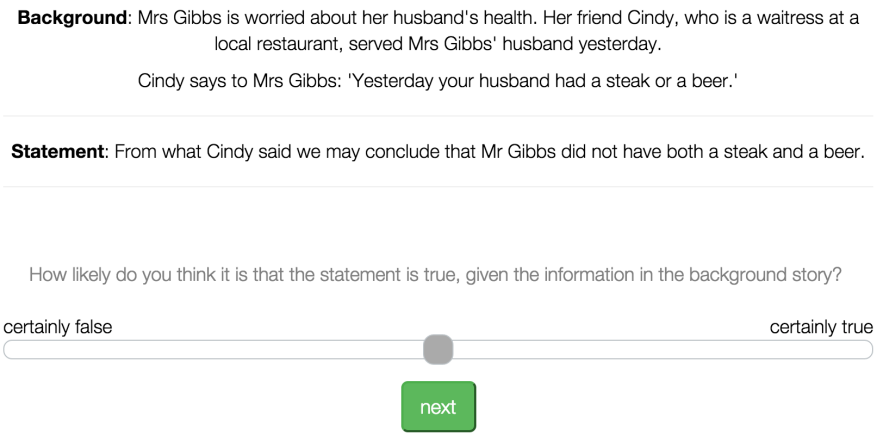

Figure 1: Example of a trial [mf: insert pic with example from main text]

agreement and maximal disagreement).[3] Consequently, data from a total of 193 (Exp. 1) and 192 (Exp. 2) participants made it into our analysis. Unfortunately, there was a mistake in the formulation of two stories, one for each experiment. We removed all data for these stories for the analysis. (See Appendix A.)

In what follows, we discuss the results of Exps. 1 and 2 in turn, starting with Exp. 1, which tested the upper-bounded construal of 'some'.

Experiment 1: Results

Controls. Ratings for control statements are unsurprising. Means, averaged over all vignettes, for ratings of false (0.22), uncertain (0.39) and true statements (0.80) are pairwise different (two-population directed t-test: $t \approx -3.83, p < 0.001$ for false vs. uncertain; $t \approx -9.98, p < 0.001$ for uncertain vs. true).

Explanatory factors. Statements that targeted relevance, competence and prior probability were rated in accordance with our intuitive classification (see Figure 2, all high/low constrasts are significantly different).

[3]The formal criterion for exclusion was having a *deviance score* greater than a fixed threshold, where the deviance score of a participant is the sum of the absolute differences between the expected answers for all control questions (0, 0.5, or 1) and the subjects' answers. We set the threshold of exclusion to the mean plus twice the standard deviation. This exclusion criterion was also used in all other experiments.

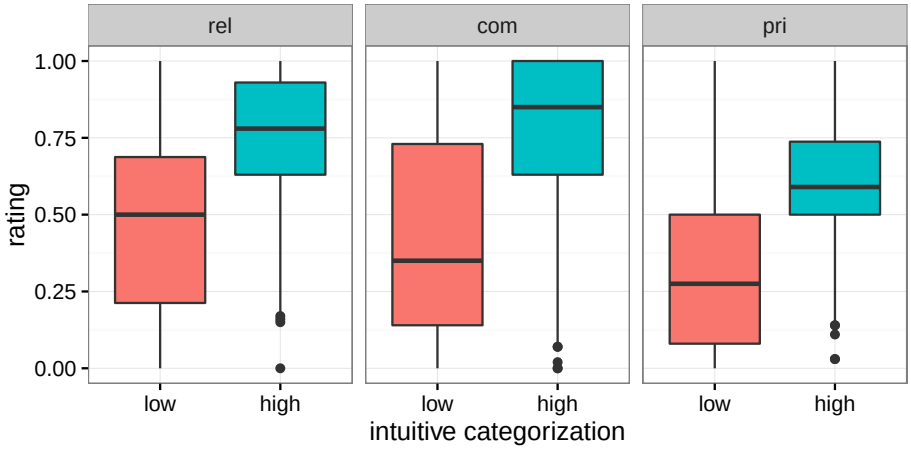

Figure 2: Ratings of statements according to intuitive pre-classification in Experiment 1

Figure 3 shows the relation of per-vignette mean implicature ratings and per-vignette mean ratings for the three explanatory factors. From visual inspection, it seems that REL and COM are unlikely good predictors of implicature strength, while low values of PRI seem to be correlated with high implicature ratings, as expected.

Main analysis. We want to explain ratings of the *Inference*-statement in terms of explanatory factors REL, PRI, and COM, which are the respective means of ratings of the corresponding statements for each vignette. Figure 4 gives the Bayes factors of regression models with our three explanatory variables as main factors. The best model only contains factor PRI and is made roughly six times more likely by the data than the two runner-ups which contain additionally REL or COM. Clearly, the data provides very strong evidence in favor of all models that include PRI, relative to those which do not.

Controls in Exp. 2. Control statements were rated as expected, indicating that participants understood the task in general and paid attention to the background stories. Means, averaged over all vignettes, for ratings of false (0.26), uncertain (0.48) and true statements (0.81) are pairwise different (two-population directed t-test: $t \approx -3.72, p < 0.001$ for false vs. uncertain; $t \approx -6.93, p < 0.001$ for uncertain vs. true).

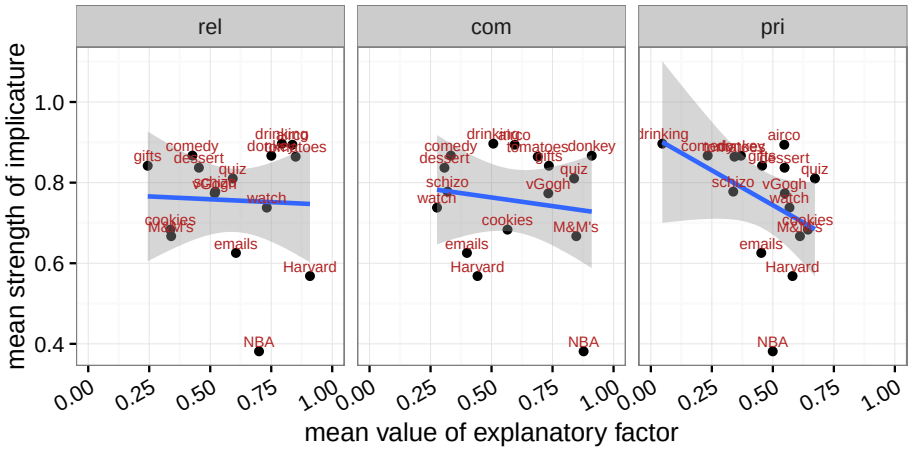

Figure 3: Per-vignette means of ratings of relevance, competence and prior statements vs. per-vignette means of implicature rating in Experiment 3

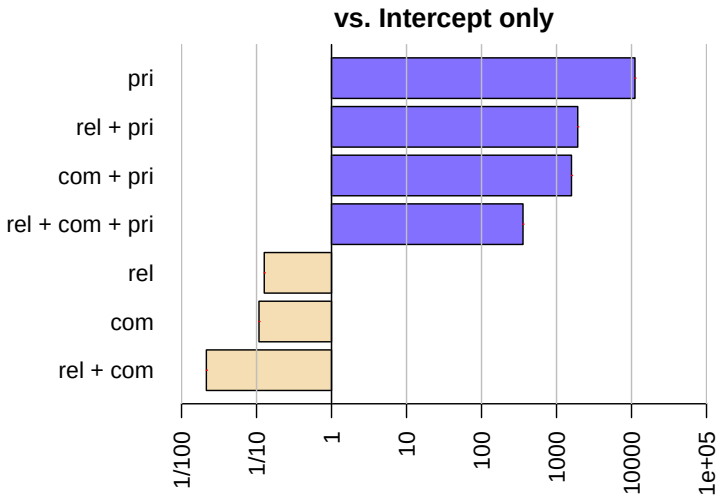

Figure 4: Bayes factor comparison of different main factor combinations, predicting the strength of scalar enrichment of *some* in Experiment 1.

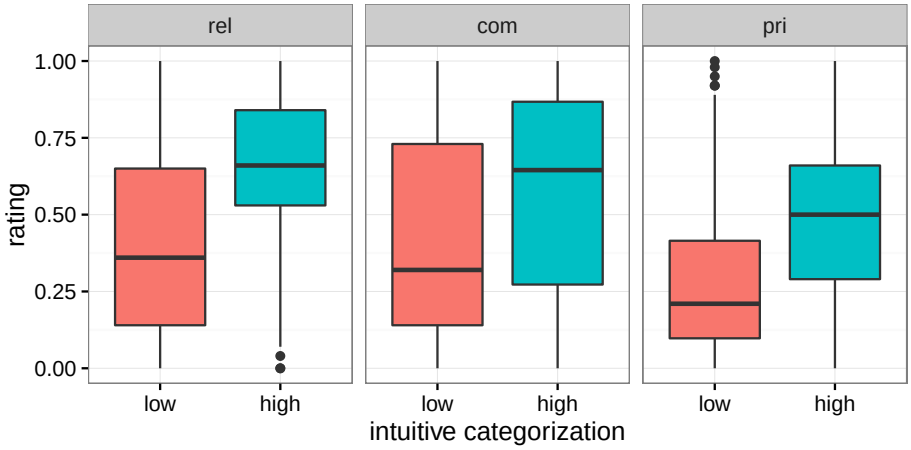

Figure 5: Ratings of statements according to intuitive pre-classification in Experiment 1

Explanatory factors. Ratings of relevant explanatory factors are not uniformly distributed across vignettes, but validate our intuitive pre-classification (see Figure 5, all high/low contrasts are significant).

Figure 6 shows the relation of per-vignette mean implicature ratings and per-vignette mean ratings for the three explanatory factors. From visual inspection, it seems that relevance and competence are not good predictors of implicature strength, while low prior plausibility seems to be correlated with high implicature ratings, as expected.

Main analysis. To check whether factors "relevance," "prior" and "competence" have an influence on the strength of exclusive readings, we compare regression models of different complexity. The dependent variable are ratings of the *Xor*-statement. Explanatory factors REL, COM, and PRI are, respectively, the means of the ratings, for each vignette, of the *Relevance*, *Competence* and *Prior* statements. We take a Bayesian approach to comparing regression models in terms of their Bayes factors (Rouder & Morey, 2012), as implemented in the *BayesFactor R*-package. This gives us a more nuanced picture of the relative evidence for models of different complexity, including information about how much, e.g., the absence of a factor in a model, is supported by our data.[4]

[4]All conclusions of theoretical relevance are also supported by more traditional, frequentist regression analyses in terms of significance of factors and model comparison by AIC.

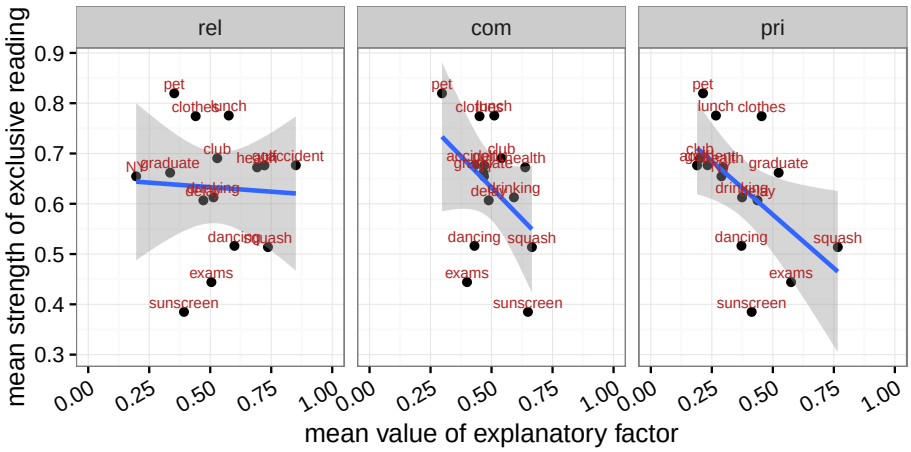

Figure 6: Per-vignette means of ratings of relevance, competence and prior statements vs. per-vignette means of implicature rating in Experiment 1

Figure 7 shows the Bayes factors of all regression models that can be built with our three explanatory variables as main factors. The graph gives the Bayes factor of each regression model, listed on the right, against the intercept-only model. A model with only REL as a main factor, for example, is roughly 8 times worse than the intercept-only model, suggesting that REL alone makes no useful contribution to harnessing the variance in *Xor*-ratings, but only makes for a more complex model. Single main factor COM does make a significant contribution, compared to the intercept-only model, but a model with single main factor PRI is more than 30 times more likely, given our data, than the model with just COM. The best model, by this standard, is a model with COM and PRI as main effects, but there is no substantial difference between this and the model with only PRI as a main factor.

The main conclusion to be drawn from this analysis is that REL is a bad, COM an unnecessary, and PRI the best predictor of strength of exclusive readings.[5] Factor REL should be omitted for reasons of parsimony (every model with it is worse than the corresponding one with it), while COM can be omitted at no substantial loss (adding COM makes models better, but not substantially so, when PRI is present). Omitting PRI leads to a substantial decline in explanatory power.

Estimates of the posterior distributions over model parameter coefficients for

[5]This general conclusion is also vindicated by more complex analyses that would take interactions and random effects for participants into account.

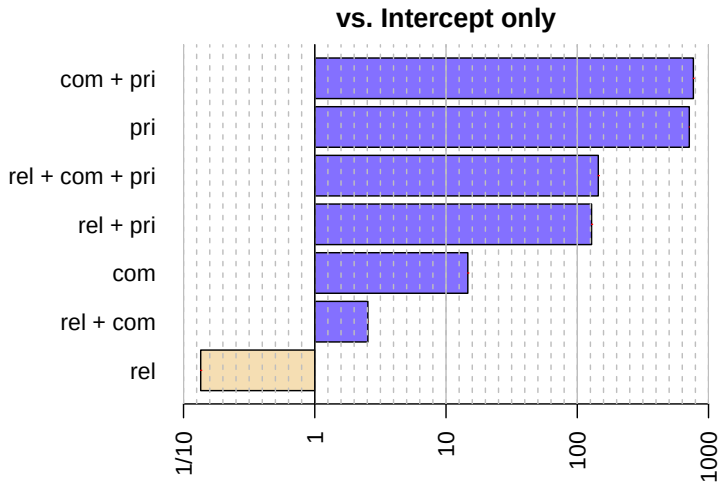

Figure 7: Bayes factor comparison of different main factor combinations. Notation like "com + pri" stands for a regression model with main factors COM and PRI.

the linear model that contains all three factors REL, COM and PRI are shown in Figure 8. Noteworthily, most credible values for coefficients for COM are negative. This is the same for all other models containing factor COM. This means that our data suggests that the more competent the speaker was felt to be, the lower the strength of the exclusive reading. This is the reverse of what we would expect from basically all pragmatic theories. In contrast, the impact of PRI is as expected: the more likely the conjunctive alternative, the less strong the exclusive reading is felt to be.

Discussion.

Prior plausibility of the conjunctive alternative seems to be the main explanatory factor of *Xor*-ratings. This is interesting since standard theories usually do not emphasize the role of prior plausibility. Moreover, it is actually surprising from the point of view of standard pragmatic theories of exclusive disjunction readings that relevance does not seem to play an explanatory role and that competence is correlated with *Xor*-strength in the "wrong direction," so to speak. [mf: more here? or rather later?]

Before drawing firm conclusions, we should address some potential worries

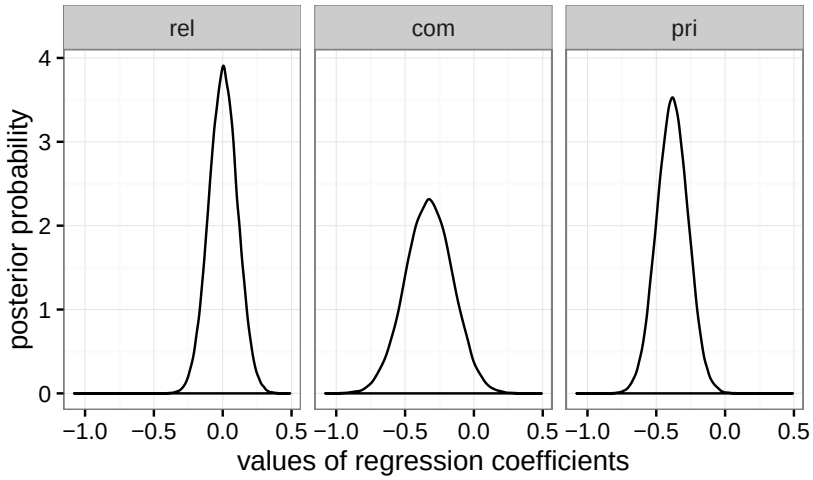

Figure 8: Density estimates of posterior over model parameter coefficients for a linear model with all three main factors.

about this design and the evidence that our results provide for or against theoretical positions. First of all, it could be objected that the way in which we measured factor PRI is inadequate. What matters, so a possible objection goes, is the prior plausibility of "*A* and *B*" not the mean of the plausibility of conditionals "if *A*, *B*" and "if *B*, *A*." Experiment 2 presents a follow-up that addresses this issue. Secondly, we should verify that our experimental measures of relevance, competence and prior do what we would like them to. In order to address this issue, Experiment 3 looks at scalar quantifier *some* in a parallel design to that of Experiment 1.

Results

Data preparation. Four participants were excluded from the analysis because they were not self-reported native English speakers. Another 6 participants were excluded for poor performance, by the same criterion as used with Experiment 1. All data from the remaining 193 participants entered into the analysis.

3.2 Discussion

We could conclude from this analysis that PRI is the key factor in our regression model comparison for predicting the strength of scalar inferences.[6] Factors REL and COM do not seem to carry extra explanatory power. This would suggest that the behavior of scalar *some* parallels that of disjunction *or* in terms of which factors seem to influence the strength of the putative implicatures.

There is, however, a particular oddity in our data. A look back at Figure 3 reveals that one vignette received a surprisingly low mean score for implicature strength, namely *NBA*. When we compare the ratings of the *notAll*-statement given for the *NBA* vignette with those given for each other vignette, we see that all of these fifteen pairwise comparisons shows a significant difference. No other vignette had that property in Experiment 3, and also no vignette from Experiment 1 was an "outlier" in this sense. Low implicature ratings for this vignette are particularly surprising, because it was intuitively classified as high relevance, high competence, and low prior. So, all explanatory factors should, by the standard theory, point towards high implicature rates. The observed ratings of these factors for this vignette accorded with intuition. Moreover, the NBA story had remarkably many participants answering that the likelihood of the implicature was exactly zero. Eight participants provided such a response; none did so for the next lowest-scoring item. Clearly, this case seems to stand out in some way.

Here is what we believe went wrong with this item. Consider the *notAll*-statement of this vignette:

> *notAll*
> From what Jason Barley said we may conclude that Greg Jones did not secure victory for his team during the last seconds of all of the decisive playoff matches.

Rather than directly modifying the noun phrase, the negation modifies the verb phrase and is separated by three constituents from the noun phrase. This may invite a reading, which was not intended, in which the negation modifies the verb rather than the noun phrase. In other words, it invites a reading of the complement of 'said' that can be paraphrased as 'Greg Jones failed to secure victory for his team during the last seconds of all of the decisive playoff matches.' We suspect

[6] This is also the case for more complex analyses that take interactions and random effects for participants into account: the model with only PRI as factor is the best, and every model that contains it is strongly favored by the data above any model that does not.

that participants arrived at this reading because of the amount of material between the negation and the noun phrase, which invited participants to instead have the negation modify the verb. Indeed, we found that ratings of exactly zero were overall much more frequent in cases of VP-negation than in cases of NP-negation. [mf: I don't understand the last sentence. Where did we find that? Should we really mention this? Should we go into the NP- vs. VP-negation thing in more detail? Otherwise, maybe drop this?]

In order to test our hypothesis that participants arrived at an unintended reading of the target statement, we conducted a small follow-up experiment in which we gathered implicature ratings for a statement that better expressed the intended reading than the one used in Experiment 3:

> From what Jason Barley said we may conclude that not all of the decisive playoff matches were secured during the last seconds by Greg Jones.

The follow-up consisted of the NBA story followed by three statements: two control statements and one target statement. The target statement was varied between the one used in Experiment 3 (see above) and the alternative one with NP-negation, and was varied between participants. 40 participants were drafted on Mechanical Turk and were paid $0.15 for their participation. They were instructed to, first, read the story carefully and, afterwards, indicate the likelihood of the corresponding statements on a seven-point Likert scale. We hypothesised that implicature ratings would be substantially higher for the modified statement than for the original one.

The normalised implicature ratings for the original statement were slightly lower than in Experiment 3 (0.25 versus 0.39). Crucially, however, the normalised implicature ratings for the modified statement were significantly higher and much more in line with what we observed for the other items (0.76, $t(29) = 4.28$, $p < .001$). Since the two statements were synonymous on their intended readings, we consider this compelling evidence that participants in Experiment 3 arrived at an unintended reading of the target sentence and sufficient reason for discarding the item from our analysis.

Consequently, we reran the regression model comparison after excluding all data from the *NBA* vignette. The results are shown in Figure 9. The best model considers only factors COM and PRI. It is more than 6 times likelier, given the data, than the second best model, which also includes REL, which in turn is about 3.5 times likelier than the third model with single factor PRI. Figure 10 shows

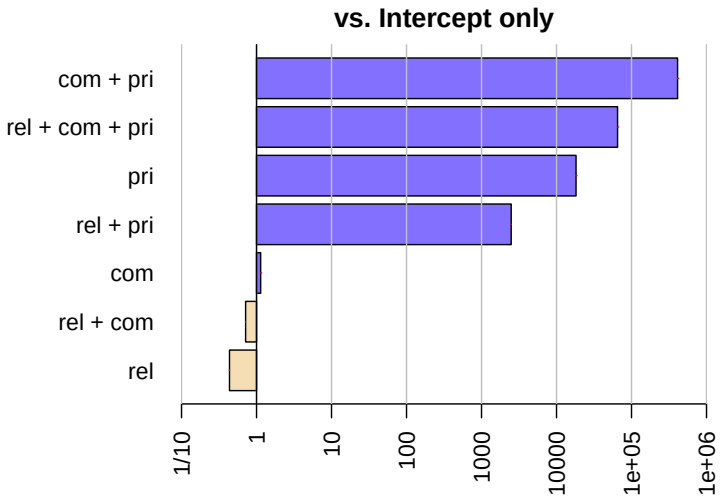

Figure 9: Bayes factor comparison of different main factor combinations, predicting the strength of scalar enrichment of *some* in Experiment 3 after excluding data from the "NBA" scenario.

posteriors over regression coefficients for that model with main effects REL, COM and PRI. Unlike for the disjunction case in Experiment 1, the effect of factor COM on strength of scalar enrichment is as expected from standard theory: the more competent a speaker is felt to be, the stronger the scalar implicature reading.

In sum, we believe that there are good reasons to exclude the *NBA* vignette from our analysis, because of an unintended ambiguity in the *notAll*-statement. Doing so, reveals that factors PRI and COM contribute most to explaining the variance in implicature strength. Just as for disjunction, relevance seems to be a superflous factor, because any model without factor REL is worse than the corresponding one where it is added. This suggests that the speaker expertise paradox [mf: terminology?] may be a real problem. While manipulations of competence do have the effects predicted by standard theories of scalar implicature for the case of *some*, this is not the case for disjunctive readings of *or*. There does seem to be a difference, which is, as such, already unexpected under standard conceptions.

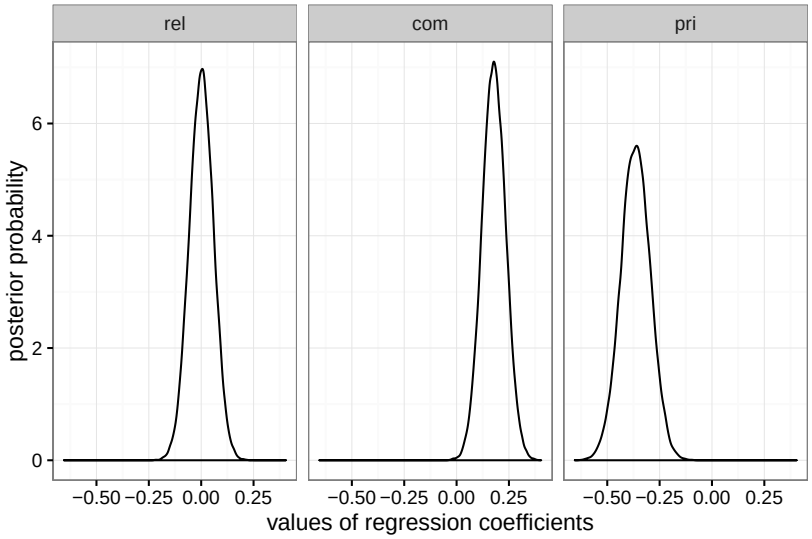

Figure 10: [mf: fill me]

4 Experiment 4

4.1 Design

Exclusive readings of disjunctions can also come about by exhaustifying individual disjuncts (see Section XYZ). This approach would predict that the strength of an exclusive disjunction reading of "*A* or *B*" should be positively correlated with the strength of exhaustive readings that statements of individual disjuncts "*A*" and "*B*" would receive in the same context. The purpose of Experiment 4 was therefore to collect data on the strength of exhaustive readings of such single-disjunct statements in the background contexts used in Experiment 1. We would then like to investigate whether strengths of exhaustive readings make for a reliable predictor of strength of exclusive readings across contexts.

4.2 Participants

Using the same selection criteria as before, 131 subjects were recruited via Amazon's Mechanical Turk and paid US$ 0.50 for participation.

4.3 Materials

Experiment 4 used the fifteen vignettes from Experiment 2 (that is excluding the erroneous "Bill's orders" scenario). For each vignette we consider the speaker's utterance of single disjuncts (see Appendix A). Concretely, where Experiment 1 had an utterance of a disjunction:

> *Utterance of disjunction*
> Jill says to Danny: 'Alex bought a racket or a pair of shoes.'

Experiment 2 had two single-disjunct utterances by the same speaker:

> *Utterance of disjunct 1*
> Jill says to Danny: 'Alex bought a racket.'
>
> *Utterance of disjunct 2*
> Jill says to Danny: 'Alex bought a pair of shoes.'

Additionally, each vignette also had corresponding statements that subjects had to rate:

> *Exh1*
> From what Alex's girlfriend said we may conclude that Alex did not buy a pair of shoes as well.
>
> *Exh2*
> From what Alex's girlfriend said we may conclude that Alex did not buy a racket as well.

4.4 Procedure

The procedure followed that of Experiment 2 very closely. After reading (slightly amended) instructions and seeing examples for the use of the slider bar, each participant was presented with six randomly sampled vignettes. Subjects read the background story, followed with an utterance of disjunct 1 or 2, randomly chosen. Subjects first rated a random control question and then rated the *Exh1* or *Exh2* statement, depending on which utterance was shown to them.

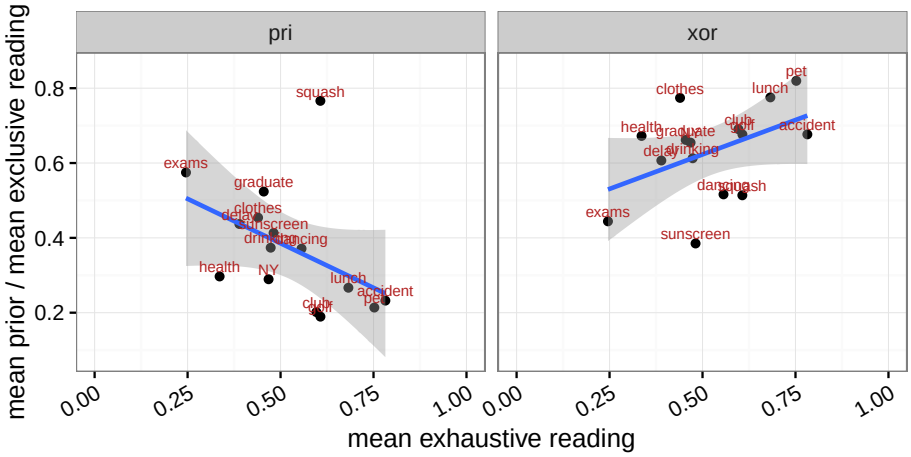

Figure 11: Means of ratings of *Exh*-statements from Experiment 4 (x-axis) vs. means of ratings of *Prior*-statements and *Xor*-statements from Experiment 1 (y-axis)

4.5 Results

Data from one subject was discarded because English was not the self-reported native language. Another four subjects were removed for bad performance on the control questions, using the same criterion as before.

Figure 11 shows the per-vignette means of the ratings of the *Exh*-statements plotted against the corresponding mean ratings of the *Prior*- and *Xor*-statements from Experiment 1. There is no significant correlation between *Prior*-ratings and *Exh*-ratings ($r \approx -0.44$, $p \approx 0.1$), suggesting that our measures of prior expectations and exhaustive strength do not coincide. Adding the per-vignette mean *Exh*-ratings as an additional explanatory factor EXH to the regression model comparison, we obtain the picture given in Figure 12. A model using single factor PRI to predict *Xor*-ratings is about 8.5 times more likely than a model using single factor EXH. That means that our data provides evidence for the assumption that prior expectations are a better explanatory factor of exclusive readings than the strength of exhaustive readings.

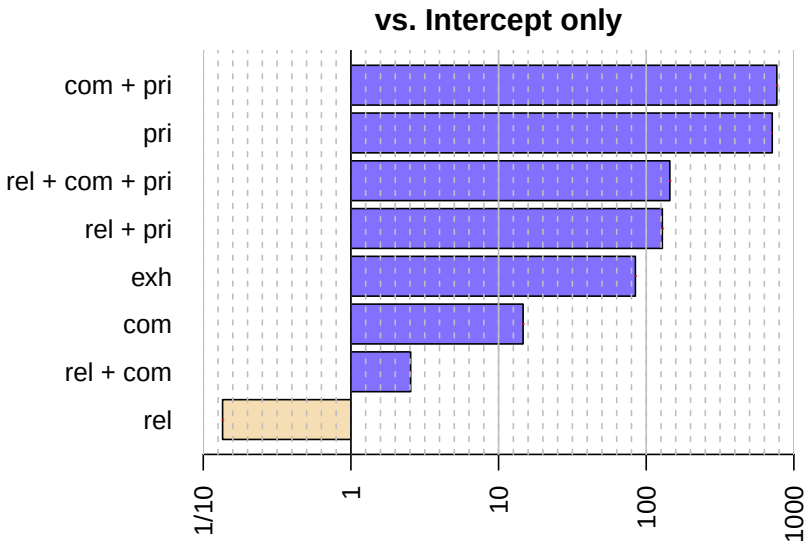

Figure 12: Bayes factor comparison of different main factor combinations, predicting the strength of exclusive disjunction readings with additional factor EXH from Experiment 4.

5 General discussion

Statements of the form '*A* or *B*' are sometimes interpreted as '*A* or *B* but not both'. The consensus in the recent literature is that these exclusive readings are scalar implicatures, which are inferences deriving from a pragmatic reasoning process about the speaker's intentions. A crucial assumption in this reasoning process is that the speaker is taken to be knowledgeable about whether the corresponding statement with '*A* and *B*' is true. In this paper, we have presented data that challenge the implicature account. In contrast with the predictions made by this account, the robustness of the exclusive reading does not increase with the competence of the speaker about '*A* and *B*'. Instead, we observed the reverse effect.

We also tested an alternative proposal, according to which the exclusive reading is derived as a scalar implicature by exhaustifying the individual disjuncts. According to this proposal, a sentence of the form '*A* or *B*' is read, in effect, as '*A* and nothing else or *B* and nothing else'. This proposal predicts that the robustness of the exclusive reading depends on the robustness of the inference from *A* to '*A* and nothing else'. Although we did find a significant effect of this factor, it was a significantly worse predictor of the robustness of the exclusive reading than the prior probability that the statement with 'and' was true.

Taken together, these findings indicate that the exclusive reading of 'or' is not a scalar implicature. Instead, we propose that exclusive readings are the result of probabilistic inferences about the state of the world. In other words, the meaning of 'or' is inclusive, but hearers may exclude the possibility that both disjuncts are true based on their knowledge of the world.

Such probabilistic inferences have been shown to influence various other parts of language. A case in point is quantifying expressions. In a series of experiments, Moxey and Sanford (1993) presented participants with statements such as:

(10) a. *Q* people found Miss Sweden attractive.
b. *Q* earthquakes occurred in California in 1951.

Q was varied between ten quantifying expressions, including 'a few' and 'many'. Participants were instructed to estimate, e.g., the number of people who found Miss Sweden attractive and the number of earthquakes in California in 1951, based on the information expressed in the sentences. For most quantifying expressions, Moxey and Sanford observed that these estimates increased as a function of the prior probability. For example, the estimates would be higher for (10a) than (10b), since the prior probability that someone finds Miss Sweden attractive is

higher than the prior probability that an earthquake occurs in California (see also Pepper & Prytulak 1974).

In a similar vein, the interpretation of gradable adjectives, such as 'big' and 'small', varies depending on our world knowledge. Kamp and Partee (1995) discuss the following minimal pair:

(11) a. My two-year old son built a really tall snowman yesterday.
b. The D.U. fraternity brothers built a really tall snowman yesterday.

It is obvious that one's prior expectations about the size of the snowman are lower in the case of (11a) than in the case of (11b), thus demonstrating that the interpretation of 'tall' varies depending on one's prior expections.

World knowledge and prior expectations thus influence the interpretation of various aspects of natural language. We propose that the interpretation of 'or' is one such aspect: although, strictly speaking, a statement of the form '*A* or *B*' leaves open the possibility that both *A* and *B* are true, prior expectations can modulate the likelihood of this possibility

We compared 'or' with 'some', which is often interpreted as 'some but not all'. This upper-bounding construal is an uncontroversial example of a scalar implicature. Unlike the exclusive reading of 'or', the robustness of the upper-bounding construal of 'some' increased with the competence of the speaker about the truth of the corresponding statement with 'and', in line with the results of Goodman and Stuhlmüller (2013).

In addition, we observed a significant effect of prior probability: the robustness of the upper-bounding construal decreased with the prior probability that the statement with 'all' was true. This effect ties in with recent experimental work from Degen et al. (2015), but speaks against observations made by Geurts (2010). Geurts discusses examples such as the following to show that the decision to derive a scalar implicature is made independently from the prior probability that the statement with 'all' is true.

(12) Cleo threw all her marbles in the swimming pool. Some of them sank to the bottom.

Here, according to Geurts' intuitions, the speaker communicates that some but not all of the marbles sank to the bottom, even though such a situation has an extremely low prior probability.

However, the tension between these two sets of observations may be more apparent than real. Our experiments, as well as the experiments of Degen and

colleagues, tested the robustness of the scalar implicature, rather than whether or not a scalar implicature was derived. So it may well be the case that, even though people tend to construe 'some' in (12) with an upper bound, they are less certain about this than if the statement were about, e.g., bread crumbs.

Unlike competence and prior probability, relevance to the hearer's interest did not turn out to have any effect on the robustness of the upper-bounding construal of 'some'. This is problematic to relevance theory, which predicts the robustness of an upper-bounding construal to be an increasing function of its importance to the hearer. It may suggest that relevance theory should adopt a more restricted notion of relevance in terms of discourse purposes (Cummins & Rohde, 2015).

An alternative explanation for the absence of an effect of relevance, however, is that we did not manipulate how relevant the upper-bounding construal was to the participant. According to this explanation, we failed to observe a significant effect because there were no differences in how relevant the upper-bounding readings were to the participants. Even though this alternative explanation warrants closer examination, there are several reasons to be sceptical about it. Perhaps the most prominent reason is that, strictly speaking, participants did not have any reason to derive the scalar implicature. The very fact that they do indicates, from a relevance-theoretic point of view, that they engaged with the story.

In summary, then, our results provide an insight in how various factors conspire in shaping differences in the robustness of scalar implicatures. It will be interesting to determine which other factors—e.g., typicality (van Tiel, 2014), prosodic and linguistic prominence (Breheny, Katsos, & Williams, 2006), and politeness (Bonnefon, Feeney, & Villejoubert, 2009)—have similar effects, and how these factors might interact. It may well be that these factors also hold the key in explaining the observation that different scalar expressions license scalar implicatures at substantially different rates (van Tiel et al., 2016).

A Stories from Experiment 1

1. Check *(Rel+ / Com+ / Pri+)*
Harry's job is to inspect hotels of an international franchise in order to check whether everything is according to the high standards of the organization. He reports to his boss Mr Jaynes after a visit to the San Diego hotel. Harry reports to Mr Jaynes: 'Some of the rooms had working air conditioning.'

2. Quiz *(Rel+ / Com+ / Pri+)*

George and Amber are playing a quiz. George gets to ask Amber ten questions on a topic of George's choice. George picked 'Star Trek' as a topic but he does not know Amber is actually a big 'Star Trek' fan. Amber just answered the final question. George says: 'You got some of the answers right.'

3. NBA *(Rel+ / Com+ / Pri–)*
Jason Barley and Richard Trellis are TV experts engaged in a live discussion of the current NBA season which is nearing its end. They are debating about whether Greg Jones should be the Most Valuable Player of the season. Jason Barley thinks so, but Richard Trellis is less convinced. Jason Barley says: 'Greg Jones secured victory for his team during the last seconds of some of the decisive playoff matches.'[7]

4. Donkey *(Rel+ / Com+ / Pri–)*
Grace and Oliver are playing a game of pin the tail on the donkey. Oliver has

[7]This item was removed from the analysis. We observed that the average robustness rating for the upper-bounding construal of 'some' for this item (39%) was substantially below the corresponding ratings for the other items (58% or more). In addition, almost half of the participants indicated that the probability of the upper-bounded reading was zero, while none did so for the next lowest scoring item. On closer examination, we noticed an ambiguity in the statement measuring the robustness of the upper-bounding construal:

(i) From what Jason Barley said we may conclude that Greg Jones did not secure victory for his team during the last seconds of all of the decisive playoff matches.

This statement may be read as implying that Greg Jones did not secure victory during the last seconds of *any* of the decisive playoff matches, which is is manifestly false. In order to determine if participants indeed succumbed to this ambiguity, we tested an alternative formulation:

(ii) From what Jason Barley said we may conclude that not all of the decisive playoff matches were secured during the last seconds by Greg Jones.

Importantly, this formulation is equivalent to the desired reading of (i) but lacks the ambiguity because the negation directly modifies the noun rather than the verb phrase.

We presented 40 participants on Mechanical Turk with the NBA story, followed by three questions. The first two were control questions. The third one was the target question. Half of the participants saw (i) and the other half (ii). The procedure was the same as in the original experiment, except that responses were made by marking a value on a seven-point scale rather than by means of a continuous slider. The results for the formulation in (i) (28%) were in line with the results of the original experiment (39%). However, the results for the formulation in (ii) were significantly higher (66%, $t(29) = 4.28$, $p < .001$), thus indicating that participants did not arrive at the desired reading of (i). Hence, we decided to remove this item from the analysis.

hung up a large number of pictures of donkeys and Grace has to pin the tail on as many of them as possible while wearing a blindfold. Grace just pinned the last tail. Oliver says: 'You put some of the tails in the right place.'

5. Watch *(Rel+ / Com– / Pri+)*
Mr Tobler works for a Swiss factory which produces luxury watches. Today he ran several tests with a new prototype in order to make sure that it meets the very high requirements on product quality that the company aims for. At this late stage in the project nobody expects any problems. But quality control is very important for the company's image. Mr Tobler's boss Mr Papenhauer is curious about the results and tries to call Mr Tobler but he has just left. Therefore Mr Tobler's apprentice answers the phone. The apprentice has only attended the first half of the testing procedure. The apprentice says: 'The prototype passed some of the tests scheduled for today.'

6. Harvard *(Rel+ / Com– / Pri+)*
Ethan is hoping to go to Harvard so he has to do well on his final exams. Fortunately he has had a month of spare time during which he studied day and night. His friend Eleanor just heard rumors from her fellow students. She says: 'You passed some of the exams.'

7. Tomatoes *(Rel+ / Com– / Pri–)*
Ted is a chef at a three star restaurant. He asks his sous chef Gilbert to check today's purchase of fresh produce. Gilbert looks at the first crate of tomatoes, and comes running back. Gilbert says: 'Some of the tomatoes are rotten.'

8. Schizo *(Rel+ / Com– / Pri–)*
Lucy has to give a talk in front of a big audience of psychologists. She is going to criticize one of the dominant theories about schizophrenia. Afterwards, Jacob, who was in the audience, chatted with his neighbours. He tells Lucy: 'Some of the people enjoyed your talk.'

9. Gifts *(Rel– / Com+ / Pri+)*
Leonard celebrated his 5th birthday with his best friends today. All of his friends brought him presents. The presents were given to him while everybody was sitting at the table after a nice piece of chocolate cake. When Suzanne's mother comes to take her daughter home from the party, Leonards's mother tells Suzanne's mother about the party. Leonard's mother says: 'Leonard unwrapped some of the presents

right away.'

10. M&M's *(Rel– / Com+ / Pri+)*
Henry is in the hospital because he broke his arm in a skiing accident. His aunt Bethany visited him yesterday and brought him some fruit and a bag of M&Ms. Later, Henry tells his father: 'I liked some of the M&Ms.'

11. Van Gogh *(Rel– / Com+ / Pri–)*
The mayor joins the opening of a new van Gogh exhibition in the city's Museum of Modern Art. The curator welcomes him warmly and gives him a private tour through the exhibition. The curator says: 'Some of van Gogh's paintings are among the most valuable paintings in the world.'

12. Drinking *(Rel– / Com+ / Pri–)*
Edward is in a crowded bar IDing the patrons who are drinking alcohol. There are new laws and if there is even one incident of underage drinking today Edward will close down the bar. Later he tells the owner of the bar: 'Some of the patrons drinking alcohol were underage.'

13. Dessert *(Rel– / Com– / Pri+)*
Jeremy is asking his friend Claire for her opinion about the newly opened gourmet restaurant on 5th Street. Claire has been there twice already and she is very enthusiastic. Claire says: 'Some of their desserts are fantastic.'

14. Cookies *(Rel– / Com– / Pri+)*
Martha received a couple of chocolate cookies even though neither she nor her husband eats chocolate. This evening the two of them are going out for dinner leaving Emily to babysit their children. Emily is a voracious eater and chocolate lover. When she comes home, Martha sees some crumbs on the couch. She says to her husband: 'Emily ate some of the cookies.'

15. Comedy *(Rel– / Com– / Pri–)*
Lennard is a very popular stand-up comedian. He is currently on tour. Both shows in Seattle were sold out. 500 tickets were sold for each show. He phones his wife after the second show in Seattle. Lennard says: 'Some of the people who came yesterday also came to today's show.'

16. Emails *(Rel– / Com– / Pri–)*

Kate has problems with her laptop but she needs to let Joshua know she will not be able to make it to their meeting today. Therefore she sends the same email numerous times from different email accounts. Later that day, Joshua phones Kate up and says: 'I received some of your emails.'

B Stories from Experiment 2

1. Health *(Rel+ / Com+ / Pri+)*
Mrs Gibbs is worried about her husband's health. Her friend Cindy, who is a waitress at a local restaurant, served Mrs Gibbs' husband yesterday. Cindy says to Mrs Gibbs: 'Yesterday your husband had a steak or a beer.'

2. Squash *(Rel+ / Com+ / Pri+)*
Danny and Alex reserved a squash court but Alex still has to buy a racket and a pair of shoes. Danny is talking to Alex's girlfriend Jill who just went to the sports store with him. Jill says to Danny: 'Alex bought a racket or a pair of shoes.'

3. Lunch *(Rel+ / Com+ / Pri–)*
During summer camp, every student was allowed at most one main dish for lunch. But some students queued up twice. Mrs Sanders was with her son Ted during the whole lunch break yesterday. A teacher approaches Mrs Sanders to find out how her son behaved during lunch. Mrs Sanders says to the teacher: 'Ted had pizza or pasta yesterday.'

4. Accident *(Rel+ / Com+ / Pri–)*
Pete was playing on the schoolyard during lunch break when he fell off the slide. His teacher drove him to the hospital immediately and called Pete's mother from there. The teacher told Pete's mother: 'Pete broke his arm or his leg.'

5. Clothes *(Rel+ / Com– / Pri+)*
Brad is notorious for his shabby clothes. His friends Gina and Mandy make frequent jokes about it, much to Brads distress. Gina heard that he finally went shopping the other day. Gina tells Mandy: 'Brad bought a sweater or a pair of jeans.'

6. Exams *(Rel+ / Com– / Pri+)*
Recently, Gigi took exams in Physics and Chemistry. She studied day and night and even paid for private tutoring. Carrie just heard some rumors from her fel-

low students. Carrie tells Gigi's best friend: 'Gigi passed her Physics exam or her Chemistry exam.'

7. Dancing *(Rel+ / Com– / Pri–)*
Many people came to Carls garden party yesterday, including his ex-wives Sue and Mary. It was terribly crowded everywhere. John helped out in the kitchen where it was much more quiet during most of the evening. Afterwards John told Carls wife: 'Carl danced passionately with Sue or Mary.'

8. Gol *(Rel+ / Com– / Pri–)*
Leo is sponsoring a golf tournament. In particular, he provides a luxurious car for any hole-in-one. When Leo arrives at the tournament, his friend Hans comes up to him to tell about the events so far. Hans tells Leo: 'I heard someone just hit a hole-in-one on the first or second hole.'

9. Order *(Rel– / Com+ / Pri+)*
Jake and Bill went jogging together in the afternoon. Afterwards they were very hungry and went to a restaurant. Jake later told Bills wife: 'Bill ordered a starter or a side dish with his main course.'

10. Sunscreen *(Rel– / Com+ / Pri+)*
Chloe is telling her father about her holidays with her two friends Susan and Amelia. Chloe forgot to bring suntan lotion. Chloe tells her father: 'Susan or Amelia brought some extra suntan lotion I could use.'

11. Club *(Rel– / Com+ / Pri–)*
Tony loves to dance to electronic music, preferably Techno. The simpler and more monotonic, the better. He contemplates going to club Macabre tonight, but he has no idea what music they play. He phones his friend Rob who occasionally works at club Macabre as a bartender. Rob tells Tony: 'They play Jazz or Metal tonight.'

12. Drinking *(Rel– / Com+ / Pri–)*
Lily is worried that her husband Eric is still drinking despite the doctor's orders. She phones up Eric's drinking buddy Bill who went to a bar with him last night. Bill tells Lily: 'Your husband drank whisky or vodka.'

13. Delay *(Rel– / Com– / Pri+)*
Kevin, Cory and Joshua are 7-year-olds from the same neighborhood. They are

friends but argue a lot. Yesterday Joshua, who has very strict parents, was later than usual to arrive at their hang-out. Kevin thinks that Joshua doesnt like to play with them anymore. Cory says to Kevin: 'Joshua had to eat lunch or do his homework.'

14. Graduation *(Rel– / Com– / Pri+)*
Nico and Rick are talking about their mutual friend Gerald. Gerald has been studying for ages but Nico heard he recently graduated. Nico says to Rick: 'His mother or his father attended the ceremony.'

15. NY *(Rel– / Com– / Pri–)*
Jimmy skipped school to go on a day-trip to New York City last week. Having to be back in the afternoon, he barely had one hour to spend in the city. Jimmy's younger brother, who is very talkative even if he doesn't know what he is talking about, told the neighbors about the trip afterwards. Jimmy's younger brother told the neighbors: 'Jimmy visited the Statue of Liberty or the Empire State Building.'

16. Pet *(Rel– / Com– / Pri–)*
Richard and Tom are talking about Harold. Harold has been wanting a pet for a long time but his girlfriend wouldn't let him. Richard heard she finally gave in. Richard says to Tom: 'Harold bought a cat or a dog.'

C Material for Experiment 3

[mf: fill me]

References

Basson, A. H., & O'Connor, D. J. (1960). *Introduction to symbolic logic*. Free Press of Glencoe.

Baum, R. (1996). *Logic* (4th edition). Harcourt Brace.

Bonnefon, J.-F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: scalar inferences in face-threatening contexts. *Cognition*, *112*(2), 249–258.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalized scalar implicatures generated by default? An online investigation into the role of context in generating pragmatic inferences. *Cognition*, *100*(3), 434–463.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Chevallier, C., Noveck, I. A., Nazir, T., Bott, L., Lanzetti, V., & Sperber, D. (2008). Making disjunctions exclusive. *The Quarterly Journal of Experimental Psychology*, *61*(11), 1741–1760.

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., Fox, D., & Spector, B. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.) *An international handbook of natural language meaning*, (pp. 2297–2332). Berlin: Mouton de Gruyter.

Copi, I. M., & Cohen, C. (2005). *Introduction to logic* (12th edition). Prentice Hall.

Crain, S. (2008). The interpretation of disjunction in universal grammar. *Language and Speech*, *51*(1-2), 151–169.

Cummins, C., & Rohde, H. (2015). Evoking context with contrastive stress: effects on pragmatic enrichment. *Frontiers in Psychology*, *6*.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: listeners revise world knowledge when utterances are odd. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.) *Proceedings of the 37th annual conference of the Cognitive Science Society*, (pp. 548–553). Austin, TX: Cognitive Science Society.

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). Houndmills: Palgrave Macmillan.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Franke, M. (2009). *Signal to act: game theory in pragmatics*. Ph.D. thesis, University of Amsterdam.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*(1), 3–44.

Geurts, B. (2006). Exclusive disjunction without implicature.

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

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

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

Horn, L. R. (1972). *On the semantic properties of logical operators in English*. Ph.D. thesis, University of California, Los Angeles. Distributed by Indiana University Linguistics Club.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*(2), 129–191.

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

McCawley, J. D. (1981). *Everything that linguists have always wanted to know about logic but were ashamed to ask*. University of Chicago Press.

Moxey, L. M., & Sanford, A. J. (1993). Prior expectation and the interpretation of natural language quantifiers. *European Journal of Cognitive Psychology*, *5*(1), 73–91.

Pepper, S., & Prytulak, L. S. (1974). Sometimes frequently means seldom: context effects in the interpretation of quantitative expressions. *Journal of Research in Personality*, *8*(1), 95–101.

Rescher, N. (1964). *Introduction to logic*. St. Martin's Press.

Roberts, C. (2012). Information structure in discourse: towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*(6), 1–69.

Rouder, J. N., & Morey, R. D. (2012). Default bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877–903.

Rubin, R., & Young, C. M. (1989). *Formal logic: a model of English*. Mayfield.

Russell, B. (2012). *Probabilistic Reasoning and the Computation of Scalar Implicatures*. Ph.D. thesis, Brown University.

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, *27*(3), 367–391.

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija*, *43*(4), 441–464.

Sperber, D., & Wilson, D. (1995). *Relevance: communication and cognition* (2nd edition). Blackwell.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167.

Storto, G., & Tanenhaus, M. K. (2005). Are scalar implicatures computed online? In E. Maier, C. Bary, & J. Huitink (Eds.) *Proceedings of Sinn und Bedeutung 9*, (pp. 431–445). Nijmegen: Nijmegen Centre for Semantics.

van Kuppevelt, J. (1996). Inferring from topics: scalar implicatures as topic-dependent inferences. *Linguistics and Philosophy*, *19*(4), 393–443.

van Tiel, B. (2014). Embedded scalars and typicality. *Journal of Semantics*, *31*(2), 147–177.

van Tiel, B., van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, *33*(1), 137–175.

Wilson, D., & Sperber, D. (2002). Relevance theory. *UCL Working Papers in Linguistics*, *14*, 249–290.

Yanal, R. J. (1988). *Basic logic*. Thomson.

Zondervan, A. (2010). *Scalar implicatures or focus: an experimental approach*. Ph.D. thesis, Utrecht University.