Title

Some but not all speakers sometimes but not always derive scalar implicatures

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Ramotowska, Sonia
Marty, Paul
van Maanen, Leendert
et al.

Publication Date

2024

Peer reviewed

Some but not all speakers *sometimes but not always* derive scalar implicatures

Sonia Ramotowska (s.ramotowska@uva.nl)
Institute for Logic, Language and Computation, University of Amsterdam

Paul Marty (paul.marty@um.edu.mt)
Institute of Linguistics and Language Technology, L-Università ta' Malta

Leendert Van Maanen (l.vanmaanen@uu.nl)
Department of Experimental Psychology, Utrecht University

Yasutada Sudo (y.sudo@ucl.ac.uk)
Department of Linguistics, University College London

Abstract

Experimental studies show that the tendency to derive Scalar Implicatures (SIs) varies considerably between individuals: some individuals accept sentences that are literally true but carry a false SI, while others systematically reject them. The question of what factors drive these differences is crucial to understanding the mechanisms involved in SIs and currently at the center of numerous discussions. To date, there is no agreement on how to quantify individual differences in SI rates. In this article, we show how a hierarchical Bayesian modelling approach can be used to quantify subjects' preferences observed in the results of a truth value judgement task that investigated intra-individual and inter-individual variability in the rates of upper-bounding and lower-bounding SIs associated with the ⟨some, all⟩-scale. The results provide further evidence that the robustness of an SI is modulated within individuals by certain linguistic features, such as the presence of negation.

Keywords: Scalar Implicatures; Hierarchical Bayesian models; Individual differences; Implicature strength

Introduction

One of the most studied pragmatic phenomena are Scalar Implicatures (SIs). Consider the sentence (1) for an illustration:

Some eagles are birds. (1)

This sentence has two possible readings. The first one follows from the semantic interpretation of the existential quantifier *some* as *at least one*. On this reading, (1) is true. The second reading of (1) is enriched by the SI 'not all'. On this *some-but-not-all* reading, (1) is thus false, since all eagles are birds. In this paper, we will refer to the first reading as **literal** reading and to the second one as **pragmatic** reading.

The core mechanism responsible for the ambiguity of (1) is considered to be the underlying competition between *some* and its stronger alternative, *all* (Horn, 1972). *All* is a competitor for *some* because both expressions come from one lexical scale ⟨some, all⟩. In the case of (1), for example, when judging the truth of this sentence, one considers alternative sentences, such as *All eagles are birds*. By the principles of communication that are commonly assumed by the interlocutors (more specifically here, the Maxim of Quantity, Grice, 1975), the listener reasons that the speaker should have used this alternative sentence whose meaning is stronger, had she believed it to be true. Since the speaker used the weaker one, the listener infers the *some-but-not-all* reading.

Diversity in scalar implicatures

The 'not all' SI associated with the quantifier *some* is arguably the most expansively investigated implicature (cf. Van Tiel, Van Miltenburg, Zevakhina, & Geurts, 2016). On the assumptions that SIs constitute a uniform class and that they are **Strength Invariant**, one would expect the results obtained with the 'not all' SI of *some* to generalize to other types of SIs (for discussion, see Van Tiel et al., 2016 and Degen, 2015). However, an increasing number of experimental studies report a high degree of between-scale and within-scale diversity in the strength of SIs, leading to new questions on what factors account for scalar diversity (Doran, Ward, Larson, McNabb, & Baker, 2012; Van Tiel et al., 2016; Degen, 2015; Gotzner, Solt, & Benz, 2018; Ronai & Xiang, 2022).

One linguistic feature considered to contribute to this diversity is boundedness (Van Tiel et al., 2016). The ⟨some, all⟩-scale, for instance, has a lower bound designated by *some* and an upper bound designated by *all*. The lower bound expression of the scale gives rise to an upper-bounding (UB) SI via negation of the upper bound expression on the same scale. The 'not all' SI of (1) is one such case. In addition to the 'not all' UB SI, there is also a lower-bounding (LB) 'some' SI associated with the ⟨some, all⟩-scale (cf. Chierchia, 2004; Cremers & Chemla, 2014). This SI arises when the upper bound expression *all* is explicitly negated as in example (2):

Not all eagles are reptiles. (2)

The same UB SI (i.e, *Some eagles are reptiles*) is found in examples like (3), where negation appears in the scope of the lower bound expression *some*:

Some eagles are not reptiles. (3)

Van Tiel et al. (2016) found that bounded scales give rise to higher rates of SIs compared to unbounded scales, such as ⟨content, happy⟩. Moreover, while many studies have demonstrated that the 'not all' SI of *some* may come at an extra processing cost (for a review, see Khorsheed, Price, & van Tiel, 2022), other experimental evidence suggests that this does not generalize to LB SIs (Cremers & Chemla, 2014; Van Tiel & Pankratz, 2021; Van Tiel, Pankratz, & Sun, 2019;

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Romoli & Schwarz, 2015). In our study, we included all three cases of SIs associated with examples (1)–(3) above to test whether UB and LB SIs are derived at a similar rate. Additionally, we examined individual differences as potential sources of within-scale diversity in SI strength (Van Tiel et al., 2016; Gotzner et al., 2018; Ronai & Xiang, 2022; Sun, Tian, & Breheny, 2024; Degen, 2015).

Individual differences in scalar implicatures

Numerous studies have explored inter-individual variability in rates of with-SI-response. Since at least Bott and Noveck (2004) and Noveck and Posada's (2003) studies, it has been observed that some participants consistently accept sentences like (1), while others consistently reject them. For this reason, it is not uncommon that, for the purposes of data analysis, participants be classified into two groups, **literal** vs. **pragmatic responders**, based on which reading is most preferred. It should be noted that such classification is often based on a coarse-grained threshold of the sort 'n% of responses are consistent with the pragmatic reading'. For instance, Dieussaert, Verkerk, Gillard, and Schaeken (2011) set a threshold of 90% of consistent answers while Politzer-Ahles, Fiorentino, Jiang, and Zhou (2013) used a 83% threshold, Hunt, Politzer-Ahles, Gibson, Minai, and Fiorentino (2013) an 80% threshold and Spychalska, Kontinen, and Werning (2016) a 70% threshold.

Several studies have also examined the consistency in the choice of reading using such classification methods. For example, Spychalska et al. (2016) classified all participants as either literal or pragmatic responders based on the above criterion. Hunt et al. (2013) found that the majority of participants were consistent, with only 3 out of 24 not having a reading preference by their criterion. Politzer-Ahles et al. (2013) found that 7 participants provided consistent answers, while 11 did not. Kursat and Degen (2020) report that 3% of their participants provided an equal number of literal and pragmatic responses; the others were classified as literal or pragmatic responders. Marty, Romoli, Sudo, and Breheny (2024) use a similar classification criterion to investigate implicature priming effects in consistent responders and found that these effects are systematically observed towards these responders' less preferred reading. Degen and Tanenhaus (2015), in turn, use 5 levels of response inconsistency and found that inconsistent responders were slower to respond than consistent responders, irrespective of the reading preference of the latter. Heyman and Schaeken (2015) offer a more principled way of classifying participants into literal and pragmatic responders by using a latent class analysis. They found that the model classifying participants into three groups had the lowest BIC score and that only 21% of participants did not provide consistent responses. Finally, some studies like Fairchild and Papafragou (2021) treat the rate of implicature-consistent responses as a continuous variable, hence setting aside the categorical distinction between responder groups.

In sum, there is strong evidence supporting the existence of inter-individual differences in rates of of with-SI responses. However, there is no agreement on how participants should be categorized into responder groups, or even explicit discussion of this issue. The lack of consistency in the criteria of classification between studies makes it difficult to come to solid conclusions about how individual differences contribute to explaining variability in SI strength. As a response to this challenge, we develop in the following a systematic and rigorous method to quantify individual differences regarding literal and pragmatic readings.

Current study

We propose to use hierarchical Bayesian modelling (Lee & Wagenmakers, 2010, see also Franke & Degen, 2016) to quantify individual differences in SI rates. Hierarchical models have several benefits. First, they allow for the operationalization of theoretical constructs, such as groups of responders, and their relationships in explicit model parameters. This allows us, for instance, to test whether the pattern of responses observed in an experiment comes from a specific distribution. Crucially for us, including a hierarchical structure makes it possible to test how the response patterns of individuals depend on the assumed response pattern of their group. Second, Bayesian models let us use a fine-grained probabilistic criterion to classify participants, instead of applying a coarse-grained threshold. Finally, as these models assume that participants' responses come from a specific distribution, they can be used even when the total number of responses per experimental condition is relatively small.

To develop this approach, we conducted an experimental study to collect data about the strength of the UB and LB SIs associated with the scalar term *some* and its scale-mate, *all*. Results from Degen (2015) challenge the Strength Invariance assumption by showing that the strength of the 'not all' SI depends on the context. Capitalising on this work, we were interested in testing a modified version of the Strength Invariance hypothesis, according to which SIs associated with the same scale have the same strength for all listeners if the overall context remains constant. While this hypothesis predicts no individual differences on the sole basis of the scale involved, it does allow for variability in strength depending on linguistic factors such as the presence or absence of negation. Thus, on this hypothesis, if one type of SIs is observed more frequently than another, this preference should be observed across participants. This hypothesis was tested by fitting hierarchical Bayesian models that made different assumptions about individual differences and comparing their fit. Specifically, we compared models that assume that all individuals had the same preference and models that allowed participants to belong to one of two groups with some probability. The outcomes of these comparisons were used to quantify individual differences in responses across SI types. Should participants provide inconsistent responses, the more complex model that allows for probabilistic classification will outperform the simpler model that does not allow for a mixture of literal and pragmatic responses for a given SI type.

Methods

Participants

95 native speakers of English, all located in the UK, were recruited on Prolific to participate in the study (62 females, 33 males; mean age 30 yrs, range 18-69). The experiment lasted approximately 9 min and participants were paid £1.35. All participants provided written informed consent. The study was approved by the UCL Ethics committee.

Material and design

The material and design were built on Bott and Noveck's (2004) classical truth value judgement studies. Participants were presented with categorical sentences like those in (1)–(3) above and asked to provide True/False judgements. Test sentences were simple sentences with *some* (SOME), sentences with *some* and negation following it (SOME NOT), and sentences with *all* with preceding negation (NOT ALL), all of which potentially give rise to SIs. The SI of SOME is UB whereas those of NOT ALL and SOME NOT are LB. For each sentence type, target and control conditions were constructed by manipulating category membership, as illustrated in Table 1. Sentences were pseudo-randomly generated from a base of 9 categories and 11 exemplars from each of these categories. Each survey included 5 examples of each condition per construction and 54 filler trials, hence a total of 99 trials.

Table 1: Example sentences used in the experiment. In the target conditions, the sentences can be considered true or false depending on whether or not the implicature is drawn.

Sentence	Condition	Example sentence	Response
SOME	Target	Some eagles are birds	?
	True	Some birds are eagles	T
	False	Some eagles are insects	F
SOME NOT	Target	Some eagles are not insects	?
	True	Some birds are not eagles	T
	False	Some eagles are not birds	F
NOT ALL	Target	Not all eagles are insects	?
	True	Not all birds are eagles	T
	False	Not all eagles are birds	F

Procedure

Participants were told that they would see English sentences, presented one word at a time, and that they would have to say whether they considered the sentences to be true or false. They were not given specific instructions on how to interpret the sentences. Participants started with two practise trials and then continued to the test phase. Each trial consisted of the presentation of a fixation point (250 ms) followed by a word-by-word presentation of the sentence. Each word remained on the screen for 250 ms, except the last one, which remained until the participant provided their response. Participants were asked to provide their response as quickly as possible by pressing one of two keyboard keys (F and J). Response keys were counterbalanced across subjects. Participants were not given feedback on whether their response was correct or not. Each session was divided into two blocks to give participants a self-timed break.

Model

We fit seven beta-binomial hierarchical Bayesian models, two basic models, four models with latent group classification, and one basic model further testing the difference in group classification between UB and LB SI (Basic Gr. SI) to the truth value judgement data. All models assumed that literal and pragmatic responders have a different distribution of 'true' responses. The probability of providing 'true' responses to a target sentence depends on which of those groups a participant belongs to. The observed response distribution in truth value judgement task is coming from one or the other group or a mixture of both groups.

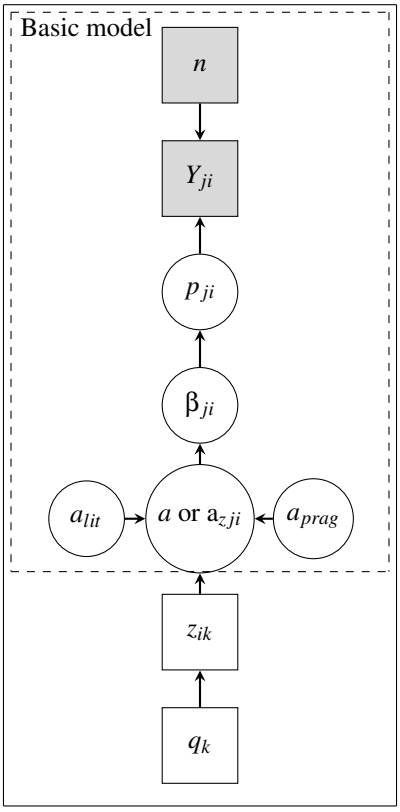

Figure 1: Graphical representation of the Basic models (in dashed rectangle) and Latent Gr. models (in solid rectangle).

First, we fit two basic models, which assume that all participants belong to either the literal (Basic lit) or pragmatic (Basic prag) responder group (see Figure 1). The model predicted the number of literal responses (Y_{ji}) for the i-th participant and j-th SI as coming from a Binomial distribution:

$$Y_{ji} \sim \text{Binomial}(n_{ji}, p_{ji})$$

with n being the number of trials per SI type and participant ($n = 5$ for all i and j because no responses were ex-

cluded from the analysis) and p_{ji} being a probability of 'true' response predicted via the following regression model with a zero intercept and coefficients for each implicature type:

$$logit(p_{ji}) = 0 + \beta_{0i} \times some_i + \beta_{1i} \times somenot_i + \beta_{2i} \times notall_i$$

SI type was contrast-coded such that each coefficient corresponded to the rate of 'true' responses for each implicature type. In the basic model, all βs (j = 0,1,2) were drawn from a Beta prior distribution

$$\beta_{ji} \sim \text{logit}(\text{Beta}(a,b))$$

where a was conceptualized as the frequency of literal responses a_{lit} or pragmatic responses a_{prag}, depending on the responder group under consideration, and assumed to come from one of the Uniform hyperprior distributions with a threshold of 50% of 'true'-responses ($n = 5$), and $b = n - a$.[1] For literal responders, this distribution was above the threshold and and, for pragmatic responders, below the threshold:

$$a_{lit} \sim \text{Uniform}(n/2, n)$$

$$a_{prag} \sim \text{Uniform}(0, n/2)$$

.

To test for individual differences in the choice of reading, we extended the basic model by introducing a latent group classification (see Figure 1). These models assume that each participant is classified as a literal or pragmatic responder with some probability; therefore, by comparing these probabilities between participants, we can assess the inter-individual differences. We can further assess the intra-individual consistency, by checking how likely each individual is classified into one group or the other.

We fit four models with different numbers of possible classifications of participants ($k = 1,2,3$). In the models with one classification ($k = 1$), participants were classified either as pragmatic or literal responders, independently of the SI type. This model assumes, therefore, that participants have a general preference to respond pragmatically or literally across SIs. In the most flexible model ($k = 3$), participants were classified for each SI type separately. This model therefore assumes that participants may have different preferences for different SIs, independently of the linguistic features of SIs. For the other two models with $k = 2$, the classification of participants also depended on the SI, but with further restrictions. In one of them, participants were classified separately for the UB and LB SIs, and in the other, for the strong quantifier (NOT ALL) and the weak quantifier (SOME, SOME NOT). These models, therefore, tested whether or not participants' reading preference is mediated by specific linguistic properties of the scalar sentences. Based on the literature reviewed above, we expected the classification based on SI type (Latent Gr. SI) to give a better fitting model than the classification based on quantifier strength (Latent Gr. STR).

In the latent group models, all coefficients β_js ($j = 0,1,2$) were drawn from an SI-specific and participant-specific beta prior distribution:

$$\beta_{ji} \sim \text{logit}(\text{Beta}(a_{zji}, b_{zji}))$$

where a_{zji} was conceptualized as the probability of literal responses provided by participant i for SI type j, coming from a Beta hyperprior distribution dependent on the group classification z, and $b_{zji} = n - a_{zji}$. The group classification priors were drawn from the Bernoulli distribution:

$$z_{ik} \sim \text{Bernoulli}(q_k)$$

where q_k was drawn from a hyperprior uninformative Beta distribution:

$$q_k \sim \text{Beta}(1,1)$$

.

For $k = 2$, we had q_{ub}/q_{lb} or q_{strong}/q_{weak}. All models were fit in R Studio and R JAGS (packages RJAGS and JAGSUI). For each model, we ran six Markov chains with 10,000 iterations and 1,000 burn-in iterations per chain. The adaptive phase included 1,000 iterations and the thin rate was 2.

Results

Materials along with the data files and code files for result analysis can be found on the OSF Platform here.

General task performance

We compared mean accuracy to the True and False control trials across sentence types to asses participants' general performance in the task (see Fig. 2, panel A). We fit generalized linear mixed-effects models with a by-participant random intercept. Adjusting the significance level for multiple comparisons (Bonferroni correction), we found that, for the False controls, the accuracy was lower for NOT ALL sentences than the other two sentence types; for the True controls, the accuracy was higher for SOME sentences than the other two sentence types; finally, for NOT ALL, the accuracy was higher for the True than for the False controls. Overall, these results align with the classical finding that, without a supporting context, negative sentences are usually more difficult to process and verify than affirmative sentences, leading to more errors (for an overview, see Kaup & Dudschig, 2020).

[1] We applied the constraints of Uniform distribution 0.001 and 4.999 to avoid convergence problems.

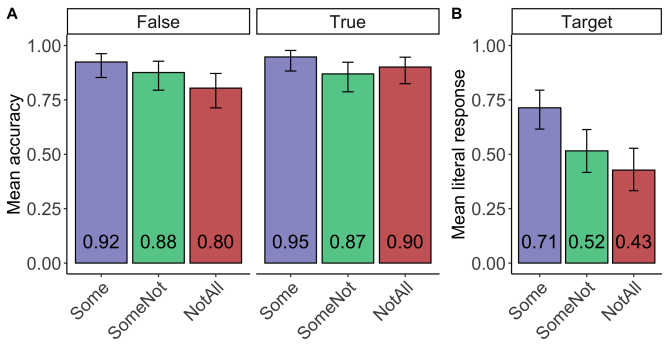

Figure 2: **A** Mean accuracy in True and False control conditions and **B** Mean literal response in Target conditions. Error bars represent 95% confidence intervals.

The means of literal response in the three target conditions are shown in Fig. 2, panel B. The mean of literal responses for SOME in our study (71%) aligns with that observed in Bott and Noveck (2004)'s Experiment 4 in the 'short lag' condition where respond speed was stressed (72%).

Modelling results

All six models converged with Rhats below 1.1. Table 2 summarizes the model comparison results. For the purposes of model comparison, predictive accuracy was measured using two information criteria: the *Deviance Information Criterion* (DIC), a Bayesian criterion for model comparison including a penalty for model complexity (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) and the *Watanabe-Akaike Information Criterion* (WAIC; Watanabe, 2010) which, unlike DIC, uses the entire posterior distribution (rather than a point estimate) and is invariant to parametrization (for discussion, see Vehtari, Gelman, & Gabry, 2017). The model classifying participants into two groups separately for LB and UB SIs (Latent Gr. SI) had the lowest DIC and WAIC values, thus the best fit to the data. In the following, we report the results of this model in more detail.

Table 2: Model comparison. Abbreviations: Resp stands for the assumed responder group, q_ks for the number of classifications, DIC for deviance information criterion, WAIC for Watanabe-Akaike information criterion, pD for the effective number of parameters and PP for posterior predictive checks.

Model	Resp.	q_ks	DIC	WAIC	pD	PP
Basic lit	a_{lit}	0	1059	1061	262	0.53
Basic prag	a_{prag}	0	1076	1073	271	0.02
Latent Gr.	Mix	1	958	956	242	0.49
Latent Gr. SI	Mix	2	**925**	**898**	242	0.49
Latent Gr. STR	Mix	2	991	954	279	0.5
Latent Gr.	Mix	3	1024	948	318	0.50
Basic Gr. SI	Mix	0	1010	1012	244	0.48

Figure 3 illustrates the model fit of the best-fitting model. To test our hypothesis about inter- and intra-individual differences in the choice of the response, we investigated the posterior distributions of by-participant parameter $z_{ik=2}$.

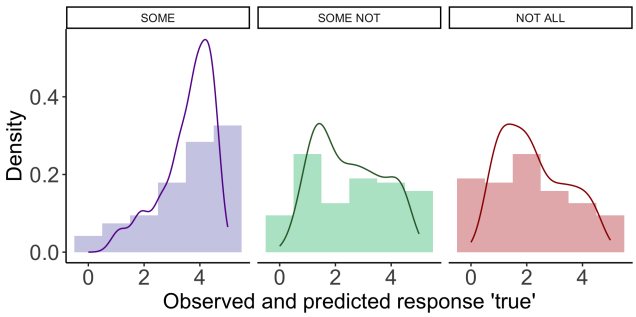

Figure 3: Observed frequencies of literal responses (histogram) vs. predictions of the best model (line).

Figure 4 shows that the probability of being classified as a literal responder was higher for UB SI (Mean $q_{ub} = 0.86$) than LB SI (Mean $q_{lb} = 0.34$). This result suggests that the UB SIs were less likely to be derived. Nonetheless, we observed substantial inter-individual variation for both types of SIs, which is not predicted by the modified Strength Invariance assumption. To test the robustness of this variation, we constrained the best-fitting model such that all participants were literal responders for *some*, and all were pragmatic responders for two other expressions. The model fit of this model (Basic Gr. SI) is as in the last row of Table 2. Should the more complex model (Latent Gr. SI) outperform this simpler one, we provide strong evidence for the individual differences. The comparison of the DIC and WAIC values speaks against the modified Strength Invariance assumption (see Table 2).

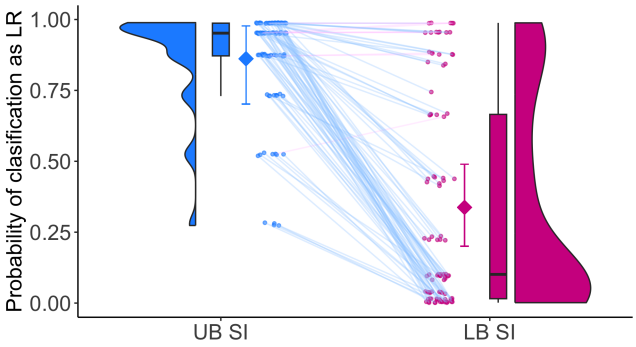

Figure 4: By-subject probability of being classified as literal responder ($z_{ik=2}$) with mean posteriors q_{lb} and q_{ub} and their 95% CIs. Lines show within-subject consistency in responses to sentences associated with UB SIs (SOME) and LB SIs (SOME NOT, NOT ALL): the steeper the line, the less consistent the responder. Violin and box plots show the z_{iub} and z_{ilb} parameters distributions.

Next, we investigated how consistent participants are in the choice of literal or pragmatic responses. Figure 4 illustrates the intra-individual differences in the probability of being classified as a literal responder for each SI type. For most participants, the probability of literal response was higher for UB SIs than for LB SIs. Finally, we conducted an exploratory analysis to test whether the individual differences we observed could be an artefact of participants' general performance in the task. This was done by looking at pairwise Pearson correlations with Holm's correction for multiple comparisons between the by-participants z_{iub} and z_{ilb} parameters and mean accuracy in all control conditions. None of these correlations was significant ($0.2 < r < 0.2$, for all tests). These results show that the intra-individual differences cannot be explained by poor task performance.

Discussion

We reported on a truth value judgement task experiment designed to test a modified version of the Strength Invariance hypothesis by comparing individual differences in the derivation of three types of SIs involving the ⟨some, all⟩-scale. To quantify intra- and inter-individual differences across SI types, we used a series of hierarchical Bayesian models allowing for probabilistic classification of participants into groups of literal and pragmatic responders. Our results show that participants had different preferences for different SIs. Specifically, while the strength of the quantifier did not affect participants' responses by itself, we found that LB SIs from negative SOME NOT and NOT ALL sentences were more likely to be derived than UB SIs from positive SOME sentences by most participants. This finding stands in contrast with previous research suggesting a consistent preference toward one reading for LB and UB SIs with ⟨some, all⟩ (Cremers & Chemla, 2014) and raises an obvious question: why would negation increase the likelihood to derive an SI?

We hypothesise that this boosting effect could be due to negation acting as a linguistic cue for retrieving a prominent Question under Discussion (QuD; cf. Roberts, 2012). Specifically, Tian and Breheny (2016) propose that, without contextual support, the most prominent QuD for a negative sentence $\neg p$ is a question where the truth of p is at issue, i.e., the question *whether p*.[2] For negative scalar sentences like *not all eagles are reptiles* or *some eagles are not reptiles*, this means that, in the absence of further cues, speakers may be biased to accommodate the question *whether all eagles are reptiles*. Crucially, accommodating such a question should increase the contextual relevance of the *none*-alternatives involved in the derivation of the *some*-SIs of interest, promoting in turn the derivation of LB SIs. As far as we can see, this hypothesis aligns with recent work emphasising the effect of QuD on speakers' interpretive preferences (a.o., Degen & Tanenhaus, 2015; Kursat & Degen, 2020; Ronai & Xiang, 2021; Marty et al., 2024) and could explain the higher rates of with-SI responses for negative sentences, compared to positive ones. It remains to be seen, however, whether the contrasts we found reproduce with other scales, stimuli sentences and experimental setup than those used in this study.

Our results also demonstrate sizeable intra- and inter-individual differences in the choice of pragmatic responses. In line with previous studies (e.g., Van Tiel, Marty, Pankratz, & Sun, 2019), we found greater variability in participants' pragmatic responses for LB SIs. Following Degen and Tanenhaus (2015) and Kursat and Degen (2020), we take this inconsistency to suggest that participants in our study were more uncertain about the intended QuD for negative than positive sentences. On our hypothesis above, this could be the case, for instance, if negation modulated participants' expectations to a different extent, biasing them more or less strongly toward the QuD *whether p*. Our results also indicate that a significant number of participants provided mixed responses for at least one type of SI; importantly, our correlation analysis shows that this inconsistency cannot be attributed to poor task performance. Taken together, these findings disconfirm the prediction of the modified Strength Invariance hypothesis.

In recent years, individual differences have gained in attention in experimental and computational pragmatics. In this paper, we introduced an advanced methodology for studying such differences in the derivation of SIs. The method we developed enables to quantify these differences and test how different linguistic features affect the availability of with-SI reading. We believe that this method offers a new perspective on the role of individual differences in accounting for scalar diversity. Future developments of this approach could also help contribute to the ongoing debate surrounding the processing of SIs. While some SIs are more difficult to process than others, some are just as fast and easy to process as literal meaning (Van Tiel et al., 2019). Additionally, some studies suggest that literal and pragmatic responders process SIs differently (Spychalska et al., 2016; Kursat & Degen, 2020) and are affected differently by different types of primes (Marty et al., 2024). Our model provides a fine-grained way of investigating such individual differences by examining the probabilities of being classified as a literal/pragmatic responder for different SI types and how these probabilities correlate with any measure of interest.

[2] A related idea commonly found in the literature is that negative sentences presuppose that there is an (implicit or explicit) expectation in the context that their positive counterparts be true (e.g., Moxey, 2006; Moxey, Sanford, & Dawydiak, 2001; Wason, 1965).

Acknowledgments

SR was supported by the NWO OC project Nothing is Logical (grant no 406.21.CTW.023). PM was supported by the Leverhulme Trust grant RPG-2018-425. YS received funding from an AHRC-DFG grant (AHRC: AH/V003526/1; DFG: EB 523/2-1, STE 958/12-1, STE 2555/3-1).

References

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language*, *51*(3), 437–457.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In *Structures and beyond* (Edited by Belletti ed., p. 39-103). Oxford University Press.

Cremers, A., & Chemla, E. (2014). Direct and indirect scalar implicatures share the same processing signature. In *Pragmatics, semantics and the case of scalar implicatures* (pp. 201–227). Springer.

Degen, J. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, *8*, 11–1.

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive science*, *39*(4), 667–710.

Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: Further evidence that scalar implicatures are effortful. *Quarterly Journal of Experimental Psychology*, *64*(12), 2352–2367.

Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. E. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*, 124–154.

Fairchild, S., & Papafragou, A. (2021). The role of executive function and theory of mind in pragmatic computations. *Cognitive Science*, *45*(2), e12938.

Franke, M., & Degen, J. (2016). Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PloS one*, *11*(5), e0154854.

Gotzner, N., Solt, S., & Benz, A. (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in psychology*, *9*, 1659.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Heyman, T., & Schaeken, W. (2015). Some differences in some: examining variability in the interpretation of scalars using latent class analysis. *Psychologica Belgica*, *55*(1), 1.

Horn, L. R. (1972). *On the semantic properties of logical operators in english*. University of California, Los Angeles.

Hunt, L., Politzer-Ahles, S., Gibson, L., Minai, U., & Fiorentino, R. (2013). Pragmatic inferences modulate n400 during sentence comprehension: Evidence from picture–sentence verification. *Neuroscience Letters*, *534*, 246–251.

Kaup, B., & Dudschig, C. (2020). Understanding negation: Issues in the processing of negation.

Khorsheed, A., Price, J., & van Tiel, B. (2022). Sources of cognitive cost in scalar implicature processing: A review. *Frontiers in Communication*, *7*, 990044.

Kursat, L., & Degen, J. (2020). Probability and processing speed of scalar inferences is context-dependent. In *Annual meeting of the cognitive science society* (p. 1236-1242).

Lee, M. D., & Wagenmakers, E.-J. (2010). A course in bayesian graphical modeling for cognitive science. *Unpublished manuscript. Retrieved from http://www. ejwagenmakers. com/BayesCourse/BayesBookWeb. pdf*.

Marty, P., Romoli, J., Sudo, Y., & Breheny, R. (2024). Implicature priming, salience, and context adaptation. *Cognition*, *244*, 105667.

Moxey, L. M. (2006). Effects of what is expected on the focussing properties of quantifiers: A test of the presupposition-denial account. *Journal of Memory and Language*, *55*(3), 422–439.

Moxey, L. M., Sanford, A. J., & Dawydiak, E. J. (2001). Denials as controllers of negative quantifier focus. *Journal of Memory and Language*, *44*(3), 427–442.

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and language*, *85*(2), 203–210.

Politzer-Ahles, S., Fiorentino, R., Jiang, X., & Zhou, X. (2013). Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification. *Brain research*, *1490*, 134–152.

Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, *5*, 6–1.

Romoli, J., & Schwarz, F. (2015). An experimental comparison between presuppositions and indirect scalar implicatures. In *Experimental perspectives on presuppositions* (pp. 215–240). Springer.

Ronai, E., & Xiang, M. (2021). Pragmatic inferences are qud-sensitive: an experimental study. *Journal of Linguistics*, *57*(4), 841–870.

Ronai, E., & Xiang, M. (2022). Three factors in explaining scalar diversity. In *Proceedings of sinn und bedeutung* (Vol. 26, pp. 716–733).

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, *64*(4), 583–639.

Spychalska, M., Kontinen, J., & Werning, M. (2016). Investigating scalar implicatures in a truth-value judgement task: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience*, *31*(6), 817–840.

Sun, C., Tian, Y., & Breheny, R. (2024). A corpus-based examination of scalar diversity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *50*(5), 808–818.

Tian, Y., & Breheny, R. (2016). Dynamic pragmatic view of negation processing. *Negation and polarity: Experimental perspectives*, 21–43.

Van Tiel, B., Pankratz, E., & Sun, C. (2019). Scales and scalarity: Processing scalar inferences. *Journal of Memory and Language*, *105*, 93-107.

Van Tiel, B., Marty, P., Pankratz, E., & Sun, C. (2019). Scalar inferences and cognitive load. In *Proceedings of sinn und bedeutung* (Vol. 23, pp. 427–442).

Van Tiel, B., & Pankratz, E. (2021). Adjectival polarity and the processing of scalar inferences. *Glossa: a journal of general linguistics*, *6*(1), 1–21.

Van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts,

B. (2016). Scalar diversity. *Journal of semantics*, *33*(1), 137–175.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, *27*, 1413–1432.

Wason, P. C. (1965). The contexts of plausible denial. *Journal of verbal learning and verbal behavior*, *4*(1), 7–11.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(116), 3571–3594.