Scalar diversity and negative strengthening[1]

Anton BENZ — *Leibniz–Centre General Linguistics*
Carla BOMBI — *Potsdam University*
Nicole GOTZNER — *Leibniz–Centre General Linguistics*

Abstract. In recent years, experimental research has demontrated great variability in the rates of scalar inferences across different triggering expressions (Doran et al. 2009, 2012, van Tiel et al. 2016). These studies have been taken as evidence against the so-called uniformity assumption, which posits that scalar implicature is triggered by a single mechanism and that the behaviour of one scale should generalize to the whole family of scales. In the following, we present an experimental study that tests negative strengthening for a variety of strong scalar terms, following up on van Tiel et al. (2016). For example, we tested whether the statement *John is not brilliant* is strengthened to mean that John is not intelligent (see especially Horn 1989). We show that endorsement rates of the scalar implicature (e.g., *John is intelligent but not brilliant*) are anti-correlated with endorsements of negative strengthening. Further, we demonstrate that a modified version of the uniformity hypothesis taking into account negative strengthening is consistent with van Tiel et al.'s data. Therefore, variation across scales may be more systematic than suggested by the van Tiel et al. study.

Keywords: Scalar diversity, scalar implicature, manner implicature, negative strengthening, inferencing task.

1. Introduction

For more than a decade, scalar implicatures haven been a core topic of experimental pragmatics. However, theoretical and experimental research has concentrated on a few scales only, most notably the scales ⟨*all, some*⟩ and ⟨*and, or*⟩. In van Tiel et al. (2016) the authors provide an overview of 29 experimental studies from 2001 to 2014. Of them, only two studies consider scales other than ⟨*all, some*⟩ and ⟨*and, or*⟩. They speculate that the underlying reason for this bias is the belief that these scales are somehow representative for scales in general, such that findings on them can be generalised to all scales. This is the so–called uniformity hypothesis. This hypothesis has received some interest in recent years. The experimental studies in Doran et al. (2009, 2012) and van Tiel et al. (2016) addressed it in a special form: they tested the hypothesis that all scales show the same capacity for generating scalar implicature. This means, in this special form the hypothesis states that there is a constant percentage *s* such that for all scales *i* about *s*% of the subjects will draw an implicature for the weak scalar alternative. The most thorough and systematic study on this hypothesis was presented by van Tiel et al. (2016). They tested 43 scales, among them 32 scales with adjectives, 6 with main verbs, 2 with auxiliary verbs, 2 with quantifiers, and 1 with adverbs. In their first experiment, they presented 25 subjects with questions of the form: *John says: She is intelligent. Would you conclude from*

[1]We would like to thank Jacopo Romoli, Alexandre Cremers, Richard Breheny, Stephanie Solt, Bob van Tiel and the audience of Sinn und Bedeutung and the annual Xprag.de meeting for helpful comments on this work. This research was supported by the Bundesministerium für Bildung und Forschung (BMBF) (Grant Nr. 01UG1411), and the Deutsche Forschungsgemeinschaft (DFG) (Grant Nr. BE 4348/4-1). Author names appear in alphabetical order. AB and NG contributed equally to writing this paper and CBF implemented the experiment.

this that, according to John, she is not brilliant? Subjects then had to choose between the answers *yes* and *no*. Here, the relevant scale is ⟨*brilliant, intelligent*⟩. If subjects answer *yes*, then they must have drawn the implicature *intelligent* $+>$ *not brilliant*. The results of the study revealed that scales show considerable variance in their ability for generating scalar implicatures. In a post-analysis of their data, van Tiel et al. (2016) found that boundedness of a scale and perceived distance between the strength of alternatives predicts implicature rates. That is, participants were more likely to derive a scalar implicature if the stronger scale-mate denotes an endpoint on the underlying measurement scale (see especially Kennedy and McNally 2005 for a scale typology) and the greater the difference in strength was rated.

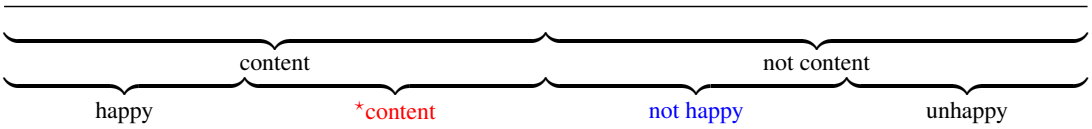
Van Tiel and colleagues also considered a number of other parameters as predictors of implicature (such the stronger term's cloze probability, relative frequency, latent semantic value, and grammatical category) but none of these parameters had an effect on implicature rates. Further, they briefly dismissed negative strengthening as a possible confounding parameter (see the discussion on page 141 in van Tiel et al. 2016).

Here, we present the results of a study based on van Tiel et al. (2016) which shows that negative strengthening is (anti-)correlated with scalar implicatures and that a modified version of the uniformity hypothesis, postulating a constant ratio between scalar implicature and negative strengthening, can be maintained. At the same time, we provide evidence that different scale types behave differently with respect to the modified uniformity hypothesis. In conclusion, our data motivate further research into the impact of scale structure on implicature derivation.

2. Negative strengthening

Negative strengthening is the phenomenon whereby the negation of the stronger scalar alternative is pragmatically strengthened to an interpretation that also negates the weaker alternative (Horn 1989, Levinson 2000, Blutner 2004, Krifka 2007). In (1) this is demonstrated for the scale ⟨*happy, content*⟩.

(1)



The second line shows the semantic extension of the adjective *content* and its negation, the third line the effect of scalar implicature and negative strengthening: the extension of *content* is shortened to ⋆*content* (*scalar implicature* SI), and that of *not happy* is strengthened such that it covers the area between *content* and *unhappy* (*negative strengthening* NegS). Negative strengthening is variously explained as R-implicature (Horn 1989), I–implicature (Levinson 2000), or blocking phenomenon (Blutner 2004, Krifka 2007). All authors agree that it arises differently from scalar implicatures, which are a special Q–implicature.

To see the relevance of negative strengthening for the experimental set up of van Tiel et al. (2016), let us consider the following item:

(2)

John says:

She is content.

Would you conclude from this that, according to John, she is not happy?

☐ Yes ☐ No

If a subject interprets *content* and *not happy* semantically, then s/he has to answer with *no* since the statement *she is content* is semantically consistent with her being not happy. If the subject narrows the meaning of *content* based on scalar implicature, the subject should answer *yes*. This is how the experiment intended to measure the rate of scalar implicature. However, if participants negatively strengthen the conclusion sentence *She is not happy* to *not content*, this interpretation is incompatible with the semantics and the scalar implicature of *content*. Hence, negative strengthening leads to a *no*–answer, whatever the subject's interpretation of *content* is. The different possibilities of reading *content* and *not happy* and the expected *yes* and *no*–answers are shown in (3).

(3)

content	*not happy*	Yes	No
semantic	semantic		✓
semantic	NegS		✓
SI	semantic	✓	
SI	NegS		✓

Hence, a *no*-answer may be based on a semantic interpretation of negative strengthening. For this reason, the lack of scalar implicature may be masked by the effect of negative strengthening. Let us now consider what this means for the *uniformity hypothesis*. As we may recall, van Tiel et al. addressed the uniformity hypothesis in a special form, namely that for all scales i the proportion of observed *yes*–answers $s_o(i)$ is equal to a fixed probability s. In this form, the uniformity hypothesis is clearly refuted by the experimental studies of Doran et al. (2009, 2012) and van Tiel et al. (2016). However, the formula $s_o(i) = s$ assumes that negative strengthening has no influence on the observed *yes*–answers. Let us assume that we can observe negative strengthening with probability $n_o(i)$ for scale i. Now consider (3). The simplest hypothesis about the relation between negative strengthening and scalar implicature is that negative strengthening of *not happy* occurs independently of drawing the scalar implicature for *content*. A *yes*–answer is given if the scalar implicature is drawn (probability s according to the uniformity hypothesis) and no negative strengthening occurs (probability $1 - n_o$). Hence, the observed proportion of *yes*–answers s_o should equal the product $s \times (1 - n_o)$. This leads to the modified uniformity hypothesis that the observed scalar implicature for scale i equals the product of a constant s and the observed probability of no negative strengthening, as formula:

(4) $s_o(i) = s \times (1 - n_o(i))$.

A peculiarity of the uniformity hypothesis is that, to our knowledge, it is a hypothesis that

no-one has ever defended. Even though its prior plausibility is low, it seems interesting to defend it for purely methodological reasons. Ultimately, we hope to gain insight into which sub–classes of scales show a uniform behaviour with respect to scalar implicature and negative strengthening.

In order to evaluate the modified uniformity hypothesis, we need an estimate of the proportion of negative strengthening $n_o(i)$. We need to know how likely it is that subjects understand, for example, *not happy* as implying *not content*. We, therefore, ran an experiment with exactly the same items and fillers as (van Tiel et al. 2016: Exp. 1), but modified the questions. For example, for the $\langle$*happy, content*$\rangle$ scale, we asked subjects *John says: He is not happy. Would you conclude from this that, according to John, he is not content?* If the answer is *yes*, this indicates that subjects negatively strengthened *not happy* to *not content*. We will see that the observed rates of *yes* answers shows similar variability between scales as the rates of *yes* answers in the original scalar implicature experiment. We show that $s_o(i)$ and $n_o(i)$ are anti–correlated, and that the anti–correlation is so strong that the modified uniformity hypothesis cannot be rejected on the basis of van Tiel et al.'s results. However, we also show that we can find sub–classes of scales that behave very differently with respect to the uniformity hypothesis, so that the paper ends with an open question: what are the parameters that determine sub–classes of scales that behave uniformly with respect to scalar implicature and negative strengthening?

3. The Experiment

3.1. Methods

3.1.1. Participants

40 participants with US IP addresses were recruited on Amazon's Mechanical Turk platform. They were further screened for their native language. In total, 40 native English speakers (mean age: 37.02, 20 female, 20 male) took part in the study.

3.2. Materials

Our task and all materials were based on the study by van Tiel et al. (2016). Participants were presented with a scenario involving two characters, Mary and John, who make a series of statements. Their task was to decide whether a strengthened interpretation follows from a given statement. For example, participants saw the statement *John is not brilliant* and were asked whether they conclude that John is not intelligent. The latter task is a measure of negative strengthening of the stronger scale-mate. Figure 3.2 presents a sample display participants saw.

If participants respond with *yes*, they have negatively strengthened *not brilliant* to *not intelligent*.

In total, each participant saw statements with 43 scales, all of which are provided in Table 3 in the Appendix, in addition to 6 filler sentences. Two versions of the survey with different orders

Mary says:

He is not brilliant.

Would you conclude from this that, according to Mary, he is not intelligent?

Yes No

Figure 1: Sample item of the negative strengthening task

were created and administered to 20 participants each.

3.3. Results

In our analysis, we used the average endorsement rates of scalar implicature provided in van Tiel et al. (2016) and the negative strengthening rates obtained from our own experiment.

On average, for all scales, 42.3% of the subjects answered *yes* in our rating of negative strengthening. Table 3 in the Appendix presents the negative strengthening ratings for all items. Selected results are shown in (5), plotting the ratings in the scalar implicature (SI) task (van Tiel et al. 2016) and the negative strengthening (NegS) ratings next to each other.

(5) Results for selected scales: % of scalar implicature (SI) from van Tiel et al. (2016), % of negative strengthening (NegS) from our study

Scale	SI	NegS
⟨*free, cheap*⟩:	100%	28%
⟨*all, some*⟩:	96%	42%
⟨*love, like*⟩:	50%	43%
⟨*finish, start*⟩:	21%	14%
⟨*exhausted, tired*⟩:	4%	69%
⟨*happy, content*⟩:	4%	92%

Scale	SI	NegS
⟨*impossible, difficult*⟩:	79%	25%
⟨*none, few*⟩:	75%	31%
⟨*unsolvable, hard*⟩:	71%	43%
⟨*unavailable, scarce*⟩:	62%	58%
⟨*unforgettable, memorable*⟩:	50%	56%

Overall, we observe a correlation between $s_o(i)$, the observed % of SIs for scale i, and $(1 - n_o(i))$, with $n_o(i)$ the % of NegS for i (Spearman's rank correlation: 0.463, $p < 0.002$).[2] That is, participants are less likely to endorse a scalar implicature if they apply negative strengthening to the stronger scale-mate. Hence, the lack of scalar implicature can, in part, be explained by the presence of negative strengthening.

We also ran a linear regression model for the negative strengthening ratings involving boundedness, semantic distance, grammatical category, frequency, cloze probability, and latent semantic values (using the values obtained in the van Tiel et al. study) as predictors of variability across

[2]We based the correlational analysis on the complement rate of the negative strengthening task $(1 - n_o(i))$, which will be explained in detail in Section 4.

scales. The results of the model are displayed in Table 1. The analysis showed that participants were more likely to apply negative strengthening if the weaker and stronger scale-mate had a strong association strength as indexed by the measure obtained in van Tiel et al.'s cloze task. Further, semantic distance (the perceived difference in strength between the statement involving the weaker and the one with the stronger term) had a negative effect on ratings. That is, the occurence of negative strengthening was less likely the closer the semantic distance between the stronger and weaker term. In our experiment, the upper boundedness of scales did not have a significant effect on negative strengthening rates.

Table 1: Predictors of negative strengthening ratings

	Estimate	SD	t-value	p-value
(Intercept)	1.01843	0.23029	4.422	0.000
Cloze probability	0.45191	0.11194	4.037	0.00028
Category	0.08695	0.104	0.836	0.4088
Frequency	0.04462	0.03086	1.446	0.15706
LSA	-0.11782	0.17463	-0.675	0.50428
Distance	-0.13364	0.04042	-3.307	0.00219
Boundedness	-0.09099	0.05843	-1.557	0.12841

3.4. Discussion

The current study showed that negated strong scalar terms give rise to varying degrees of inferences negating their weaker scale-mates. Such negative strengthening is traditionally thought of as a manner implicature, arising from a different principle than scalar implicatures (Horn 1989, Levinson 2000). In our analyis, we showed that participants' endorsement of scalar implicatures was anti-correlated with the degree of negative strengthening of the stronger scale-mate.

Van Tiel et al. (2016) discussed negative strengthening as a possible confound in their results (p. 144) but dismissed this possibility with the argument that their data show that scales containing a negative element generate high rates of implicature, although these scales are known for showing a robust tendency towards negative strengthening (Horn 1989, Krifka 2007). Table 5 on the right side shows the results for negative scales. Contrary to expectations, negative scales in our study were not particularly strong triggers of negative strengthening.

It should be noted that the numerical correlation we observed was not perfect. Hence, it is not the case that negative strengthening takes away all the variance observed in the scalar implicature task. Further, previous studies by Doran et al. (2009, 2012) demonstrated a similar amount of variation across scales as the van Tiel et al. study and their paradigm did not involve a negation as part of the instructions. In that study, participants were presented with a dialogue and a fact. Their task was to judge whether the answer was true or false given the fact. For example, Sam said *Gus ate most of the birthday cake* and the fact was that Gus had eaten the entire cake. In this verification task, the rates of scalar inferences were comparable to the ones by van Tiel et al. (2016) and there was considerable variation across adjectival scales and quantifiers.

It remains to be established whether the correlation between scalar implicature and negative strengthening we observed here is an artefact of the inferencing task, that is, because the negation of the stronger scale-mate was mentioned in the conclusion sentence. Rather than assuming that the interaction between scalar implicature and negative strengthening is merely a task effect, we might expect this interaction to be of broader importance. While the two kinds of implicature arise from different conversational principles, Levinson (2000) and Horn (1989) assume that the Q and R principle govern each other in conversation (see also Blutner 2004, Krifka 2007). Therefore, whether or not hearers derive a scalar implicature may also be influenced by the availability of other types of inferences.

4. The uniformity hypothesis: A modified version

The studies of Doran et al. (2009, 2012) and van Tiel et al. (2016) convincingly show that the uniformity hypothesis $s_0(i) = s$ for a constant s is false. However, the question arises whether the assumption of a uniform constant can be maintained if the effect of negative strengthening is factored in. Given the anti-correlation between $s_o(i)$ and $1 - n_o(i)$, the simplest reformulation of the uniformity hypothesis (UH) is to postulate a constant ratio between these values, i.e. that there is a constant s such that for all scales i $s_o(i)/(1 - n_o(i)) = s$, or, equivalently, that $s_o(i) = s - s\,n_o(i)$, see (4). The constant s can be fitted to the data. Using the data from van Tiel et al.'s scalar implicature task and our negative strengthening task, an optimal value of $s = 0.77$ was found.[3] Figure 2 shows $s_o(i)$ over $n_o(i)$ for all scales i. A simple linear regression was calculated to predict $s_o(i)$ (*yes*-answers in van Tiel et al.'s SI task) based on $n_0(i)$ (*yes*-answers in our NegS task). A significant regression equation was found ($F(1,41) = 7.80$, $p < .01$), with an adjusted $R^2 = 0.14$. The proportion $s_o(i)$ of *yes* answers in the SI task is equal to $0.68 - 0.55\,n_o(i)$. The regression line (blue) is also shown in Figure 2, together with its 95% confidence interval. The green line is the regression line predicted by the modified uniformity hypothesis with $s = 0.77$, i.e. $s_o(i) = 0.77 - 0.77\,n_o(i)$. As can be seen from Table 2, the line predicted by the modified uniformity hypothesis lies within the 95% confidence interval of the calculated linear regression line. Hence, the predicted regression line does not significantly differ from the calculated one, and can, therefore, not be rejected. In this sense, the modified uniformity hypothesis is *consistent* with the results found by van Tiel et al.

Clearly, to defend a hypothesis by showing that it cannot be refuted by some statistics is not an argument to accept the hypothesis. However, the modified uniformity hypothesis is nevertheless interesting because it establishes a numeric relation between a scale's propensity to trigger two different types of implicature, in this case a quantity implicature (SI) and an I/M-implicature (NegS). In the following we will see that the modified uniformity hypothesis can be a useful tool for distinguishing different classes of scales that support or do not support it. As previously noted, the presence of an upper bound and semantic distance between scalar alternatives are significant predictors of *yes* answers in van Tiel et al.'s SI task. In Section 3.3, we have seen that semantic distance and cloze probability are significant predictors of negative strengthening. We

[3]We used the form of the modified uniformity hypothesis as stated in (4) and chose the s that minimizes $|((1 - n_o(i))\,s - s_o(i))_i|$. Choosing the mean of ratios $s_o(i)/(1 - n_o(i))$ leads to a slightly higher value for s but doesn't change the conclusions. The reason for not choosing ratios is that some of them are greater than 1. As s is supposed to represent the proportion of subjects answering *yes* in the van Tiel et al. task, values higher than 1 are empirically meaningless.

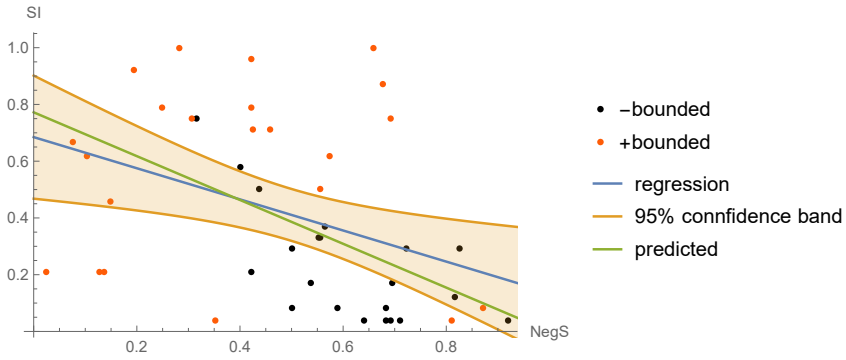
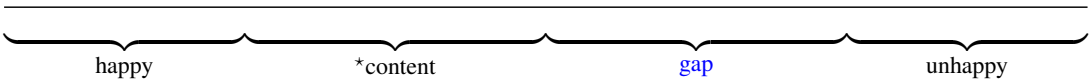


Figure 2: Fit of modified uniformity hypothesis, $s = 0.77$

now introduce a distinction of scale types that is primarily motivated by research on negative strengthening (Blutner 2004, Krifka 2007), and show that the new scale types behave very differently with respect to the modified uniformity hypothesis.

The distinction that we introduce is that between L-scales and M-scales. A prototypical L-scale would be the ⟨*all, some*⟩ scale. If we consider the underlying measurement scale that reaches from proportions of 0% to 100%, then the Horn scale ⟨*all, some*⟩ starts from the *lower* end of the measurement scale. This means that the weak scale mate *some* covers the whole measurement scale except for 0%, that is the lower end. The contrary *none* of *all* is also the contradictory of *some*. In contrast, M-scales are scales that start somewhere in the *middle* of the underlying measurement scale. Examples are the ⟨*happy, content*⟩ scale, and the ⟨*hot, warm*⟩ scale. In both cases, there is a gap between the weaker scale mate and the contrary of the stronger scale mate. In other words, the contradictory of the weaker scale mate is not the contrary of the stronger one. For ⟨*happy, content*⟩ this means that there is a gap between the meaning of *content* and the contrary of *happy*, namely *unhappy*, see (6) and (1). Likewise, for ⟨*hot, warm*⟩ there is a gap between the meaning of *warm* and the contrary of *hot*.

(6)



In Blutner (2004) and Krifka (2007), negative strengthening is explained as a blocking phenomenon. In their models, marked expressions narrow their meanings as they are blocked from referring to certain meanings *m* by the existence of less marked expressions that are better candidates for referring to *m*. This means that, for example, *not happy* is blocked from referring to states covered by *content*, as *content* is less marked. Likewise, *not happy* is blocked from referring to the extreme end of the unhappiness side because of the less marked expression *unhappy*. Hence, the meaning of *not happy* is narrowed down to the gap between *content* and *unhappy*. If this explanation for negative strengthening is correct, then L-scales should not give rise to negative strengthening as there is no gap which can be filled by the negation of the stronger scale mate. The observed rates of *yes*-answers in our NegS task would then have to be explained as random noise. This also means that we should expect M-scales to better conform to the modified uniformity hypothesis than L-scales. In the following, we test this prediction.

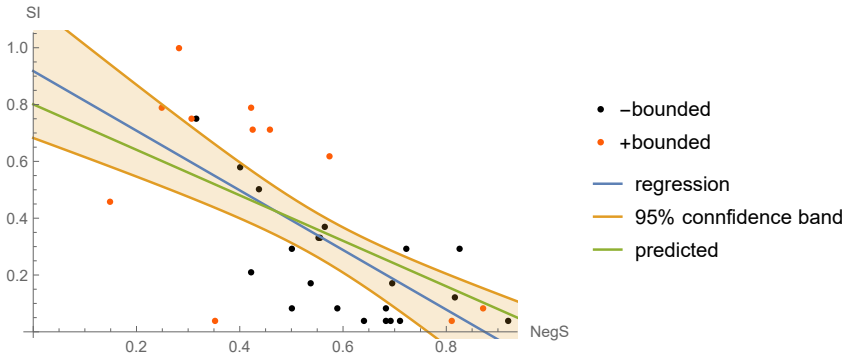


Figure 3: Fit of modified uniformity hypothesis for M-scales, $s = 0.80$

In Table 3 in the Appendix we provide an annotation of the different scale types. There are 32 M-scales among the 43 scales considered by van Tiel and colleagues. A simple linear regression was calculated to predict $s_o(i)$ (*yes*-answers in van Tiel et al.'s SI task) based on $n_0(i)$ (*yes*-answers in NegS task) for M-scales i. A significant regression equation was found ($F(1,30) = 28.27$, $p < .0001$), with an adjusted $R^2 = 0.47$). The proportion $s_o(i)$ of *yes* answers in the SI task is equal to $0.92 - 1.05\, n_o(i)$. The regression line (blue) is shown in Table 3, together with its 95% confidence interval. The green line is the regression line predicted by the modified uniformity hypothesis with $s = 0.80$, i.e. $s_o(i) = 0.80 - 0.80\, n_o(i)$. As can be seen from Figure 3, the line predicted by the modified uniformity hypothesis lies within the 95% confidence band of the calculated linear regression line.

Further, the statistical parameters show that the correlation between $s_o(i)$ and $1 - n_o(i)$ is much stronger in the case of M-scales than for all scales taken together.

There are 11 L-scales among the 43 scales considered by van Tiel et al. A simple linear regression was calculated to predict $s_o(i)$ (*yes*-answers in van Tiel et al.'s SI task) based on $n_0(i)$ (*yes*-answers in NegS task) for M-scales i. A marginally significant regression equation was found ($F(1,9) = 5.02$, $p = .052$), with an adjusted ($R^2 = 0.29$). The proportion $s_o(i)$ of *yes* answers in the SI task is equal to $0.40 + 0.67\, n_o(i)$. The regression line (blue) is shown in Table 2, together with its 95% confidence interval. The green line is the regression line predicted by the modified uniformity hypothesis with $s = 0.73$, i.e. $s_o(i) = 0.73 - 0.73\, n_o(i)$. As can be seen from Figure 2, the line predicted by the modified uniformity hypothesis does not lie within the 95% confidence interval of the calculated linear regression line; rather, it follows a completely different pattern.

As we can see, there is no significant positive correlation between $s_o(i)$ and $1 - n_o(i)$; to the contrary, there is a marginal negative correlation between them for L-scales. There is also a considerable visual difference between the calculated regression line and the predicted regression line. We conclude that the modified uniformity hypothesis does not explain the pattern L-scales adhere to.

As we mentioned before, the uniformity hypothesis is peculiar in that it has, to our knowledge, not been defended by anyone. It was merely put forward as a likely explanation for why

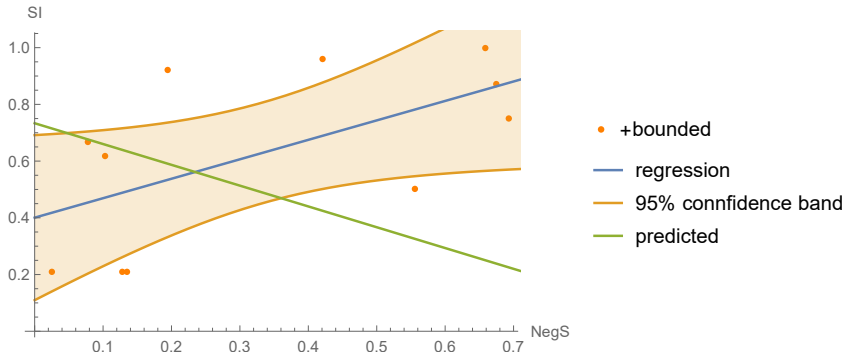


Table 2: Fit of modified uniformity hypothesis for L-scales, $s = 0.73$

previous experimental research concentrated on a few scales, most notably the ⟨*all, some*⟩–scale. The distinction between L– and M–scales may provide another reason for concentrating on this scale. As it is an L–scale, it should be affected by negative strengthening to a lower extent.[4] It may, therefore, be better suited to test scalar implicature.

Our analysis of L-scales and M-scales is intended as a demonstration of the usefulness of the modified uniformity hypothesis as a tool for establishing interesting distinctions among scale types. However, one should not over-estimate what we have achieved here. The same contrast between L-scales and M-scales that we find in Figures 2 and 3 exists between bounded and unbounded scales, as well as between non-adjectival and adjectival scales. This means that M-scales, unbounded scales, and adjectival scales conform better to the modified uniformity hypothesis than all scales taken together, and for L-scales, bounded scales, and non–adjectival scales, it has to be rejected. Even this result should be taken with caution. There is a considerable overlap between M-scales, unbounded scales and adjectival scales in the sample collected by van Tiel et al. such that it remains an open issue which of them causes scales to conform or not to conform to the modified uniformity hypothesis.

The issue about the predictors of uniformity carries over to the issue of predictors of *yes*–answers in van Tiel et al.'s paradigm. Van Tiel et al. found that boundedness and semantic distance are significant predictors of *yes*–answers. Due to the overlap between bounded, non–adjectival, and L-scales, however, the significant correlation between boundedness and *yes*–answers vanishes once the effect of being an M–scale or being an adjectival scale is taken into account.

In a similar vein, McNally (2017) argues that the methods used by van Tiel et al. were too crude to (i) detect certain implicatures and (ii) detect effects of the parameters explaining variation across scales tested. Essentially, the problem McNally discusses is that adjectives are polysemous, and in the absence of a context participants may construct a meaning on the fly and not think of the intended pair as scale-mates. This criticism also applies to the current study and it stresses the need to present test sentences within a conversational context. Our analysis showed that it is not entirely clear at this point which predictors of variability are crucial in explaining

[4]In fact, it has been argued that negation of the stronger scale-mate leads to scale reversal, i.e. that *not all* implicates *some but not all* (see e.g. Levinson 2000: p. 80ff, with references to previous literature).

diversity. Hence, our investigation motivates further research into the impact of scale structure on implicature derivation. Comparing how a large variety of scales behave within an enriched communicative context has to be left to future research. One experimental paradigm which might be useful for this endeavour is the action-based task by Gotzner and Benz (2018), and its interactive version (Benz et al. 2018), which has been implemented for the quantifier *some* and the determiner *or* (Benz and Gotzner 2017). The advantage of this paradigm is that utterances are embedded in a communicative situation and candidate readings are made relevant. The current study indicates that for future experiments on scalar diversity, a balanced set of items varying in scale structure is needed.

5. Conclusion

In the current study, we demonstrated an interaction between two kinds of implicature: scalar implicatures which are Q–based, and negative strengthening, which is I– or M–based. In particular, there was an anti-correlation between the endorsement rates of scalar implicatures and the degree of negative strengthening of the stronger scale-mate. We showed that a modified version of the uniformity hypothesis is consistent with the data presented by van Tiel et al.'s study. We also provided evidence that the correlation between scalar implicature and negative strengthening may be sensitive to general scale structure. This shows that a more fine-grained typology of scales can be motivated by numerical analysis. However, the most interesting outcome of our study is the questions that it raises. What are the true predictors of scalar implicature and negative strengthening for different types of scales? Can a classification based on structural properties of scales be established such that all members of a class have the same propensity for triggering different types of implicature? Which other types of conversational implicature are sensitive to scale structure, besides scalar implicature and negative strengthening? Can conversational context make scales behave uniformly? For example, do all scales reliably trigger scalar implicatures if the meaning differences are made contextually relevant? Is there an experimental paradigm which allows the measuring of scalar implicature without negative strengthening or typicality effects as confounding factors? In conclusion, the present paper highlights the importance of further research into the impact of scale structure on scalar implicature.

References

Benz, A. and N. Gotzner (2017). Embedded disjunctions and the best response paradigm. In R. Trueswell and H. Rohde (Eds.), *Proceedings of Sinn und Bedeutung 21*, University of Edinburgh.

Benz, A., N. Gotzner, and L. Raithel (2018). Embedded implicature in a new interactive paradigm. In M. Krifka, M. Grubic, U. Sauerland, S. Solt, and M. Zimmermann (Eds.), *Proceedings of Sinn und Bedeutung 22*, ZAS & University of Potsdam.

Blutner, R. (2004). Pragmatics and the lexicon. In L. Horn and G. Ward (Eds.), *The Handbook of Pragmatics*, pp. 488–514. Oxford: Blackwell Publishing.

Doran, R., R. E. Baker, Y. McNabb, M. Larson, and G. Ward (2009). On the non-unified nature of scalar implicature: an empirical investigation. *International Review of Pragmatics 1*, 1–38.

Doran, R., G. Ward, Y. McNabb, M. Larson, and R. E. Baker (2012). A novel paradigm for distinguishing between what is said and what is implicated. *Language 88*, 124–154.

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

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

Kennedy, C. and L. McNally (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language 81*, 345–381.

Krifka, M. (2007). Negated antonyms: Creating and filling the gap. In U. Sauerland and P. Stateva (Eds.), *Presupposition and Implicature in Compositional Semantics*, pp. 163–177. Houndmill: Palgrave Macmillan.

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

McNally, L. (2017). Scalar alternatives and scalar inference involving adjectives: A comment on van Tiel, et al. (2016). In J. Ostrove, R. Kramer, and J. Sabbagh (Eds.), *Asking the Right Questions: Essays in Honor of Sandra Chung*, UC Santa Cruz Previously Published Works, pp. 17–27.

van Tiel, B., E. van Miltenburg, N. Zevakhina, and B. Geurts (2016). Scalar diversity. *Journal of Semantics 33*(1), 107–135.

6. Appendix

Table 3: Weak and strong scale-mates, their negative strengthening rates obtained in our experiment, scalar implicature rate from van Tiel et al. (2016, reprinted with permission from Oxford University Press), scale type, upper bound (B = bounded, NB = non-bounded) and category (open vs. closed class) and part of speech (Adj = adjective, V = verb, Det = determiner, Adv = adverb)

Weak/Strong	NegS	SI	Scale Type	Boundedness	Category	PoS
adequate/good	0.72	0.29	M	NB	O	Adj
allowed/obligatory	0.08	0.67	L	B	O	Adj
attractive/stunning	0.59	0.08	M	NB	O	Adj
believe/know	0.13	0.21	L	NB	O	V
big/enormous	0.54	0.17	M	NB	O	Adj
cheap/free	0.28	1	M	B	O	Adj
content/happy	0.92	0.04	M	NB	O	Adj
cool/cold	0.55	0.33	M	NB	O	Adj
dark/black	0.35	0.04	M	B	O	Adj
difficult/impossible	0.25	0.79	M	B	O	Adj
dislike/loathe	0.83	0.29	M	NB	O	V
few/none	0.31	0.75	M	B	C	Det
funny/hilarious	0.64	0.04	M	NB	O	Adj
good/excellent	0.56	0.37	M	NB	O	Adj
good/perfect	0.15	0.46	M	B	O	Adj
hard/unsolvable	0.43	0.71	M	B	O	Adj
hungry/starving	0.56	0.33	M	NB	O	Adj
intelligent/brilliant	0.5	0.08	M	NB	O	Adj
like/love	0.44	0.5	M	NB	O	V
low/depleted	0.46	0.71	M	B	O	Adj
may/have to	0.69	0.75	L	B	C	V
may/will	0.68	0.87	L	B	C	V
memorable/unforgettable	0.56	0.5	L	B	O	Adj
old/ancient	0.69	0.17	M	NB	O	Adj
palatable/delicious	0.4	0.58	M	NB	O	Adj
participate/win	0.03	0.21	L	B	O	V
possible/certain	0.19	0.92	L	B	O	Adj
pretty/beautiful	0.68	0.08	M	NB	O	Adj
rare/extinct	0.42	0.79	M	B	O	Adj
scarce/unavailable	0.58	0.62	M	B	O	Adj
silly/ridiculous	0.68	0.04	M	NB	O	Adj
small/tiny	0.81	0.04	M	NB	O	Adj
snug/tight	0.82	0.12	M	NB	O	Adj
some/all	0.42	0.96	L	B	C	Det
sometimes/always	0.66	1	L	B	O	Adv
special/unique	0.87	0.08	M	B	O	Adj
start/finish	0.14	0.21	L	B	O	V
tired/exhausted	0.69	0.04	M	NB	O	Adj
try/succeed	0.1	0.62	L	B	O	V
ugly/hideous	0.71	0.04	M	NB	O	Adj
unsettling/horrific	0.5	0.29	M	NB	O	Adj
warm/hot	0.32	0.75	M	NB	O	Adj
wary/scared	0.42	0.21	M	NB	O	Adj