STRENGTH AND SIMILARITY OF SCALAR ALTERNATIVES[*]

NATALIA ZEVAKHINA
Radboud University of Nijmegen

Introduction

If a speaker uttered (1), a hearer may reason about whether the speaker believes that all of the books on the shelf have pictures. The hearer assumes that the speaker is cooperative and, therefore, he infers that the speaker doesn't believe that all of the books on the shelf have pictures. Assuming the speaker is competent whether the state of affairs holds or doesn't hold, the hearer makes a stronger inference that the speaker believes that not all of the books on the shelf have pictures (see 1A). The same reasoning, mutatis mutandis, goes for the derivation of (1B).

(1) Some of the books on the shelf have pictures.
A. $\rightsquigarrow$ Not all of the books on the shelf have pictures.
B. $\rightsquigarrow$ Most of the books on the shelf don't have pictures.

The inferences like (1A) or (1B) are called *scalar inferences* (SIs) in neo-Gricean pragmatics due to the generally adopted term *scale*, i.e. a unit of linguistic expressions, which are ordered by logical entailment (e.g., ⟨*some, most, all*⟩). The idea is that the derivation of SIs involves reasoning about stronger alternatives in the same scale. The stronger alternatives for (1) are *All of the books on the shelf have pictures* and *Most of the books on the shelf have pictures*.

Traditional theoretical work on implicatures (Horn (1972), Levinson (2000), Chierchia (2004) and others) assumes that all SIs are derived in a similar way. That is, given the scale ⟨*palatable, tasty, delicious*⟩, the derivation for (2) is supposed to be analogous to (1).

(2) The cake Charlie baked was palatable.
A. $\rightsquigarrow$ The cake Charlie baked was not tasty.
B. $\rightsquigarrow$ The cake Charlie baked was not delicious.

Zevakhina and Geurts (2011) showed with a series of experiments that this assumption is not correct since there was significant variation *between* scales. For example, quantifier, modal, and

[*] I would like to thank Bart Geurts, Bob van Tiel, Noor van Leusen, Rob van der Sandt, and Corien Bary for fruitful discussions. Many thanks to Bart Geurts, Jennifer Spenader, and Galit Sassoon for the comments on the first draft of the paper. This research was supported by a grant from the Netherlands Organisation for Scientific Research (NWO), which is gratefully acknowledged.

adjectival scales were derived at significantly different rates. Furthermore, there was significant variation between adjectives. The explanation for this we proposed is differential availability of scalar alternatives[1].

It is quite natural now to pose the question whether we can observe variation *within* scales. Such a variation is logically and intuitively expected. Are the inferences in (1A) and (1B) or inferences in (2A) and (2B) derived in a similar way? Intuitively, it is clear that (1A) is more robust than (1B), whereas it's not easy to draw a similar conclusion for (2A) and (2B).

We can assume that such difference(s) might reflect the *strength* of alternatives: since the alternatives for the (A) inferences are informatively stronger than the alternatives for the (B) inferences, we can say that the former are more likely to be derived than the latter. The current paper investigates this hypothesis and gives evidence for several underlying factors. One is that only available scales show significant differences in strength between their alternatives. Amongst those scales, one can distinguish between scales with an end-point and those without it. The strongest alternative that denotes an end-point of a scale tends to be negated at a higher rate than the others. This does not occur in scales without end-points. The other factor, being true for available scales, is the *semantic similarity* of the scalar alternatives. The results of the current paper show that the less semantically similar alternatives are, the more likely an SI is derived.

The paper is structured as follows. In section 1, I present the traditional view, according to which, the safest bet is to negate the strongest alternative. Sections 2 and 3 discuss two experiments that investigated the strength hypothesis with an explanation in terms of semantic similarity of scalars. Section 4 summarizes the main findings of the paper.

1 Strength: theoretical predictions

To the best of my knowledge, the idea that SIs of one scale might not be alike was first noted by Horn (Horn, 1972:90). One suggestion concerned the specificity of the strongest alternative for scales with end-points. It states that the negation of the strongest alternative *must* be inferred by the hearer, whereas the negation of a weaker (non-strongest) alternative *may* be inferred but need not be. To illustrate this, (1A) must be inferred, while (1B) may be inferred. The other suggestion was that the stronger an alternative is, the more likely an SI is derived. In other words, it is safer to conclude (1A) than to conclude that not many of the books on the shelf contain pictures since σ[most] entails σ[many] and, therefore, σ[most] is stronger than σ[many], where σ is an utterance that contains a scalar item.

Chierchia in (Chierchia, 2004:16) made a similar statement with respect to the first Horn's suggestion. Considering examples in (3) he states, "... we often do not consider all of the possible values of a given scale" (ibid.). For example, (3) clearly has (3B) as an implicature. Whether it also implicates (3A) will depend on what is specifically at issue in the context.

(3) Some students will do well.
A. Not many students will do well.
B. Not every student will do well.

(3A) is selected only if it becomes contextually most relevant out of the scale ⟨*some, many, most, every*⟩. As Chierchia states, in absence of information to the contrary, it is often safest to imply negation of the strongest alternative, e.g. (3B).

[1] Availability is understood as retrieval of a word from the mental lexicon.

The other more general assumption that the stronger an alternative is, the more likely an SI is derived may be formulated as the *strength hypothesis* in the following way:

Given a set of alternative assertions a speaker could have made (which contain scalar expressions $\langle\phi, \chi, \psi\rangle$, where $\phi\supset\chi\supset\psi$), if she chooses an assertion with ϕ, a hearer is more likely to infer that an assertion with ψ does not hold than to infer that an assertion with χ does not hold.

Section 2 presents an experiment designed to test this hypothesis.

2 Experiment 1

The study was designed as an inference task with a 10-points Likert scale. To clarify, the inference paradigm used in psychology of reasoning presents participants with statements followed by questions (see Figure 1). According to Zevakhina and Geurts (2011), this sort of a paradigm is not a panacea for making scalar alternatives relevant and, consequently, for boosting the rates for SIs, as it has been assumed in Geurts and Pouscoulous (2009). We received very low rates for some of the adjectival predicates and concluded from this that they were not sensitive to the paradigm. However, there are several reasons for why I employed the same task in the current study. The first reason is that a well-known alternative, i.e. verification paradigm[2], is not appropriate to test a relatively considerable amount of scalars (I tested 14 scales in this experiment). The other reason is that instead of “yes”/“no” dichotomy in Zevakhina and Geurts (2011), I used a 10-points Likert scale, which might give different results in combination with the very same paradigm. Last but not least, the paradigm allows me to be consistent with Zevakhina and Geurts (2011) and to compare results of both studies.

2.1 Participants, materials and procedure

Participants

I posted questionnaires for 99 subjects (mean age: 34; range: 18-76; 58 females) on Amazon Mechanical Turk (henceforth, AMT). Respondents were paid $0.6 for their participation. Three participants were excluded from the analysis, either because they were not native speakers of English or because they returned incorrect responses for over 75% of the filler trials.

Preliminary study

The materials were preliminarily selected. The goal of the preliminary study was to select entailment scales. Critical sentences asserted a stronger expression and negated a weaker one (e.g., *The weather is hot but not warm*). 75 native speakers of English recruited via AMT had to judge how natural the sentences sounded, choosing one of the 5 possible ratings from “1" (“very unnatural") to “5" (“very natural"). If participants evaluated sentences as unnatural or very unnatural, it suggests that the items constitute a scale because it is logically inconsistent to assert a subset and negate a superset. There were 3 surveys (with 25 participants assigned to each) and the critical materials didn’t overlap. Fillers were either definitely true or definitely false sentences (e.g., “The grass is green but not red”, which is true, and “Six plus nine is eighteen but not fifteen”, which is false). Each participant answered a total of 55 sentences (16 critical and 39 filler sentences).

[2] The verification paradigm used, for example, in Bott and Noveck (2004) and Larson et al. (2009) presents participants with statements that have to be evaluated according to the world knowledge or a given context.

The results were split into two groups on the basis of the overall mean for the critical items (*M*=2.12). If the mean of a pair was below this value, the pair was accepted. That is, for example, the means for the three adjectival pairs out of the scale ⟨*pretty, beautiful, gorgeous*⟩ were lower than this value and, hence, were accepted, whereas the means for the three adjectival pairs out of the scale⟨*competent, talented, gifted*⟩ were higher than that and were rejected. Using this method, 14 critical items were selected for the main experiment.

Materials

Figure 1 gives an example of a critical item. In each trial, a statement uttered by a character John contained a scalar expression. A question that followed the statement contained a stronger scalar expression. Participants had to choose one of the 10 possible options, answering the question about the statement. The range of possible answers (from “1" to “10") enabled participants to make their choice certain to a lesser or greater extent.

John says:
The water in the lake is cool.

Would you infer from this that,
according to John, the water in the lake is not cold?

	○	○	○	○	○	○	○	○	○	○	
definitely not	1	2	3	4	5	6	7	8	9	10	definitely

Figure 1: Example of an item used in Experiment 1

The study tested 6 triples of adjectives, 1 triple of nouns, 2 triples of verbs, 3 triples of quantifiers, and 2 triples of epistemic modals. Figure 2 contains all the critical items used in the study.

I included 26 filler contexts designed identically to the critical items. Fillers comprised sentences and questions about them, which were either clearly valid or not valid. The following examples illustrate the two cases:

(3) The book is interesting. ⇒ The book is not boring.
(4) Peter has at least five jeans. ⇏ Peter doesn’t have more than five jeans.

Procedure

Participants received the following instructions:

In the following you will see 40 sentences and questions about what can be inferred from them. We ask you to give one of the ten possible ratings for a question. The ratings range between “I would definitely infer this” and “I would definitely not infer this”. For example, if somebody says that he bought more than 5 books, then it is quite likely that he bought 6 books. We are interested in your first intuition, so please don’t think too long about it.

Three questionnaires were counterbalanced and contained different pairs of scalar items. For example, for the triple scale ⟨*cool, cold, freezing*⟩, ⟨*cold, freezing*⟩ occurred in the first questionnaire, ⟨*cool, cold*⟩ in the second questionnaire, and ⟨*cool, freezing*⟩ in the third one. 33 participants were assigned to each questionnaire. Each participant responded to a total number of 40 sentences (14 critical items and 26 fillers). The experiment took approximately 8.5 minutes to complete.

2.2 Results

Figure 2 shows the results. It basically distinguishes between three classes of scales. Quantifiers and modals form the first class and are characterized with high rates for SIs with the strongest scale-mates and significant differences between the scalar pairs. Predicates (adjectives, nouns, verbs) constitute the other two classes and are distinguished from the above group on the basis of lower, if not very low, rates for SIs and lesser significant differences between the scalar pairs, if at all.

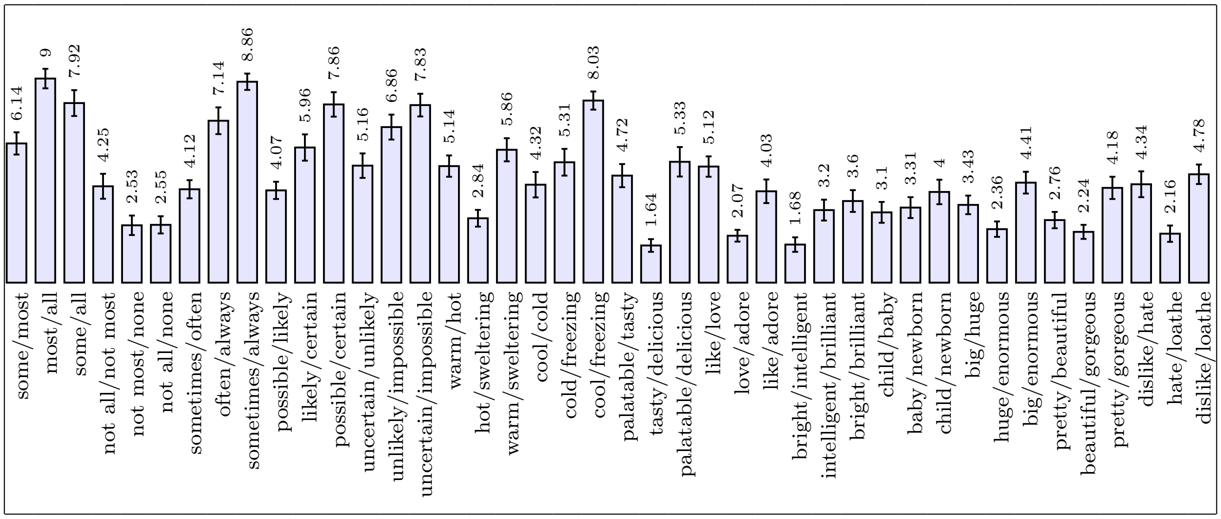

Figure 2: Rates for the critical items in Experiment 1

Statistical analysis supports dividing items into three classes. Starting out with the first class of quantifiers and modals, Tukey's HSD post hoc test showed that

- the mean for ⟨*some, most*⟩ (*M*=6.14, *SD*=0.49) was significantly lower than the mean for ⟨*some, all*⟩ (*M*=7.92, *SD*=0.57) and the mean for ⟨*most, all*⟩ (*M*=9, *SD*=0.43), all *p*'s<.05 (*F*(2,78)=8.775, *p*<.001);
- the mean for ⟨*sometimes, often*⟩ (*M*=4.12, *SD*=0.4) was significantly lower than the mean for ⟨*sometimes, always*⟩ (*M*=8.86, *SD*=0.36) and the mean for ⟨*often, always*⟩ (*M*=7.14, *SD*=0.59), all *p*'s<.05 (Welch's *F*(2,51)=37.92, *p*<.001[3]).

However, there were no significant differences between the negative quantifiers. A possible explanation might be double syntactic negatives of the forms "... not not all..." and "... not not most...", which were used in the questions. They might have yielded difficulties for participants, many of which complained about them in the comments after the questionnaire. As for the modals,

[3] Welch's ANOVA is reported if the within-group variances are significantly different on Levene's test.

- the mean for ⟨*possible, likely*⟩ (*M*=4.07, *SD*=0.38) was significantly lower than the mean for ⟨*possible, certain*⟩ (*M*=7.86, *SD*=0.54) and ⟨*likely, certain*⟩ (*M*=5.96, *SD*=0.58), which, in its turn, was significantly lower than the mean for ⟨*possible, certain*⟩, all *p*'s<.05 (*F*(2,79)=14.838, *p*<.001);
- the mean for ⟨*uncertain, unlikely*⟩ (*M*=5.16, *SD*=0.54) was significantly lower than the mean for ⟨*uncertain, impossible*⟩ (*M*=7.83, *SD*=0.5), *p*<.05 (*F*(2,79)=6.2, *p*<.01).

The second group comprises temperature, taste adjectives, and *like* verbs:

- the mean for ⟨*hot, sweltering*⟩ (*M*=2.84, *SD*=0.37) was significantly lower than the mean for ⟨*warm, sweltering*⟩ (*M*=5.86, *SD*=0.51) and the mean for ⟨*warm, hot*⟩ (*M*=5.14, *SD*=0.47), all *p*'s<.05 (*F*(2,79)=11.158, *p*<.001);
- the mean for ⟨*cold, freezing*⟩ (*M*=5.31, *SD*=0.59) was significantly lower than the mean for ⟨*cool, freezing*⟩ (*M*=8.03, *SD*=0.4), which, in its turn, was significantly higher than the mean for ⟨*cool, cold*⟩ (*M*=4.32, *SD*=0.56), all *p*'s<.05 (Welch's *F*(2,51)=16.607, *p*<.001);
- the mean for ⟨*tasty, delicious*⟩ (*M*=1.64, *SD*=0.28) was significantly lower than the mean for ⟨*palatable, tasty*⟩ (*M*=4.72, *SD*=0.51) and the mean for ⟨*palatable, delicious*⟩ (M=5.33, SD=0.65), all *p*'s<.05 (Welch's *F*(2,45)=22.183, *p*<.001);
- the mean for ⟨*love, adore*⟩ (*M*=2.07, *SD*=0.26) was significantly lower than the mean for ⟨*like, love*⟩ (*M*=5.12, *SD*=0.45) and the mean for ⟨*like, adore*⟩ (*M*=4.03, *SD*=0.52), all *p*'s<.05 (Welch's *F*(2,47)=18.881, *p*<.001).

Finally, there were no significant differences between the remaining items, i.e. between pairs out of the scales ⟨*bright, intelligent, brilliant*⟩, ⟨*pretty, beautiful, gorgeous*⟩, ⟨*big, huge, enormous*⟩, ⟨*child, baby, newborn*⟩, and ⟨*dislike, hate, loathe*⟩.

2.3 Discussion

The results of the study showed that the strength hypothesis was only moderately confirmed. Quantifier and modal scales received significantly higher rates for $\langle \phi, \psi \rangle$ than for $\langle \chi, \psi \rangle$, where $\phi \supset \chi \supset \psi$. However, it is noteworthy that scalar pairs $\langle \phi, \psi \rangle$ contained the strongest alternatives. Therefore, probably the difference might be due to the specialty of the strongest alternatives rather than the degree of strength itself. Hence, the strongest alternatives fit the predictions of Chierchia and Horn except for that they *tend* to be negated rather than *must* be.

Other scales, except for ⟨*cool, cold*⟩ and ⟨*cool, freezing*⟩, did not show significant results and, thus, did not confirm the strength hypothesis. In other words, the rates for $\langle \phi, \psi \rangle$ were comparable to the rates for $\langle \phi, \chi \rangle$, where $\phi \supset \chi \supset \psi$. Unlike quantifiers and modals, all these scales lack end-points and, consequently, do not have strongest alternatives. One could not name an end-point of temperature, taste, or size scale. The fact that the strength hypothesis was not confirmed by the scales without end-points supports the idea that the finding observed in quantifiers and modals was due to the specialty of the strongest alternatives rather than the strength hypothesis.

The other important issue concerns the results of some predicates: temperature, taste adjectives, and ⟨*like, love, adore*⟩ verbs. There were remarkable differences between the pairs of intermediate and stronger alternatives and the pairs of weak and stronger alternatives. To formalize, the rates for $\langle \phi, \psi \rangle$ were significantly higher than the rates for $\langle \phi, \chi \rangle$, where $\phi \supset \chi \supset \psi$. A likely explanation for that is *semantic similarity* of the items. It seems that participants took into account how close, or how similar, the meanings of the items are to each other. For example, in case of ⟨*hot, sweltering*⟩, the meaning of *sweltering* has to be close to the meaning of *hot*. Since there are no other reasons to deny the alternative with *sweltering*, the hearer merely accepts it, that is, he does not derive an implicature. On the contrary, *warm* and *sweltering* are much less similar. Taking that into account, the hearer derives an implicature. This may be formulated as the *similarity hypothesis* in the following way:

The hearer takes into consideration semantic similarity, or closeness, of scalars' meanings. If they tend to be close, an SI is less likely to be derived. If they tend to be distant, an SI is more likely to be derived.

The similarity hypothesis predicts a negative correlation between the derivation of SIs and similarity of scalar items. This is what the next experiment is aimed at.

Before moving to the discussion of the second experiment, one peculiar observation needs to be discussed. Strength and similarity seem to affect scales only if the scale is available. Highly available scales seem to lead to high rates of SIs. Intuitively it is very clear: if scalar alternatives are available, the strength of them and/or similarity between them might be more easily detected.

Quantifiers and modals ⟨*some, most, all*⟩ received high rates for SIs: most of the rates for them were higher than “6". Their scalar alternatives seem to be relevant and, therefore, available. One possible reason for this might be the closed and small set of quantifier expressions in natural languages, which allows alternatives of roughly the same complexity. For example, such alternatives for an English sentence containing *some* σ[some] are, basically, σ[many], σ[most], and σ[all]. Therefore, if a speaker uttered σ[some], it's not cognitively difficult to consider a few alternatives. I believe the same reasoning applies for modals.

Temperature and taste adjectives as well as ⟨*like, love, adore*⟩ verbs constitute the second group. I found lower rates of SIs derivation: most of the rates were lower than “5”. The alternatives derived in this group seem to be less relevant and only relatively available since the items seem to constitute open sets rather than closed ones.

The third group of scalars received very low rates and seem to be unavailable. It did not show any significant differences between the items. Because of their unavailability, one cannot observe neither the strength of alternatives, nor the semantic similarity between them.

3 Experiment 2

The goal of the second experiment was to test the similarity hypothesis. The prediction was that there is a negative correlation between similarity of scalars and the derivation of SIs: the more similar two scalars are, the less likely the SI will be derived. The study was designed as a comparative task with the 10-points Likert scale.

3.1 Participants, materials and procedure

Participants

99 subjects (mean age: 32.7; range: 18-88; 57 females) were recruited via AMT. Respondents were paid $0.5 for their participation. Eleven participants were excluded from the analysis, either because they were non-native speakers of English or because they submitted incorrect responses for over 75% of the fillers.

Materials

The materials were identical to those used in the previous study.

Figure 3 illustrates the critical pair ⟨*love, adore*⟩. Participants had to judge how similar the meanings of two words are, choosing one of the 10 possible rates which ranged from "very dissimilar" to "very similar".

love, adore

	○	○	○	○	○	○	○	○	○	○	
very dissimilar	1	2	3	4	5	6	7	8	9	10	very similar

Figure 3: Example of a critical item (experiment 2)

Fillers were designed similarly to critical items but differed substantially. They were split into several groups in order to cover a whole range of possible rates. The first group comprised the words that have the same hyperonym (e.g., *football, tennis*). The second group consisted of the words with the place/institution-person relation (e.g., *university, professor*). The third group was constituted by the words with the part-whole relation (e.g. *pocket*, *coat*). The three filler groups triggered associations between the items. Therefore, they had to be evaluated positively (from "6" to "10").

The fourth group was formed by items, which are associated with each other, though the associations are not so solid as in the previous groups. For example, *train* and *caterpillar* were characterized as long entities with respect to the length standards of transport and insects but not with respect to each other. The prediction was that generally, they would be evaluated negatively (from "1" to "5"). The last group comprised the words that showed no similarity at all (e.g. *chair* and *carrot*)[4]. Such pairs were predicted to receive very low rates.

Procedure

Participants received the following instructions:

In the following you will see 40 pairs of words. In each case, we ask you to judge how similar two words are, choosing one of the possible ten ratings. The ratings range from "the two words are very dissimilar" to "the two words are very similar". For example, a circle is pretty similar to an oval but rather dissimilar to a triangle. We are interested in your first intuition, so please don't think too long about it.

Three questionnaires were constructed in a similar way as in experiment 1. 33 participants were assinged to a questionnaire. Each participant responded to a total number of 40 pairs of words (14 target items and 26 fillers). The experiment took approximately 4 minutes to complete.

[4] This pair does begin with the same letter but I think people did not pay much attention to that because the task to evaluate the similarity of words means, first of all, semantic similarity rather than structural or phonetic one.

3.2 Results

Figure 4 shows the results of the second study. For quantifiers and modals, Tukey's post hoc test showed significant differences between the pairs of weak and strongest scalars and the pairs of weak and intermediate ones:

- the mean for ⟨*some, all*⟩ (*M*=4.06, *SD*=2.43) was significantly lower than the mean for ⟨*some, most*⟩ (*M*=5.9, *SD*=2.29), $p<.05$ ($F(2,84)=16.346$, $p<.01$);
- the mean for ⟨*sometimes, always*⟩ (*M*=3.76, *SD*=2.75) was significantly lower than the mean for ⟨*sometimes, often*⟩ (*M*=6, *SD*=2.45), $p<.05$ ($F(2,85)=17.504$, $p<.01$);
- the mean for ⟨*possible, certain*⟩ (*M*=5.72, *SD*=2.25) was significantly lower than the mean for ⟨*possible, likely*⟩ (*M*=8.13, *SD*=1.9), $p<.05$ ($F(2,85)=9.652$, $p<.01$);
- the mean for ⟨*uncertain, impossible*⟩ (*M*=5.23, *SD*=2.66) was significantly lower than the mean for ⟨*uncertain, unlikely*⟩ (*M*=8.34, *SD*=1.84), $p<.05$ ($F(2,85)=20.054$, $p<.01$).

Therefore, if there is a scalar triple $\langle\phi,\ \chi,\ \psi\rangle$ where $\phi\supset\chi\supset\psi$, the meaning of χ is closer, or more similar, to the meaning of ϕ than the meaning of ψ is to the one of ϕ.

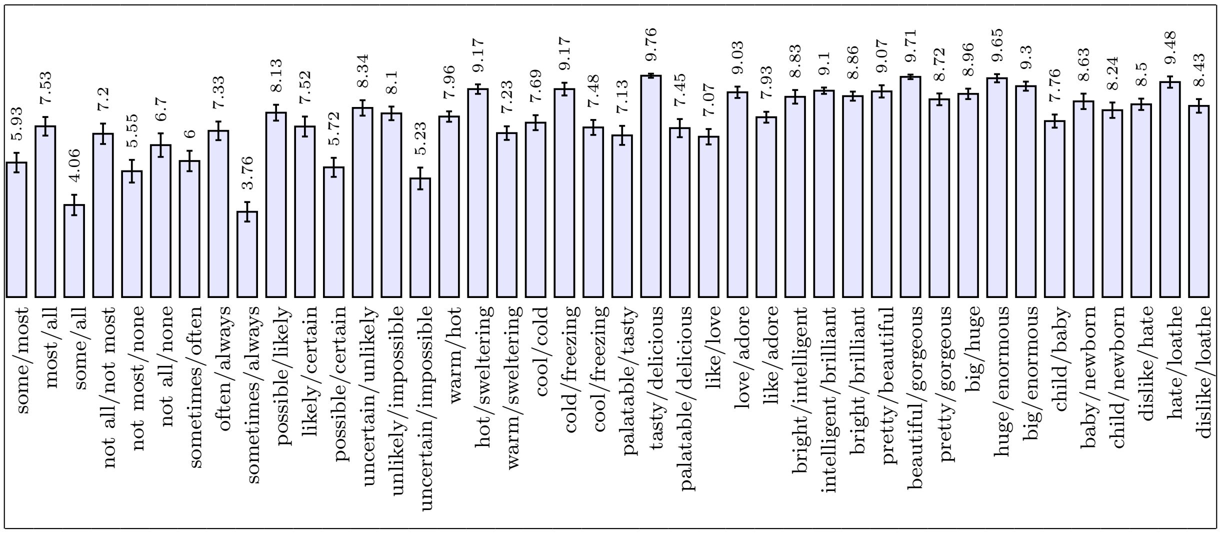

Figure 4: Rates for the critical items in experiment 2

Temperature, taste adjectival and *like*, *dislike* verbal predicates did not reveal significant differences between the pairs of weak and stronger scalars and the pairs of weak and intermediate ones. However, they did show significant differences between other pairs:

- the mean for ⟨*hot, sweltering*⟩ (*M*=9.17, *SD*=1.17) was significantly higher than the mean for ⟨*warm, hot*⟩ (*M*=7.96, *SD*=1.24) and the mean for ⟨*warm, sweltering*⟩ (*M*=7.23, *SD*=1.69), all *p*'s$<.05$ ($F(2,85)=14.572$, $p<.01$);
- the mean for ⟨*cold, freezing*⟩ (*M*=9.17, *SD*=1.46), was significantly higher than the mean for ⟨*cool, cold*⟩ (*M*=7.69, *SD*=1.79) and the mean for ⟨*cool, freezing*⟩ (*M*=7.48, *SD*=1.76), all *p*'s$<.05$ ($F(2,85)=8.87$, $p<.01$);

- the mean for ⟨*tasty, delicious*⟩ (*M*=9.76, *SD*=.51) was significantly higher than the mean for ⟨*palatable, tasty*⟩ (*M*=7.13, *SD*=2.33) and ⟨*palatable, delicious*⟩ (*M*=7.45, *SD*=2.18), all *p*'s<.05 (Welch's *F*(2,42)=31.477, *p*<.01);
- the mean for ⟨*love, adore*⟩ (*M*=9.03, *SD*=1.35) was significantly higher than the mean for ⟨*like, love*⟩ (*M*=7.07, *SD*=1.8) and the mean for ⟨*like, adore*⟩ (*M*=7.93, *SD*=1.31), all *p*'s<.05 (*F*(2,84)=12.301, *p*<.01).

Finally, the rest of the predicates showed no significant differences between the items.

3.3 Semantic similarity and availability

Surprisingly, similar to experiment 1, one can distinguish between three classes of items with respect to the rates they received and significant differences between them.

The first class of quantifiers and modals showed relatively low rates for the pairs of weak and strongest scalars (on average "4"-"5") and relatively high rates for other pairs (on average "7"-"8"). The pairs significantly differed from each other. Therefore, the scalars may not be quite similar to each other.

Unlike them, the rest of the items received much higher rates for all the pairs (on average "7"-"9"). Therefore, the meanings of their scalars are much more similar to each other than those of quantifiers and modals. Amongst them, there are predicates (temperature, taste adjectives, and *like*, *dislike* verbs) that showed significant differences between the pairs. They resemble quantifiers and modals with this respect. The rest of the predicates did not show significant differences.

A possible explanation for observing the same patterns in both studies might be availability of scalars. Significant differences between the scalars suggest that scalars are available and participants could evaluate the degree of similarity between items. The lack of significant differences on a par with high rates suggest that items are unavailable and, thus, participants were unaware of what exactly they had to compare.

3.4 Correlation test

The results of the second study confirm the similarity hypothesis. For most cases, the more similar the items are, the less likely an SI will be derived. Vice versa, the less similar the items are, the more likely an SI will be derived. The prediction of an inverse relationship between the derivation of SIs and the similarity between items was fulfilled. Being relatively strong, Pearson correlation coefficient supports the finding: *r*=-.65, *p*<.001.

There are some outliers for the general trend, which are visible in the figure. For example, the rate for ⟨*most, all*⟩ was high and, consequently, the scalar items *most* and *all* seem to be very similar to each other. However, in the first study, the rate for the SI was high, as well. The described and few other cases are consequences of the intervening 'end-point' factor. The alternative with *all* is the strongest alternative for the utterance which contains *most*. According to the results of the first study, the strongest alternative tends to be negated to a much higher rate than others, regardless how semantically similar the scalars are to each other.

4 Conclusion

The paper showed that the variation within scales exists empirically. However, differential strength of alternatives, which has been acknowledged by the traditional view, seems to be a moderate factor at best. There are two other factors that can affect variation within scales. One is that if a scale has an end-point, there is a strong preference for negating the strongest alternative. The other is semantic similarity of scalars, which also influences the derivation of SIs due to the highly significant and strong correlation. It was evidenced that scale's availability is a necessary condition for both factors to be fulfilled. This agrees with Zevakhina and Geurts (2011).

The results of the experiments contribute to the general theory of Q-implicatures. First, they allow to reveal a fine-grained connection between SIs and other Q-implicatures in terms of availability. According to (Geurts, 2010:128), Q-implicatures might be split into two categories: Q_c- and Q_o-implicatures. The former involves a closed set of alternatives whereas the latter does not have a determinate set of alternatives. As a consequence, Q_c-alternatives are available and play a key role in the derivation of implicatures, whereas Q_o-alternatives are less available, if at all. The data presented in this paper argue for that SIs can be either Q_c- or Q_o-implicatures. Modals and quantifiers fall into Q_c-implicatures whereas adjectives, nouns, and verbs belong to the class of Q_o-implicatures.

Second, the results suggest to adopt *information preservation principle* (IPP):

Among the available alternatives choose the one that allows you to preserve information (if there is no other reason not to do this).

IP captures the specificity of the strongest alternatives and semantic similarity of scalars. The strongest alternatives are only revealed in the Q_c-class and their negation allows to preserve maximal information. For example, in case of quantifiers, the negation of the strongest alternative that contains *all*, preserves more information than the negation of other quantifier alternatives with *most* or *many*.

However, the Q_o-class lacks strongest alternatives but can involve semantic similarity that also preserves information, if a scale is relatively available. For example, in case of adjectives ⟨*palatable, tasty, delicious*⟩, *palatable* is not very similar to *tasty* and *delicious* whereas the latter ones are similar to each other. Thus, for an utterance with *palatable* the implicature with *tasty* and the implicature with *delicious* are derived at comparable rates since the same amount of information is preserved in both cases. It does not hold for the *delicious* and *tasty*. Since they are very similar to each other, there seems no sense to derive the implicature with *delicious* for the statement with *tasty*.

Finally, if a scale of the Q_o-class is unavailable, it doesn't show semantic similarity (cf. ⟨*big, huge, enormous*⟩).

References

Bott, Lewis, and Ira A. Noveck. 2004. Some utterances are underinformative: the onset and time course of scalar inferences. *Journal of memory and language* 51:437–457.

Chierchia, Gennaro. 2004. Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In *Structures and beyond*, ed. Adriana Belletti, 39–103. Oxford University Press.

Geurts, Bart. 2010. *Quantity implicatures*. Cambridge University Press.

Geurts, Bart, and Nausicaa Pouscoulous. 2009. Embedded implicatures?!? *Semantics and pragmatics* 2:1–34.

Horn, Laurence R. 1972. On the semantic properties of the logical operators in English. Doctoral Dissertation, University of California at Los Angeles.

Larson, Meredith, Ryan Doran, Yaron McNabb, Rachel Baker, Matthew Berends, Alex Djalali, and Gregory Ward. 2009. On the non unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics* 1:211–248.

Levinson, Stephen C. 2000. *Presumptive meanings*. Cambridge, Massachusetts: MIT Press.

Zevakhina, Natalia, and Bart Geurts. 2011. Scalar diversity. *Paper submittion* .