Is *good* better than *excellent?* An experimental investigation on scalar implicatures and gradable adjectives

Andrea Beltrama (University of Chicago)
Ming Xiang (University of Chicago)

Abstract – In this study we look at the computation of scalar implicatures with adjectives and modals, and show that scalar inferences are not equally triggered by these scales. In Experiment 1, we test implicatures on three-place adjective scales such as <*decent, good, excellent*> and find that only adjectives at the low end of the scale generate implicatures, while those in the middle fail to do so. In Experiment 2 we test implicatures on the modal scale <*possible, likely, certain*>, and show that, contrary to what happens to adjectives, both modals at the low end and in the middle of the scale generate implicatures. We interpret these results as suggesting that adjective scales are qualitatively different from modal scales. What distinguishes them, in particular, is the nature of the extreme term on the scale. While extreme adjectives do not constitute a real scalar maximum and make use of different semantic representations from their non-extreme counterparts, extreme modals are real scalar maximums and are semantically homogenous with respect to the other terms on the scale. We conclude by providing two alternative explanations of how the properties of extreme adjectives might inhibit the calculation of the inferences.

Keywords. Scalar Implicatures, adjectives, gradability, modals, scales, Italian

1. Introduction

Scalar implicatures (SIs) represent a well-known example of how pragmatic inferences can affect the semantic interpretation of an utterance. However, the way in which such inferences are licensed by different kinds of scale structures is not fully understood. In this study we demonstrate that not all scales are equally likely to generate SIs. In particular, we show that adjectives trigger SIs less easily than other scales, especially with respect to the SI inferences on *extreme adjectives* (also see 2.2). As an explanation, we suggest that extreme adjectives are crucially different from the extreme terms on other scales. We test this hypothesis by designing two experiments. In Experiment 1 we test implicatures on three-place adjective scales and, consistently with previous investigations, find that adjectives in the middle of scale fail to generate scalar implicatures. In Experiment 2 we compare implicatures on adjective scales with implicatures on modal scales, and show that, contrary to what happens to adjectives, modals in the middle of the scales do generate implicatures. We argue that these results support the idea that extreme adjectives are qualitatively different from other extreme scalar terms, and we conclude by providing two alternative explanations of how extreme adjectives might inhibit the calculation of the inferences.

2. Background

Implicatures triggered by the Quantity Maxim, also known as *scalar implicatures (SIs),* have received a significant amount of attention in the literature. An example is reported in (1).

(1) Some people left the party

In the sentence, the logical content merely says that at least (and possibly all) some people left the party. What is inferred, however, corresponds to an enriched, upper-bounded interpretation, paraphrasable as "some people left the party, but some people didn't". The inference that not all people left is based on the assumption that the speaker is being cooperative and is respecting the conversational maxims, including the Quantity Maxim ("make your contribution as informative as is required", Grice 1967:26). As a consequence, the speaker is expected not to withhold information that he would be in a position to provide. For example, if the speaker *knows* that all people left the party, he is expected to *say* that all people left the party. By the same token, if the speaker utters (1), it is inferred that he cannot truthfully say that all people left the party, and therefore that not all people left the party. Hence, the interpretation of the quantifier is more restrictive.

A particularly influential model to capture the mechanism underlying this kind of inference has been proposed by Horn (1984). According to him, scalar implicatures are triggered by *entailment scales,* which are total orderings of terms in which each expression entails the weaker terms and implicates the negation of the stronger terms. Such a mechanism provides a unified account to capture the observation that many different kinds of scales in natural language give rise to SIs. Examples of such scales include quantifiers (<*some, many, all*>), adverbs (<*sometimes, often, always*>), modals (<*possible, likely, certain*>) and gradable adjectives (<*warm, hot*>), as in (2-4).

(2) It's warm.
(3) It's possible that it will rain tomorrow.
(4) Sometimes, the weather is sunny down here.

For (2), the logical content is that "it is warm and possibly *hot*", whereas the upper-bounded interpretation added by the implicature is that it is warm *and not* hot. For (3), the logical content is that "it's possible and perhaps certain that it will rain tomorrow", whereas the upper-bounded interpretation is that it is possible *and not* certain that it will rain tomorrow. For (4), the logical content is that "the weather is sometimes, and possibly always, sunny down here", whereas the reading generated by the implicature is that only sometimes, *and not* always, is the weather sunny.

While Horn's theory predicts that the calculation of implicatures happens uniformly across different scales, recent experimental investigation suggests that not all scales induce these inferences in the same way. In particular, scalar inferences on adjectives turned out to be much harder to trigger than those on other scales (Doran et al. 2008, Zevakhina and Geurts 2011). The aim of the current study is to cast further light on this asymmetry and to provide an explanation to it.

2.1 Previous experiments

Doran et al. (2008) looked at whether different kinds of scales – cardinal numbers, quantifiers, ranked orderings and gradable adjectives - would be processed in the same way with respect to scalar inferences. One of the major experimental manipulations was whether

alternative values on a scale were explicitly provided in the context[1] or not. As far as adjectives are concerned, the stimuli consisted of triplets of terms from the same scale, but of different strength, such as *<huge, big, average>*. The subjects were asked to read dialogues between two characters: Irene and Sam, as in (5), in which Irene asks Sam two different kinds of questions. In condition (a) Irene asked a neutral question without evoking any scalar term; whereas in condition (b) Irene's question explicitly provided the three scalar terms as possible answers. In his response to both questions, Sam always utters an under-informative description of the fact by using a middle adjective from a given scale, (e.g., *big* here). Sam's answer is always under-informative because the ultimate FACT, which is given after this mini-conversation, portrays an "extreme" situation that requires a stronger term from the scale, rather than the middle one. In this particular example, given the fact that Jeremy cannot fit into an airplane seat, a supposedly more adequate description of his size should be *He is huge*, rather than *He is big*.

(5) *<average, big, huge>*

Irene: a) What size is Jeremy? / b) Is Jeremy average, big or huge?
Sam: He's big.
FACT: Jeremy can't fit in an airplane seat.
Is Sam telling the truth?

Given this scenario and the FACT, participants were asked to perform a Truth Value Judgment Task by answering the question "Is Sam telling the truth?". Given that the description is supposed to be under-informative, responding "no" (i.e. saying that Sam's answer is false) would signal that participants have calculated the SI *big but not huge*. A "yes" response, on the other hand, would indicate the lack of an SI. The results show that overall implicatures on adjectives were calculated significantly less frequently than for other scales. Moreover, contrary to what happened with other scales, the rate of inferences associated with adjectives significantly increased when alternative values were evoked (version (b) of Irene's question), showing that the interpretation of adjectives was significantly affected by the discourse condition manipulation. No other scale tested in the experiment showed this kind of sensitivity. The authors explain these findings by suggesting that adjective scales, compared to other entailment scales, are domain-dependent and have fuzzy boundaries. As a consequence, "non-maximal values can range indefinitely higher, without excluding stronger scalar values" (p. 228), making implicatures unnecessary. For example, *big* can range indefinitely high on a scale and extend to the region of the scale covered by *huge,* discouraging an upper-bounded reading (*big* but not *huge)*.

In a more recent investigation, Zevakhina and Geurts (2011) addressed similar issues with a different experimental paradigm. They tested implicatures across different scales via an inference paradigm, which have been shown in previous studies to give rise to higher levels of implicatures than truth value judgment tasks (Geurts and Pouscoulous 2009). In their procedure, a fictional character (John) made a statement containing a scalar term (e.g. *some*).

[1] A further manipulation was the perspective to be taken for answering the True/False question. Participants had to answer by taking their own point of view or the point of view of an imaginary character (Literal Lucy) that was famous for interpreting everything in a literal way. This further manipulation is not discussed here.

Participants were asked whether, according to John, the statement would have been false if *some* had been replaced with a stronger scale-mate, such as *all*. A "yes" answer would indicate that participants have generated a SI and assigned the scalar term an upper-bounded interpretation. A "no" answer, on the other hand, was taken to indicate that the implicature was not generated. This study tested quantifiers, modals and gradable adjectives. An example of gradable adjectives is given below (the target expressions have been italicized by us):

(6) <*good, excellent*>
John says: The movie was *good*.
"Would you infer from this that, according to John, the movie was not *excellent*?"

A comparison between quantifiers, modals and gradable adjectives suggests that adjectives trigger implicatures much less frequently than the other two types of scalar expressions. Rather than attributing this to a category specific effect, Zevakhina and Geurts (2011) argued that the degree to which SIs are calculated is determined by how accessible the (stronger) scalar alternatives are for any given scalar expression. In a follow up experiment, they showed that scalar expressions that trigger fewer SIs are also the ones to which it is more difficult for participants to provide scalar alternatives in a cloze procedure[2]. In their analysis, the difficulty of deriving SIs for adjectives is not adjective-specific, but rather due to the fact that scalar alternatives are in general less accessible for adjectives than for other scalar expressions (although it is not a perfect correlation, as shown by cases like "warm/cold" in this experiment). This explanation is crucially different from the one proposed by Doran et al., according to whom the low rate of implicatures on adjectives is due to two special properties of adjective scales: lack of clear boundaries between the terms, and lack of an upper boundary at the top of the scale.

2.2 Extreme adjectives

It is important to note that for both of these studies, what is essentially being tested on adjectives is whether the negation of a strong term on an adjective scale is (easily) available to speakers (e.g. *good* implicates *not excellent*; *big* implicates *not huge*). If we look at the expressions that are targeted by this inference, e.g. *excellent* or *huge*, we notice that they belong to the category of *extreme adjectives,* a natural class that has drawn a lot of attention in the semantics literature (Cruse 1986, Paradis 2001, Morzycki 2009 and 2011). This class generally includes expressions that seem to be located on the highest region of a scale, such as *excellent, huge, extraordinary, horrible,* and so on. As Morzycki (2009, 2011) points out, all these expressions display a series of particular properties that, critically, are not shared by non-extreme adjectives. First, they take their own class of modifiers (*simply, flat-out, downright, balls-out*, and more), which are not as acceptable with other non-extreme gradable adjectives (as in (7)). Second, they are resistant to explicit comparative constructions, as (8) shows. Third, they are resistant to intensification with *very* and other degree modifiers (9). Finally, compared to other extreme scalar terms, they are not ontological scalar maximums, in that they belong to scales that do not have an upper-bound. Because of this, it is always

[2] Subjects were presented a target sentence containing a scalar term, such as "The movie was good", and were requested to answer to the question "Which words could have occurred instead of the highlighted one?" This procedure provides a measure about the likelihood of a particular scalar-mate (e.g. "excellent" in this case) being considered by the participants.

possible to find individuals that rank even higher on the scale, while this cannot be done with quantifiers or modals, as in (10).

(7) a. simply gigantic/??big
b. just gorgeous/??pretty
c. flat-out crazy/??sane
d. flat-out excellent/??good

(8) a.? Godzilla is more gigantic than Mothra.
b.? Monkeys are less marvelous than ferrets.

(9) a. very ??excellent/ ??marvelous/ ??fantastic/ good
b. very ??gigantic/ big
c. How ??excellent/good is this performance?
d. How ??gorgeous/pretty is that girl?

(10) This thing is gigantic, but it could be bigger
?? All students came to class, but more of them could have come
?? It is certain that the event will took place, although it could be more certain

These observations show that extreme adjectives seem to have a somewhat different nature from other adjectives on the same scale (example (7)-(9) above) and from maximum terms on other scales (example (10)).

A natural question, at this point, is whether the properties in (7-10) might impact the behavior of extreme adjectives with respect to scalar inferences, and whether they could ultimately be responsible for the low rates of adjectival implicatures discussed in the experimental literature. We suggest two possible explanations, on which we will elaborate more in section 4. One possibility is that, contrary to the common view, only non-extreme adjectives are lexically gradable, while extreme adjectives are not. This would have crucial consequences for implicature calculation. Since, under this view, extreme adjectives would no longer be scale mates with non-extreme ones, it would follow that they would no longer be a valid scalar alternative to them. As a consequence, implicatures from non-extreme adjectives to extreme adjectives (such as *good* → *Not excellent*) are simply not computed.

The alternative explanation would be that extreme adjectives, while on the same scale as non-extreme ones, might just be particularly hard to access as alternatives for the computation. Their low accessibility could be due to the fact that, as Doran et al. pointed out, there are no definite boundaries on the scale between non-extreme and extreme adjectives. This makes it possible for the non-extreme adjectives to stretch all the way up to the top of the scale, making the inference unnecessary. We will elaborate more on these two possibilities in Section 4.

2.3 Current study

Although previous studies have consistently shown that SIs are in general less accessible for adjectives, a few issues remain unresolved. The first observation is that previous studies have

only looked at whether the negation of the extreme gradable adjectives can be computed based on those adjectives in the "middle", as in the case of computing an SI from *good* to *excellent* on a scale with multiple scalar terms <*decent, good, excellent*>. The current study extends the investigation to other scalar terms on the scale. Specifically, we will look at not only the SIs based on the middle-adjectives, but also those based on the lower terms on the same scale. The purpose of this manipulation is two-fold. First, we are interested in replicating the previous results that SIs from "middle" to "extreme" adjectives are difficult, even when some methodological concerns have been controlled (see below); and second, the current study also aims at addressing the question of whether scalar distance correlates with the size of the SI effect, i.e. whether triggers lower on the scale generate implicatures more often than triggers in the middle on the same scale.

Second, a methodological consideration is in order. Albeit in slightly different ways, both studies discussed above elicited SIs by providing explicit alternatives. Doran et al tested the *big* → *not huge* implicature by asking whether a certain individual was indeed "big" or "huge" (see (5b)). Zevakhina and Geurts also explicitly asked the subjects whether they would infer *not excellent* from *good* (see (6)). However, it is rarely the case that such alternatives are given in natural occurrences of the scalar terms. Providing them explicitly, therefore, could have forced the inferences to be drawn in an unnatural way (see also Zevakhina and Geurts 2011: p. 14 on this point). In our design, we therefore choose not to provide explicit alternatives under any condition.

Third, the impact of context on the calculation of implicatures was only tangentially taken into consideration in previous experiments. Therefore the role of context in SI calculation has not been carefully addressed. If calculating implicatures is about understanding the informativity of an utterance, then speakers should be sensitive to a context that reinforces or weakens the informativity of a statement (Horn 1984). In our experiment, we aim to verify whether manipulations of informativity and discourse context have an effect on SI calculations.

Finally, none of the previous studies explicitly discussed the category of *extreme adjectives* and its peculiar characteristics. However, in both studies the scalar alternatives that were supposed to be negated via implicatures were indeed instances of extreme adjectives, and displayed all the properties discussed in section 2.2. We suggest that such properties could be responsible for the low rates of SIs on adjectives. To provide further arguments for this, we will compare the behavior of extreme adjectives to the behavior of extreme expressions on other kinds of scales.

3. Experiment 1

As mentioned earlier, adjective scales are often composed of multiple terms with increasing strength and ordered along a continuum. As such, they contain at least two distinct triggers for scalar inferences. Given a three-place scale such as <*decent, good, excellent*>, the two potentially available inferences would be in (11).

(11) (i) *decent* (weakest trigger) → *not good/not excellent*
(ii) *good* (middle trigger) → *not excellent*

In Experiment 1, we test whether these two SI-triggers (i.e. the two lower terms on a given adjective scale) indeed both generate SIs. In addition, we will also test whether triggers lower on the scale (i.e. *decent*) generate implicatures more often than triggers in the middle of the scale (i.e. *good*). This would allow us to verify if there is an effect of scalar distance. Finally, Experiment 1 also assesses the interaction between SI computation and context effect—in particular, the extent to which context manipulations might alter the computation of implicatures. To achieve this, we presented our stimuli in two different discourse contexts, one of which was designed to stress the under-informativity of the adjective at stake (see 3.2.3).

3.1. Materials

Two factors were manipulated in the design: adjective strength and discourse context. As far as adjective strength is concerned, we created twenty-four adjective scales in Italian, each of which contained three adjectives of different strength: 1) a weak adjective; 2) a middle adjective; 3) a strong adjective. An example of such a scale from the experiment, already mentioned above, is <*decent, good, excellent*>. Other examples are given in (12).

(12) <*acceptable, satisfactory, heavenly*>
<*mediocre, bad, horrific*>
<*pretty, beautiful, gorgeous*>
<*edible, tasty, delicious*>
<*doable, easy, ridiculous*>
<*getting by, rich, billionaire*>

For each scale a two-sentence scenario was created. The first sentence was a statement containing an adjective in one of the three conditions ("weak", "middle", or "strong"). The second one portrayed a situation linked to the first one via the causative connector *that's why* (*per questo* in Italian). The second sentence is the same across all conditions. It always described a rather "extreme" situation. Its function was to force a construal of the first sentence as a cause, creating a context for which only the strongest adjective would be adequately informative, whereas the middle and the weak ones would result in under-informativity. For instance, in the example in (11), the second sentence in the scenario encourages the inference that since Harvard is very difficult to get into, Mark needs to be an excellent student in order to be accepted. If the SI "not excellent" is computed on "decent" or "good" in the first sentence, we expect to see degraded judgment for the overall scenario, and potentially increased reading times (RTs) at the critical word "Harvard", since this is the first word that would create difficulty for comprehension if the relevant SIs were computed earlier. Crucially different from previous studies, the current design did not explicitly provide scalar alternatives for people to choose from, nor did we ask for conscious judgments about the target SI inferences. If participants have spontaneously computed SIs at the adjectives in the first clause of each scenario, this should affect how they interpret the second clause. Finally, for each triplet of adjectives, we also added an adjective with the opposite polarity (e.g. *bad*, for the "goodness" scale). This last adjective always gives rise to contradiction in our stimuli, and we use it as a baseline condition to assess whether participants are doing the task correctly. An example is shown below in (13a-d).

The second factor we manipulated is discourse context. In four additional conditions that are otherwise identical to (11), we added an additional context sentence *The competition for entering top schools is very tough* (in Italian: *la competizione per entrare nelle scuole più prestigiose è estremamente serrata*). The purpose of this new information was to build a more constrained context for the interpretation of the target sentence (i.e. *That is why…*). This new context sentence was meant to reinforce the participants to reassess the informativity of each scenario more carefully, and to raise the threshold of informativity required to have an adequate description. The ultimate goal was to highlight the fact that any adjective different than the strongest one (that is, *decent* or *good*) would make the general scenario under-informative.

With the two factors, there are a total of 8 conditions for each scale, resulting from a 2 (context) x 4 (adjective strength) design. The eight conditions for each adjective scale were distributed into eight different lists with a Latin Square design, so that each list had 24 experimental sentences (together with an additional 30 fillers). A full example of an experimental item is given in (13):

(13)
Without the strengthened context:
a.Mark è uno studente **discreto**. Per questo è stato preso ad *Harvard* per un dottorato. (weak)
Mark is a ***decent*** student. That's why he has been accepted to ***Harvard*** for a Ph.D

b.Mark è uno studente **buono**. Per questo è stato preso ad *Harvard* per un dottorato. (middle)
Mark is a ***good*** student. That's why he has been accepted to ***Harvar*****d** for a Ph.D

c.Mark è uno studente **eccellente**. Per questo è stato preso ad *Harvard* per un dottorato. (strong)
Mark is an ***excellent*** student. That's why he has been accepted to ***Harvard*** for a Ph.D

d.Mark è uno studente **scarso**. Per questo è stato preso ad *Harvard* per un dottorato. (contradictory)
Mark is a ***bad*** student. That's why he has been accepted to ***Harvard*** for a Ph.D

With the strengthened context:

e. La competizione per le migliori università è molto dura. Mark è uno studente **discreto**. Per questo è stato preso ad *Harvard* per un dottorato. (weak)
The competition for entering top programs is very tough. Mark is a ***decent*** student. That's why he has been accepted to ***Harvard*** for a Ph.D

f. La competizione per le migliori università è molto dura. Mark è uno studente **buono**. Per questo è stato preso ad *Harvard* per un dottorato. (middle)
The competition for entering top programs is very tough. Mark is a ***good*** student. That's why he has been accepted to ***Harvar*****d** for a Ph.D

g. La competizione per le migliori università è molto dura. Mark è uno studente **eccellente**. Per questo è stato preso ad *Harvard* per un dottorato. (strong)

The competition for entering top programs is very tough. Mark is an ***excellent*** student. That's why he has been accepted to ***Harvard*** for a Ph.D

h. La competizione per le migliori università è molto dura. Mark è uno studente **scarso**. Per questo è stato preso ad *Harvard* per un dottorato. (contradictory)
The competition for entering top programs is very tough. Mark is a ***bad*** student. That's why he has been accepted to ***Harvard*** for a Ph.D

3.2 Participants, procedure and predictions

Forty-two native speakers of Italian between 18 and 40 years old participated in the experiment. Twenty-two subjects were graduate or undergraduate students affiliated with the University of Chicago, and twenty were students affiliated with an Italian university or already in possession of a B.A. diploma. Subjects were recruited through announcements posted online, personal email communications and word-of-mouth advertising. Subjects tested in Chicago were paid $ 10, while subjects tested in Italy were paid 5 Euros.

The experiment combined an acceptability rating task with a self-paced reading task and was run with Eprime. In each trial the first or the first two sentences (when there is an extra context sentence) appeared on the screen as a whole chunk, whereas the last sentence – the target sentence starting with "that's why" – was read word by word on the next screen at participants' own pace (i.e. participants pressed the space bar to move from one word to the next). After reading the last word of the target sentence, subjects were asked to provide an acceptability rating on the plausibility of the whole scenario. They were prompted to give their rating by the following instruction "How sensible is the second sentence given the first one?" (In Italian: *Quanto è sensata la seconda frase rispetto alla prima?*). On a 1 to 5 scale, 1 represents "completely nonsensical", and 5 "perfectly sensible". Participants were told at the practice session to evaluate each scenario on the basis of general common sense, and leave aside markedly stylistic interpretations.

As far as judgments are concerned, we anticipate degraded acceptability for scenarios that contain an under-informative adjective. That is, we predict the degree of implicatures to be inversely proportional to the acceptability rating. The more subjects draw scalar inferences on under-informative adjectives (weak and middle adjectives), the less acceptable these sentences should be. This output is relevant for assessing (i) whether both weak and middle adjectives are perceived as under-informative, and therefore triggered implicatures with respect to the strong ones; (ii) whether weak adjectives are perceived as more under-informative, and therefore trigger more implicatures than middle adjectives. Furthermore, if participants are sensitive to the enhanced under-informativity due to additional context, we expect to see even larger degradation in acceptability for the weak and middle adjective scenarios.

Since calculating SIs involves extra steps of processing, it is possible that SI computation might evoke extra processing cost. Whether such processing cost should emerge immediately in online measures is still a question under debate. There has been evidence to support both an immediate online effect (Nieuwland, Ditman & Kuperberg 2010) and a delayed effect (Bott and Noveck 2004). We predict that, if SIs are incrementally processed online, we

should observe longer reading times at the critical word (e.g. *Harvard* in example (13)) in sentences containing under-informative scalar terms.

3.3 Results

Out of the total of 42 subjects, data from three subjects were excluded, due to problems in comprehending the task. For the rest of the participants, as expected, both of the contradiction conditions (condition *d* and *h* in (13)) are judged significantly lower than the other conditions (ps<0.001), suggesting participants were successfully detecting the difference between contradiction and under-informativity. We therefore focused our analysis only on the rest of the conditions. The results reported below did not include conditions *d* and *h* (but they are still plotted in Figure 1). For the rest of the conditions, we carried out two-way ANOVAs to assess the effects of the two independent variables - strength of the adjectives and presence/absence of the context sentence. When a higher-level main effect of interaction is significant, we also carried out paired-comparisons between relevant conditions.

3.3.1 Acceptability judgments and online RTs

The acceptability results are plotted in Figure 1. There was a main effect of adjective strength (by-subject $F_1(2,76) = 31.7, p < .0001$; by-item $F_2(2,46) = 20.0, p < .001$) and a main effect of context (by subject $F_1(1,38) = 41.7$, p <.001; by item $F_2(2,46) = 28.2, p < .001$). There was no interaction between the two factors ($F_1(2,76) = 0.6, p > 0.5$; $F_2(2,46) = 0.6, p > 0.5$).

For the effect of adjective strength, follow-up paired comparisons showed that scenarios with weak adjectives were rated lower than those with middle adjectives, both without the context sentence ($t(38) = 5.2$, $p < .001$) and with the context sentence ($t(38) = 4.9$, $p < .001$). Scenarios with weak adjectives were also rated considerably lower than those with a strong adjective, both without the context sentence ($t(38) = 6.1$, $p < .001$) and without the context sentence ($t(38) = 4.8$, $p < .001$). On the other hand, no significant difference was found for the corresponding comparisons between middle and strong adjectives (all *ps* >.05). For the effect of context, paired t-tests showed that for all conditions, scenarios with context were significantly less acceptable than scenarios without context (for weak adjectives, $t(38) = 3.8$, $p < .0001$, for middle adjectives $t(38)=4.0$ and $p<.0001$, for strong ones, $t(38) = 5.4$, $p<.0001$.

The reading times at the critical word (i.e. “Harvard”) were plotted in figure 2. We found no significant main effect of strength or context, and no interaction effect between them (all *ps* > .1). No effect was found on the spill-over word either. The current data therefore does not lend any support to the idea that SI computations are carried out immediately online. But since no conclusions can be made based on a null result, we will not discuss this point further in this paper. More future research is called for to understand the online process of SI computations.

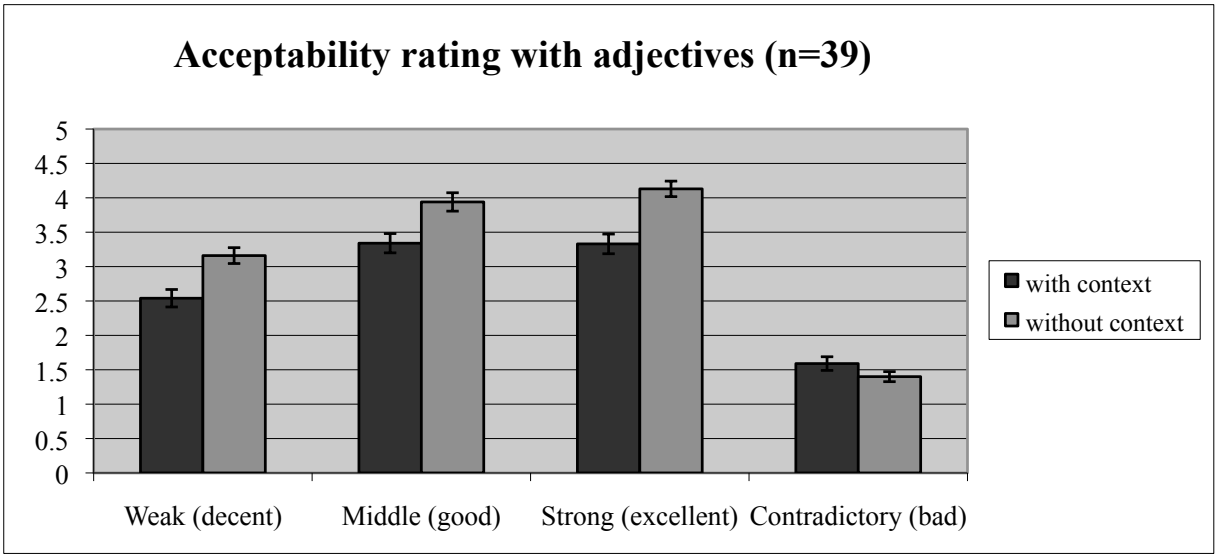

Figure 1: Acceptability judgments on a 1-5 scale for Experiment 1. The Y-axis indicates the average 1-5 acceptability ratings. The X-axis plots the four different adjective conditions, with an example enclosed in brackets. Dark grey bars stand for scenarios without the context sentence, and light grey bars stand for scenarios with the context sentence. Error bars indicate standard errors.

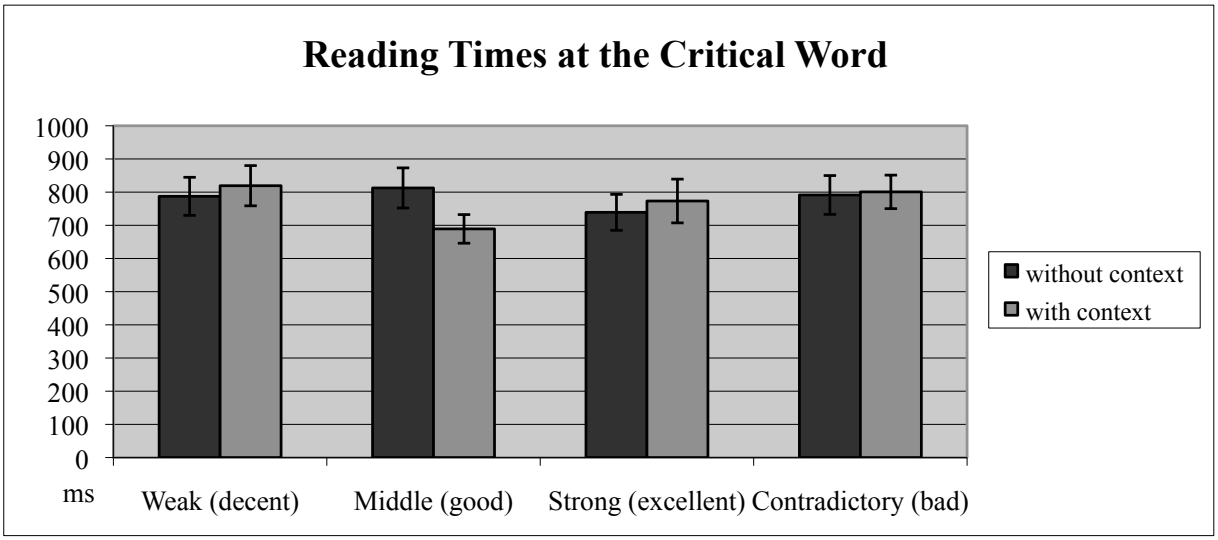

Figure 2: Reading times at the critical word for Experiment 1. The Y-axis indicates the average time in milliseconds. The X-axis plots the four different adjective conditions. Dark grey bars stand for scenarios without the context sentence, and light grey bars stand for scenarios with the context sentence. Error bars indicate standard errors.

3.3.3. Discussion

The most robust result that we observe is that, for the adjective scale tested here, there is a clear divide between weak adjectives on the one hand and middle/strong adjectives on the other hand, and the latter two are not distinguishable from each other. Weak adjectives were consistently rated lower than the other two, both with and without the additional context

sentence. We take this to be evidence that they were perceived as under-informative, and that they triggered scalar implicatures to the negation of the stronger expressions on the scale.[3] On the contrary, the behavior of middle adjectives, which constitute the other potential trigger for implicature, turned out to be basically identical to that of strong adjectives, suggesting that no scalar inference was drawn on these middle terms with respect to the stronger terms. We also note that extra context had an across-the-board effect to lower acceptability judgments, for all three adjective types.

Based on these observations, we conclude that, first of all, there seems to be a clear scalar distance effect: only weak adjectives trigger implicatures, whereas middle ones never do under any condition. Second, we observe that although context does have a strong effect, it seems to have made participants more conservative in their judgments across the board, rather than differentiating middle adjectives from the strong ones.

Different from previous studies, our results showed that it is not adjectives per se that are "hard triggers" of SIs. Lower terms on an adjective scale, in fact, trigger SIs in a normal fashion. Instead, only middle adjectives seem to be particularly resistant to triggering the inference, as their behavior is practically indistinguishable from that of extreme adjectives. There are two alternative possibilities to account for these facts. On the one hand, the absence of implicatures on middle adjectives could be explained by appealing to the peculiar nature of extreme adjectives, which would prevent them from being a salient alternative to middle ones, failing to generate the implicature. On the other hand, it would also be possible to explain these data with a more general, across the board "scalar distance" effect, according to which a scalar term must be distant enough from the alternative to generate the implicature. Applying the principle here, the data suggest that weak adjectives are distant enough from strong adjectives, and therefore trigger the inference. Middle ones, instead, are too close, failing to generate the inference. In Experiment 2 we aim to tease apart these two alternatives.

4. Experiment 2

In order to distinguish between the possibilities outlined above, we turned our attention to SIs generated by modal scales such as *<possible, likely, certain>*. While modals share some properties with adjectives, they are crucially different from them in other respects, and therefore constitute an interesting basis of comparison to cast further light on the results of Experiment 1. On the one hand, modals resemble adjectives in that they also consist of multiple terms ordered along a weak-middle-strong strength continuum with decreasing distance from the top of the scale. In this sense, for both kinds of scale the weak term (e.g. *decent* for adjectives, *possible* for modals) is maximally distant from the strong term, whereas the middle term (e.g. *good* for adjectives and *likely* for modals) is closer.

On the other hand they differ from adjectives in two main respects. First, they consistently trigger scalar implicatures, as widely shown in the literature (see, among others, Zevakhina and Geurts 2011). Second, following Lassiter (2011), they all share the same semantic

[3] Crucially, their ratings were still significantly higher than those for the contradictory condition. This showed that having a contradictory adjective made the whole scenario completely unacceptable, as we expected, and that speakers captured the difference between mere contradiction, which made the scenario completely unacceptable, and under-informativity, which only made it "less-than-perfectly-fine", but still potentially consistent with the facts.

structure and representations across every part of the scale, as their core meaning is built around a fully closed ratio scale of probability. According to the author, this is shown by three facts: a) modals can all combine with proportional modifiers; b) modals all equally support degree modification and comparatives and c) the extreme term on a modal scale (i.e. *certain*) marks a "real" upper boundary on the scale. These properties make them crucially different from extreme adjectives, which, as pointed out in section 2.2, do not easily combine with degree modifiers and comparatives and are not scalar maximums. In light of these facts, they might be using different semantic representations from non-extreme adjectives altogether.

In order to establish a comparison between the behavior of modals and adjectives as implicature triggers, we tested the calculation of implicatures on a three-place modal scale: <*possibile, probabile, certo*> (<"possible, likely, certain"> in English). We anticipate that the crucial data point will be the behavior of the middle modal *probabile* ("likely") with respect to the extreme modal *certo* ("certain"). If a difference between middle and extreme modals emerges, it would suggest that extreme adjectives have a different status from extreme modals, and that their peculiarity is responsible for the results obtained in Experiment 1. On the contrary, if the results pattern like those in Experiment 1, this could be taken as evidence that the relevant issue is the scalar distance effect across the board, and not the specific difference between extreme modals and extreme adjectives. We tested these predictions by comparing the acceptability ratings between different modals. The paradigm of Experiment 2 is similar to Experiment 1 (see below), but since Experiment 1 didn't reveal any online RT difference, we took out the self-paced reading part of the task in Experiment 2.

4.1 Materials

In order to keep the design and materials as similar as possible to those of Experiment 1, we used the same scenarios as before. Regarding the strength manipulation, modals came in three different conditions: weak (*possible*), middle (*likely*) and strong/extreme (*certain*). Since the results of Experiment 1 clearly showed the difference between contradiction and under-informativity, we decided not to keep the contradictory condition this time. Regarding the context manipulation, we maintained the context sentence used in Experiment 1. The target modal expression was always inserted in the last sentence of the scenario, while the previous statement always contained a weak adjective. A complete item, with conditions (a-f), is reproduced in (13).

(14)
No context sentence:
Mark is a decent student. a. That's why it's possible that she will get into Harvard (weak)
b. That's why it's likely that she will get into Harvard (middle)
c. That's why it's certain that she will get into Harvard (strong)
With context sentence:
The competition for entering top programs in the US is incredibly tough.
Sofia is a decent student. d. That's why it's possible that she will get into Harvard (weak)
e. That's why it's likely that she will get into Harvard (middle)
f. That's why it's certain that she will get into Harvard (strong)

4.2 Participants, procedure and predictions

34 native speakers of Italian (age 18-39) participated in the study. None of them had participated in Experiment 1. Monetary compensation (5 Euros) was offered in exchange for participation. Within each scenario, one sentence at a time was visualized on the screen. Subjects could move to the next sentence by pressing the space bar. Like in Experiment 1, at the end of each scenario subjects were asked to judge how sensible the last sentence was in light of what they had read before by providing an acceptability score ranging from 1 (completely nonsensical) to 5 (perfectly sensible). Given the simultaneous presence of a weak adjective in the previous sentence, if participants draw implicatures on weak and middle modals (e.g. *not certain* on *possible* or *likely*), the scenarios containing these expressions should be significantly more acceptable than those containing *certain*. If, on the contrary, people do not exclude *certain* when processing *possible* or *likely*, acceptability ratings on the scenarios containing weak and middle modals should be low as well.

4.3 Results

The averaged acceptability ratings were plotted in Figure 3. A two-way ANOVA revealed a main effect of adjective strength (by-subject F_1 (2,66) = 5.7, p <.01; by-item F_2(2,46) = 4.8, p < .01. A main effect of context was also found (by-subject F_1(1,33) = 21.5, p < .01; by-item F_2(1,23) = 28.7, p < .01). We found no interaction between the two factors (by-subject F_1(2,66) = 0.5, p > 0.5; by-item F_2(2,46) = 0.6, p > 0.5).

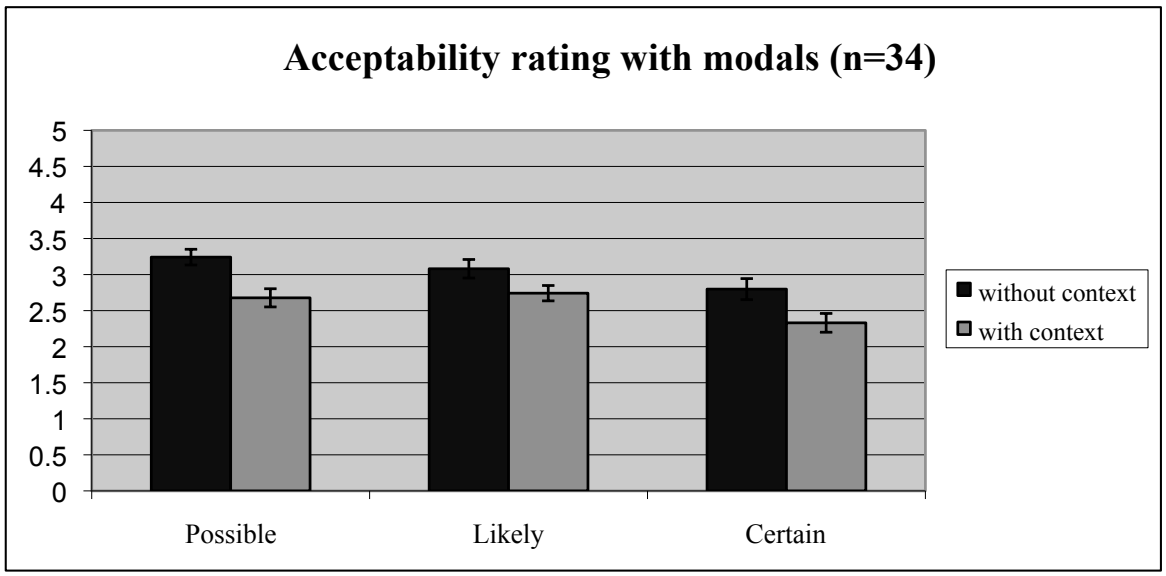

Figure 3: Acceptability judgments on a 1-5 scale for Experiment 2. The Y-axis indicates the average 1-5 acceptability ratings. The X-axis plots the three different modal conditions. Dark grey bars stand for scenarios without the context sentence, and light grey bars stand for scenarios with the context sentence. Error bards indicate standard errors.

For the effect of modal strength, paired-comparisons showed that weak modals were judged significantly higher than strong ones (t(33) = 2.0, p < .05 without context; t(33) = 4.9, p < .001 with the context sentence). By contrast, middle modals turned out to be significantly more acceptable than the strong ones only when the context sentence was present (t(33) =

3.2, $p < .05$). Without the context sentence, they are not significantly different from either weak or strong modals (all $ps > .05$).

For the effect of context, paired t-tests showed that for all conditions scenarios with context was significantly less acceptable than scenarios without context (Weak modals: $t(33) = 5.1$, $p < .0001$; middle modals: $t(33) = 2.25$, $p < .01$; strong modals: $t(33) = 3.1$, $p < .0001$).

4.4 Discussion

Without the additional context sentence, scenarios containing weak modals were considered to be significantly more acceptable than scenarios containing strong modals. This is similar to what we observed in Experiment 1 and shows that, once again, inferences to the exclusion of the strongest term on the scale were consistently computed on the weakest one. However, scenarios with *likely* do not seem to pattern with either of the other two conditions. On the one hand they are not different from scenarios with *possible*. On the other hand they are not different from scenarios with *certain* either. This constitutes a difference with respect to Experiment 1, in which middle adjectives were significantly different from weak ones and pattern together with strong ones.

With the context sentence, scenarios with weak modals are rated significantly higher than those with strong modals, unsurprisingly. However, a clear pattern for middle modals finally emerges, as scenarios containing *likely* turn out to be significantly more acceptable than those containing *certain* and not significantly different from scenarios containing *possible*. This shows that the context sentence triggered SIs to the negation of *certain* for middle modals, generating the upper bound reading (*likely but not certain*). This constitutes a crucial difference with respect to Experiment 1, in which middle adjectives never generated SIs to the negation of strong ones, regardless of whether they came with or without the additional context sentence.

In sum, Experiment 2 showed that modal scales behave differently from adjective scales with respect to the computation of SIs. While no difference between the interpretation of middle and strong adjectives ever emerged, modal scales exhibit a boundary between the middle and the strong term, and a proper context made the distinction even sharper. Given that both scales are composed of multiple terms ordered along a continuum with decreasing distance from the top, we can conclude that scalar distance could not be the main factor in accounting for the results of Experiment 1. If the reason why middle adjectives failed to generate implicatures was because they were not distant enough from the top of the scale, then we should have expected a similar pattern with middle modals too. Instead, the fact that middle modals behaved differently suggests that there is something peculiar to extreme adjectives and adjective scales that accounts for the resistance of middle adjectives to generate SIs. In the next section we spell out this idea in greater detail, outlining two possible theoretical explanations as to why middle and extreme adjectives were always interpreted in the same way.

5. General discussion

We suggest that the lack of implicatures from middle adjectives to the negation of extreme ones is compatible with two different theoretical explanations. The first possibility is that extreme adjectives are not on the same lexical scale as weak and middle adjectives, and because of this they are not considered to be salient alternatives for the computation of SIs. The second possibility is that extreme adjectives are indeed on the same scale as the others, but middle ones are simply flexible enough to extend to the upper region of the scale, making the implicature unnecessary. This second possibility would be substantially similar to the account proposed by Doran et al. in their investigation.

We start by discussing the first proposal. At first sight, the proposal that *good* and *excellent* are not on the same scale seems hardly plausible. There is a clear intuition that both adjectives are, roughly speaking, talking about the same thing, namely some qualitative evaluation of "goodness". Moreover, we get the solid entailment pattern according to which asserting *excellent* necessarily entails *good,* as the contradictory nature of (15) shows. This observation also seems to prove that the two terms must share some kind of common scalar dimension.

(15) # This is *excellent* but not *good*

Our suggestion is that, while the two expressions are indeed measuring the same dimension in a broader sense, they nonetheless do not share the same *lexical* scale. While middle adjectives such as *good* behave as fully gradable predicates and encode a degree argument in their denotation, extreme adjectives such as *excellent* are simply *not* gradable, and express simple properties from individuals to truth values. Their different denotation is reflected in (16) and (17), where only (17) encodes a degree argument in its meaning. Note that positing this difference would also provide a straightforward explanation for the facts discussed in (7-10) in 2.2. If no degree argument is present for extreme adjectives, then it follows naturally that they do not support comparison or degree modification of other sorts.

(16) [|excellent|] = λx. Excellent (x)
(17) [|good]] |] = λx λd. x is d-good [4]

Modals, on the contrary, might all be considered to be lexically gradable and to share the same semantic representation in their denotation. Lassiter (2011) offers various arguments in support for this claim, including the fact that all modals can be modified by proportional modifiers and can be inserted in comparative constructions. On the grounds of these facts, he argues that *likely* and *certain* are situated at different points along a shared closed scale of "degrees of probability" ranging from 0 (scalar minimum) to 1 (scalar maximum). As, such they both have a degree argument in their denotations. What changes is the specific degree of probability that sets the standard for the truth conditions of each of the two expressions. On

[4] We are aware that many different implementations of degree semantics are available in the literature, and committing to a particular one goes beyond the scope of this paper. The denotation in (16) has been borrowed from Morzycki 2010, but a different implementation of degree semantics would have been equally adequate for our purposes. The crucial element is the presence of a degree argument with middle adjectives, and the absence of it with extreme ones.

the one hand, the truth conditions for *likely* are satisfied if an object ϕ is associated with a degree of probability (prob(ϕ)) that exceeds a contextually variable standard S-prob comprised between 0 and 1. On the other hand, the truth conditions for *certain* are satisfied if the degree of probability associated with ϕ coincides with the scalar maximum. In brief:

(18) ϕ is likely/probable is true iff prob(ϕ) > S-poss.
(19) ϕ is certain is true iff prob(ϕ) = 1.

We suggest that this difference in the semantics can account for the contrast between modals and adjectives in the experimental data. In a situation in which no alternative is explicitly evoked in the context (as it was Experiment 1 and 2), it could be the case that only members of the same lexical scale can be salient alternatives, and therefore take part in the process of computing SIs. Since middle and extreme modals are indeed lexical scale mates, *certain* is an available alternative to *likely,* and implicatures from *likely* to *certain* therefore arise. Since middle and strong adjectives, instead, are not lexical scale mates, *excellent* is not an available alternative to *good,* and SIs from *good* to *excellent* are not computed. [5]

The second possibility that we want to outline is that middle and extreme adjectives are full-fledged scale mates, just like middle and extreme modals, but that the strongest term is not easily accessed as an alternative, blocking the implicature. This happens because the semantics for a middle adjective like *good,* combined with the lack of an upper-boundary on the scale, allows for the adjective to stretch indefinitely high on the scale, overlapping with the area covered by the corresponding *extreme* adjective without being perceived as under-informative. In this way, *good* becomes potentially equivalent to (and not only compatible with) *excellent*. For this reason, giving *good* an interpretation that excludes the assertion of the stronger term is no longer necessary, making *excellent* inaccessible as an alternative. On the other hand, since modal scales do have a real scalar maximum, there is always a clear boundary between middle and strong modals. Since *certain* corresponds to the real maximum point on the scale (where "probability" = 1), there is no circumstance in which *likely* can index the very same degree as *certain,* and therefore be truth conditionally equivalent to it. This necessary gap between the two modals, which does not exists for adjectives, makes *certain* an accessible alternative to *likely* under any circumstance, creating the grounds for the computation of the implicature.

6. Conclusion

In this investigation we looked at how SIs are computed on different scales. In Experiment 1 we tested the inference on triplets of adjectives that are traditionally considered to be on the same scale. Only the lowest adjective on the scale seems to have generated implicatures, whereas the middle term on the scale did not. We further show in Experiment 2 that scalar members on a modal scale, both the lowest and the middle modal terms, generated SIs. This suggested that the low rate of implicatures triggered by middle adjectives is not due to a

[5] Only by presenting *excellent* as an explicit alternative, like Doran et al. did in their study, the contrast between the middle and the strong adjective can be retrieved and activated, finally allowing for an upper-bound reading on the middle adjective. However, quoting Zevakhina and Geurts' words, this would be a "brute force" imposed inference, and would hardly count as an implicature.

general property of scalar distance, but to the particular status of extreme adjectives with respect to gradability, and the accessibility of extreme adjectives when SIs are calculated.

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.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Doran, R., R. Baker, Y. McNabb, M. Larson, and G. Ward (2008), On the Non-Unified Nature of Scalar Implicature: An Empirical Investigation. *International Review of Pragmatics* 1, 1-38.

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

Grice, P. (1975). Logic and conversation. In H. P. Grice (Ed.) (1989). *Studies in the Way of Words*. Harvard University Press. 22-40.

Horn, L. (1984). A new taxonomy for pragmatic inference: Q-based and R-based implicature. D. Schiffrin (Ed.) (1985), *Meaning, Form and Use in Context*. Georgetown University Press. 11-42.

Horn, L. (2005). Implicature. L. Horn and Ward, G. (2006) (Eds.), *The handbook of pragmatics*. Wiley-Blackwell. 3-29.

Lassiter, D. (2010). Gradable epistemic modals, probability, and scale structure. Li, N. and D. Lutz (2011) (Eds.), *Semantics and Linguistic Theory (SALT) 20*. CLC Publications, 197-215.

Morcyzki, M. (2009). Degree Modification of Extreme Adjectives. Bochnak, R., N. Nicola, P. Klecha, J. Urban, A. Lemieux, C. Weaver. (2011) (Eds.) *The Proceedings of the Chicago Linguistic Society 45,* 471-475.

Morzycki, M. (2010). Adjectival extremeness: Degree modification and contextually restricted scales. *Natural Language and Linguistic Theory* 30(2), 567-609.

Nieuwland, M.S., T. Ditman, and G.R. Kuperberg, (2010). On the incrementality of pragmatic processing: An ERP investigation of underinformative scalar sentences. *Journal of Memory and Language* 63, 324–346.

Paradis, C. (2001). Adjectives and boundedness. *Cognitive linguistics* 12(1), 47-65.

Zevakhina, N. and B. Geurts. (2011). *Scalar diversity*. Manuscript.