Proceedings of SALT 22: 226–246, 2012

Granularity shifting: Experimental evidence from degree modifiers *

Galit Sassoon
Hebrew University

Natalia Zevakhina
National Research University Higher School of Economics

Abstract This paper argues that modeling granularity and approximation (Krifka 2007; Lewis 1979) is crucial for capturing important aspects of the distribution and interpretation of adjectives and their modifiers, modulo certain differences between modified adjectives and numerals. In addition, the paper presents supporting experimental results with minimizers like *slightly* and maximizers like *completely*.

Keywords: semantic, adjective, granularity, degree, degree modifier, inference, approximation

1 Introduction

Part 1 of this paper presents an account of degree modifiers that involves granularity shifting. Various aspects affect the likelihood that a speaker would derive an inference from statements with a modified adjective to similar statements without the modifier, as in *x is slightly dirty* vs. *x is dirty*, or vice versa. The following sections discuss some of these aspects, including rules for categorization under adjectives (1.1), the semantics of degree modifiers like *slightly* and *completely* (1.2), the relevance of the notions of granularity, approximation and level of fit (1.3-1.5), and factors affecting the choice between upper-bounded and upper-open interpretations (1.6). Part 2 presents two studies that test predictions of the developed account.

1.1 Scale-based categorization rules

Assume a λ-categorial language (Heim & Kratzer 1998) and an analysis of gradable adjectives G as denoting relations between entities x and degrees d based on a mapping g of entities to their degree in G. Let $s(g)$ stand for the membership standard of an adjective G: x is G is true in a context c iff x is G to at least degree

* We thank Manfred Krifka, Chris Kennedy, Louise McNally, Bart Geurts, and Emmanuel Chemla. Any mistake is solely ours.

$s(g)$, i.e. $g(x) \geq s(g)$. In short, $G(s(g))(x)$ in c. For example, *x is tall* is true iff x's height exceeds the contextual height standard.

On scale structure theory (Rotstein & Winter 2004; Kennedy & McNally 2005; Kennedy 2007; McNally 2011) gradable adjectives are classified by their scale type, namely as lower-closed (+min), upper-closed (+max), both, or neither. Endpoints, when they exist, function as standards (Kennedy & McNally 2005; Kennedy 2007). Thus, adjectives are also classified by their standards, as follows.

(1) a. G is partial iff G's standard minimally exceeds the scale zero: $s(g) > \min(g)$.

b. G is total iff G's standard is the maximum on G's scale: $s(g) = \max(g)$.

c. G is relative otherwise.

(1) states that partial adjectives have a minimum standard; for example, one stain suffices for a shirt to count as dirty, meaning that *dirty* is a partial adjective. Total adjectives have a maximum standard; for example, to count as clean a shirt has to be completely free of dirt, meaning that *clean* is a total adjective. The standard of relative adjectives like *tall* is a midpoint on their scale, which varies with context.

Absolute adjectives (i.e., total and partial ones) typically allow for a wider range of modifiers than relative ones. They license relative modifiers as in *very clean/dirty*, but they co-occur significantly more often than relative adjectives with absolute modifiers, including minimizers as in *slightly dirty*, and maximizers as in *completely clean*. Minimizers are viewed as referencing scale minima. Their distribution is thought to be restricted to +min adjectives. Maximizers are viewed as referencing scale maxima. Their distribution is thought to be restricted to +max adjectives.

(2) a. *slightly* $\Leftrightarrow \lambda G_{+\min} \lambda x . \exists d > \min(g), G(d)(x)$.

b. *completely* $\Leftrightarrow \lambda G_{+\max} \lambda x . \exists d = \max(G), G(d)(x)$.

According to this account, absolute modifiers are semantically vacuous. Yet, they function as cues for an adjective's scale type and thereby indicate how its standard is to be set (Syrett & Lidz 2010). Moreover, they function as slack regulators (Lasersohn 1999) signaling precision level. The level of precision of utterances in ordinary discourses is not maximal. For example, by default, statements like *The garage is clean* may be accepted as "true enough" even when they are actually false, e.g., there are stains on the garage floor. This is so because they are normally precise enough for all practical purposes.

According to Lasersohn (1999), precision level is determined by pragmatic halos. Within context any expression is associated with a pragmatic halo, i.e., a set of alternative denotations of the same type as its actual denotation which differ only in some pragmatically ignorable respect. Their use counts as acceptable, even if it

leads to a strictly speaking false proposition. For instance, the pragmatic halo of *is empty* in *The theater is empty tonight* includes properties that are true of objects that are a bit less than empty. How much less is determined by context, e.g., by non-linguistic factors such as the size of the theater (Kennedy & McNally 2005).

1.2 A new analysis of degree modifiers

The slack regulator account of modifiers works well for maximizers, but not for minimizers. Assume, for example, that *slightly* signals a high standard of precision. This implies that normally when we say *x is dirty* we speak loosely, referring to objects that are almost dirty but actually clean, while when we say *x is slightly dirty* we signal reference to objects that are strictly speaking dirty. This is highly unintuitive. Thus, assume that *slightly* signals a low standard of precision. This implies that normally when we say *x is slightly dirty* we speak loosely, referring to objects that are almost dirty but actually clean. This is not intuitive either. Thus, an account in terms of precision level alone does not do.

Minimizers are better accounted for in terms of granularity. Sassoon (2012) proposes the following analysis. Adjectives are interpreted relative to a coarse granularity level g. When using statements such as *The car is dirty/clean* it is normally appropriate to ignore almost invisible dirt. By contrast, modified adjectives are interpreted relative to a fine granularity level g_p. While using statements such as *The car is completely clean/slightly dirty*, almost invisible dirt specks are rendered relevant in judging degree of cleanliness. Formally,

(3) $\forall g, g_p \in D_{xd}$, g_p is finer-grained than g iff (Lewis 1979; van Rooij 2011)
$\exists x,y \in D_x, ((g(x) = g(y)) \wedge (g_p(x) \neq g_p(y))) \bigwedge$
$\bigwedge \nexists x,y \in D_x, ((g_p(x) = g_p(y)) \wedge (g(x) \neq g(y)))$.

For example, two glasses filled with an amount of wine differing in but few drops might be indistinguishable relative to g, but distinguishable relative to g_p. A total adjective like *full* denotes maximally full entities presupposing coarse granularity g. By contrast, the maximized total adjective form *completely full* denotes maximally full entities, presupposing finer granularity g_p. Formally,

(4) a. $[G_{\text{total}}]_g = \lambda x \in C : g(x) = \max(g)$.
b. $[\text{completely } G]_g = \lambda x \in C : g_p(x) = \max(g_p)$, for g_p finer than g.

Since g_p is finer than g, it follows that g_p assigns fewer entities the same degree ($=_g \supset =_{g_p}$). Thus, (4b) is stronger than (4a).

Similarly, a partial adjective like *dirty* denotes minimally dirty entities, presupposing coarse granularity g. Thus, objects covered with a few specks of dirt are

considered to be as clean as objects which are completely free of dirt. By contrast, the minimized partial adjective form *slightly dirty* denotes minimally dirty entities, presupposing finer granularity g_p. Formally,

(5) a. $[G_{\text{partial}}]_g = \lambda x \in C : g(x) > \min(g)$.

b. $[\text{slightly } G]_g = \lambda x \in C : g_p(x) > \min(g_p)$, for g_p finer than g.

Since g_p is finer than g, more distinctions are made, i.e., g_p assigns more entities different degrees $> _g \subset > _{gp}$). Thus, (5b) is weaker than (5a).

On this account, the distribution of modifiers is not sensitive to scales with endpoints, but to classification rules based on identity to a degree as in (4a) vs. an external threshold as in (5a). Thus, definition (5b) can be replaced with (6).

(6) $[\text{slightly } G]_g = \lambda x \in C : g_p(x) > d_s$, for g_p finer than $g : > _g \subset > _{gp}$, where d_s represents a threshold external to the denotation, but not necessarily the scale zero.

This account captures the use of *slightly* with relative and total adjectives. In relative adjectives *slightly* is licensed only if an external threshold is made salient, for instance due to excessive- or *for*-phrase modification, as in *slightly too tall* and *slightly tall for her age*. With total adjectives, *slightly* triggers minimal shifting to a non-maximal standard, $d_s < \max(g_p)$, which can therefore function as an external threshold for denotation members to exceed. Hence, *slightly full* implies *rather full*.

The predictions of this analysis differ from those of the scale structure analysis in (2). The standard shifting involved in modification of total adjectives according to this analysis renders combinations of *slightly* with total adjectives less compatible than combinations with partial adjectives. Indeed, felicity judgments and corpus frequencies of *slightly* support sensitivity to standard types (partial$\gg$total) rather than mere existence vs. absence of scale minimum (Sassoon 2012).

In addition, according to scale structure theory, statements with minimized total adjectives do not entail the same statements without the minimizer, e.g., $\neg$(*The city square is slightly full* $\Rightarrow$ *The city square is full*). The former only requires that the argument's degree minimally exceed the scale zero, while the latter requires it to reach the scale maximum. However, intuitively, *The city square is slightly full* conveys something equivalent to 'the city square is rather full/more full than empty'. This is unexpected given a semantics for *slightly full* that merely requires that the degree of instances exceed the zero on *full*'s scale. Instead, the extension of *slightly full* seems to cover degrees above a threshold d_s slightly below the maximum.

Rotstein & Winter (2004) argue that the status of absolute scale minima and maxima as standards of absolute adjectives is based on a default rule which may be overridden in context. However, Rotstein & Winter (2004) do not specify in what

ways contexts affect the selection of standard. McNally (2011) completely abandons the notion of endpoints, arguing that, e.g., total predicates may pick out regions on a scale that are not actual maxima as in *This wine glass is full*. When ordering a glass of wine in a restaurant, one cannot demand one's money back if served a glass filled up to half of its capacity. This suggests that a convention exists for a standard located away from the maximum of a fine scale.

We propose that granularity is a crucial factor in accounting for these facts. In most ordinary discourse contexts, scales of coarse granularity are employed. Therefore, even endpoint standards, as they are based on coarse scales, cover more than the absolute endpoints of fine-grained scales.[1]

In sum, a theory of modifiers that includes granularity levels and granularity shifting improves on one that abstracts away from these parameters in that it extends to minimizers. It reveals a semantic core that degree modifiers share.

Another advantage is that a granularity shifting theory applies in other domains of grammar, too. For example, Krifka (2002, 2007) observed a coupling between (non-)round numbers and (non-)round interpretations, and argued more generally that simple expressions align with imprecise interpretations based on coarse scales, while complex or long expressions align with precise interpretations based on fine scales. Viewed in this light, adjectives and their modified forms are merely subcases of simple vs. complex expressions and are therefore expected to trigger the use of different default levels of granularity.

Thus, the next section is aimed at presenting a formal account of granularity, precision, and their effects on the interpretation of numerals, as developed in Krifka 2007, and a detailed comparison with adjectives and their modifiers.

1.3 Granularity and approximation in numerals vs. adjectives

When we count, we map pluralities of individuals into numbers representing their cardinalities. This mapping depends on the contextual setting of a granularity parameter, namely the selection of a set of alternative numerals. A set-inclusion relation between alternative sets affects the derivation of vague vs. precise interpretations.

(7) a. ... one hundred, two hundred, three hundred, ...

b. ... ninety, one hundred, one hundred and ten, one hundred and twenty ...

For example, the alternative set in (7a) triggers interpretations relative to a coarse scale, whereas the set in (7b) triggers interpretations relative to a finer scale. In addition, the average complexity of expressions as measured by, for example, number

1 McNally (2011) suggests that standards tend to be based on factors such as proportionality, perceptual salience, and purposes. Probably these factors affect the choice of granularity level as well.

of syllables, is greater in alternative sets corresponding to finer scales, e.g., in (7b) as compared to (7a).

Krifka (2007) derives the facts concerning the most likely interpretation of an ambiguous or vague form such as that of numerals by using the principles of strategic communication (Parikh 2000; Benz, Jäger & van Rooij 2006), as follows. First, it is hypothesized that addressees assume the speaker intends the most likely interpretation and speakers assume the addressees select the most likely interpretation, given the choice of expressions and a priori likelihood of a message.

Second, to adopt this framework, approximate interpretations of a numeral n are represented via an interval $[n-n/s,\ n+n/s]$ interpreted as a normal distribution with mean n (the precise interpretation) and standard deviation s (the level of approximation). The range within n/s above and below the central tendency is used as an acceptable level of fit, corresponding to a pragmatic halo in (Lasersohn 1999). With a smaller $1/s$, the interpretation gets more precise. At a level of approximation of, e.g., $1/10$, *ten* stands for the interval $[\text{ten}]_{1/10} = [9, 11]$. At a level of approximation of $1/\infty(\approx 0)$, *ten* is interpreted in a precise way: $[\text{ten}]_0 = \{10\}$, see (Dehaene 1997).

Third, Krifka (2002, 2007) assumes a maxim of brevity whereby alternative sets with shorter expressions are preferred. Hence, the use of an expression α is assumed to introduce the most coarse-grained alternative set possible for it into the pragmatic evaluation. Expressions with similar interpretations like *thirty eight* and *forty* given $s = 10$ are assumed to be *interpretatively equivalent*, because their precise values, 38 and 40, are within each other's ranges: $[\text{forty}]_{1/10} = 40 \pm 4$ and $[\text{thirty-eight}]_{1/10} = 38 \pm 3.8$.

By virtue of being wider, approximate interpretations are more probable than precise ones. Assume equal a priori probabilities p_v of scale values, as well as equal a priori probabilities p_a of approximation levels. The probability of the precise interpretation ($s = 0$) is $p([\text{forty}]_0) = 1p_v$, and that of the approximate interpretation ($s = 1/10$) is $p([\text{forty}]_{1/10}) = p([36, 44]) = 9p_v$. The total probability $p(1/10) \cdot p([forty]_{1/10}) = 9p_vp_a$ is bigger than the total probability $p(0) \cdot p([\text{forty}]_0) = p_vp_a$. Thus, the approximate interpretation of *forty* is more probable. The same holds for *thirty eight*, but the latter is more complex than the interpretatively equivalent *forty*. Hence, by brevity, it cannot be used to refer to the preferred approximate interpretation. Thus, by default it is associated with a more precise interpretation.

Finally, assuming equal a priori probabilities p_s for scales, resort to coarser ones is more probable. For example, *forty minutes* invokes the more coarse-grained scale (9a) rather than the fine-grained scale (9b), as its approximate interpretation in the former is wider and therefore more probable (8):

(8) $p([40]_{9a}) = p([35, 45]) = 9p_v > p([40]_{9b}) = p([39.5, 40.5]) = 1p_v.$

(9) a. ...-30————————————-40————————————-50-...

b. ...-30-31-32-33-34-35-36-37-38-39-40-41-42-43-44-45-46-47-48-49-50-...

Like numerals, (un)modified adjectives create alternative sets, too.

(10) a. clean, dirty.

b. completely clean, very clean, clean, pretty clean, slightly clean, slightly dirty, pretty dirty, dirty, very dirty, completely dirty.

We hypothesize in analogy with numerals that adjectives and their modifiers are associated with a probability distribution over the scale ranges they denote, representing level of fit of reference to each scale point. Peaks represent typical contexts of use. For example, the denotation of *clean* overlaps with (is a superset of) that of *completely clean*. Yet, we typically use *clean* to refer to lower degrees than *completely clean* the peak of which is located at the maximum. The denotation of *slightly dirty* overlaps with (is a superset of) that of *dirty*. Yet, we typically use only the latter to refer to high degrees. Moreover, the sum of scale ranges covered by modified adjectives (e.g., *slightly dirty*, *pretty dirty*) on a fine scale aligns to the scale range covered by the adjective they modify (e.g., *dirty*) on a coarser scale.

Assume that the precise interpretation of the n$^{\text{th}}$ member in an alternative set is the n$^{\text{th}}$ point on the scale, and its approximate interpretation per threshold s is $n \pm n/s$. Suppose $f_n : D_x \to \{1, \ldots, 10\}$ is a function from entities to degrees on a scale like $[1, 2, 3, \ldots, 10]$. For the alternative set (10b):

(11) a. $f_{10,s=0}(\text{dirty}) = 8$; $\quad f_{10,s=3}(\text{dirty}) = [8 \pm 8/3] = [5, \ldots, 10]$.

b. $f_{10,s=0}(\text{slightly dirty}) = 6$; $f_{10,s=3}(\text{slightly dirty}) = [6 \pm 7/3] = [4, \ldots, 8]$

This means that this approximate interpretation of *dirty* partially overlaps with that of *slightly dirty*. Similarly, the interpretations of *slightly dirty* and *slightly clean* overlap, etc.

Approximate interpretations are more probable than precise ones. Assuming equal a priori probabilities p_v for scale values and p_a for approximation levels, the precise interpretation of *clean* per (10b), is $f_{10,s=0}(\text{clean}) = 3$, with total probability $p(0) \cdot p([\text{clean}]_0) = p_a p_v$. The total probability of the approximate interpretation for, e.g., $s = 3$, is greater: $p(3) \cdot p[\text{clean}]_{1/3} = 3p_v p_a$, so *clean* is interpreted approximately.

Furthermore, interpretations relative to coarse scales are more probable than fine ones. For example, assuming equal scale probabilities p_s, on the coarse scale (10a), the widest range of *dirty* is $[\text{dirty}]_c = [5, 10]$, whereas on the fine scale (10b), the widest range is $[\text{dirty}]_f = [7.5, 8.5)$. Thus, the probability of the coarse scale, $p([\text{dirty}]_c) = 5p_v p_s$, is bigger than the probability of the fine scale, $p([\text{dirty}]_f) = 1p_v p_s$. Therefore, *dirty* invokes a coarse scale rather than a fine one. By contrast,

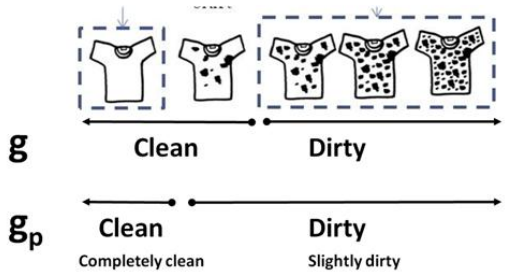

Figure 1 The coarse and fine interpretation of partial adjectives like *dirty*.

completely dirty does not denote any scale point on the coarse scale, so it cannot be interpreted at this level. The first scale it can be interpreted at is a fine one, where its default precise interpretation captures points in [9.5, 10]. Thus, this analysis predicts that *dirty* is interpreted in a less precise way than *completely dirty*.

With these results at hand, we can now turn to modifying our assumptions regarding adjectives by incorporating insights from scale structure theory (Kennedy & McNally 2005; Rotstein & Winter 2004; Kennedy 2007) and so acknowledge differences between numerals and adjectives.

First, an adjective like *dirty* is associated with a mapping of entities to degrees representing amount of dirt, while *clean* is associated with a converse function. Moreover, unlike numerals, the precise interpretation of (modified) adjectives is often a scale range rather than a point, and they appear to denote monotonic scalar functions (Heim 2000); for example, the range denoted by *very+A* includes besides its most probable value, also the range denoted by *completely+A*. Therefore, the interpretation of the n^{th} expression in an adjectival alternative set is better described by means of an interval [min, *n*] or [*n*, max], where min/max are either scale ends or infinity (see also sections 1.1-1.2).

Second, the size of the scale ranges denoted by antonyms or by their modified forms are not all equal. Unlike numerals, the precise interpretation of partial adjectives like *dirty* on a coarse scale covers a wide range as their semantics is based on a minimum-standard rule (cf. section 1.1). It is even wider on a fine scale, where it is identical to the interpretation of *slightly dirty* (cf. section 1.2). The precise interpretation of total adjectives like *clean* is based on a maximum-standard rule and so on a fine scale it equals that of *completely clean*, and might be extended further down only under a non-zero level of approximation (Kennedy 2007). Thus, upon shifting to a finer granularity, the interpretation of *clean* shrinks, as hardly visible dust grains become relevant. Such narrowing is prevalent in numerals upon a shift to a finer granularity level. Note, however, that the excluded bits move into the interpretation of *dirty*, which thereby widens, as figure (1) illustrates. Widening

upon a shift to finer granularity never occurs in numerals.

Hence, the precise interpretations of partial adjectives on coarse scales are narrower than on fine ones. Thus, in order not to wrongly predict that they tend to be interpreted on fine scales, we have to assume that their approximate interpretations on coarse scales are at least as wide as on fine scale. A reason for that can be that precise partial denotations on fine scales are too large to be widened, whereas on coarse scales they can be widened. With this auxiliary assumption, Krifka's theory captures the facts. Since approximate interpretations are larger they are still the preferred choice for unmodified adjectives, and so is the coarse scale.

The situation is different with modified adjectives, which only denote on fine scales, and, by brevity, their precise interpretations are preferred. Therefore, bare adjectives are predictably interpreted more coarsely than modified ones.

Finally, intuitively, the higher the degree of an entity in an adjective, the higher its level of fit as an argument of the adjective. Thus, the best level of fit of an adjective typically includes extreme points (the *very/completely* range) on coarse scales. Only upon a shift to finer scales, the peak is intuitively located lower, namely between the ranges of *pretty* and *very*. The next section supports these claims.

1.4 Level of fit

On a scale structure + slack regulator account of imprecise uses of total adjectives like *clean* (see section 1.1), reference to the maximum of the finest scale has the highest level of fit. The level of fit reduces as the distance increases between an argument's degree and the maximum. Moreover, any level of fit other than the highest is considered imprecise. A difficulty to extend this account to capture modification of partial adjectives as in *slightly dirty* emerges in part because the precise interpretation of partial adjectives like *dirty* consists of almost the whole scale range. It is therefore more difficult to characterize loose use of *dirty*. We propose that the notion of level of fit does extend to cover also these cases, as follows.

Scale structure theory predicts that statements with partial adjectives do not entail their maximized versions: $\neg$(*The city square is dirty* $\Rightarrow$ *The city square is completely dirty*). However, intuitively, the adjective *dirty* has a higher level of fit when applied to an argument denoting a very dirty or even completely dirty entity than to an argument denoting an entity covered by few hardly visible dirt specks. In other words, upon hearing *dirty* we typically have in mind something dirtier than just slightly dirty things. Thus, people might agree that by default *The city square is dirty* conveys the content of *The city square is completely dirty* or some other proposition slightly weaker, but still significantly stronger than just 'the city square is covered by a slight amount of dirt'. This is unexpected given a mere "more than zero dirt" categorization rule for *dirty*. It supports a representation of level of fit with a high

peak away from the beginning of the scale range depicted by a minimum-standard categorization rule.

Similarly, in scale structure theory, statements like *The table is dirty* and *The table is slightly dirty* are predicted to be equivalent. However, there is an intuitive difference between them, which is captured by the assumption that in total and partial adjectives alike, the highest level of fit lies far above that of minimized total/partial adjectives. A theory of scale structure combined with a granularity (cf. section 1.3) can capture these effects of level of fit on judgments of inferences. Assuming that peaks of antonyms are near opposite scale endpoints, even partial adjectives like *dirty* are predicted to typically refer to *rather/completely dirty* points, not to *slightly dirty* ones.

This prediction aligns with experimental evidence (Paradis & Willners 2006). In this study, participants had to grade arguments of adjectives in Swedish on an 11-point scale. Thus, for example, arguments of the total adjective *dead* and its negation *not dead* were associated with points 1 and 6, respectively, whereas arguments of the partial antonym *alive* and its negation *not alive* were associated with points 8 and slightly above 1, respectively. In sum, the scale range denoted by the total adjective appears to be narrower than that of the partial one, as scale structure theory predicts. The former consists of at most 5 points (1—5), while the latter might consist of up to 10 points (2—11). However, the default level of fit of the partial adjective appears to be the high degree 8, not 2.[2]

In sum, intuitively, adjectives have significantly higher peaks, especially on coarse scales, than minimized ones have on fine scales. As a result, inferences between, e.g., *dirty/clean* and *completely dirty/clean* should be easier than inferences between *dirty/clean* and *slightly dirty/clean*. Generally, we expect inferences between statements with and without a maximizer to be stronger than inferences between statements with and without a minimizer: "If maximized A, then A" $\gg$ "If minimized A then A" (**Prediction 1**).[3]

2 Moreover, Anderson (1996) tested the hypothesis that modifiers have fixed locations on a quantity scale, providing evidence for our assumptions concerning the locations of peaks of modified adjectives. For example, judgments of how likable a person is described by an adjective like *sincere/insincere* form a linear increasing/decreasing curve with *slightly* at the most extreme point, followed by *somewhat, fairly/rather, pretty, quite/very*, and *extremely*. An unmodified adjective is located in between *pretty* and *quite/very* (Anderson 1996: 398-404).

3 In line with this view, note also that maximizers express agreement with a previously uttered bare adjective (e.g., *My shirt is dry/wet – Yes/??No, it's completely dry/wet*), while minimizers serve as polite markers of disagreement (e.g., *My shirt is dry/wet – Yes/No, it's slightly dry/wet*). Similarly, an adjective following its maximized form hardly expresses discontent (e.g., *My shirt is completely dry/wet – Yes/??No, it's dry/wet*), while a bare adjective following its minimized form conveys disagreement with the use of the hedge (e.g., *My shirt is slightly wet/dry – Yes/No, it's dry/wet*).

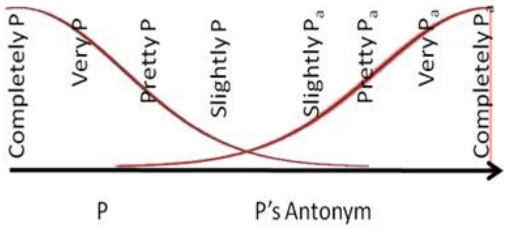

Figure 2 Level of fit in adjectives P and their antonyms.

1.5 Licensed and unlicensed granularity shifts in discourse

Lewis (1979) argues that a shift from default to finer granularity and/or precision level is a natural discourse move, but not vice versa. We may state that the Netherlands is flat, presupposing coarse granularity, and then point out that it is actually a bit bumpy by shifting to a finer criterion, taking as evidence bumps we previously ignored. However, we cannot state that the Netherlands is bumpy and smoothly proceed to say that it is actually flat, thereby ignoring bumps we previously regarded relevant. Thus, once an expression of a fine alternative set is uttered like *slightly dirty*, the interpretation of an expression like *dirty* is forced to be relativized to a fine scale.

Two studies of inferences with modified and bare adjectives provide experimental evidence for the granularity shifting account of degree modifiers presented above. Following Lewis, we hypothesize that inferences from modified to unmodified adjectives are easier to agree with than ones from unmodified to modified adjectives, because the former involve an irreversible shift to finer granularity due to which the interpretation of an adjective is rendered similar to that of a modified one. Study 1 tests this hypothesis for combinations of minimizers with partial adjectives and maximizers with total adjectives: "If M A then A" $\gg$ "If A then M A" (**Prediction 2**).

The question whether this pattern holds for maximized partial adjectives and minimized total adjectives was tested in study 2. However, since the interpretation of total adjectives remains different from the interpretation of their minimized forms even when interpreted precisely following an irreversible granularity shift, we predicted positive answers to "If M A then A" and "If A then M A" except for combinations with minimizers and total adjectives (**Prediction 3a**) and "if minimized partial A then A" $\gg$ "if minimized total A then A" (**Prediction 3b**).

Before we proceed to describe the studies, the next section discusses shortly the notions of alternative sets and level of fit in relation to upper-bounded vs. upper-open interpretations in numerals vs. adjectives.

1.6 Upper-bounded vs. Upper-open readings

Recently it has been convincingly demonstrated (Breheny 2005; Geurts 2009) that numerals have upper-bounded interpretations (e.g., *five* means 'exactly five'). In contrast, experimental evidence suggests that upper-bounded readings of bare adjectives, such as a reading of *cool* as 'not cold', are rarer than upper-bounded readings of numerals, quantifiers (*some* as 'not all'), and modals (*probably* as 'not definitely') (Doran, Baker, McNabb, Larson & Ward 2009; Zevakhina & Geurts 2011).

In the inference task in the study of Zevakhina & Geurts (2011), participants rarely admitted *yes* responses to texts like *John says: The sand was warm. Would you infer from this that, according to John, the sand was not hot?*. Thus, they did not derive upper-bounded readings for adjectives. An additional task, where participants had to list 3 alternatives for an adjective in a sentence such as *The sand is warm*, showed that an increased degree of availability of an alternative set increases the probability of an upper-bounded interpretation.

The verification study in Doran et al. (2009) also demonstrated that setting an adjective within an alternative set increases the probability of an upper-bounded interpretation. Given the fact *Jeremy can't fit in an airplane seat*, the multi-condition question *In terms of size would you say that Jeremy is average, big or huge?* significantly increased the likelihood of a scalar implicature compared to the none condition *What size is Jeremy?* and the one condition *In terms of size would you say that Jeremy is huge?*.

In sum, adjectives tend to be interpreted along coarse-grained scales, with high approximation levels, therefore their denotation range is extended, possibly up to the high end of the scale (an upper-open interpretation). However, contexts in which richer alternative sets are salient trigger interpretations along finer scales, with increased probability of upper-bounded interpretations. In addition, an overall low rate of upper-bounded readings with unmodified total or partial adjectives might be due to an effect of level of fit, namely due to the fact that entities with high degrees have the highest level of fit even in adjectives with low standards of membership.[4]

The studies described above did not test semantic relations and inferences in statements with modified adjectives. We predict that in these cases too upper-bounded readings will be rare since the alternative set is vague. No fixed set of lexical items and modifiers is salient outside context. Evidence that, e.g., the modified adjective *slightly dirty* has an 'at least slightly dirty and possibly dirty' reading comes from the fact that the typical intonation contour of utterances such as *The floor is not slightly dirty, it is very dirty* is characteristic of metalinguistic negation. This suggests that the logical negation of a minimized adjective cannot be used to refer to high degrees, as these are part of the minimized adjective denotation

4 In that, the role of level of fit resembles the role of typicality in van Tiel (2012).

(e.g., *slightly dirty*)[5,6].

2 Experimenting with modified and unmodified adjectives

2.1 Study 1: Granularity shifting

2.1.1 Methods

Participants were recruited using Amazon Mechanical Turk (AMT). Every hit was answered by 25 participants. All in all, 295 participants answered an average of 17 questions per participant ($SD = 34$). *The lexical items* consisted of 34 adjectives divided to by their standard type, with 17 partial adjectives (*dirty, sick, wet, bent, open, transparent, difficult, longer, awake, visible, dangerous, worried, bumpy, inaccurate, impure, needed*, *uncertain*) and 17 total adjectives (*clean, healthy, dry, straight, closed, opaque, full, empty, asleep, invisible, safe, unworried, flat, accurate, pure, unneeded*, *certain*), arbitrarily chosen from a pool of widely accepted standard-based typologies (Rotstein & Winter 2004; Kennedy & McNally 2005; Kennedy 2007).

Each adjective A occurred in 6 versions of a text involving a modifier M, *slightly, somewhat*, or *a bit* for the partial adjectives, and *completely, entirely*, or *perfectly* for the total ones. The texts consisted of an agreement task with either a coarse-fine inference ("Nick thinks that x is A. Nick's mother thinks that x is M A. Would Nick agree that x is M A?"), or a fine-coarse inference ("Nick thinks that x is M A. Nick's mother thinks that x is A. Would Nick agree that x is A?"). The fillers consisted of similar texts, except that instead of "x is A" and "x is M A" they included statements such as "x is more A1 and A2" and either "x is more A2", or "x is less A2". Half of the fillers were likely to be answered positively and half negatively.

This design resulted in 204 target sentences (3 modifiers $\times$ 34 adjectives $\times$ 2 inferences), and 272 fillers. The filler and target texts were counterbalanced into 34 lists of 14 texts each, including 8 fillers and 6 targets. Participants were asked to provide an answer on a scale ranging from 1 (certainly not) to 5 (certainly yes). The predictions are listed in sections 1.4-1.5.

	SOMEWHAT	SLIGHTLY	A BIT	Total
If A, M A	4.09 (.29)	3.89 (.23)	4.07 (.28)	4.01 (.28)
If M A, A	4.29 (0.2)	4.22 (0.23)	4.33 (0.16)	4.28 (.20)
Total	4.19 (0.27)	4.05 (0.28)	4.2 (0.26)	**4.14 (.28)**
	ENTIRELY	COMPLETELY	PERFECTLY	Total
If A, M A	4.22 (.19)	4.28 (.25)	4.22 (.18)	4.24 (.21)
If M A, A	4.85 (0.09)	4.85 (0.08)	4.8 (0.16)	4.83 (.11)
Total	4.54 (.35)	4.56 (0.34)	4.51 (0.34)	**4.54 (.34)**

Table 1 Averaged agreement ratings plus standard deviations for experiment 1.

2.1.2 Results and discussion

Table 1 and figure 3 present the averages over 17 target items per condition of the averages over 25 participants per item. The 136 positive and 136 negative fillers had averaged ratings of 4.85 (.12) and 1.29 (.22), respectively. Thus, all the averaged scores of the target items fall within the positive range (points 3.8-5). This suggests that the range denoted by maximizers and minimizers alike begins at the denotation minimum (not scale minimum). It also shows that upper-bounded readings are relatively minor (in the given types of context). For example, in the given texts, *slightly dirty* is not interpreted as 'at most slight amount of dirt' (*dirty* only given fine granularity: $[\text{dirty}]_{gp}$ and $[\neg\text{dirty}]_{g}$), and *clean* is not interpreted as 'at most clean' (*clean* only given coarse granularity: $[\text{clean}]_{g}$ and $[\neg\text{clean}]_{gp}$).

A 2-way factorial ANOVA for 2 samples (partial adjectives + minimizers and total adjectives + maximizers), with 2 repeated measures (inference forms "If A, M A" and "If M A, A") yields a significant effect for adjective + modifier type ($F(1,50) = 155.6$, $p < .0001$), and inference type ($F(1,50) = 312.67$, $p < .0001$), along predictions 1 and 2, and also a significant interaction ($F(1,50) = 45.33$, $p < .0001$). The interaction value is due to the fact that the difference between inference types is more pronounced with the maximizers than with the minimizers, and the effect of adjective + modifier is more pronounced with the inference pattern "If M A, A" (M=4.56, SD=3.2) than with "If A, M A" (M=4.13, SD=.27).

Thus, prediction 2 of the granularity shifting account is borne out by these findings. The pattern "If A, M A" involves moving from coarse to fine granularity.

5 Moreover, negating minimized adjectives, as in ?*not slightly dirty*, is a bit odd, except in the presence of a marker of unexpectedness, as in *not even slightly dirty*, which is perfectly felicitous. The reason is arguably that minimized adjectives refer to a wider range than non-minimized ones. It is especially unlikely for an entity to fall outside this wide range.

6 Notice, however, that our studies only test for inferences between statements of the form "x thinks that y is A/ M A" and "x agrees that y is M A / A", where A is an adjective, and M a modifier.

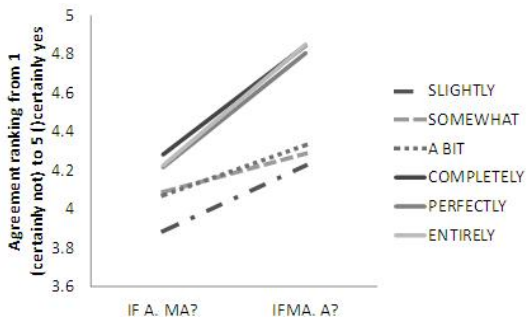

Figure 3 Averaged agreement ratings for experiment 1.

Speakers directly infer from an utterance of *The floor is dirty* the entailment that the floor is slightly dirty. However, since the floor could be dirtier than a floor that Nick would typically call slightly dirty, the answers for the inference pattern "If A, M A" are only weakly positive, because there is room for doubt; a "certainly yes" answer would suggest that Nick uses the words in an atypical way.

By contrast, in the pattern "If M A, A", since shifting back from fine to coarse granularity is not acceptable, *slightly* triggers an irreversible shift to finer granularity. Hence, in support of Lewis (1979), *dirty*, affected by the shift, is interpreted on a fine-grained scale, as equivalent to its minimized form. So speakers infer from *The floor is slightly dirty* that the floor is dirty with greater certainty than vice versa, i.e. the answers for "If M A, A" are significantly more positive than for "If A, M A". Similar reasoning holds for, e.g., *clean* and *completely clean*.

As the tests above show, the results for total adjectives + maximizers ($M = 4.54$, $SD = .34$) are significantly more positive than for partial adjectives + minimizers ($M = 4.15$, $SD = .28$). This suggests that the default interpretation of unmodified total adjectives is closer to that of their modified form than the default interpretation of unmodified partial adjectives is close to that of their modified form. However, the design of this experiment makes it impossible to say whether this effect is due to modifier type (prediction 1) or adjective type. Study 2 aims to answer this question.[7]

7 The results of Wilcoxon signed-ranks tests further show that all of the maximizers behave alike, while the minimizer *slightly* is distinguished from the other minimizers, mainly with respect to the pattern "If A, M A" where it induces the least positive results.

	If A, completely A	If A, slightly A	If completely A, A	If slightly A, A
Total	6.03 (.35)	4.52 (.37)	6.73 (.19)	4.69 (.38)
Partial	5.62 (.30)	5.09 (.21)	6.72 (.18)	5.70 (.37)
Relative	5.58 (.34)	5.25 (.46)	6.75 (.08)	5.44 (.48)
TOTAL	5.77 (.38)	4.89 (.45)	6.73 (.16)	5.24 (.61)

Table 2 Averaged agreement ratings plus standard deviations for experiment 2.

2.2 Study 2: Level of fit and modifier vs. adjective type effects

2.2.1 Methods

Participants were recruited using AMT. All in all, 161 participants answered on average 19 questions per participant ($SD = 29$). *The lexical items* consisted of 30 different adjectives divided to 5 types by their standard and scale type: total doubly-closed (*full, closed, empty, invisible, correct, opaque*), total upper-closed (*clean, healthy, dry, straight, safe, flat*), partial doubly-closed (*open, transparent, visible, wrong, incorrect, unclear*), partial lower-closed (*dirty, sick, wet, bent, dangerous, bumpy*), and relative (*long, short, big, small, happy, sad*), arbitrarily chosen from a pool of widely accepted standard- and scale-based typologies (Rotstein & Winter 2004; Kennedy & McNally 2005; Kennedy 2007).

Each adjective occurred in 4 versions of a text involving either the modifier *slightly* or the modifier *completely*, and the same inference patterns as in study 1. This design resulted in 120 target texts (5 adjective types $\times$ 6 items $\times$ 2 modifiers $\times$ 2 inference types). The fillers were target items of other experiments, consisting of the same texts as the targets, except that instead of statements of the form "x is (M) A", 120 fillers involved statements with round vs. precise numerals (as in "x has nine/ten shirts"), and 136 fillers involved statements with adjectives in the comparative form ("x is more attentive/caring than y"). The participants were asked to provide an answer on a scale ranging from 1 (certainly not) to 7 (certainly yes).

2.2.2 Results and discussion

Based on the averages over 25 participants per text, table 2 and figure 4 present the averages over texts with partial, total and relative adjectives. The results of experiment 1 were replicated. A 2-way factorial ANOVA for 3 blocks (12 total, 12 partial, and 6 relative adjectives) and 4 repeated measures (2 inference types with *slightly* vs. *completely*), yields a significant difference between inference types ($F(3,29) = 192$, $p < .0001$), and between adjective types ($F(2,29) = 11.4$, $p < .0003$), as well as a significant interaction ($F(6,29) = 13.8$, $p < .0001$).

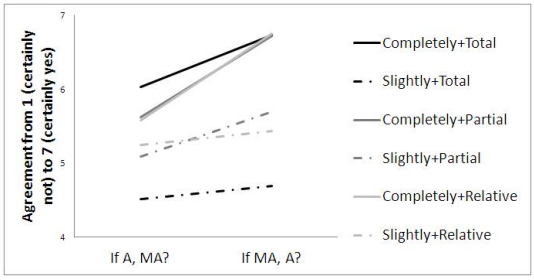

Figure 4 Averaged agreement ratings for experiment 2.

Regarding inference type, a Wilcoxon signed-ranks test yields that ranking of agreement is generally significantly higher for "If M A, A" than for "If A, M A" ($W = -866$, $n_{s/r} = 43$, $z = -5.23$, $p < .0001$), and this is also the case for *completely* ($W = 465$, $n_{s/r} = 30$, $z = 4.78$, $p < .0001$), and *slightly* ($W = 330$, $n_{s/r} = 30$, $z = 3.39$, $p < .0007$) considered separately. As in study 1, this supports the hypothesized facilitation of inference from modified to unmodified adjectives (fine to coarse interpretations) than vice versa (prediction 2).

Regarding modifier type, a Wilcoxon test yields that ranking of agreement is generally significantly higher for *completely* than for *slightly* ($W = 1790$, $n_{s/r} = 60$, $z = 6.59$, $p < .0001$), and this is the case for "If M A, A" ($W = 465$, $n_{s/r} = 30$, $z = 4.78$, $p < .0001$) and for "If A, M A" ($W = 425$, $n_{s/r} = 30$, $z = 4.37$, $p < .0001$). Thus, inferences are easier with maximizers than minimizers, even when maximized partial adjectives and minimized total ones are considered (prediction 1).

Furthermore, a 2-way factorial ANOVA for 2 repeated measures (inference type: "If slightly A, A" vs. "If A, slightly A"), yields a significant difference between inference types ($F(1,29) = 20.67$, $p < .0002$), and adjective types ($F(2,29) = 23.83$, $p < .0001$), and a significant interaction ($F(2,29) = 3.78$, $p < .04$). The interaction is due to the fact that the difference between inference types is not significant in inferences with *slightly* and total adjectives. In addition, certainty is lower with minimized total than relative or partial adjectives. Thus, predictions 3a-b are borne out.

Agreement was predicted to be higher with partial adjectives than total ones, due to the fact that the interpretation of minimized total adjectives differs from the fine and precise interpretation of unmodified total adjectives. The former is based on shifting of the standard to a degree lower than the maximum and its best level of fit is at that lower point. Thus, it is not even included in the maximum (the denotation of the unmodified total adjective, as illustrated in figure 5).[8]

8 A 2-way factorial ANOVA for 2 repeated measures (inference type: "If completely A, A" vs. "If A, completely A"), yields a significant difference between inference types ($F(1,29) = 195.29$, $p <$

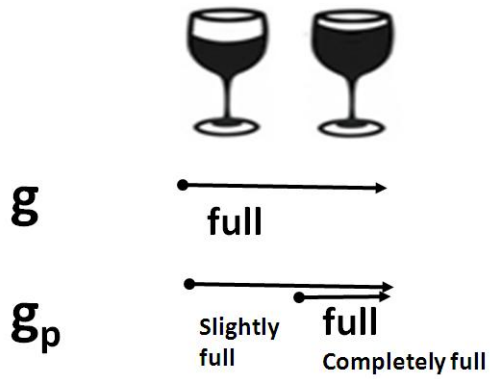

Figure 5 The coarse and fine interpretation of total adjectives like *full*.

As for more fine-grained distinctions, significant differences were found between adjectives differing in their standard type, but not scale type. Doubly-closed partial and doubly-closed total adjectives differ significantly in the patterns "If A, completely A" ($M_p = 4.3$, $M_t = 8.7$; $U = 31$, $z = -2$, $p < .05$), "If slightly A, A" ($M_p = 8.9$, $M_t = 4.1$, $U = 32.5$, $z = 2.24$, $p < .03$), and "If slightly A, A" ($M_p = 9.3$, $M_t = 3.8$, $U = 34.5$, $z = 2.56$, $p < .011$). Two additional differences were found between adjectives differing in their scale type but not standard type. Considering the inference "If A, completely A", a Mann-Whitney test yields significantly more positive average responses for doubly-closed total adjectives ($n = 6$, $M = 8.6$) than only upper-closed total ones ($n = 6$, $M = 4.4$; $U = 5.5$, $z = 1.92$, $p < .055$). By contrast, considering the inference "If slightly A, A", a Mann-Whitney test yields significantly more positive average responses for only lower-closed partial adjectives ($n = 6$, $M = 8.5$) than doubly-closed partial ones ($n = 6$, $M = 4.5$; $U = 6$, $z = 1.84$, $p < .07$).

To summarize, according to the granularity shifting account, adjectival modification triggers a granularity shift, rendering relevant, e.g., hardly visible dirt specks that are by default ignorable in judging cleanliness. This analysis correctly predicted lower agreement rates in coarse to fine patterns than in fine to coarse patterns ("If M A then A" $\gg$ "If A then M A"). This effect was predictably absent in total adjectives modified by *slightly*. The reason is that the precise interpretation of, e.g., *full* is different from that of *slightly full*. The fine and precise interpretation of *full*, $[\text{full}]_{gp}$, consists of maximally full entities, where any tiny amount of content missing (e.g.

.0001), and adjective types ($F(2,29) = 4.57$, $p < .02$), and a significant interaction ($F(2,29) = 4.71$, $p < .02$), mainly due to a higher certainty in total than relative or partial adjectives concerning the inference "If A, completely A", but equal certainty concerning "If completely A, A", in line with scale structure theory.

two drops in a glass) render an object non-maximally full. By contrast, in addition to triggering a shift to g_p, *slightly* also triggers a standard shift such that a non-maximal external threshold is set for denotation members to exceed. This means that *slightly full* does not entail *full*, although it implies *rather full* or *fuller* (it doesn't entail *not full* either).

3 Conclusions: Toward a general theory of granularity

The results of the studies presented support the hypotheses based on Lewis's constraints on granularity shifting (1979), and an extension to adjectives of Krifka's representation of granularity in numerals (2007), modulo semantic differences between adjectives and numerals. This supports the view that general principles govern the setting of a granularity level and its shifting within discourse, principles which apply in different domains of grammar, including round and complex numerals and unmodified and modified adjectives.

A crucial difference between numerals and adjectives is that on fine interpretations different numerals denote different points, while modified and unmodified adjectives often denote highly overlapping intervals. This fact might be connected with the fact that upper-bounded readings are minor in adjectives (Doran et al. 2009). Future work should address this issue, as well as the connections between level of fit and scalar implicatures, and their respective role in explaining inference data.[9]

References

Anderson, Norman H. 1996. *A Functional Theory of Cognition*. New Jersey: Lawrence Erlbaum Associates.

Benz, Anton, Gerhard Jäger & Robert van Rooij (eds.). 2006. *Game Theory and Pragmatics*. Oxford: Palgrave Macmillan.

Breheny, Richard. 2005. Some scalar implications really aren't quantity implicatures but 'some"s are. In Emar Maier, Corien Bary & Janneke Huitink (eds.), *Sinn und Bedeutung (SuB) 9*, 40–64. Ithaca, NY: CLC Publications.

Dehaene, Stanislas. 1997. *The Number Sense: How the Mind Creates Mathematics*. Oxford: Oxford University Press.

Doran, Ryan, Rachel Baker, Yaron McNabb, Meredith Larson & Gregory Ward.

9 The reader is referred to Sassoon & Zevakhina (2012) for a presentation of the results pertaining to texts with round and precise numerals (the fillers of study 2 of this paper). In addition, Sassoon & Zevakhina (2012) used texts with modified adjectives, but with an entailment task, rather than an agreement task. The general structure of the texts was "If A, does it follow that M A" and vice versa. Such a task was predicted to minimize level of fit effects, and maximize the role of precise interpretations. The predictions were borne out.

2009. On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics* 1(1). 211–248.

Geurts, Bart. 2009. Scalar implicature and local pragmatics. *Mind and Language* 24(1). 51–79.

Heim, Irene. 2000. Degree operators and scope. In Brendan Jackson & Tanya Matthews (eds.), *Semantics and Linguistic Theory (SALT) 10*, 40–64. Ithaca, NY: CLC Publications.

Heim, Irene & Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Oxford: Blackwell.

Kennedy, Christopher. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjective. *Linguistics and Philosophy* 30(1). 1–45.

Kennedy, Christopher & Louise McNally. 2005. Scale structure and the semantic typology of gradable predicates. *Language* 81(2). 345–381.

Krifka, Manfred. 2002. Be brief and vague! And how bidirectional optimality theory allows for verbosity and precision. In David Restle & Dietmar Zaefferer (eds.), *Sounds and Systems. Studies in Structure and Change. A Festschrift for Theo Vennemann*, 439–458. Berlin: Mouton de Gruyter.

Krifka, Manfred. 2007. Approximate interpretation of number words: A case for strategic communication. In Joost Zwarts Gerlof Bouma, Irene Krämer (ed.), *Cognitive Foundations of Interpretation*, 111–126. Amsterdam: Mouton de Gruyter.

Lasersohn, Peter. 1999. Pragmatic halos. *Language* 75(3). 522–551.

Lewis, David. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic* 8(1). 339–359.

McNally, Louise. 2011. The relative role of property type and scale structure in explaining the behavior of gradable adjectives. In Rick Nouwen, Robert van Rooij, Uli Sauerland & Hans-Christian Schmitz (eds.), *ESSLLI 2009 Workshop on Vagueness in Communcation*, 151–168. Berlin: Springer.

Paradis, Carita & Caroline Willners. 2006. Antonymy and negation: The boundedness hypothesis. *Journal of Pragmatics* 38(7). 1051–1080.

Parikh, Prashant. 2000. Communication, meaning and interpretation. *Linguistics and Philosophy* 23(2). 185–212.

van Rooij, Robert. 2011. Vagueness and linguistics. In Giuseppina Ronzitti (ed.), *The Vagueness Handbook*, 123–170. Dordrecht: Springer.

Rotstein, Carmen & Yoad Winter. 2004. Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. *Natural Language Semantics* 12(3). 259–288.

Sassoon, Galit W. 2012. A slightly modified economy principle: Stable properties have non stable standards. In *Proceedings of the Israel Association of Theoretical Linguistics 27*, Bingley, MA: MIT working Papers in Linguistics.

Sassoon, Galit W. & Natalia Zevakhina. 2012. Granularity shifting: Experimental evidence from numerals vs. adjectives. In *Proceedings of the Israel Association of Theoretical Linguistics 28*, Bingley, MA: MIT working Papers in Linguistics.
Syrett, Kristen & Jeffrey Lidz. 2010. 30-month-olds use the distribution and meaning of adverbs to interpret novel adjectives. *Language Learning and Development* 6(4). 258–282.
van Tiel, Bob. 2012. Embedded scalars and typicality. Unpublished MS. Radboud University Nijmegen.
Zevakhina, Natalia & Bart Geurts. 2011. Scalar diversity. Unpublished MS. National Research University and Nijmegen.

Galit Sassoon
Mount Scopus
The Hebrew University, LLCC
91905 Jerusalem, Israel
galitadar@gmail.com

Natalia Zevakhina
Khitrovsky lane 2/8
National Research University
Higher School of Economics
Department of Linguistics
109028 Moscow, Russia
natalia.zevakhina@gmail.com