GRANULARITY SHIFTING: EXPERIMENTAL EVIDENCE FROM NUMERALS

GALIT SASSOON[1]
NATALIA ZEVAKHINA[2]
[1]*Hebrew University,* [2]*National Research University Higher School of Economics*

1 Introduction

Simple/short expressions like the numeral *ten* tend to have coarser and more approximate interpretations than complex/long ones like *nine point three* which tend to have fine and precise interpretations (Krifka 2002, 2007). Krifka models this phenomenon by means of a representation of granularity. In addition, Lewis (1979) discusses constraints on licensed types of shifting between different granularity levels within discourse. This paper aims to test empirically the validity of Lewis's (1979) constraints in the domain of numerals (experiment 1). In addition, Sassoon and Zevakhina (2012) report the results of a similar investigation in the domain of adjectives and their modifiers. This paper compares the two sets of results, and using a new type of task, it reports a replication of some of the results with adjectives and their modifiers (experiment 2).

1.1 Granularity: A Krifka-Lewis Analysis of Numerals

When we want to express certain meanings, we select a particular set of expressions – the *expression-choice space*, out of which we choose the expressions that identify the semantic objects. Here are several examples:

(1) a. … one thousand, two thousand, three thousand, …
b. … nine hundred, one thousand, one thousand one hundred, …
c. … nine hundred ninety, one thousand, one thousand and ten,…
d. … nine hundred ninety-nine, one thousand, one thousand and one, …

(2) a. dirty, clean;
b. dirty, clean, sterile;
c. completely dirty, very dirty, pretty dirty, slightly dirty, slightly clean, pretty clean, very clean, completely clean, sterile, completely sterile.

(3) a. hot, cold
b. very hot, hot, warm, tepid, cool, cold, very cold
c. …, 10^{o}, 15^{o}, 20^{o} …
d. …, 14^{o}, 15^{o}, 16^{o}, …

(4) a. early, late
b. very early, early, slightly early, on time, slightly late, late, very late
c. …, 10 minutes late, 15 minutes late, 20 minutes late …
d. …, 14 minutes late, 15 minutes late, 16 minutes late, …
c. …, 10 seconds late, 20 seconds late, 30 seconds late …
d. …, 14 second late, 15 seconds late, 16 seconds late, …

Importantly, once we choose a certain set (say 1b) and we use an expression of this set (say *nine hundred*), the expressions of the set that we do not choose function as alternatives, so that the fact that they were not chosen may lead to pragmatic implicatures (e.g., an utterance such as *I have nine hundred stickers,* may give rise to the inference that I do not have thousand stickers). Expressions that are not in the set do not count in the derivation of implicatures. They do affect, however, the derivation of vague vs. precise interpretations, according to Krifka's proposal, as explained shortly.

Notice that there is a systematic relation between these expression-choice spaces: The upper ones are *coarser-grained*, the lower ones are *finer-grained*. We can model this relationship by set inclusion of sets of expressions. In addition, the average complexity of expressions, as measured by, e.g., number of syllables, is greater in finer-grained expression spaces than in less fine-grained ones.

Krifka (2002) accounts for the fact that short expressions have a preference for vague interpretations, and long expressions have a preference for precise interpretations, using Bidirectional Optimality Theory (Blutner 1998, 2000; Dekker & van Rooy 2000), and assuming pragmatic tendencies for short expressions (5a) and vague interpretations (5b). Given (5a), the use of an expression α introduces the most coarse-grained expression-choice space possible for it into the pragmatic evaluation.

(5) a. BREVITY: Expression-choice spaces with shorter, less complex expressions are preferred over expression-choice spaces with longer, more complex ones.
b. VAGUENESS: Measurement terms are preferably interpreted in a vague way.

In classical Optimality Theory (OT), the input to a rule expressed in a tableau of ranked or unranked constraints is a set of expressions, and the output is the one expression, or the set of expressions, that violate the constraints the least. In semantic and pragmatic applications of Bidirectional OT, the inputs are pairs of objects – an Expression and its Interpretation, E, I; constraints are independently specified for the members of the pair, and the output are those pairs that violate the constraints the least. In addition, in Jäger's (2002) *bidirectional super-optimality*

comparison of pairs, a pair ⟨E, I⟩ of a set of candidate expressions GEN is super-optimal iff there is no super-optimal ⟨E', I⟩ ∈ GEN such that either ⟨E', I⟩ $>>_E$ ⟨E, I⟩ or there is no super-optimal ⟨E, I'⟩ ∈ GEN such that ⟨E, I'⟩ $>>_I$ ⟨E, I⟩. The notion of super-optimal pairs ⟨E, I⟩ is thus restricted to those that have no competitor on the expression level or on the interpretation level that is itself super-optimal.

It follows that sometimes violating one constraint allows for the violation of other constraints, with no punishment. For example, considering that the distance between A and V is 965 km, and the Expression-interpretation pairs in (6), we get the comparisons in (7).

(6) a. < The distance between A and V is *one thousand* kilometers, **vague** >
b. < The distance between A and V is *one thousand* kilometers, **precise** >
c. < The distance between A and V is *nine hundred sixty-five* kilometers, **vague** >
d. < The distance between A and V is *nine hundred sixty-five* kilometers, **precise** >

(7) a. (6a) $>>_I$ (6b)
b. (6a) $>>_E$ (6c)
c. (6c) $>>_I$ (6d)
b. (6b) $>>_E$ (6d)

Notice that (6a) and (6d) cannot be compared directly, as they differ both in their expression component and in their interpretation component, and *bidirectional super-optimality comparison* only allows for comparison between pairs that are identical in one component. Clearly, (6a) is a super-optimal pair. There is no super-optimal pair ⟨E, I⟩ such that either ⟨E, I⟩ $>>_E$ (6a) or ⟨E, I⟩ $>>_I$ (6a). Hence, we predict that brief measure expressions are preferably interpreted in a vague way. The pairs (6b) and (6c) clearly are not super-optimal, because they have a better super-optimal competitor, (6a). But (6d) is also a super-optimal pair, for the only competitors (6b) and (6c) are not themselves super-optimal. Hence we predict that complex measure expressions are preferably interpreted in a precise way. For example, in the evaluation of an expression such as *The distance between Amsterdam and Vienna is nine hundred sixty five kilometers*, we consider a vague vs. precise interpretation, (6c) and (6d). In general, vague interpretations are preferred, so we are inclined to impose a vague interpretation, (6c). But under a vague interpretation, we could express the same truth conditions with *The distance between Amsterdam and Vienna is one thousand kilometers*, (6a). The pair (6a) should be preferred over (6c) because it is shorter. Since the speaker has not used it, the vague interpretation is rejected in favor of the precise one (6d). This is how violation of one constraint (brevity) allows for violation of another (vague interpretation).[1,2]

Krifka (2007) provides an alternative account in terms of the principles of **strategic communication** (Parikh 1991, 2000, Benz, Jäger, and van Rooij 2006), whereby an addressee assumes the speaker intends the most likely interpretation and a speaker assumes the addressee selects the most likely interpretation, given the choice of expressions and the a-priori likelihood of a message. The basic question for interpretation is therefore: Given the a-priori-likelihood of

[1] 965 can have a vague interpretation only in relation to a finer choice space (one that includes expressions such as 963.5 and 967.5). In other words, the choice between vague and precise interpretations depends on the choice space.

[2] Lasersohn (1999) developed a theory of imprecise interpretations in terms of pragmatic "slack" of expressions. But the relation to the complexity of expressions is not addressed in his paper.

the communicated information and general interpretation strategies, what is the most likely interpretation of an ambiguous or vague form?

Within this framework, the approximate interpretations of a numeral *n* are represented by an interval [*n – n/s, n + n/s*], and a function representing goodness of fit of each point in this interval. This function is thought to have the bell shape of a normal distribution with mean *n* and standard deviation *s*. At a level of approximation of, for example, s = 10, a reported value *Ten* would stand for [9, 11], whereas a reported value of *100* would stand for [90, 110].The quotient of standard deviation *s* and mean *n* indicates the coarseness of the interpretation; with smaller *n/s*, the interpretation gets more precise. At a level of approximation of 0 (s = ∞), numbers are interpreted in a precise way (cf. Dehaene 1997).

Assuming equal probabilities for scale points, it follows that approximate interpretations are more probable than precise ones (because they include more points), and interpretations relative to coarse scales are more probable than fine ones (for the same reason). Assuming a preference for short expressions, the shortest among any set of numerals with roughly the same distribution is associated with the most probable (coarse and approximate) interpretation, while the longer ones must, therefore, be interpreted more precisely.

However, besides these results, the theory can explain apparent exceptions to the preference for short expressions; e.g. why is *forty five minutes* interpreted in a less precise way than *forty minutes*? The explanation proposed is as follows. The most frequent type of scale in the western culture is the one based on the powers of ten, but other dominant representations exist, such as the 15-point and 5-point time scales in (8b,c):

(8) a. 0----------60----------…----------120-…
b. 0----------30----------60----------90----…-----120-…
c. 0------15----------30-------45----------60----------75-------90----…-----120-…
d. 0-5-10-15-20-25-30-35-40-45-50-55-60-65-70-75-80-85-90-95-…--120-…

The use of *forty five minutes* invokes the coarser-grained scale (8c) rather than the fine-grained scale (8d). With respect to level (8c), the scale point 45 represents the times in [38, 52]. With respect to level (8d), the scale point 45 represents the times in [43, 47].[3] Let the a-priori probability for a report of each time in the range be the same, say *r*. Then the probability that the measured time is in [43, 47] is 5r, and the probability that the measured time is in [38, 42] is 15r. Let the a-priori probability that one of the scales (8c) or (8d) are used be the same, say *s*. Then the probability that the finer-grained scale (8d) is used is 5rs, and the probability that the more coarse-grained scale (8c) is used is bigger, 15rs. Hence, a hearer will assume the coarser-grained scale (8c). By contrast, *forty* does not denote any scale point on (8c), so it cannot be interpreted at this level. The first scale it can be interpreted at is (8d), where it captures time in [37, 42]. By similar reasoning, this is the level at which *forty* will be interpreted, and not the even finer-grained level of single minutes with scale points like 38, 39, 40, 41 etc.

Finally, Lewis (1979) argues that a shift from default to finer granularity and/or precision level is a natural discourse move, but the opposite shift is not. For example, we may state that the Netherlands is flat, presupposing default coarse granularity *g*, and then point out that it is actually a bit bumpy by shifting to a finer criterion g_p, taking as evidence bumps we previously ignored. However, we cannot state that the Netherlands is bumpy and smoothly proceed to say

[3] On the assumption that 45 is never used to convey a point closer to the precise interpretation of 44 or 43 than to that of 45.

that it is actually flat, thereby ignoring bumps that we previously regarded as relevant evidence. This means that once an expression of a fine alternative set is uttered like *nine hundred sixty five*, the interpretation of an expression like *one thousand* is forced to be relative to a finer granularity scale than the default for this expression.

Putting it all together, the simple numeral *10* in (9a) is expected to have a default coarse and approximate interpretation 'about ten', say, $[\text{ten}]_g$ = [9.5, 10.5), and *9.5* is expected to have a precise interpretation, $[9.5]_{gp}$ = [9.5]. Since the latter is part of the former, $9.5 \in [9.5, 10.5)$, in (9a) B's utterance is understood as confirming A's utterance. By contrast, we expect that following an utterance of a complex numeral, e.g., *9.5* in (9b), the interpretation of the simple numeral *10* would be finer and more precise than the default, $[\text{ten}]_{gp}$ = [10]. Since $10 \neq 9.5$, and since numerals tend toward upper-bounded or 'exact' readings (Breheny 2005; Geurts 2009; e.g. *9.5* and *10* are understood as conveying 'exactly 9.3' and 'exactly 10', in particular when interpreted along precise criteria), in (9b) B's utterance is understood as contradicting A's. Therefore, we predict higher agreement of speaker B to speaker A's utterance in coarse-to-fine contexts like (9a) than in fine-to-coarse ones like (9b).

(9) a. A: Stock Exchange fell by ten percent. B: Yes (#No), it fell by nine point five percent.
b. A: Stock Exchange fell by nine point five percent. B: No (#Yes), it fell by ten percent.

Experiment 1 tested similar predictions of a Krifka-Lewis analysis of numerals.

2 Experiment 1: Inference Relations between Statements with Round vs. Precise Numerals

2.1 Methods

Participants were recruited using Amazon Mechanical Turk (AMT) – an online labor market place where workers are paid small amounts of money to complete small tasks named HITs (Human Intelligence Tasks). It has been shown that AMT provides a quick and relatively cheap method to acquire high-quality experimental results that do not differ significantly in performance from standard experimental settings (Buhrmester et al. 2011). The hits were not accessible to workers outside the USA. The reward per hit was \$0.3 cents. All in all, 156 participants answered on average 19 questions per participant (*SD* = 30), with Average Hourly Rate of \$5.7.

The stimuli consisted of sixty texts containing statements with numerical expressions. The texts were divided to 15 conditions by the type of round number (*10, 100* or *1000*) and precise number (e.g., for *10* there were *9, 9.3, 8.7, 9.33* and *8.67*). The precise numbers differed in terms of their granularity (coarse, fine and very fine, as in *9, 9.3* and *9.33* for *10*), depending on the number of non-zero digits, and in terms of their distance from the round numbers (short, middle or long, as is the distance of *9.3, 9* and *8.7* from *10*, respectively). Granularity and distance were combined into five condition types all in all (see Appendix 1). In addition, each item occurred in two different versions corresponding to two inference types ("If round, precise", "If precise, round").

The following examples illustrate the five resulting conditions for the round number *100* in coarse-fine contexts ("If round, precise"):[4]

(10) a. *Middle distance, coarse granularity* ("If 100, 99"):
Nick thinks there are 100 balloons in the sky. Nick's mother thinks there are 99 balloons in the sky. Would Nick agree that there are 99 balloons?
b. *Short distance, fine granularity* ("If 100, 99.2"):
Nick thinks the table's width is 100 cm. Nick's mother thinks it is 99.2 cm. Would Nick agree that it is 99.2 cm?
c. *Long distance, fine granularity* ("If 100, 98.8"):
Nick thinks Sara jumped a distance of 100 cm. Nick's mother thinks it was 98.8 cm. Would Nick agree that it was 98.8 cm?
d. *Short distance, very fine granularity* ("If 100, 99.23"):
Nick thinks George's height is 100 cm. Nick's mother thinks it is 99.23 cm. Would Nick agree that it is 99.23 cm?
e. *Long distance, very fine granularity* ("If 100, 98.77"):
Nick thinks Bill's weight is 100 kilos. Nick's mother thinks it is 98.77 kilos. Would Nick agree that it is 98.77 kilos?

This design resulted in 30 conditions (five granularity & distance levels × three round numbers × two inference forms). With four items per condition it yielded all in all 120 target contexts (see Appendix 2).

The fillers consisted of two types of texts belonging to two different experiments. The first type consisted of 120 fillers involving inference relations between statements with modified and unmodified adjectives, and the second type consisted of 136 fillers involving inferences between statements with two different adjectives in the comparative construction. The following examples (11a,b) and (12a,b) illustrate these two types of fillers, respectively.

(11) a. Nick thinks that he is completely healthy. Nick's mother thinks that he is healthy. Would Nick agree that he is healthy?
b. Nick thinks that the road is bumpy. Nick's mother thinks that it's slightly bumpy. Would Nick agree that it's slightly bumpy?

(12) a. Nick thinks that his cousin is more attentive than his aunt. Nick's mother thinks that his cousin is more caring than his aunt. Would Nick agree that his cousin is more caring than his aunt?
b. Nick says that the side effects of the old medicine are more severe than those of the new one. Nick's mother says that they are more numerous than those of the new one. Would Nick agree that they are more numerous than those of the new one?

In all target and filler contexts, participants were asked to provide an answer to whether Nick would agree with his mother on a scale ranging from 1 (certainly not) to 7 (certainly yes).

Each text was presented as a separate hit. The target and filler contexts were presented in a randomized order except that each two target contexts were separated by at least two fillers of

[4] The precise numbers occurring with *1000* were exactly 10 times bigger (so the round number × precise number ratios were exactly the same).

different types. Each hit began with the following instructions: **Notice that this HIT is for English native speakers only!** *Read a short text about Nick and his mother and determine to what extent you think Nick agrees to a statement on a scale from 1 (certainly not) to 7 (certainly yes). For example, if the context says that Nick has 100 books and Nick's mother has 200 books and you are asked whether Nick would agree that his mother has more books, you will probably think he will certainly agree (answer 7). However, if you are asked whether Nick would agree that his mother has fewer books, you will probably think he would certainly not agree (answer 1).* The instructions were followed by one context and a scale for ranking.

2.2 Predictions

The foremost aim of the experiment was to test Lewis's granularity shifting hypothesis, namely the hypothesis that shifting from coarse to fine granularity is preferable over shifting from fine to coarse granularity. The main prediction is that in the context of an utterance with a fine expression, agreement to a subsequent similar utterance with a coarse expression, as in (9a), will be higher than agreement to an utterance with a fine expression given a similar prior utterance with a coarse expression, as in (9b). We can reformulate this prediction in the following way: agreement ratings in coarse-fine contexts should be significantly higher than ratings in fine-coarse ones (**Prediction 1**). A second null hypothesis is that the granularity shifting effect depends neither on the particular round number used (**Prediction 2a**) nor on the particular precise number (**Prediction 2b**). Another expectation is that, in the two context types, agreement rates will be inversely related to the distance between the round and precise numbers (**Prediction 3a**), and perhaps also to the extent to which the precise number is finer than the round one (**Prediction 3b**).

2.3 Results: Numerals

Table 1 presents averages over 25 participants for each one of the sixty items in the two inference types ("If round, precise", "If precise, round"). The means are near the mid of the seven-point scale, which is "3.5", ranging from "2.7" to "5.5". More importantly, as Table 1 shows, the means in the coarse-fine contexts are higher than the means in the fine-coarse contexts. A Wilcoxon signed-ranks test applied to two-sample of 60 matched pairs yields that agreement in contexts of shifting from coarse to fine granularity is greater than in contexts of shifting from fine to coarse granularity (z = -6.47, $p < .0001$).This result confirms **Prediction 1**, i.e., Lewis's granularity shifting hypothesis, which states that shifting from coarse to fine granularity is a natural discourse move, whereas shifting from fine to coarse one is not. This idea is also illustrated by Figures 1 and 2.

	10 coarse-fine	**10 fine-coarse**	**100 coarse-fine**	**100 fine-coarse**	**1000 coarse-fine**	**1000 fine-coarse**	**Total**
SHORT FINE	4.64 (0.69)	3.77 (0.59)	4.91 (0.55)	4.32 (0.5)	4.87 (0.48)	4.24 (0.71)	**4.46 (0.24)**
SHORT VERY FINE	4.37 (0.47)	3.66 (0.1)	4.86 (0.7)	3.86 (0.38)	4.76 (0.07)	3.9 (0.29)	**4.24 (0.16)**
MID COARSE	3.71 (0.14)	2.74 (0.4)	5.46 (0.39)	4.62 (0.17)	4.89 (0.5)	4.24 (0.18)	**4.28 (0.13)**
LONG FINE	3.88 (0.49)	2.97 (0.6)	4.95 (0.22)	3.85 (0.46)	4.81 (0.34)	3.71 (0.54)	**4.03 (0.19)**
LONG VERY FINE	3.41 (0.29)	2.83 (0.41)	4.88 (0.13)	4 (0.6)	4.6 (0.3)	3.58 (0.17)	**3.88 (0.14)**
Total	**4 (0.2)**	**3.28 (0.2)**	**5 (0.2)**	**4.13 (0.2)**	**4.77 (0.16)**	**3.93 (0.19)**	4.19 (0.08)

Table 1: Averaged agreement ratings (and standard deviations) for experiment 1.

The data, involving three types of round numbers and five types of distance × granularity levels, allow drawing additional conclusions. First, as Figure 1 clearly shows, agreement rates differed by the type of round number, but the inference effect was identical for the three tested round numbers. A two-way factorial ANOVA for two randomized blocks of matched items (type of inference: coarse-fine vs. fine-coarse) and a repeated measure (type of round number: *10* vs. *100* vs. *1000*) yields a significant difference between the inference types ($df = 1$, $F = 83.9$, $p < .0001$) and between the round number types ($df = 2$, $F = 40.8$, $p < .0001$), though no significant interaction ($df = 2$, $F = 0.54$, $p > .05$). Thus, **Prediction 2a** was borne out.

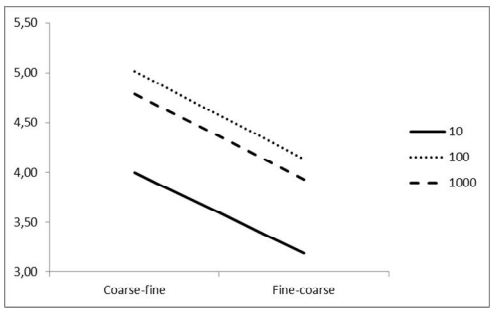

Figure 1: Averaged agreement ratings for the 3 round number conditions in the two inference types.

Second, as Figure 2 illustrates, agreement rates differed by the type of precise number, but the inference effect was identical for the five types of precise numbers. A two-way factorial ANOVA for two randomized blocks of matched items (type of inference: coarse-fine vs. fine-coarse) and a repeated measure (distance × granularity level) yields a significant difference between inference types ($df = 1$, $F = 52.6$, $p < .0001$) and between the five distance × granularity levels ($df = 4$, $F = 2.9$, $p < .05$), though no significant interaction ($df = 4$, $F = 0.2$, $p > .05$). Thus, **Prediction 2b** was borne out.

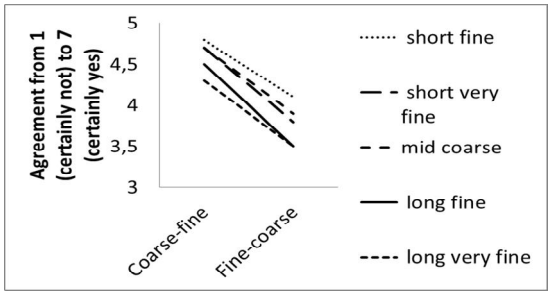

Figure 2: Averaged agreement ratings for the 5 precise number (distance × granularity) conditions in the two inference types.

Notice that the low agreement rates for the *10* conditions are probably due to the bigger distance between the round and precise numbers in this case, and as such they cannot be taken as evidence for an inherent difference between *10* and the bigger round numbers. The *relative error*, namely the value $(x - y) / x$, where x is a round number and y is a precise number, is much greater for *10* and *9* than for *100* and *99*, and for *1000* and *990*. Only the latter two are equal.

(13) a. (10 – 9) / 10 = 1/10.
b. (100 – 99) / 100 = 1/100.
c. (1000 – 990) / 1000 = 1/100.

In accordance, for standard deviation s = 20, the precise interpretation of, e.g., *9* or *8.9* is simply not within the vague interpretation of *ten*, [*9.5, 10.5*). By contrast, the answers for *100* and *1000* are all very positive because the differences between the precise and round numbers are small enough. For 1/s = 1/20, the precise interpretation of all of the precise numbers used in the case of *100* (e.g., *98.77*) is within the vague interpretation of *100*, [*95, 105*). The same holds for *1000*.

As Figures 1 and 3a-b demonstrate, all the conditions with *10* received much lower ratings than all those with *100* or *1000*. A two-way factorial ANOVA for 3 randomized blocks of matched items (types of round number: *10* vs. *100* vs. *1000*) and a repeated measure (granularity × distance levels) yields a significant difference between the three types of round numbers ($df = 2$, $F = 27.9$, $p < .0001$) and between the five granularity × distance levels ($df = 4$, $F = 3.2$, $p < .05$), with a significant interaction ($df = 8$, $F = 2.2$, $p < .05$).

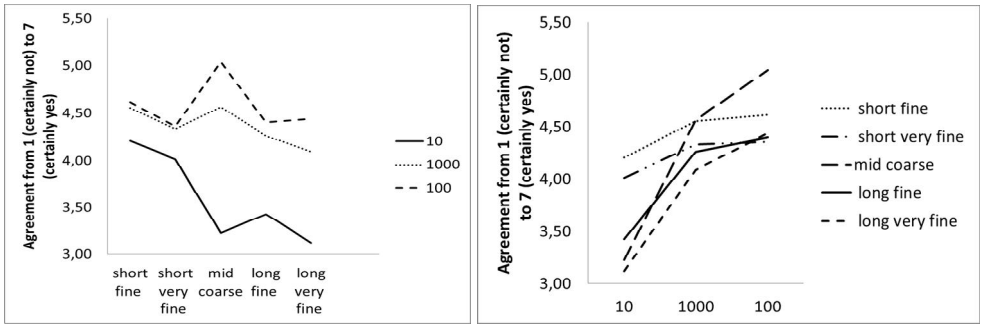

Figure 3a-b: Averaged agreement ratings for the 5 distance × granularity conditions in the 3 round number types.

Finally, considering conditions *100* and *1000*, **prediction 3a** was **not** confirmed in that the difference between the long- and short-distance conditions (fine and very fine granularity) is not significant. Figure 3b clearly illustrates the fact that agreement rates in the short & very fine condition were almost the same as in the long & fine condition. However, a Mann-Whitney test yields a significant difference between the long distance (n = 32, MR = 21.4) and mid distance (n = 16, MR = 30.8) conditions (U = 356, z = -2.18, p < .03), and a significant difference at the confidence level of 90% between the short distance (n = 32, MR = 22.1) and mid distance (n = 16, MR = 29.3) conditions (U = 332, z = -1.65, p < .1, p_1 < .05)[5]. Additional research is necessary in order to establish whether these are systematic patterns and in order to uncover their origins.

Moving on from distance to granularity (coarse, fine, and very fine), **prediction 3b** is partially confirmed. A Mann-Whitney test yields a significant difference in fine-coarse contexts between the very fine (n = 16, MR = 9.4) and coarse (n = 16, MR = 18.8) conditions (U = 114, z = -3.03, p < .003), and between the fine (MR=10.8) and coarse (MR=16) conditions (U =92, z = -1.68, p < .1) and no difference between the very fine and fine conditions (p > .3). It yields a similar difference in coarse-fine contexts, between the very fine (n = 16, MR = 10.2) and coarse (n = 16, MR = 17.1) conditions (U = 101, z = -2.24, p < .03), and between the fine (MR=10.9) and coarse (MR=15.6) conditions (U =89, z = -1.5, p < .134, p_1 < .07), with no difference between the very fine and fine conditions (p > .5). Thus, in both fine-coarse and coarse-fine contexts, a precise number of a relatively coarse granularity triggers higher agreement ratings than one of either fine or very fine granularity, which, in turn, have the same effect. Again, additional research is needed to explain these effects.

2.4 Discussion

The results of the current study align with the expectations of a granularity theory for numerals (Krifka 2007 combined with Lewis 1979). The reason is that the precise interpretations of two subsequent numerals are non-overlapping. For example, consider the numerals *99* and *100*. When one speaker thinks *99* and another speaker thinks *100*, the former probably does not agree with the latter even if the latter is his mother and an authority, because having thought of such a precise number, one must be knowing precisely what the facts are and refuse to either agree that they are different, or to shift to coarser granularity, in support of David Lewis. In other words, the averaged response for "If 99, 100" is 4 on a seven-point scale, because *99* is interpreted precisely and relative to a fine scale, and so is *100* following the mention of *99*. Assuming an upper-open interpretation, evidence for 'at least 99' is consistent with 'at least 100', but assuming an upper-bounded interpretation, which is dominant for numbers, evidence for 'exactly 99' is evidence against 'exactly 100'.

However, if one thinks *100*, one would more probably agree that *99*, assuming his mother either knows more or is more pedant. In other words, the expression *100* is interpreted

[5] Focusing on fine-coarse contexts, a Mann-Whitney test yields a significant difference (at a 90% confidence level) between the long distance (n = 16, MR = 13.5) and short distance (n = 16, MR = 19.5) conditions (U = 80, z = 1.77, p < .08), in contrast to an absence of difference in coarse-fine contexts. Put differently, distance seems to reflect degree of disagreement, but not degree of agreement. The mid distance, though, is characterized by the highest agreement rates in coarse-fine contexts compared to the long (U = 100, z = -2.17, p < .05, MR_{mid} = 17 vs. MR_{lone} = 10.3) and short distance (U = 90, z = -1.56, p < .15, MR_{mid} = 15.8 vs. MR_{short} = 10.9), in fine-coarse contexts compared to the long (U = 111, z = -2.85 p < .005, MR_{mid} = 18.4 vs. MR_{long} = 9.6) and short distance (U = 95, z = -1.87, p < .07, MR_{mid} = 16.4 vs. MR_{short} = 10.6).

approximately and the use of *99* creates a shift to a finer scale and is interpreted more precisely. Evidence for *around 100* is consistent with *99* as *99* is likely to be part of the approximate interpretation of *100*, so unless a speaker has a reason to think an addressee is mistaken, she accepts that *99*. The averaged response is 5 on a seven-point scale, and not higher (i.e., speakers are not completely certain that Nick would agree with his mother), because it is not strictly entailed that '99'. The validity of this account can be tested in the future by using a mere inference task ("Does it follow…").

Notice that we have not used in our experiment cases in which the precise number exceeded the round number (as in "If 100, 101?", and "If 101, 100?"). For such cases, we strongly believe, are always understood as corrections. At any rate, future work can test this assumption, too.

In sum, as expected, the simple numeral *10* is expected to have a default coarse and approximate interpretation 'about ten' (e.g., $[ten]_g = [9.5, 10.5)$), and *9.5* is expected to have a precise interpretation ($[9.5]_{gp} = [9.5]$). Since the latter is part of the former ($9.5 \in [9.5, 10.5)$), in dialogue (14a) B's utterance is understood as confirming A's utterance. By contrast, following an utterance of a complex numeral, e.g., *9.5* in (14b), the interpretation of a simple numeral would be finer and more precise than the default ($[ten]_{gp} = [10]$). Since $10 \neq 9.5$, and since numerals tend toward upper-bounded readings, B's utterance is understood as contradicting A's. Therefore, agreement rates in coarse-to-fine contexts are higher than in fine-to-coarse ones.

(14) a. A: Stock Exchange fell by 10 percent.
B: Yes (#No), it fell by 9.5 percent.
b. A: Stock Exchange fell by 9.5 percent.
B: No (#Yes), it fell by 10 percent.

2.5 Numerals vs. Adjectives

Sassoon and Zevakhina (2012) hypothesized that the granularity shifting effect in bare vs. modified adjectives is opposite to the one hypothesized for numerals in experiment 1 of this paper. The reason is that, from a semantic point of view, modified adjectives, resembling precise interpretations of numerals, are still quite different from them. The crux is that the precise interpretations of two subsequent numerals are always non-overlapping, unlike the precise interpretations of adjectives and their modified forms. Only minimized and non-minimized total adjectives (for instance, *slightly closed* and *closed,* respectively) resemble numerals in this respect. Let us explain this with some detail.

We assume, following Kennedy and McNally (2005), that adjectives divide to *partial* and *total* ones (e.g., *dirty* and *clean*, respectively). Partial adjectives have a minimum standard; e.g., one stain suffices for a shirt to count as *dirty*. Total adjectives have a maximum standard; e.g., to count as *clean* a shirt has to be completely free of dirt (maximally clean).

Moreover, intuitively, adjectival modification as in *slightly dirty* and *completely clean* triggers a granularity shift, rendering relevant, e.g., hardly visible dirt specks that are by default ignorable in judging cleanliness. Thus, we further assume, following Sassoon (2012), that the use of modifiers such as *slightly* and *completely* triggers a shift to finer scales, namely, to scales representing more distinctions (Lewis 1979; van Rooij 2009). Formally, for any two degree functions $g,g_p \in D_{xd}$, g_p is finer-grained than g, $g \subset g_p$ iff $\exists x,y \in D_x$, $(g(x) = g(y)) \ \& \ \neg(g_p(x) = g_p(y))$, but not v.v. $\neg\exists x,y \in D_x$, $(g_p(x) = g_p(y)) \ \& \ \neg(g(x) = g(y))$. For example, two similar glasses

filled with an amount of wine differing in but few drops are indistinguishable relative to g (they fall under $[\text{equally full}]_g$), but distinguishable relative to g_p (they fall under $[\text{fuller}]_{gp}$).

A total adjective like *full* denotes maximally full entities *presupposing coarse granularity level* g, as stated in (15a). A few missing drops in a full glass are ignorable – such glasses are considered to be as full as they can be. By contrast, the maximized total adjective form *completely full* denotes maximally full entities, *presupposing finer granularity* g_p, as stated in (15b). A few missing drops in a glass render it less full than it can be, and thus *not full*. Since g_p is finer than g, it follows that g_p assigns fewer entities the same degree, e.g. fewer entities are mapped to a degree identical to the maximum ($=_g \supset =_{gp}$). Thus, (15b) is stronger than (15a). It becomes harder for an object to count as full as it can be.

(15) a. $[G_{total}]_g = \lambda x.\ g(x) = \max(g)$
b. $[\text{completely G}]_g = \lambda x.\ g_p(x) = \max(g_p)$, for g_p finer than g
c. $[G_{partial}]_g = \lambda x.\ g(x) > \min(g)$
d. $[\text{slightly G}]_g = \lambda x.\ g_p(x) > \text{standard}(g_p)$, for g_p finer than g

Similarly, a partial adjective like *dirty* denotes minimally dirty entities, presupposing coarse granularity g, as stated in (15c). 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 dirty entities, *presupposing finer granularity* g_p, as stated in (15d). A few dirt specks turn an object dirtier than dirt free entities, and thus *dirty*. Since g_p is finer than g, more distinctions are made, i.e., g_p assigns more entities different degrees ($>_g \subset >_{gp}$). Thus, (15d) is weaker than (15c). It is easier to exceed the minimum threshold.

This account captures the use of *slightly* with total adjectives, where *slightly* tends to trigger 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*. In other words, combinations of minimizers with doubly closed total adjectives such as *slightly full* refer to the minimum in the denotation, not scale, namely, to relatively high degrees, rather than very low degrees. This consequence is intuitively correct, e.g., an utterance of a sentence like *The city square is slightly full* implies that the city square is rather full; it is more full than empty. Moreover, if there are but few people in the square we cannot possibly describe the situation using this sentence. These facts are unexpected given a scale-minimum analysis of *slightly*.

Returning to Lewis's (1979) granularity shifting principle, the prediction is that utterances with a modified adjective such as *completely full* or *slightly dirty* involve an irreversible shift to fine granularity and, therefore, a subsequent utterance of a bare adjective, e.g., *full* and *dirty,* respectively, is interpreted on a fine scale. However, the non-default fine and precise interpretation of *dirty* following *slightly dirty*, as in (17), is identical to that of *slightly dirty*. It includes instances with slight amount of dirt, so B's utterance is understood as confirming A's. By contrast, the coarse interpretation of *dirty* in (16) excludes slightly dirty entities. Slight dirt is ignorable. So B and A's utterances are conceived as a disagreement. Thus, we predict lower agreement rates in coarse-to-fine contexts like (16) than in fine-to-coarse ones like (17), opposite to the prediction for numerals (fine-coarse >> coarse-fine).

(16) A: The table is dirty.
B: No (?Yes), it's slightly dirty.
(17) A: The table is slightly dirty.
B: Yes (?No), it's dirty.

With these analyses at hand, Sassoon and Zevakhina (2012) predicted that agreement ratings for "If x is M A, x is A" contexts should be significantly higher than ratings for "If x is A, x is M A" contexts, where M is a modifier and A is an adjective. The results of their study confirmed this prediction, revealing significant effects of inference ("If A, M A" << "If M A, A", $p < .0001$), modifier type (*slightly* << *completely,* $p < .0001$), adjective type (total << partial, $p < .0003$), and an interaction: the inference effect is stronger with *completely* than *slightly*, and is, predictably, not significant in *slightly* + total adjectives. The reason is that the precise interpretation of, e.g., *full* is different from that of *slightly full,* as is the case with round numbers following precise ones (e.g., "If 9.5, 10"), and unlike the case with partial adjectives and their minimized forms illustrated in (17) above with *slightly dirty* and *dirty*. Indeed, agreement rates were higher for "If minimized partial A, then A" than for "If minimized total A, then A".

In sum, the main results suggest that speakers tend to agree that, e.g., "If x is slightly bumpy, x is bumpy" to a larger extent than they tend to agree that "If x is bumpy, x is slightly bumpy", and they tend to agree that, e.g., "If x is completely flat, x is flat" to a larger extent than they tend to agree that "If x is flat, x is completely flat".[6] Moreover, as predicted, the agreement rates were very high (significantly higher than those associated with a set of clearly false fillers), except for combinations of minimizers and total adjectives.[7]

One surprising result was that the modifier effect (the higher agreement rates for *completely* than *slightly*: "If maximized A, then A" >> "If minimized A, then A") extended to partial adjectives. To illustrate, the agreement ratings for inferences from, e.g., *dirty* to *completely dirty* were higher than for inferences from *dirty* to *slightly dirty*. Sassoon and Zevakhina (2012) proposed that this result is due to a *level of fit effect*. Recall that according to Krifka's (2007) model, the interpretation of a numeral *n* is represented by an interval [*n – n/s, n + n/s*], and a function representing goodness of fit of each point in this interval. Sassoon and Zevakhina (2012) proposed that also the interval denoted by a minimum-standard adjective such as *dirty* is associated with a function representing goodness of fit of each point. Thus, for example, entities covered with some but not much dirt are dirty, but their level of fit with respect to *dirty* is not as high as the level of fit of entities covered with much dirt. Put informally, when speakers think of dirty entities, they think of very dirty ones, not of slightly dirty ones. This effect of level of fit was thought to be manifested by the higher agreement rate for "If x is dirty, x is completely dirty" than for "If x is dirty, x is slightly dirty".

Given this surprising result, the second goal of this paper is to test semantic relations between statements with modified and non-modified adjectives, using an entailment task, rather than an agreement task, with the goal of minimizing level of fit effects. Thus, experiment 2 aimed to test

[6] Although the second effect is stronger, the first one is more important, for it is not captured by mere entailment.

[7] Some differences between numerals and modifiers may stem from the fact that with numerals truth conditions are based on fact, e.g., number of objects/measures in the world, and, relatedly, often interpretations are upper-bounded. By contrast, for adjectives and their modifiers there are no objective facts on which to base truth value evaluations. The standard of membership in a denotation is so context and speaker dependent that with adjectives the shift to a fine granularity is not interpreted as "knowing better the facts". Mainly considerations of how pedantic one is in his criteria of application play a role. The precise implications of this difference are yet to be explored.

whether the Lewis granularity shifting effect survives when an entailment task is used, whereas the level of fit effect disappears.

3 Experiment 2: Inference Relations between Statements with Modified vs. Unmodified Adjectives

3.1 Methods

Participants were recruited using AMT. The reward per hit was \$0.3 cents. All in all, 120 participants (American workers) answered on average 21 questions per participant (*SD* = 21), with average hourly rate of \$5.7. 68% of the participants had some knowledge of logic and probability, 7% had deep knowledge, and 24% had no knowledge.

The *lexical items* consisted of 8 partial adjectives (*open, transparent, visible, wrong, incorrect, unclear, dirty, sick*) and 8 total adjectives (*full, closed, empty, invisible, correct, opaque, clean, healthy*). Rather than an agreement task, the task this time involved an entailment judgment. Thus the general structure of the texts used in this experiment was "If S_1, does it follow that S_2?". Each adjective occurred in 8 versions of a text involving either the modifier *slightly* or the modifier *completely*, and the inference patterns "If x is A, x is M A" (cf. (18a) and (19a)), "If x is M A, x is A" (cf. (18b) and (19b)), as well as "If x is M A, x is not A" (cf. (20a) and (21a)), and "If x is not M A, x is not A" (cf. (20b) and (21b)).

(18) a. *If A, slightly A*:
If x is dirty, does it follow that x is slightly dirty?
b. *If slightly A, A*:
If x is slightly dirty, does it follow that x is dirty?

(19) a. *If A, completely A*:
If x is dirty, does it follow that x is completely dirty?
b. *If completely A, A*:
If x is completely dirty, does it follow that x is dirty?

(20) a. *If slightly A, not A*:
If x is slightly dirty, does it follow that x is not dirty?
b. *If not slightly A, not A*:
If x is not slightly dirty, does it follow that x is not dirty?

(21) a. *If completely A, not A*:
If x is completely dirty, does it follow that x is not dirty?
b. *If not completely A, not A*:
If x is not completely dirty, does it follow that x is not dirty?

This design resulted in 128 target texts, including 2 randomized blocks (adjective types) of 8 items each, occurring in 8 versions (repeated measures across the blocks) differing by the modifier and inference type.

The 128 fillers consisted of 8 versions of 16 texts (per 16 adjective pairs) presenting entailment tasks with conjunctions and disjunctions. The entailments had the general forms "If x is A and B, x is B" (cf. (22a)), "If x is B, x is A and B" (cf. (22b)); "If x is A or B, x is B", and "If x is B, x is A or B". Also, similar entailments were included with conjunctive and disjunctive comparatives: "If x is more B than y, x is more A and B than y" (cf. (22c)), "If x is more A and B than y, x is more B than y" (cf. (22d)); "If x is more A or B than y, x is more B than y", and "If x is more B than y, x is more A or B than y".

(22) a. If chicken is nutritious and digestible, does it follow that it's digestible?
b. If chicken is digestible, does it follow that it's nutritious and digestible?
c. If chicken is more digestible than beef, does it follow that it's more nutritious and digestible than the beef?
d. If chicken is more nutritious and digestible than beef, does it follow that it's more digestible than the beef?

In all cases participants were asked to provide an answer on a scale ranging from 1 (certainly not) to 5 (certainly yes). The instructions, which were very similar to those of experiment 1 (see section 2.1) clarified that the survey is intended for English speakers. The targets and fillers were counterbalanced into 16 lists of 8 targets and 8 fillers each, presented in a randomized order.

3.2 Results

Table 2 presents averages over 25 participants for 2 modifiers with 2 adjective types (8 adjectives per type). All the averages for "If A, M A" and "If M A, A" are higher than the mid of the five-point scale, i.e., "2.5". This is comparable to the averages in the agreement task reported in Sassoon and Zevakhina (2012: 241). In accordance, the averages for inferences containing negation (in particular, the pattern "If M A, not A") were often below the mid.

Modifier + adjective type	**If A, MA**	**If MA, A**	**If MA, NOT A**	**If NOT MA, NOT A**	Total
COMPLETELY + PARTIAL	2.93 (.42)	4.91 (.09)	1.36 (.49)	2.74 (.48)	**2.99 (1.34)**
COMPLETELY + TOTAL	3.76 (.70)	4.95 (.05)	1.15 (.16)	4.06 (.55)	**3.48 (1.50)**
SLIGHTLY + PARTIAL	3.29 (.35)	4.04 (.48)	1.64 (.29)	2.99 (.45)	**2.99 (.96)**
SLIGHTLY + TOTAL	2.95 (.44)	2.63 (.55)	2.85 (.93)	3.43 (.34)	**2.96 (.65)**
Total	**3.23 (.58)**	**4.13 (1.02)**	**1.75 (.85)**	**3.30 (.67)**	3.11 (1.17)

Table 2: Averaged agreement ratings (and standard deviations) for experiment 2.

First, focusing on the inference types "If M A, A" and "If A, M A", a Wilcoxon signed-ranks test applied to two samples of 32 matched pairs yields that agreement rates were higher in the condition "If M A, A" than in the condition "If A, M A" (W = -413, $n_{s/r}$ = 32, z = -3.86, $p < .0001$), as expected given the granularity shifting constraint of Lewis (1979) and a semantics for partial and total adjectives as denoting entities above the minimum and at the maximum of the scale they employ in the context, respectively. The inference effect is more pronounced within the sample of judgments with *completely* than with *slightly* modifying partial adjectives, and is only reversed in the sample of judgments with *slightly* modifying total adjectives. This is

also illustrated in Figure 4. A two-way factorial ANOVA for 2 randomized blocks (*slightly* vs. *completely*) and a repeated measure (inference type: "If A, M A" vs. "If M A, A") yields a significant modifier effect (*completely* >> *slightly*; $df = 1$, $F = 27.73$, $p < .0001$), and inference effect ($df = 1$, $F = 51.44$, $p < .0001$), as well as an interaction ($df = 1$, $F = 30.24$, $p < .0001$) due to the greater acceptance of inferences containing *completely* than inferences containing *slightly*. It is easy to see in Figure 4 that the significance of the modifier effect (*completely* >> *slightly*) is mainly due to judgments for "If M A, A". Still, a Wilcoxon test yields that the modifier effect is significant in "If M A, A" with partials ($W = 36$, $n_{s/r} = 8$, $p < .01$) and totals ($W = 36$, $n_{s/r} = 8$, $p < .01$), and in "If A, M A" with totals for one-tailed p ($W = 26$, $n_{s/r} = 8$, $p_1 < .05$),[8] but not with partials. Thus, as predicted an entailment task preserves the Lewis effect but eliminates what we called *level of fit* effects. The apparently reversed modifier effect with the latter (*slightly* >> *completely*) is not significant ($W = -19$, $n_{s/r} = 8$, $p > .05$).

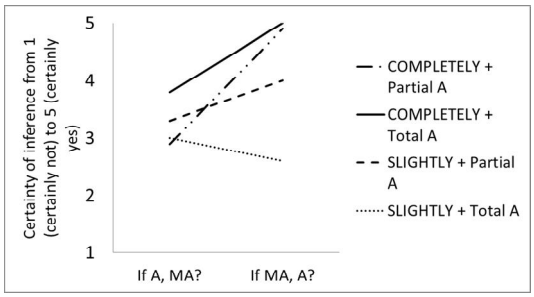

Figure 4: Averaged entailment ratings for the 2 inference types ("If A, M A" and "If M A, A") in the 4 modifier + adjective conditions.

Moreover, a two-way factorial ANOVA for 4 blocks (*slightly* vs. *completely* with partial vs. total adjectives) and a repeated measure (inference type: "If M A, A" vs. "If M A, NOT A") yields a significant modifier + adjective effect ($df = 3$, $F = 3.67$, $p < .024$) and inference effect ($df = 1$, $F = 312.45$, $p < .0001$), as well as a significant interaction ($df = 3$, $F = 46.69$, $p < .0001$). As predicted, the inference effect is absent in judgments with *slightly* modifying total adjectives. This is also illustrated in Figure 5. The figure illustrates very clearly that speakers tend to infer from "M A" that "A", and they clearly do not tend to infer that "not A". On the same line, a Wilcoxon signed-ranks test for two-matched 8-item samples yields that the ratings for "If M A, NOT A" are significantly lower than for "If NOT M A, NOT A" in conditions with *completely* modifying partials ($W = -36$, $n_{s/r} = 8$, $p < .01$) and totals ($W = -36$, $n_{s/r} = 8$, $p < .01$), and *slightly* modifying partials ($W = -36$, $n_{s/r} = 8$, $p < .01$), but not totals ($W = 20$, $n_{s/r} = 8$, $p > .05$). Figure 6 shows these effects, as well.

[8] Interestingly, the outliers are the two only-upper-closed adjectives. Without them the effect is two-tailed significant ($W = 21$, $n_{s/r} = 6$, $p < .05$).

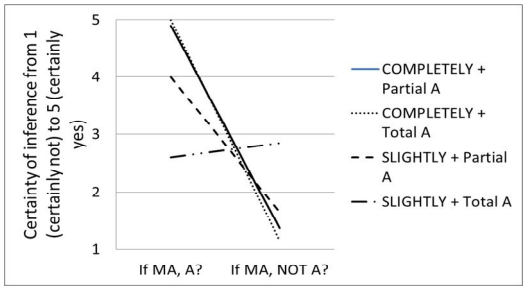

Figure 5: Averaged entailment ratings for the 2 inference types ("If M A, A" and "If M A, not A") in the 4 modifier + adjective conditions.

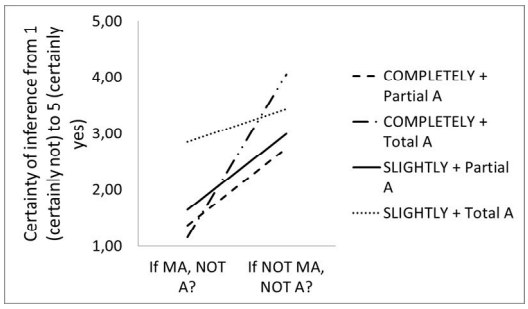

Figure 6: Averaged entailment ratings for the 2 inference types ("If M A, not A" and "If not M A, not A") in the 4 modifier + adjective conditions.

3.3 Discussion

The results support the hypotheses based on the granularity shifting account. According to this account, adjectival modification triggers a granularity shift, rendering, e.g., hardly visible dirt specks that are by default ignorable relevant in judging cleanliness. The non-default fine and precise interpretation of *dirty* following *slightly dirty* as in (23b) is identical to that of *slightly dirty* (including instances with slight amount of dirt). By contrast, the coarse interpretation of *dirty* in (23a) excludes slightly dirty entities (slight dirt is ignorable). This results in an inference effect ("If M A, A" << "If A, M A") opposite to the one found in numerals.

(23) a. If the table is dirty, it's slightly dirty.
b. If the table is slightly dirty, it is dirty.

However, this effect was predictably reversed in total adjectives modified by *slightly*. The reason is that the precise interpretation of, e.g., *full* is actually different from that of *slightly full.* The fine and precise interpretation of *full,* $[\text{full}]_{g_p}$, consists of maximally full entities, where any tiny amount of content missing (e.g. 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, cf. Figure 5). Thus, inferences from minimized total adjectives to non-minimized ones exhibit the inference effect typical of numerals.

Moreover, speakers do not tend to agree that, e.g., *if something is dirty, it is completely dirty* to a larger extent than they tend to agree that *if something is dirty, it is slightly dirty*, thus the modifier effect is lost in partial adjectives as we move from an agreement task to an entailment task. This result supports the view that it is the location of level-of-fit peak high in the scale of partial adjectives that affects inference in an agreement task. Naturally, it has a reduced effect in an entailment. If an entity x falls under a partial A, speakers tend to agree that x falls under *completely A*, but neither does it certainly follow that it is *completely A*, nor does it follow that it is *not completely A*.

Finally, the results presented in Figure 5 support the view that the ratings for "If {M A, A}" are generally in the positive range (with the exception of *slightly* + total adjectives), as the average certainty for "If M A, A" is 4–5 on a five-point scale, while that for "If M A, not A" is 1–2. By contrast, the ratings for *slightly* + total adjectives are at the uncertainty zone 2.85, and are relatively balanced in inferences to A and to "Not A".

4 Conclusion

To conclude, these results support the hypotheses based on Lewis's (1979) constraints on granularity shifting, Krifka's (2007) representation of granularity in numerals, and its extension to adjectives. Thus, they support the view that general principles govern the setting of a granularity level and its shifting within discourse in different domains of grammar, including simple and complex numerals and unmodified and modified adjectives. A crucial fact underlying the different inference effect in adjectives and numerals is that on fine interpretations different numerals denote different points, while modified and unmodified adjectives often denote highly similar intervals. Relatedly, upper-bounded readings are dominant in numerals but appear to be minor in adjectives (see Doran et al 2009). Future work should address the connections between level of fit and scalar implicatures and their respective role in explaining inference data.

Appendix 1

Granularity	**Distance**	**10**	**100**	**1000**	**10 - n**	**100 - n**	**1000 - n**
Coarse	Mid	9	99	990	1	1	10
Fine	Short	9.3	99.2	992	0.7	0.8	8
	Long	8.7	98.8	988	1.3	1.2	12
Very fine	Short	9.33	99.23	992.3	0.67	0.77	7.7
	Long	8.67	98.77	988.7	1.33	1.23	11.3

Appendix 2

The item set of the coarse-fine inference conditions of experiment 1:

1. Nick thinks John's library contains 1000 books. Nick's mother thinks it contains 990 books. Would Nick agree that it contains 990 books?
2. Nick thinks Sue's stamp collection contains 1000 stamps. Nick's mother thinks it contains 992 stamps. Would Nick agree that it contains 992 stamps?
3. Nick thinks the auditorium contains 1000 seats. Nick's mother thinks it contains 988 seats. Would Nick agree that it contains 988 seats?
4. Nick thinks the document contains 1000 lines. Nick's mother thinks it contains 992.3 lines. Would Nick agree that it contains 992.3 lines?
5. Nick thinks he got 1000 points on this test. Nick's mother thinks he got 987.7 points. Would Nick agree that he got 988.7 points?
6. Nick thinks Sven has 1000 CDs. Nick's mother thinks Sven has 990 CDs. Would Nick agree that Sven has 990 CDs?
7. Nick thinks Ron's old car is worth 1000 dollars. Nick's mother thinks it is worth 992 dollars. Would Nick agree that it is worth 992 dollars?
8. Nick thinks he sold 1000 balls. Nick's mother thinks he sold 988 balls. Would Nick agree that he sold 988 balls?
9. Nick thinks the mountain is 1000 meters tall. Nick's mother thinks it is 992.3 meters tall. Would Nick agree that it is 992.3 meters tall?
10. Nick thinks he drove 1000 kilometers last week. Nick's mother thinks he drove 987.7 kilometers. Would Nick agree that he drove 988.7 kilometers?
11. Nick thinks the theatre office sold 1000 tickets for the performance. Nick's mother thinks it sold 990 tickets. Would Nick agree that it sold 990 tickets for the performance?
12. Nick thinks 1000 people live in the town. Nick's mother thinks 992 people live in the town. Would Nick agree that 992 people live in the town?
13. Nick thinks there are 1000 cars in the parking lot. Nick's mother thinks there are 998 cars. Would Nick agree that there are 998 cars?
14. Nick thinks the trip costed 1000 euros. Nick's mother thinks it costed 992.3 euros. Would Nick agree that it costed 992.3 euros?
15. Nick thinks he read 1000 pages of the book. Nick's mother thinks he read 988.7 pages. Would Nick agree that he read 988.7 pages?
16. Nick thinks the airport can land 1000 planes a day. Nick's mother thinks it can land 990 planes a day. Would Nick agree that it can land 990 planes a day?
17. Nick thinks that the athlete participated in 1000 competitions. Nick's mother thinks he participated in 992 competitions. Would Nick agree that he participated in 992 competitions?
18. Nick thinks that the distance is 1000 meters long. Nick's mother thinks it's 988 meters long. Would Nick agree that it's 988 meters long?
19. Nick thinks Kate ran 1000 meters. Nick's mother thinks Kate ran 992.3 meters. Would Nick agree that Kate ran 992.3 meters?
20. Nick thinks that Susan flew 1000 kilometers by plane. Nick's mother thinks Susan flew 988.7 kilometers by plane. Would Nick agree that Susan flew 988.7 kilometers by plane?
21. Nick thinks there are 100 houses in the street. Nick's mother thinks there are 99 houses in the street. Would Nick agree that there are 99 houses?
22. Nick thinks the table's width is 100 cm. Nick's mother thinks it is 99.2 cm. Would Nick agree that it is 99.2 cm?
23. Nick thinks Sara jumped a distance of 100 cm. Nick's mother thinks it was 98.8 cm. Would Nick agree that it was 98.8 cm?
24. Nick thinks George's height is 100 cm. Nick's mother thinks it is 99.23 cm. Would Nick agree that it is 99.23 cm?
25. Nick thinks Bill's weight is 100 kilos. Nick's mother thinks it is 98.77 kilos. Would Nick agree that it is 98.77 kilos?

26. Nick thinks there are 100 balloons in the sky. Nick's mother thinks there are 99 balloons in the sky. Would Nick agree that there are 99 balloons?
27. Nick thinks the child's length is 100 cm. Nick's mother thinks it is 99.2 cm. Would Nick agree that it is 99.2 cm?
28. Nick thinks Sara gained back 100 percents of the costs. Nick's mother thinks she gained 98.8 percents of the costs. Would Nick agree that she gained 98.8 percents of the costs?
29. Nick thinks the distance to the tower is 100 meters. Nick's mother thinks it is 99.23 meters. Would Nick agree that it is 99.23 meters?
30. Nick thinks the bag weighs 100 grams. Nick's mother thinks it weighs 98.77 grams. Would Nick agree that it weighs 98.77 grams?
31. Nick thinks that the garden has an area of 100 square meters. Nick's mother thinks that it has an area of 99 square meters. Would Nick agree that it has an area of 99 square meters?
32. Nick thinks that the distance between the house and the shop is 100 meters. Nick's mother thinks it's 99.2 meters. Would Nick agree that it's 99.2 meters?
33. Nick thinks that the cake contains 100 kilocalories. Mick's mother thinks it contains 98.8 kilocalories. Would Nick agree that it contains 98.8 kilocalories?
34. Nick thinks that the speedometer showed 100 kilometers. Nick's mother thinks it showed 99.23 kilometers. Would Nick agree that it showed 99.23 kilometers?
35. Nick thinks that the curtains are 100 centimeters long. Nick's mother thinks they are 98.77 centimeters long. Would Nick agree that they are 98.77 centimeters long?
36. Nick thinks that there were 100 people with green hair at the party. Nick's mother thinks there were 99 people with green hair at the party. Would Nick agree that there were 99 people with green hair at the party?
37. Nick thinks Bill earned 100 dollars. Nick's mother thinks he earned 99.2 dollars. Would Nick agree that Bill earned 99.2 dollars?
38. Nick thinks that the speed of the train was 100 kilometers per hour. Nock's mother thinks it was 98.8 kilometers per hour. Would Nick agree that it was 98.8 kilometers per hour?
39. Nick thinks John breathed 100 seconds in the water. Nick's mother thinks John breathed 99.23 seconds in the water. Would Nick agree that John breathed 99.23 seconds in the water?
40. Nick thinks that Kate has done sports for 100 hours. Nick's mother thinks Kate has done sports for 98.77 hours. Would Nick agree that she's done sports for 98.77 hours?
41. Nick thinks he has 10 shirts. Nick's mother thinks he has 9 shirts. Would Nick agree that he has 9 shirts?
42. Nick thinks he has found 10 euros. Nick's mother thinks he has found 9.3 euros. Would Nick agree that he has found 9.3 euros?
43. Nick thinks Mary's toy costs 10 euros. Nick's mother thinks it costs 8.7 euros. Would Nick agree that it costs 8.7 euros?
44. Nick thinks the temperature is 10 degrees Celsius. Nick's mother thinks it is 9.33 degrees Celsius. Would Nick agree that it is 9.33 degrees?
45. Nick thinks the theatre show costs 10 euros. Nick's mother thinks it costs 8.67 euros. Would Nick agree that it costs 8.67 euros?
46. Nick thinks he has 10 beers in his fridge. Nick's mother thinks he has 9 beers in his fridge. Would Nick agree that he has 9 beers in his fridge?
47. Nick thinks he drank 10 liters during the long race. Nick's mother thinks he drank 9.3 liters. Would Nick agree that he drank 9.3 liters?
48. Nick thinks Julie arrived 10 minutes late. Nick's mother thinks Julie arrived 8.7 minutes late. Would Nick agree that Julie arrived 8.7 minutes late?
49. Nick thinks the size of the box is 10 inches. Nick's mother thinks it is 9.33 inches. Would Nick agree that it is 9.33 inches?
50. Nick thinks he got a discount of 10 euros. Nick's mother thinks the discount was of 8.67 euros. Would Nick agree that the discount was of 8.67 euros?
51. Nick thinks that the house has 10 rooms. Nick's mother thinks it has 9 rooms. Would Nick agree that the house has 9 rooms?
52. Nick thinks that Bill immersed to a depth of 10 meters. Nick's mother thinks he immersed to a depth of 9.3 meters. Would Nick agree Bill immersed to a depth of 9.3 meters?
53. Nick thinks that John has slept for 10 hours. Nick's mother thinks John has slept for 8.7 hours. Would Nick agree that John has slept for 8.7 hours?

54. Nick thinks that the temperature outside is 10 degrees Celsius. Nick's mother thinks it's 9.33 degrees Celsius. Would Nick agree that it's 9.33 degrees Celsius?
55. Nick thinks that the width of the pool is 10 meters. Nick's mother thinks it's 8.67 meters. Would Nick agree that it's 8.67 meters?
56. Nick thinks he ate 10 apples. Nick's mother thinks he ate 9 apples. Would Nick agree that he ate 9 apples?
57. Nick thinks the height of the building is 10 meters. Nick's mother thinks it's 9.3 meters. Would Nick agree that it's 9.3 meters?
58. Nick thinks the distance between the house and the gate is 10 meters. Nick's mother thinks it's 8.7 meters. Would Nick agree that it's 8.7 meters?
59. Nick thinks Bill spent 10 dollars on the dinner. Nick's mother thinks he spent 9.33 dollars on the dinner. Would Nick agree Bill spent 9.33 dollars on the dinner?
60. Nick thinks that the jewelry weighs 10 grams. Nick's mother thinks it weighs 8.67 grams. Would Nick agree it weighs 8.67 grams?

References

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

Blutner, Reinhard. 1998. Lexical Pragmatics. *Journal of Semantics* 15, 115-162.

Blutner, Reinhard. 2000. Some Aspects of Optimality in Natural Language Interpretation. *Journal of Semantics* 17(3), 189-216.

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.

Michael Buhrmester, Tracy Kwang and Samuel D. Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6(3), 3-5.

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

Dekker, Paul and Robert van Rooy. 2000. Bi-Directional Optimality Theory: An application of Game Theory. *Journal of Semantics* (17), 217-242.

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

Jäger, Gerhard. 2002. Some notes on the formal properties of bidirectional Optimality Theory, *Journal of Logic, Language and Information* 11, 427-451.

Kennedy, Christopher and 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.

Parikh, Prashant. 1991. Communication and strategic inference. *Linguistic and Philosophy* 14(5), 473 – 514.

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

Sassoon, Galit W. 2012. A slightly modified economy principle: Stable properties have non stable standards. In E. Cohen and A. Mizrachi (eds.), *Proceedings of the Israel Association of Theoretical Linguistics 27*. MIT working Papers in Linguistics.

Sassoon, Galit W. and Natalia Zevakhina. 2012. Granularity shifting: Experimental evidence from degree modifiers. In Proceedings of SALT 22: 226-246.

van Rooij, Robert, 2009. Vagueness and semi-orders. In *Predication and truth: Proceedings of the Vth Navarra Workshop on Vagueness*. University of Navarra, Pamplona, Spain.