# SCALAR ALTERNATIVES AND SCALAR INFERENCE INVOLVING ADJECTIVES: A COMMENT ON VAN TIEL, ET AL. (2016)*

Louise McNally
Universitat Pompeu Fabra
louise.mcnally@upf.edu

Doran, et al. 2009, 2012 and Van Tiel, et al. 2016 have recently presented experimental evidence that gradable adjectives (e.g. *cool*) yield scalar inferences (e.g. to not *cold*) significantly less often than do numerals, quantifiers and modals do and that, moreover, there is also considerable variability *within* the class of adjectives in the frequency with which inferences are likely to be generated. Van Tiel, et al. consider two hypotheses to explain this variation: differences among pairs of adjectives in what they call *scale distinctness* (how easy or difficult it is to differentiate the adjectives), and differences in *scale availability* (how salient the stronger scalar alternative is); they argue that only scale distinctness plays a role, and a rather small one at that; they leave most of the variation unexplained. In this comment, I argue that their measures of scale availability were too crude to detect a role for contextual variation in the scalar alternatives that subjects consider, but that this latter variation is, in fact, a very plausible part of the explanation for the results. I discuss some specific ways adjectives might yield different scalar alternatives in different contexts, under sufficiently sophisticated assumptions about their lexical semantics and about the rhetorical structure of discourse.

## 1. Introduction

The phenomenon of scalar inference figures prominently in debates within formally-oriented semantics and philosophy of language over where one should draw the line between semantics and pragmatics. The recent history of this debate arguably begins with Grice's (1975) famous proposal that a conversational maxim of Quantity (specifically, "Make your contribution as informative as required (for the purposes of the conversational exchange") induces us, as a rule, to infer that an utterance involving a claim of any given quantitative or scalar strength implicates the denial of claims of greater quantitative or scalar strengths, as exemplified in (1).

(1) a. Sandy bought three pairs of Chi pants. (Sandy did not buy four pairs of Chi pants.)
    b. Some of the arguments were dubious (Not all of the arguments were dubious.)
    c. You can take High St. to campus. (You are not required to take High St. to campus.)
    d. The weather in Santa Cruz is cool. (The weather in Santa Cruz is not cold.)

In the first systematic experimental study of scalar inference patterns across syntactic categories, Doran, et al. 2009, 2012, observed that gradable adjectives (e.g. as in (1d)) yield scalar inferences

* Thanks to Bart Geurts for giving me the original impetus to work on this material, to the Festschrift organizers for giving me the impetus to finish it, and to Bob van Tiel and audiences at Radboud University Nijmegen and Utrecht University for comments. Many thanks, Sandy, for your support and interest in my work and career over the years. Congratulations on your 30th anniversary at UCSC!

significantly less often than do numerals, quantifiers and modals. In follow-up experiments using a different task, Van Tiel, et al. (2016; hereafter 'VT+') confirmed this result for adjectives, and moreover found that there is also considerable variability *within* the class of adjectives in the frequency with which inferences are likely to be generated.

VT+ consider various factors in explaining this result, none of which account for all, or indeed even a significant portion, of the data. They tentatively conclude that a combination of idiosyncratic statistical variation between adjectives in rate of scalar inferences, and past experiences with such inferences, also plays a role, but the tone of their conclusions suggests pessimism as to whether anything interesting can be said about these factors. In this comment, I suggest that the high degree of variation they found is exactly what we should expect from scalar inference with adjectives, and that it is a function of entity- and context-dependent assumptions subjects make about the number and nature of the specific scalar alternatives under consideration. There is already evidence for the role of scalar alternatives in influencing inference from studies such as Krifka 2002. Krifka observed that the utterance of a sentence like (2a) is likely to implicate (2b), whereas the utterance of (2c), where the numeral given is likely to be understood on a much coarser level of granularity, is not likely to implicate (2d).

(2) a. The Linguistics Department at UCSC has 15 faculty members.
b. The Linguistics Department at UCSC does not have 16 faculty members.
c. UCSC has 15000 undergraduate students.
d. UCSC does not have 15001 students.

There is no a priori reason to expect that adjectives should be deeply different from numerals in this respect. The principal difference between numerals and adjectives which has masked this similarity is, I will show, numerals' lack of polysemy.

In section 2, I provide a summary of VT+'s study and their analysis of the results. In section 3, I discuss the ways in which contextually-determined scalar alternatives could play a role. Finally, in section 4, I conclude by briefly discussing some of the larger implications of their study and the comments presented here.

## 2. Van Tiel, et al. 2016

VT+ carried out two main experiments. Experiment 1 consisted of questions such as in (3), in which the crucial statement appeared in italics. Subjects were asked whether it was possible to infer from this statement another one that contained a stronger adjective on the same scale:

(3) John says:

*She is intelligent.*

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

☐ Yes ☐ No

In this experiment, all of the sentences containing adjectives had a simple predicative structure and a pronoun as subject. Experiment 2 had exactly the same design with the same set of adjective pairs (e.g. <*intelligent*, *brilliant*>), except that the italicized test items contained full noun phrases in subject position (e.g., *That professor is intelligent*).[1]

Understanding how VT+ selected their test items is crucial to evaluating their results. The test materials consisted of 32 pairs of adjectives.[2] In order to be sure that these pairs formed true scalar alternatives, they were chosen by first searching the Corpus of Contemporary American English (COCA, Davies 2008), other corpora, and the internet for expressions such as 'X if not Y', 'X or even Y' and 'not just X but Y', where X and Y were adjectives; the expressions make explicit the fact that X and Y stand in a scalar relation, with Y stronger than X. From an initial set of candidate pairs, they made sure to select both pairs in which the weaker adjective on the scale was more frequent than the stronger adjective and vice versa, where frequencies were drawn from word counts in COCA.

VT+ selected the head nouns for the full noun phrases that appeared as the subjects of predication for each pair of adjectives based on a cloze pre-test administered to 10 subjects. These subjects were presented with sentences containing the two adjectives on each scale, as in (4); they had to provide three completions per item.

(4) The _____ is intelligent but she isn't brilliant.

Among the results for each sentence, VT+ selected three nouns. Whenever possible they ensured semantic variety among the nouns (avoiding the choice of e.g. both *singer* and *actress*, which are semantically more similar than e.g. *singer* and *nurse*), and whenever possible they chose two relatively frequent completions among those generated by the pre-test, and one infrequent completion, although the frequencies reported for each test item in their Appendix A suggest that this latter criterion was difficult to satisfy in a very meaningful way: only 20 out of 96 nouns used in Experiment 2 were mentioned more than 3 times in the cloze pre-test. The full list of adjective pairs and noun subjects used in the test items are listed in Table 1,[3] in order of descending frequency with which they generated a scalar inference on Experiment 2 (e.g. an example like *The task is difficult* licensed the conclusion *The task is not impossible* more frequently than e.g. *This child is content* licensed the conclusion *The child is not happy*).

As the results of Experiments 1 and 2, if not identical, were highly correlated, and as Experiment 1 leaves more room for speculation in terms of what subjects might be imagining when completing the task than does Experiment 2, I will focus on the latter. The rate at which scalar inferences were generated in both experiments varied greatly, and in a fairly smooth fashion, from 96% of the time in the case of <*difficult*, *impossible*> in Experiment 2, to 4% of the time in the case of <*content*, *happy*> in the same experiment. The largest drop in rate of inference was only 11%, occurring between <*possible*, certain> (inference in 93% of the cases) and <*allowed*, *obligatory*> (82%).

---

[1] The other aspects of their methodology are not relevant to this comment, so I will not discuss them here. See their paper for details.

[2] They also tested other syntactic categories (verbs, determiners, and adverbs), but I will not discuss these here.

[3] All nouns written in lower case in Table 1 appeared with a definite article or demonstrative determiner.

VT+ considered two main factors that could lie behind these results: the availability of the stronger scalar alternative (hereafter, *scale availability*) and the distinctness of the two alternatives (*scale distinctness*). They carried out additional experiments to operationalize these notions so that they could be included in their statistical model, and conclude that only scale distinctness had a significant role to play in the results of Experiment 2. However, the sum total of these two factors

| **Scalar pair** | **Nouns used in Experiment 2** | **Overall rate of inference in Exp. 2 (%)** |
|---|---|---|
| <*difficult*, *impossible*> | *task/journey/problem* | 96 |
| <*cheap*, *free*> | *water/electricity/food* | 93 |
| <*possible*, *certain*> | *Happiness/Failing/Success* | 93 |
| <*allowed*, *obligatory*> | *Copying/Drinking/Talking* | 82 |
| <*rare*, *extinct*> | *plant/bird/fish* | 79 |
| <*low*, *depleted*> | *energy/battery/gas* | 79 |
| <*hard*, *unsolvable*> | *problem/issue/puzzle* | 71 |
| <*warm*, *hot*> | *weather/sand/soup* | 64 |
| <*palatable*, *delicious*> | *food/wine/dessert* | 61 |
| <*scarce*, *unavailable*> | *recording/resource/mineral* | 57 |
| <*memorable*, *unforgettable*> | *party/view/movie* | 54 |
| <*cool*, *cold*> | *air/weather/room* | 46 |
| <*good*, *perfect*> | *layout/solution/answer* | 39 |
| <*old*, *ancient*> | *house/mirror/table* | 36 |
| <*good*, *excellent*> | *food/movie/sandwich* | 32 |
| <*adequate*, *good*> | *food/salary/solution* | 32 |
| <*funny*, *hilarious*> | *joke/play/movie* | 29 |
| <*dark*, *black*> | *fabric/sky/shirt* | 29 |
| <*hungry*, *starving*> | *boy/dog/elephant* | 25 |
| <*unsettling*, *horrific*> | *movie/picture/news* | 25 |
| <*small*, *tiny*> | *room/car/fish* | 25 |
| <*big*, *enormous*> | *elephant/house/tree* | 21 |
| <*snug*, *tight*> | *shirt/dress/glove* | 21 |
| <*attractive*, *stunning*> | *nurse/model/singer* | 21 |
| <*ugly*, *hideous*> | *wallpaper/sweater/painting* | 18 |
| <*wary*, *scared*> | *dog/victim/rabbit* | 14 |
| <*special*, *unique*> | *dress/painting/necklace* | 14 |
| <*silly*, *ridiculous*> | *song/joke/question* | 14 |
| <*tired*, *exhausted*> | *quarterback/runner/worker* | 14 |
| <*pretty*, *beautiful*> | *model/lady/girl* | 11 |
| <*intelligent*, *brilliant*> | *assistant/professor/student* | 7 |
| <*content*, *happy*> | *child/homemaker/musician* | 4 |

**Table 1**. Adjective pairs and test items used in VT+ 2016 (see their Table 3 and Appendix A).

as they operationalized them accounts for only 22% of the variance in their data; by including test item and participant as random variables, they account for only another 30% of the variance. They conclude (2016: 168) that "[i]n the absence of more successful candidates [for explaining the results, LMcN], we are forced to conclude that a major part of the observed variance was unsystematic." VT+ may of course be correct, but here I explore the possibility that the way they chose to operationalize and test for scale availability, in particular, was not optimal.

Scale distinctness was operationalized in two ways. First, they separated the list of pairs in Table 1 into those that associated with bounded scales and whose stronger member represented an endpoint (e.g. *impossible* or *free*) from those for which this was not the case (e.g. *brilliant* or *happy*). Second, they carried out an experiment (their Experiment 4) in which they asked subjects to measure on a 7-point Likert scale the difference in strength (or *semantic distance*) between statements involving the two adjectives in a given pair. Of these two, scale boundedness proved to be far more important, accounting on its own for 10% of the variance in the data, while semantic distance accounted for about 3% of the variance. Note, by way of illustration, that 9 of 11 pairs for which a scalar inference was generated over 50% of the time have a stronger member that marks the endpoint of a scale.

Scale availability was operationalized in four ways, none of which proved to make a significant contribution to accounting for inference patterns.

- **Grammatical category:** "The availability of a lexical scale <α, β> is greater if α and β are from a closed grammatical class." (2016: 157).[4]
- **Word frequency:** "the availability of a lexical scale <α, β> is an increasing function of the frequency of β relative to that of α" (2016: 157), or of the absolute frequency of β; for this measure VT+ used the COCA corpus.
- **Semantic relatedness**, as measured by Latent Semantic Analysis (Landauer and Dumais 1997): "The availability of a lexical scale <α, β> is an increasing function of the semantic relatedness of α and β" (2016: 158). This measure of relatedness consists in comparing the similarity of the distribution of α and β within a corpus.
- **Association strength:** The availability of a lexical scale <α, β> is an increasing function of the strength of association of β with α (2016: 151).

VT+ measured association strength in their Experiment 3 through a cloze test in which they asked subjects to generate the first three alternatives to the underlined word that came to mind when they read a sentence such as *She is <u>angry</u>*. Half of the subjects saw sentences like those in Experiment 1, with pronominal subjects; half saw sentences like those in Experiment 2, with full noun phrase subjects. In addition, half of each group received instructions that potentially facilitated the generation of a stronger scalar alternative, while the other half did not. For example, half were told (for the example in the instructions with *angry*):

> Which words could have occurred instead of the highlighted one? Some of the alternatives that may come to mind are *beautiful*, *happy*, *married*, and so on. We ask you to tell us the first three alternative

[4] VT+ considered grammatical category because their Experiments 1 and 2, as noted above, contained both open and closed class items.

> words that occur to you when you read these sentences. We are interested in your spontaneous responses, so don't think too long about it. (2016: 152-153)

The other half were given *furious* as the first sample alternative, instead of *beautiful*. However, it turned out that subjects provided stronger scalar alternatives only 27% of the time with pronominal subjects, and only 22% of the time with full noun phrase subjects; these figures barely changed when the instructions facilitated the generation of a stronger scalar alternative, though they did raise to 49% and 44%, respectively, under a more permissive coding of the data on which credit was given for the production of a scalar alternative stronger than the test item but weaker than the target scalar alternative (e.g., *probable* for pair <*possible*, *certain*>). In any case, VT+ found no clear correlation between a positive result on the scale association test and scalar inference.

It is not surprising that VT+ did not find the first two of these measures of scale availability useful in accounting for their original experimental results. The degree of variability *within* the class of adjectives – precisely what we are trying to explain – directly rules out grammatical category as a useful measure. Similarly, though relative or absolute word frequency might be a plausible factor in predicting general scale availability, it is not obvious what role it should play in the context of Experiments 1 and 2, which explicitly provided the scalar alternatives and asked subjects to make an offline judgment about them.

More interesting to consider are the measures of semantic relatedness and association strength. Readers unfamiliar with Latent Semantic Analysis should be aware that it is a fairly coarse technique for capturing semantic relatedness, because the distributions constructed for words generally do not take into account disambiguations beyond basic morphosyntactic category. Adjectives are particularly variable in interpretation depending on the nouns with which they combine. To give just one example, though *warm* and *hot* are scalemates for ascribing temperature, *hot*, but not *warm*, is used for popularity (*a hot/??warm product*), temper, and sex appeal (*a hot/??warm body*); while *warm*, but not *hot*, is used for friendliness or empathy (*a warm/??hot personality*). It is therefore to be expected that *warm* and *hot* emerged as only moderately semantically related on this measure.[5] The fact that *warm* yielded the inference *not hot* 64% of the time – more often VT+ would predict, given the semantic relatedness score of the pair – is also no surprise: The context of the task disambiguated the adjectives and rendered irrelevant the other uses that each has and that contributed to lowering the semantic relatedness score.

Finally, although operationalizing association strength via the results of the cloze test might seem reasonable a priori, the procedure VT+ used has two important weaknesses. In the version of the task with pronominal subjects (e.g. finding substitutions for *old* in *It is old*), subjects had virtually no information about the entity that the adjective was being used to describe. This information, as will be explained in the next section, is essential. However, when a full noun phrase subject was provided (e.g. *That house is old*), VT+ note that stronger scalar alternatives were

[5] The specific measure used in Latent Semantic Analysis is cosine similarity. Words are represented as vectors of co-occurrences with other words (perhaps restricted to a given window) in a corpus or "semantic space". The higher the cosine between the vectors for two words, the more similar the distributions of those two words. VT+ report that the cosine for *warm* and *hot* was 0.51 (with 0 indicating totally orthogonal distributions, and 1 indicating fully overlapping distributions) in the semantic space defined as "General Reading up to 1st year college" available at http://lsa.colorado.edu. See that site for further details.

generated even less often than with pronominal subjects. They suggest that was due to the fact that the full noun phrase (e.g. *That house*) could facilitate cloze fillers having nothing to do with the house's age (e.g. *beautiful*, *large*, *white*); in the case of the pronoun, the only facilitator of cloze fillers (other than the subjects' imaginations) was the adjective. In other words, the cloze test was not quite able to hone in on the extent to which a given stronger scalar alternative was likely to be considered for a given choice of subject.

## 3. Scale alternatives and scalar inference

As noted in the introduction, VT+ conclude that much of the variation in the results may be due to idiosyncratic statistical variation between adjectives in rate of scalar inferences, and subjects' past experiences with such inferences. Even if this is so, I am reluctant to further conclude that nothing interesting can be said about this variation. Rather, I take the variation to point to a particular view of lexical meaning and interpretation, one that Geurts 2011 appeals to for selective cases of scalar inference, as we will see momentarily, but which does not seem to be assumed in VT+ (nor is it widely assumed by formal semanticists more generally). This view of lexical meaning renders the variation less surprising.

The view consists of three assumptions. First, lexical entries for open class words are as a rule polysemous and are assigned a specific content only under contextual specification (see Bosch 1995 for a particularly clear presentation of this approach to the lexicon, though the idea has antecedents in his work going back to the early 1970s; see e.g. Hogeweg, to appear, for recent experimental evidence in its favor). Second, property predication can be fruitfully understood as a categorization task; this task can be carried out in some cases according to rule-based criteria, as is standardly assumed in classical entailment-based approaches to predication, while in others, notably with so-called relative gradable adjectives whose standard is context-dependent, it can be modeled as clustering based on similarity (McNally 2011; see Gärdenfors 2000 for a general framework in which this view can be modeled). Finally, the choice of categories under consideration for any predication can vary considerably from one class of entities to another, and, even within a class of entities, from one context to another. For example, in the case of temperature, for some purposes *hot* or *warm* vs. *cold* will be sufficient (e.g., when categorizing *types* of soups, say *caldo de Nadal* vs. *gazpacho*), whereas in others (e.g., when one wants to inform one's guests about temperature the soup they are about to eat), a finer-grained set of distinctions including also *warm* and *cool* might be called for. The set of category alternatives under consideration in any given context will influence how an entity is classified and what inferences are drawn about it.[6]

These assumptions support an analysis of variable scalar inference as follows. Imagine that a subject sees the sentence *The weather is warm*. We of course have no idea of what was in the subject's minds when deciding whether from this to infer that the weather was not hot; however, given that the warm/hot distinction is very often used for weather, it would certainly be unsurprising if the subject made that inference. Notice that this inference depends on the understanding that there is a stronger alternative than *warm* (namely *hot*) that is not being chosen. Inference from *warm* to *not hot* might also be expected in the case of *The sand is warm* because

[6] This last assumption is very much in the spirit of Bidirectional Optimality Theoretic and related probabilistic approaches to semantics. See e.g. Zeevat 2011 and Lassiter and Goodman 2015.

that difference in categorization correlates with other information that is likely to matter to the hearer: hot sand can burn. In contrast, in the case of *The soup is warm*, the subject could have opted equally for an interpretation involving contrast only with cold soups as a type of dish, or for an interpretation focused specifically on (ambient) temperature. Only in the second case would the inference to *not hot* be called for. If only two categories, *cold* and *warm*, are under consideration, there is no stronger alternative than *warm* available and no scalar inference will be possible.

In this scenario, assuming that 100% of speakers behaved uniformly in the first two cases and that speakers divided 50-50% in the third, we would expect a scalar inference 75% of the time. They observed a 64% inference rate is not terribly far from this.

Interestingly, this sort of analysis is very similar to that defended in Geurts 2011 for sentences like (5) (see also Geurts 2009 for earlier discussion of these examples):

(5) If it's warm, we'll lie out in the sun, but if it's VERY warm, we'll go inside and sit in front of the air-conditioner. (Geurts 2011: 140; example originally due to Horn 2006: 27)

Geurts observes that (5) cannot be truthfully uttered if *warm* does not, in the context, entail *not very warm*. He proposes that in this case, the lexical semantics of *warm* is contextually narrowed. This narrowing, though pragmatic in nature, affects what is said in the Gricean sense, that is, the propositional content of the sentence. In other words, if we infer *warm but not very warm* in cases like (5), the scalar inference is due not to a quantity implicature – it is not defeasible – but rather to a logically prior decision about how to carve up the categorization space.

Geurts notes, correctly, that taking polysemy and the context dependence of lexical meaning into account does not undermine a classical pragmatic theory of quantity implicature:[7] it is simply the case that, in order for scalar reasoning to be applied, the set of alternatives under discussion must first be fixed. Thus, returning to VT+'s results, for those pairs Table 1 for which scalar inferences were infrequently generated, we can hypothesize that the categorization space for most subjects and for most test items did not include the stronger alternative in the pair. Of course, VT+'s four tests for scale availability were intended precisely to test to whether or not this was in fact the case, but arguably none of them provided sufficient context sensitivity to do so.

Space precludes an item-by-item discussion of all of the adjective pairs in Table 1. I will instead finish the section by mentioning two additional factors beyond general polysemy that can affect the number or nature of the alternative categories under consideration in any given context, and that could have played a role in reduced scale availability for some of VT+'s pairs, in the hope that future experiments might be designed to control for them.

The first is subjects' assumptions about the (implicit) question(s) under discussion in the test items. VT+ (2016: 166-167) briefly consider the possibility that the question under discussion could have influenced their results. They note that van Kuppevelt 1996 argued that scalar implicatures for numerals arise when the numeral is part of the information focus of the sentence, as in (6), but not when it is part of the background, as in (7):

[7] Among other things, it does not force one to assume that lexical items are *ambiguous*, nor does it require one to adopt a logical-syntactic approach to implicature (see Chierchia, Fox and Spector 2012). See Geurts 2011 for extensive discussion.

(6) A: How many years has Sandy worked at UCSC?
B: Sandy has worked there for [thirty]F years.

(7) A: Who has worked at UCSC for thirty years?
B: [Sandy]F has worked at UCSC for thirty years.

VT+ dismiss this sort of explanation because there is no evidence of any such difference in information structure between their test items. Though this is of course correct insofar as it goes, the assumption that the question under discussion can be determined on the basis of information structure alone is too crude: A richer notion of discourse structure (as modeled in e.g. Mann and Thompson's (1988) Rhetorical Structure Theory or Asher and Lascarides' (2003) Segmented Discourse Representation Theory) must also be taken into account.

To see this, consider the pair <*adequate*, *good*>. *Adequate* is indeed a scalemate of *good*, but adequacy and goodness are generally evaluated with respect to some purpose (what Kagan and Alexeyenko 2010 call, for slightly different sorts of cases, a *functional standard*). For example, food may be adequate or good in quantity or quality, a salary may be judged as adequate to meet expenses, good insofar as it surpasses that of many peers, etc. VT+'s experimental design did not oblige subjects to choose the same standard for the two adjectives in any given pair. For example, a subject reading *The salary is adequate* might imagine this as addressing an implicit question about whether the salary meets someone's needs or not. When then asked whether that means that the salary is not good, the subject might hold the standard constant and infer that it is not. However, she could just as easily infer that no such conclusion is warranted, because there simply is no reason to think that a salary adequate to meet someone's needs could not be considered good with respect to some other standard. In other words, there is no reason the subject could not imagine that the local question under discussion shifts between the initial presentation of the test item and the subsequent question about the second adjective in the pair. Such a shift would be compatible with a rhetorical structure in which the main question under discussion might be something like *What is your salary like?*, with elaborating sub-questions such as *Does it meet your needs?*, *Are you better off than most of your peers?*, etc.[8] But if the local questions under discussion shift as the subjects move from consideration of one adjective to another, it is entirely likely that the alternative categories under consideration will also shift. In other words, what counts as a scalar alternative is potentially highly dynamic. VT+ do not contemplate this possibility.

A second factor that can affect the set of scalar alternatives under consideration is the overall semantic complexity of the adjectives in question. There are pairs, such as <*wary*, *scared*>, which do not differ only, or perhaps even primarily, in strength. Recall that VT+ chose scalar alternatives by looking for occurrences in phrases of the form 'X if not Y', 'X or even Y' and 'not just X but Y.' One might find examples of *wary if not scared* because wariness is often motivated by emotions related to fear. But wariness is a property of behavioral disposition, while fear is an emotional state that merely implies, but does not entail, particular behavioral dispositions. Wariness might be provoked by an emotion that one would not quite describe as fear (such as anxiety), and higher degrees of fear might be correlated with dispositions that are more extreme

[8] See Hunter and Abrusán, to appear, for discussion of how rhetorical structure and the notion of question under discussion can be related.

than wariness (such as a complete inability to act). For this reason, *wary* and *scared* might well serve as alternatives in some contexts. However, wariness can be provoked by feelings distinct from fear, notably mistrust. It is therefore unsurprising that *scared* was not a salient alternative to *wary* in the vast majority of cases. VT+'s method for choosing test items did not control for this factor.

## 4. General discussion

The discussion in the preceding section has identified three factors not explicitly acknowledged by VT+ that could have played a role in the highly varying degree to which adjectives yielded scalar inference in their experiments: 1) the polysemy of the adjectives (in contrast to e.g. numerals and quantifiers), 2) their semantic complexity, and 3) the nature of the question under discussion. Once these are taken into account, there does not seem to be any reason to think that adjectives, as a category, interact any differently with the general theory of scalar inference than do numerals or other sorts of expressions for which scalar inference has been more widely studied, just as Geurts 2011 concluded. The fact that scalar alternatives have been shown to vary even for numerals, depending on the granularity of the information that is considered pertinent in the context (which is a function, at least in part, of the question under discussion), suggests that the difference in the variability of scalar inference with adjectives vs. numerals and quantifiers is due to a difference in their lexical richness.

This is a welcome conclusion, but it also carries with it some important implications. Preserving a fairly simple and elegant account of scalar inference crucially depends on taking scalar alternatives to be defined in a highly local, context-dependent fashion. However, any hope of developing a theory of when and how these alternatives arise depends, in turn, on incorporating well-developed theories of polysemy and rhetorical structure into the analysis of interpretive phenomena for which they have been, to date, largely ignored. The gain in insight into the special scalar inference behavior associated with adjectives should serve as an argument for changing this situation and extending the interest in these aspects of language among formally-oriented semanticists and pragmaticists.

## References


Asher, Nicholas and Alex Lascarides (2003). *Logics of conversation*. Cambridge: Cambridge University Press.

Bosch, Peter. 1995. Meaning and contextual concepts. *Semantic and conceptual knowledge. Papers from a joint workshop of the Max Planck Arbeitsgruppe 'Strukturelle Grammatik' and the IBM Institute for Logic and Linguistics*, eds. M. Bierwisch and P. Bosch, 79-99. Arbeitspapiere des Sonderforschungsbereichs 340, Vol. 71, Universität Tübingen.

Chierchia, Gennaro, Danny Fox, and Benjamin Spector. 2012. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. *Handbook of semantics*, vol 3, eds. C. Maienborn, K. von Heusinger, and P. Portner, 2297-2331. Berlin: Mouton de Gruyter.

Davies, Mark. 2008. The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

Doran, Ryan, Rachel E. Baker, Yaron McNabb, Meredith Larson, and Gregory Ward. 2009. On the nonunified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics* 1: 211-248.

Doran, Ryan, Gregory Ward, Meredith Larson, and Yaron McNabb. 2012. A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language* 88: 124-154.

Gärdenfors, Peter. 2000. *Conceptual spaces*. Cambridge, MA: MIT Press.

Geurts, Bart. 2009. Scalar implicature and local pragmatics. *Mind and Language* 24: 51-79.

Geurts, Bart. 2011. *Quantity implicatures*. Cambridge: Cambridge University Press.

Grice, H. Paul. 1975. Logic and conversation. *Syntax and semantics 3: Speech acts*, ed. P. Cole, 41-58. New York: Academic Press.

Hogeweg, Lotte. To appear. Suppression in interpreting adjective noun combinations and the nature of the lexicon. *Journal of Semantics*.

Horn, Laurence R. 2006. The border wars: A neo-Gricean perspective. *Where semantics meets pragmatics*, eds. K. von Heusinger and K. Turner, 21-48. Amsterdam: Elsevier.

Hunter, Julie and Márta Abrusán. To appear. Rhetorical relations and QUDs. *New frontiers in artificial intelligence: JSAI-isAI Workshops LENLS, JURISIN, KCSD, LLLL revised selected papers*. Berlin: Springer.

Kagan, Olga, and Sascha Alexeyenko. 2010. Degree modification in Russian morphology: The case of the suffix *-ovat*. *Proceedings of IATL 26*.

Kennedy, Christopher. 2001. Polar opposition and the ontology of 'degrees'. *Linguistics and Philosophy* 24: 33-70.

Krifka, Manfred. 2002. Be brief and vague! And how bidirectional optimality theory allows for verbosity and precision. *Sounds and systems. Studies in structure and change: A festschrift for Theo Vennemann*, eds. D. Restle and D. Zaefferer, 439-458. Berlin: Mouton de Gruyter.

van Kuppevelt, Jan. 1996. Inferring from topics: Scalar implicatures as topic-dependent inferences. *Linguistics and Philosophy* 19: 393-443.

Landauer, Thomas K., and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104: 211-240.

Lassiter, Daniel and Noah Goodman. 2015. Probabilistic semantics and pragmatics: Uncertainty in language and thought. *Handbook of contemporary semantic theory*, second edition, eds. C. Fox and S. Lappin, 143-167. Oxford: Wiley Blackwell.

Mann, William C. and Sandra A. Thompson (1988). Rhetorical Structure Theory: A theory of text organization. *Text* 8: 243-281.

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

van Tiel, Bob, Emiel van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33: 137-175.

Zeevat, Henk. 2011. Bayesian interpretation and Optimality Theory. *Bidirectional Optimality Theory*, eds. A. Benz and J. Mattausch, 191-220. Amsterdam: John Benjamins.