preprint of 2019 Journal of Semantics article
doi:10.1093/jos/ffy020

Vagueness in implicature:
The case of modified adjectives [*†]

Timothy Leffel;[1,✉] Alexandre Cremers;[2] Nicole Gotzner;[3] Jacopo Romoli[4]

[1]NORC at the University of Chicago
[2]ILLC, Universiteit van Amsterdam
[3]Leibniz-ZAS, Humboldt University
[4]Ulster University

Abstract

We show that the interpretation of sentences like *John is not very* ADJ depends on whether ADJ is vague. We argue that this follows from a constraint on the interaction between vagueness and conversational implicature, a domain that has not been studied extensively. The constraint states that implicatures are not drawn if they lead to "borderline contradictions" (see Ripley 2011; Alxatib & Pelletier 2011; a.o.), a natural extension of the idea that implicatures should not contradict assertions (Fox 2007; Fox & Hackl 2006; a.o.). Experiment 1 establishes that *not very* ADJ gives rise to the implicature ADJ for the non-vague absolute adjective *late*, but not for the vague relative adjective *tall* (in the terminology of Kennedy & McNally 2005a). Experiment 2 generalizes this result to three relative adjectives in the positive form (*tall*, *hot*, *fast*), against those same adjectives in their (non-vague) comparative forms (*taller/hotter/faster than the average X*). We also constructed quantitative meaning representations for complex predicates of the form ADJ $\wedge \neg$ *very* ADJ, using fuzzy logic to model the contribution of boolean connectives and our experimental data to represent the meanings of adjectives. The results of these analyses suggest that strengthening *not very* ADJ with ADJ leads to a more contradictory interpretation when ADJ is vague than when it is not, as expected on our theory. While our results apply directly to only a specific set of lexical items, we hypothesize that they reflect a more general pattern among gradable predicates. This motivates more systematic investigation into the role that vagueness can play in the derivation of conversational implicatures.

[*]**Acknowledgments:** We would like to thank David Barner, Emmanuel Chemla, Julian Grove, Emily Hanink, Chris Kennedy, Yaron McNabb, Stephanie Solt, Ming Xiang, and four anonymous reviewers, as well as audiences at the LSA Annual Meeting in 2016, UCL, University of Chicago, and SIASSI for helpful comments on various stages of the project. This work was supported by the Alexander von Humboldt-Stiftung and the Andrew W. Mellon Foundation as part of the SIAS Summer Institute "The Investigation of Linguistic Meaning," by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.31361, by the Deutsche Forschungsgemeinschaft (DFG) as part of the Xprag.de Initiative (Grant Nr. BE 4348/4-1), and by the Netherlands Organisation for Scientific Research (NWO).

[†]**Supplemental Materials:** Experimental data and associated analysis and visualization code can be found at this repository: https://github.com/lefft/not_very_adj

1 Introduction

Vagueness and conversational implicature are both sources of interpretive uncertainty: with vagueness, there is uncertainty about whether a predicate applies to an object (e.g. what heights count as *tall*?); with implicature, there is uncertainty about whether a candidate inference was intended by the speaker (does *some* imply *not all*?). Both topics have been studied extensively from psycholinguistic, model-theoretic, and Bayesian approaches to natural language semantics and pragmatics, but potential interactions between them have not been investigated. In this paper we show that a collection of related contrasts involving gradable expressions are quite naturally explained as direct consequences of an interaction between the semantic property of vagueness and general pragmatic principles regulating conversational implicatures.

As a running example, we will use a contrast between the (vague) relative adjective *tall* and the (non-vague) absolute adjective *late*: *not very late* in (1) is most naturally interpreted with an inference to the positive form (*late*), whereas *not very tall* in (2) usually has no corresponding inference (qualifications to follow; see fn3. for discussion of *late*'s status w.r.t. vagueness).

(1) John was not very late.
$\rightsquigarrow$ John was late.

(2) John is not very tall.
$\not\rightsquigarrow$ John is tall.

We argue (in §2.3) that the inference of *late* from *not very late* arises just as other Manner implicatures like [*John didn't vote for Nader*] $\rightsquigarrow$ [*John voted*] do: the predicate *not very late* is strictly weaker than the simpler structural alternative *not late*, and hence (given some Gricean reasoning) implicates its negation—namely *late*. Our main proposal about the more interesting part of the puzzle—why *not very tall* does not behave similarly—is that certain "vague implicatures" fail to be drawn because when combined with a vague assertion, they can result in truth-conditions that have properties of "borderline contradictions" (sentences like *John is tall and not tall*, terminology from Ripley 2011). Combined with the structural implicature associated with restrictive modifiers (like *very*) in downward-entailing contexts, this explains the difference between (1) and (2): it is easy to find a time that satisfies the strengthened meaning [*late* $\wedge$ $\neg$[*very late*]] (just go a tiny distance beyond the threshold for *late*), but because the threshold for *tall* is never a specific, identifiable value, there is considerable uncertainty about which heights (if any) would satisfy the strengthened meaning [*tall* $\wedge$ $\neg$[*very tall*]] in a context. See §4 for details.

If something resembling our proposed mechanism is indeed responsible for the patterns discussed in this paper, we believe that this establishes strong motivation for further investigation into the role that vagueness and scale structure can play in the derivation of conversational implicature and pragmatic inference more broadly.

The paper is structured as follows: after some brief background on vagueness and scale structure (§2.1), we describe *not very* ADJ in more detail, illustrating how it constitutes a genuine semantic puzzle (§2.2). In §2.3 we propose a straightforward derivation of the implicature from *not very* ADJ to ADJ, and then sketch a hypothesis that would prevent the application of this mechanism with vague predicates like *tall*. §2.4 shows that the pattern of inference extends beyond *not very* ADJ to a range of related constructions. §2.5 motivates the experiments to follow, comparing some concrete predictions of our hypothesis to those of an alternative. In §3 we present an experiment designed to quantify the empirical difference with respect to implicatures between *tall* and *late* in the *not very* ADJ frame (Experiment 1). In §4 we develop a (partial) theory of the interaction between vagueness and implicature calculation, based on the idea that implicatures are blocked if they lead

to borderline contradictions when conjoined with the literal meaning. We then provide evidence for the proposal based on reconstructed interpretations of logically complex predicates from the response data from Experiment 1. §5 presents a second experiment that generalizes the result of Experiment 1: we compare the interpretive profiles of three relative gradable adjectives (*tall*, *hot*, *fast*) in the *not very* ADJ frame to those of the same adjectives in the comparative form (*taller*, *hotter*, *faster*) using the frame *not much* ADJ-*er than average*. Comparison of the response curves for *not (very)* ADJ and *not (much)* ADJ-*er than average*, as well as further analysis of reconstructed predicates, support our theory of the data. §6 concludes.

2 A pattern of interaction between vagueness and implicature

2.1 Background on vagueness and gradability

Vagueness and gradability are related but distinct concepts. Gradable expressions are (roughly) those that hold of objects to varying degrees. Vagueness can be characterized in terms of inherent uncertainty about whether an expression applies to an object.

Vagueness arises with diverse categories of linguistic expressions, including (some) gradable adjectives. Gradable adjectives can be subdivided into those with "relative" standards—like *tall* and *fast*—and those with "absolute" standards—like *full* and *late*. While the exact nature of this distinction remains a matter of debate, it is often taken for granted that relative gradable adjectives are inherently vague in a way that other adjectives are not.[1]

Semantics for gradable adjectives are typically stated in terms of scales—ordered sets of reified "degrees"—and threshold values on those scales ("degree semantics," Bartsch & Vennemann 1972; Cresswell 1976; von Stechow 1984; Bierwisch 1989; Heim 1985,2000; Kennedy 1999,2007; Solt & Gotzner 2012; a.o.). In this framework an adjective ADJ applies to an object x iff x's position on the ADJ-scale exceeds a contextually or compositionally designated value called the "threshold of application" (or "standard of comparison"), written 'θ_{ADJ}'.

One popular degree-based implementation of semantic composition in gradable adjectives (Kennedy 2007) holds that a positive-form adjective simply introduces a measure function μ_{Adj} from individuals to degrees, and thus is itself not inherently vague or non-vague. Adjectives can then compose with a **pos** morpheme to form a predicate which holds of individuals whose Adj-measure (e.g. height) exceeds whatever the context determines is the threshold θ_{Adj} on the appropriate degree scale (e.g. the minimum height required to count as 'tall'). θ_{Adj}'s value is constrained contextually (e.g. by comparison classes), but also by lexical properties of individual adjectives.

Analyses of the relative/absolute distinction differ most substantively in how they constrain possible (or likely) values of θ_{Adj}. In the tradition of Kennedy (1999), the underlying scale structure of an individual adjective dictates whether it will have minimum-, maximum-, or relative-standard semantics. For example since *tall* has an open scale, θ_{tall} could in principle be located anywhere along the height scale; this leads to inherent uncertainty about the location of θ_{tall}, and hence relative semantics. But since *late* has a lower-closed scale (*slightly late*/#*tall*), the lower endpoint is a possible choice for θ_{late}; this leads to minimum-standard semantics as a natural default.[2]

More recent approaches based on Bayesian inference model the resolution of θ_{Adj} using proba-

[1] Although some have suggested that all gradable adjectives are vague (e.g. Lassiter & Goodman 2014), they at least agree that there are significant differences in the amount of uncertainty introduced by different adjectives.

[2] The mechanism by which this apparent default arises is a highly interesting question in its own right; see Kennedy's (2007) discussion of "interpretive economy," and Potts's (2008) game-theoretic perspective.

bility distributions over degree scales, whose shapes can be affected by explicit comparison classes, unconscious prior knowledge, or context-specific information. Vagueness arises as the result of a complex interaction between the measure function denoted by an adjective, the comparison class determined by a context and utterance, and language users' priors (see Lassiter & Goodman, 2014, 2017; Qing & Franke, 2014a for two concrete implementations).

Following arguments and discussion from Kennedy & McNally (2005a), Kennedy (2007), Solt (2015), Qing & Franke (2014a) and others, we will assume that the relative/absolute distinction can be reduced to approximately the following: relative adjectives are associated with open scales (those without endpoints), and tend to have uniform (or possibly normal) prior distributions; absolute adjectives are associated with closed scales (those with a maximum or minimum boundary or both), and their thresholds tend to be located at a scalar endpoint because the relevant measure functions map significant probability mass to those endpoints.[3]

In what follows we assume familiarity with these and related distinctions from degree semantics (see Schwarzschild 2008 for a review).

2.2 A puzzle about intensified gradable adjectives

The contrast between *tall* and *late* in (1)-(2) exemplifies a broader pattern: in the frame '*not very* ADJ', minimum standard absolute adjectives tend to pattern like (1), as shown in (3), while relative standard adjectives tend to pattern like (2), as shown in (4).[4]

(3) a. The antenna is not very bent.
$\rightsquigarrow$ The antenna is bent.

b. The table isn't very dirty.
$\rightsquigarrow$ The table is dirty.

c. Mary isn't very sick.
$\rightsquigarrow$ Mary is sick.

(4) a. John isn't very smart.
$\not\rightsquigarrow$ John is smart.

b. The supermarket is not very far.
$\not\rightsquigarrow$ The supermarket is far.

c. The line is not very long.
$\not\rightsquigarrow$ The line is long.

That the interpretation of intensified adjectives under negation could be sensitive to the relative/absolute distinction has, to our knowledge, not been suggested in the literature. Bolinger (1972) and Horn (1989) pointed to a *not*-Adj interpretation of examples like (2), and proposed that it involves a form of euphemism or is related to politeness. Horn in particular refers to *not very* ADJ as a kind of "negative understatement," suggesting that its meaning is derived from a "conventionalized strengthening rule" that interprets '*not* intensifier ADJ' as '*rather un*-ADJ' " (Horn 1989:353-4). The intuition here is well-illustrated by example: in order to avoid asserting that someone is not smart, one can say they are *not very smart*, the literal meaning of which leaves open the possibility that they are indeed smart. This is quite similar in spirit to Krifka's (2007) analysis of "negated

[3] Absolute adjectives are also context-dependent, but arguably in a different way: clearly what it means to be *late* or *full* varies by situation (consider *full* for a wine glass versus a beer glass), but this alone does not imply that these adjectives are vague. Instead, many theories of gradability model such non-endpoint oriented readings as a form of imprecision (e.g. as in approximative readings of numerals), which introduces some flexibility into the interpretation (see especially Kennedy 2007 and Lasersohn 1999). The exact nature of non-endpoint thresholds in absolute adjectives is a theoretical debate that we do not wish to enter here (but see Lassiter & Goodman 2014; Qing & Franke 2014a; Aparicio et al. 2016; Leffel et al. 2016 for some recent studies).

[4] Contraction of negation (*isn't* versus *is not*) does not seem to make a difference, at least in our judgments. A broader and more important point is that these kinds of judgments are inherently gradient. The specific factors affecting these judgments are an important topic that deserves systematic investigation in future research.

antonyms" like *not unhappy*, which seems to express a state that is slightly too low on the happiness scale to count as 'happy'.

Horn cites pairs like *happy*/*sad* and *smart*/*stupid* to argue that the 'not ADJ' interpretation is euphemistic: in the frame '*not* intensifier ADJ', the "positive" (in the evaluative sense) predicates of these pairs have the 'not ADJ' reading, while the "negative" predicates seem to be more neutral. However, while an explanation in terms of euphemism feels plausible for evaluative adjectives (*smart*/*stupid*) or adjectives that encode a (potentially) desirable property (*tall*), it is unclear how it could extend to cases of purely dimensional relative adjectives like *far* in (4b). In fact, looking at the pair of absolute antonyms *early/late* and the pair of relative antonyms *close/far* suggests that there is a double dissociation between "positivity" and the availability of the 'not ADJ' interpretation. Indeed, both absolute adjectives *early* and *late* give rise to an 'ADJ' inference, while both relative adjectives *close* and *far* are compatible with the '*not* ADJ' interpretation. Because they are antonyms, an explanation based on euphemism would predict differences *within* pairs, but the observed contrast is *between* pairs instead, aligning with the relative/absolute distinction.

The observation that underlying lexical semantics systematically affects whether or not an adjective in the '*not very* ADJ'-frame is associated with the positive 'ADJ' inference shows that Horn's (1989) characterization of the construction is incomplete. More broadly, this variation between adjective subclasses is problematic for *any* general explanation that does not make reference to the lexical semantics of particular adjectives. This is not to say that euphemism is irrelevant to the pattern, nor to say that other properties of lexical items do not affect how this construction is interpreted—they surely do. It is instead to say that scale structure has been an overlooked determinant in previous discussions of the phenomenon (see also §6 below).[5]

Importantly, while Horn (1989) aimed to explain the existence of inferences to the *negation of the positive form*, in this paper we aim primarily to explain the *presence or absence* of inference to the *positive form*. Failure of inference to ADJ is a necessary condition for inference to ¬ADJ, but not a sufficient one—this is especially clear when ADJ is gradable. In the following discussion, we show that standard mechanisms of implicature easily derive inferences of the form [*not very* ADJ $\rightsquigarrow$ ADJ], but do so in an unconstrained fashion. In §2.5, we advance a hypothesis about why implicatures with this shape feel less "attractive" with vague predicates than with non-vague ones.

2.3 Explaining positive inferences from *not very* ADJ

Grice's Maxim of Manner states (roughly) that a speaker will not use more words than are necessary to get her point across (Grice 1975 and much subsequent work). In combination with the Maxim of Quality, this can lead to implicatures when a speaker uses an utterance which is both more complex and less informative than a readily available alternative.

Inferences generated in this fashion have received considerable attention in the literature and have been treated as "Manner implicatures" or (quasi)-presuppositions (Simons 2001/2013; Schlenker 2008), or simply as quantity implicatures given a theory of alternatives which considers all sentences

[5] The status of maximum standard absolute adjectives in the '*not very*'-frame is less clear, in part because their thresholds are located (by default) at scalar maxima: how could *very* shift a threshold higher if it is already at the scale maximum? To the extent that *very* occurs felicitously with maximum standard adjectives (e.g. *very full*), a kind of pragmatic weakening seems to be involved whereby *full*'s threshold is relaxed from the strict endpoint (Lasersohn 1999), in effect coercing it into a relative adjective (Kennedy & McNally 2005a suggest that *very* is ungrammatical with absolute adjectives except when used in a "relative-like, imprecise" way). Because of these complications, we do not investigate maximum standard adjectives in this paper. The comparison between relative and minimum standard is sufficient for our purposes since they differ semantically from one another with respect to vagueness.

that are structurally simpler than the asserted one (Katzir 2007,2014; Fox & Katzir 2011; see also Matsumoto 1995; Sauerland 2004; Chemla 2009). A hallmark case of these implicatures involves restrictive modifiers in negative and downward-entailing environments. For example Simons (2001/2013) notes that sentences like (5a) strongly imply that the unmodified alternative sentence (5b) is false—i.e. that John did indeed vote (for someone other than Nader).

(5) a. John didn't vote for Nader.
b. John didn't vote.
c. John voted.

The inference of (5c) from (5a) can be shown to follow from standard (neo-)Gricean reasoning. The following informal sketch is adapted from Simons 2001/2013 and Katzir 2007.[6]

(6) (*i*) You, the speaker, assert (5a) (=*John didn't vote for Nader*).
(*ii*) I, the addressee, observe that (5b) (=*John didn't vote*) is a "better" alternative to (5a), since (5b) asymmetrically entails (5a) (Quantity) and contains a proper subset of words (Manner).[7]
(*iii*) I observe that you chose not to assert (5b) despite this fact.
(*iv*) By Quality, I therefore conclude that you must not believe (5b), because if you did, you would have uttered it instead.
(*v*) I assume that if you know John didn't vote for Nader, then you probably know whether or not he voted (you are an "opinionated authority" in Sauerland (2004) and others' terms).
(*vi*) Given (*iv*) and (*v*), it follows that you probably believe (5c) (=*John voted*).

Implicatures like the one in (5) arise in syntactic environments other than matrix negation, as well. This is because, as Katzir (2007) and others discuss, a modified sentence generally asymmetrically entails its unmodified version in upward-entailing contexts. For example *Someone signed their name with a pencil* entails *Someone signed their name*.

In downward-entailing contexts, entailment relations are reversed, so in such environments we expect an implicature to the negation of the unmodified version from the assertion of the modified one. For instance *No one signed their name with a pencil* naturally gives rise to the negation of a simpler alternative as an implicature (i.e. ¬[*No one signed their name*]). And because (classically) $\neg\neg\exists \Rightarrow \exists$, this results in a potential inference to *At least one person signed their name*.

We propose that the puzzle about *not very* in (1)-(4) is partially explained by Manner-based pragmatic reasoning exactly parallel to cases like (5). The account runs as follows: a modified sentence like (7a) entails its unmodified counterpart (7b). In other words, *very* is a restrictive modifier on gradable adjectives, so anything that's 'very ADJ' necessarily exceeds the threshold for ADJ.

[6] A reviewer notes that derivations based on this kind of reasoning are susceptible to the so-called "symmetry problem:" in the case of *John was not very late*, since *late* is a briefer alternative to *not very late* (even briefer than *not late*), why shouldn't we negate *it* instead, to derive the implicature ¬*late*? We view the symmetry problem as an important foundational challenge for all accounts of implicatures. But we do not believe it has any *special* relevance to the connection between vagueness and implicature, the issue we aim to illuminate here. In our specific case, we suspect that the issue could be resolved by a suitably constrained system for generating structural alternatives, which would count *not late* but not *late* as an alternative to *not very late*. For some recent work on the symmetry problem as it relates to structural alternatives, see Trinh & Haida 2015; Breheny et al. 2016; and also Katzir 2007.

[7] However one wants to measure utterance length or structural complexity, it should be clear that (5b) is "briefer" and "simpler" than (5a). See Katzir 2007; Fox & Katzir 2011 for some relevant formalizations of these notions.

(7) a. John was very late.
b. John was late.

In a downward entailing environment like matrix negation, the entailment relation is reversed, so (8a) is entailed by (8b). And crucially (8b) is structurally less complex than (8a) (uncontroversially so, we believe). Therefore, given reasoning like that in (*i-vi*) above, (8a) should be associated with an implicature to the negation of (8b)—*John was late*.

(8) a. John was not very late.
b. John was not late.

And as we pointed out above, (8a) does seem to suggest that John was late. Similarly, cases like those in (9) conform to this expected pattern.

(9) a. The kitchen is not/isn't very dirty. $\rightsquigarrow$ The kitchen is dirty.
b. The chair didn't get/isn't very warped. $\rightsquigarrow$ The chair is warped.
c. The antenna isn't very bent. $\rightsquigarrow$ The antenna is bent.

A key premise of the analysis is that the presence of *very* leads to consideration of simpler structural alternatives (and ultimately implicature), due to its function as a restrictive modifier. It is therefore expected that other restrictive modifiers in this syntactic frame should give rise to corresponding inferences. And indeed the pattern appears to extend to some modifiers beyond *very*, including *super* and *really* (see Bolinger 1972; Horn 1989 for discussion). For instance our intuition is that *not super late* tends to imply *late*. Additionally, (10) shows that the same inference pattern applies in downward-entailing contexts beyond just matrix negation, further strengthening the parallel between intensified adjectives and implicatures from phrasal modifiers (cf. *I don't think John passed with an A* $\rightsquigarrow$ *I think John passed*). This generality across different adverbs and types of downward-entailing environment casts doubt on the idea that *not very* ADJ is a frozen expression with idiosyncratic meaning.

(10) a. I don't think John was very late. $\rightsquigarrow$ I think he was late.
b. None of the students were very late. $\rightsquigarrow$ At least some of the students were late.

Of course, the simple Manner-based account as stated above does not solve the entire puzzle. Because nothing in the derivation proposed for (7)-(8) above depends on the scale structure of *late*, one should expect *all* sentences of the form *X is not very* ADJ to be associated with implicatures of the form *X is* ADJ. As already illustrated for *not very tall*, this does not extend to relative gradable adjectives: without focus intonation on *very*, the sentences in (11) all seem to lack the '*X* be ADJ' inferences observed in (9).

(11) a. John is not very tall. $\not\rightsquigarrow$ John is tall.
b. The bed is not very comfortable. $\not\rightsquigarrow$ The bed is comfortable.
c. Bill is not very happy. $\not\rightsquigarrow$ Bill is happy.
d. John's lifestyle is not very responsible. $\not\rightsquigarrow$ John's lifestyle is responsible.

The positive inference also appears to be absent with some other intensifiers, e.g. as in *not super tall*.[8] And the same is true for certain other downward-entailing environments, e.g. *I don't think*

[8] The facts are less clear with stronger intensifiers (*extremely*, *astoundingly*), with many speakers endorsing the positive-form inference from sentences like *John is not extremely tall*. See §6 for discussion of intensifier strength

John is very tall, *None of the students are very tall*.[9]

The fact that the inference from *not very* ADJ to ADJ targets a specific subclass of gradable adjectives strongly suggests that the underlying cause of the asymmetry is somehow related to scale structure—and consequently, to vagueness. Next, we briefly review several related syntactic domains that appear to exhibit a similar relationship between scale structure and the attractiveness of potential inferences.

2.4 Similar patterns in related domains

Additional support for the role of scale structure and vagueness in the meaning of *not very* ADJ comes from parallel contrasts in other syntactic environments and with other syntactic categories. In (12)-(16) are examples of contrasts with similar properties to *not very tall/late*, each with a slightly different syntactic nature. The (a)-examples involve predicates with vague, relative-like meanings; the (b)-examples involve predicates with more absolute, minimum standard-like meanings. In each case, our intuition is that inference to the positive form is much more attractive in the (b)-examples than in the (a)-examples. For instance (14a) does not seem to implicate that many people came to the party; whereas (14b) seems to imply that more than ten did.

(12) a. This is not a very long stick. *attributive gradable adjectives*
b. This is not a very bent stick.

(13) a. John cannot stay at the party very long. *gradable adverbials*
b. John cannot come to the party very early.

(14) a. Not very many people came to the party. *quantificational determiners*
b. Not a lot/much more than ten people came to the party.

(15) a. I am not a huge fan of mayonnaise. *scalar nouns*
b. Digging here does not pose a huge risk.

(16) a. John is not very tall (for an American man). *positive/comparative gradable adj.*
b. John is not much taller than the average American man.

Again, these judgments are gradient and heavily dependent upon contextual factors. But even so, at least for the specific cases in (12)-(16), the sentences involving a predicate with a determinate minimum standard (*bent*, *early*, *more than ten*, *risk*, *taller than average*) seem to imply their positive form more strongly than the corresponding examples involving predicates with inherently variable thresholds (*long*, *many*, *fan*, *tall*).

Of particular relevance to the analysis of *not very* ADJ is example (16): abstracting away from syntactic details, (16a) and (16b) are both instances of an intensified gradable predicate under negation—[*not* [*very tall*]] and [*not* [*much taller than average*]], respectively. While both are demonstrably gradable, a clear difference between the two predicates is that the former is *vague* in a way that the latter is not: given knowledge of the average height and of John's height, one can know with relative certainty whether John is *taller than average*, but not whether he is *tall*. Put differently, both *very* and *much* introduce interpretive uncertainty, but *tall* does so in a way that

and the role of focus.

[9] Intuitions seem less robust in the restrictor of *every*, e.g. it is not obvious whether there is a difference in inferences to the positive form between sentences like *Everyone who was very late was sent to the principal's office* and *Everyone who is very tall will be considered for the basketball team*. We leave this as an open question for future research.

taller than average does not.

If, as Horn (1989) suggests, the reason why *tall* is not an attractive inference from *not very tall* has to do with euphemism, then one should expect the inference of *taller than average* from *not much taller than average* to be similarly unattractive—given the presumably similar evaluative profiles of *taller than average* and *tall*. Our judgments are that (16b) quite strongly implicates that John is taller than average, whereas (16a) does not. If correct, this would provide evidence in favor of an explanation in terms of scale-structure, and against one based purely on euphemism. In Experiment 2, we evaluate this claim directly by measuring judgments of agreement involving *tall* and *taller than average* in a range of different contexts (people's heights) and syntactic configurations.

2.5 Empirical predictions and motivation for experiments

We have argued that: (*i*) the inference from *not very* ADJ to ADJ—when it exists—is licensed by Manner-based Gricean reasoning. We additionally hypothesize that (*ii*): a set of cases where the inference is unattractive (e.g. with *tall*) can be explained in part by their *vagueness*: the strengthened meaning [*tall and not* [*very tall*]] has vague extension boundaries "on both sides," whereas (e.g.) the set of times counting as [*late and not* [*very late*]] is vague only on its upper boundary (due to the presence of *very*). We flesh out this idea in more detail in §4 below.

In the following sections we present two experiments designed to quantify the interpretations of vague and non-vague gradable predicates, including in the *not very* ADJ frame. In Experiment 1, we collect gradient judgments about predications involving the relative adjective *tall* and the absolute adjective *late*, across a range of scale points (e.g. if John is 6ft 2in would you agree that *John is (not) (very) tall*?). In Experiment 2, we compare judgments involving three relative gradable adjectives in their positive forms (*tall*, *hot*, *fast*) against those same adjectives in a comparative structure (*taller/hotter/faster than*), again across a range of relevant scale points.

The statement in (*i*) derives *late* as an implicature of *not very late* on the basis of structural alternatives. This directly predicts that *not late* is a strictly stronger alternative to *not very late*, and hence that the two forms should compete. In the context of Experiment 1, this means that there should be no regions of the lateness scale where *not late* and *not very late* are simultaneously deemed true.[10] Additionally, if the positive form inference is indeed generated as an implicature of *not very* ADJ, then we expect to see more uncertainty—in the form of greater variance—in judgments of truth within contexts that make the literal reading ("less than very late") true but the strengthened reading ("late but not a lot") false—since implicatures are themselves gradient and not always computed. Because the derivation relies on *not late* being strictly more informative than *not very late*, we also expect to find differences in the opposite direction on higher scale points (i.e., times which can count as *not very late*, but not as *not late*).

If we are correct that *not very tall* does not trigger an implicature to the positive form, then we don't expect to see its acceptability drop when *not tall* becomes maximally acceptable, nor do we expect greater variance in this region. A theory like Horn's (1989) takes *not tall* to be a conventionalized inference from *not very tall*, and thus (without further qualification) makes no clear predictions about differences across scale regions or adjective types.

Combined with our hypothesis (*ii*), which we develop in more detail in §4, our proposal also predicts that for vague relative adjectives, phrases of the form *not much* ADJ-*er than average* should

[10] Deemed true to some reasonably high degree, at least. By analogy, while logically speaking *all* entails *some*, situations where *all* is true are less likely to elicit judgments that *some* is true, due to its natural implicature to *not all*. This is the essence of the "competition" interpretation of alternatives for implicature.

give rise to an implicature in a way that *not very* ADJ should not. As discussed in §2.4 above, this contrasts with the "negative understatement" view. Without further qualification, this theory leads to no expectation about an inference to the positive form. It also leads to a natural expectation that *not very tall (compared to average)* and *not much taller than average* should have similar interpretations.

Another goal of Experiment 1 is simply to assess our own intuitions about the patterns, by measuring whether language users do indeed interpret *not very tall* differently from *not very late* in the relevant way. Experiment 2 broadens the empirical scope by extending the paradigm to a wider set of cases.

3 Experiment 1

3.1 Design

We estimated speakers' interpretations of gradable expressions by creating continua of ordered degrees on scales and then eliciting gradient judgments of (dis)agreement with statements about objects located at various scale positions (e.g. is a 6ft-tall man "tall"?).[11] We investigated *tall* and *late*, which for current purposes we took to be representative of the classes of relative standard gradable adjectives and minimum standard absolute gradable adjectives, respectively. Again, this distinction is crucial because relative adjectives' thresholds are underspecified in a way that absolute adjectives' are not. Because the two adjectives required different background contexts, adjective type was a between-subjects factor.

3.2 Methods

3.2.1 Participants

Experimental participants were recruited via Amazon Mechanical Turk in two survey versions. 35 participants took part in the *tall*-version, and 36 in the *late*-version (age range: 19–60). One participant in the *late*-version was excluded from the analyses because their self-reported native language was not English. We checked whether participants understood the task by looking at their responses to ADJ and *not* ADJ at extreme scale points, which should have been uncontroversial. Responses on the "wrong" side of the slider were considered errors (e.g., below 50% agreement to "John is tall" when John's height is maximal on the scale, here 6ft 10in). No participant performed worse than chance according to this measure, so no additional participants were excluded.

3.2.2 Materials and procedure

In each trial, a participant was presented with a fact concerning the height or arrival time of a different person (in the *late*-version, initial instructions provided global context by specifying that newly hired employees were expected to arrive at 9am for their first day at work). The fact was paired with a statement, uttered by a character named Mary, involving one of seven different adjective constructions. The participant's task was to indicate whether they agreed or disagreed with Mary by adjusting a slider whose position was encoded as a value ranging from 0% to 100%, where 100% is interpreted as complete agreement/acceptance, 0% as complete disagreement, etc.; see Figure 1 for sample displays.

[11] The paradigm employed here is quite similar to that of Hersh & Caramazza (1976), who studied a number of vague adjectives using comparable methods. We became aware of this research after all data had been collected.

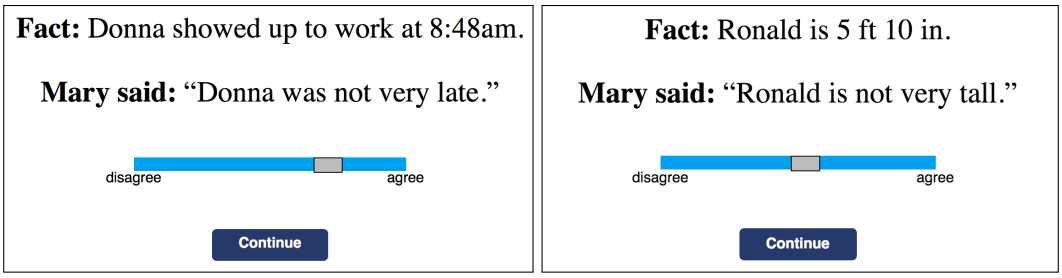

Figure 1: Sample displays for *late* (left panel) and *tall* (right panel)

In the *tall*-version, participants evaluated Mary's statements about men of different heights.[12] We tested the following seven constructions: "*X* is tall," "*X* is not tall," "*X* is very tall," "*X* is not very tall," "*X* is short," "*X* is not short," and "*X* is neither tall nor short," paired with 13 different heights ranging from 5ft 3in (160cm) to 6ft 10in (208cm).

In the *late*-version, Mary made statements about newly hired employees, who were all expected at 9:00am for their first day of work. Participants evaluated the following seven constructions: "*X* was late," "*X* was not late," "*X* was very late," "*X* was not very late," "*X* was early," "*X* was not early," and "*X* was on time," at 13 arrival times ranging from 8:39am to 9:48am.

Each participant saw all possible combinations of scale points and constructions twice (with a different name in each repetition), totaling in 182 trials per experiment version. The complete scales for *tall* and *late* are presented in Table 1. Trials were presented in a different random order for each participant.

tall	Relevant fact (not presented to participants): the average American adult male is 5ft 10in (178cm)												
scale point	5ft 3in	5ft 6in	5ft 8in	5ft 9in	5ft 10in	5ft 11in	6ft	6ft 1in	6ft 2in	6ft 3in	6ft 5in	6ft 7in	6ft 10in
deviation from 5ft 10in	−7	−4	−2	−1	0	1	2	3	4	5	7	9	12
late	From the instructions: "[new employees] are all expected to be in at 9:00am"												
scale point	8:39	8:48	8:54	8:57	9:00	9:02	9:05	9:08	9:14	9:21	9:27	9:36	9:48
deviation from 9:00	−21	−12	−6	−3	0	2	5	8	14	21	27	36	48

Table 1: The scales used in Experiment 1, with each point's distance from the average height (for *tall*) or the expected arrival time of 9am (for *late*). Only the latter was presented directly to participants in Experiment 1.

3.3 Results

Figure 2 shows judgments averaged over participants and items for all constructions, for the *late*-experiment (right panel) and the *tall*-experiment (left panel). The curves associated with *tall* and *late* can be seen as visualizations of scale structure: acceptance of the predicate *tall* follows an approximately sigmoidal shape, increasing as height increases for nearly the entire interval of heights. This reflects uncertainty about the exact value of the tallness threshold. For *late*, acceptance remains relatively constant and close to zero prior to 9am, and then shows a rapid increase as time moves beyond 9am, eventually leveling off near 100% acceptance. This reflects the virtual certainty that the threshold for lateness is located at 9am.[13]

[12] We used only unambiguous male names to keep the comparison class for tallness as homogeneous as possible.

[13] In fn3 we mentioned that although scalar endpoints serve as default thresholds for absolute adjectives, in reality thresholds vary marginally across contexts. And indeed we observe that 2min late is not judged as perfectly 'late', with mean agreement at 62% ($sd = 26$). However, note that this uncertainty only extends to times *after* 9am, and crucially not to times before 9am (with the reverse pattern for *early*). This suggests that the underlying mapping from arrival times to degrees of lateness treats all times earlier than 9am as $\mathbf{min}_{\mathbf{late}}$. In other words, even though there is no lower endpoint on the "physical scale" (time), the abstract degree scale for *late* is indeed lower-bound so that

Under the assumption that agreement reflects (degree of) truth,[14] entailment relations between expressions can be visualized by relative height on the coordinate system: if two curves in Fig. 2 have similar shapes but one has consistently lower agree-%, then it follows that whenever the lower one is true (to a certain degree), the higher one is true (to the same degree or higher). In other words, when a curve dominates another one across the scale, the lower one *entails* the higher one. For example *very tall* has a quite similar shape to *tall*, but *very tall* has consistently lower values than *tall* across all heights: whenever someone counts as *very tall*, they a fortiori count as *tall*. This reflects the fact that the former entails the latter. Similar remarks apply to *late* versus *very late*. In general, the results depicted in Fig. 2 are consistent with pre-theoretical expectations about the respective meanings of the constructions involved.

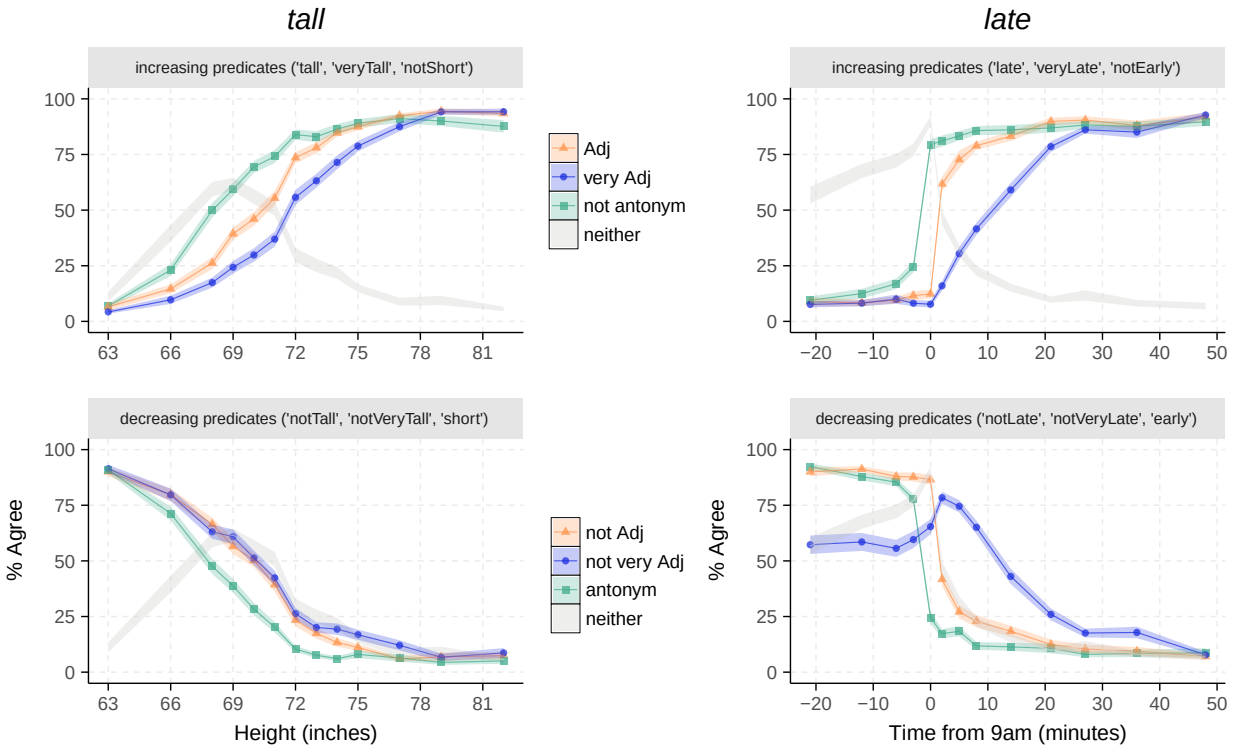

Figure 2: Mean "agreement %" by construction across degrees, for the predicates *tall/late*, *very tall/late*, *not short/early*, and *neither tall nor short/on time* (top), and the predicates *not tall/late*, *not very tall/late*, *short/early*, and again *neither tall nor short/on time* for reference (bottom). Left panel: *tall*. Right panel: *late*. Ribbon width is +/- one bootstrapped standard error of the mean at each scale point (here and for all subsequent figures).

The comparison of interest for determining the meaning of *not very* ADJ is between the mean

here, $\theta_{\textbf{late}} = \textbf{min}_{\textbf{late}} =$ 9am. This important but subtle point is discussed further in fn23. A related but somewhat independent observation is that the "acceptability curve" is not continuous at zero. This corresponds to Burnett's (2014) observation that absolute adjectives and their negations differ in tolerance: if someone is late, then we can always find a small enough difference such that a person who arrived that much earlier is also late; but if someone is not late, any small difference can lead to being late.

[14] This assumption does not hold if other factors affect acceptability. Two possible deviations comes to mind: ungrammaticality (which should not be an issue here, as all constructions involved are presumably well-formed), and pragmatic effects. An example of the latter can be seen with *not very late*, where we argue that a drop in acceptability reflects the violation of an implicature rather than semantic effects.

agree-% curve of *not very* ADJ and that of *not* ADJ. Figure 3 plots participants' mean judgments for only the critical constructions *not very* ADJ and *not* ADJ, across varying degrees of height or arrival time.

As can be seen from the plots, *not late* shows a sharp decline in acceptance in the region just beyond the threshold (9am), the very same region in which *not very late* shows the highest acceptance rate. Meanwhile, *not very late* is degraded in the region of the scale where *not late* is the most acceptable (times prior to 9am).

On the other hand, *not very tall* displays virtually the same response profile as *not tall*. While acceptance of the predicate *tall* increases with height (Fig. 2, top left), a roughly mirror-image pattern is observed for *not very tall* (Fig. 3, left). This approximates the expected relationship between an expression and its negation (cf. *tall*/*not tall* or *late*/*not late* in Fig. 2).

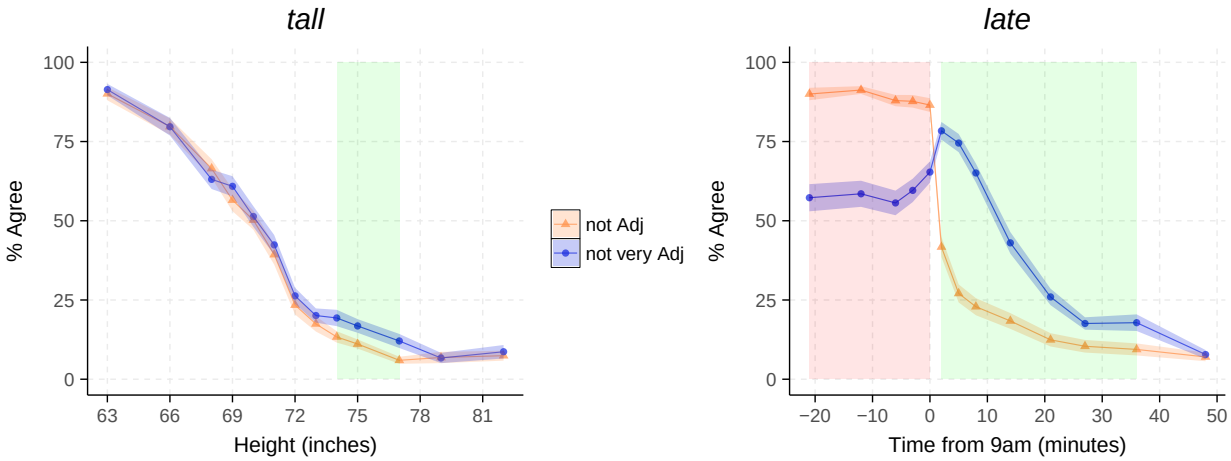

Figure 3: Mean % agree by construction across degrees. Left panel: *not late* (orange) and *not very late* (blue). Right panel: *not tall* (orange) and *not very tall* (blue). Significant clusters are indicated in green when *not very* ADJ is above *not* ADJ and in red otherwise.

To analyze the responses for *not very* ADJ and *not* ADJ statistically, we used a non-parametric cluster permutation test to identify significant differences in agreement-% along the scales. Direct comparison at each scale point would run into serious multiple comparisons problems, while applying a Bonferroni correction would be too conservative since responses to adjacent scale points tend to be highly correlated. We therefore adopted a more sophisticated method traditionally used in eye-tracking and EEG studies, treating the curves along our scales as a "signal."

We adapted the cluster analysis proposed by Maris & Oostenveld (2007), treating the degree scales as the continuous dimension (instead of time). The first step is to look for clusters of pointwise significant differences between two sets of curves (before correction for multiple comparisons) and to attribute to each cluster a size (typically, the sum of the relevant test statistic for each point in the cluster). In our case, we ran a mixed-effects model at each scale point with construction as a fixed effect and a random subject intercept, and took the t-value associated with the effect of construction as our representative statistic. We recorded clusters of t-values above 2 or below -2 and defined cluster size as the sum of absolute t-values (including "clusters" of only a single value). When no such values could be found, we simply recorded the maximum absolute t-value as the "largest cluster size" (this is useful for estimating p-values for non-significant clusters).

After extracting clusters of differences between the two sets of curves, the second step was to evaluate how likely we would be to find similar clusters by partitioning the combined set of

curves randomly rather than by construction. For this purpose, responses from each participant were grouped into two curves for each construction (corresponding respectively to the first and second repetition of that construction at each scale point). We then randomly re-assigned the labels *not very* ADJ and *not* ADJ to each curve, extracted clusters following the same procedure as with the actual data, and saved the size of the largest cluster. This sampling procedure was repeated 10,000 times in order to obtain a precise distribution for cluster sizes, from which a p-value could be computed for each of the clusters found in the actual data (see Ernst 2004 for an accessible introduction to permutation testing techniques).

This analysis yielded two significant clusters for "late": from 21min early to exactly 9:00am, *not very late* was significantly degraded compared to *not late* (mean agree 59% ($sd = 32$) versus 89% (16), respectively); and from 2min late to 36min late, we observed the opposite (46% (34) versus 20% (23); see Fig. 3, left panel). Both clusters were larger than anything obtained from the 10,000 random permutations; hence $p < .0001$.

For "tall," acceptability of the constructions *not very tall* and *not tall* coincided in most intervals. But the analysis revealed a small significant cluster ranging from 6ft 2in (188cm) to 6ft 5in (196cm), during which *not very tall* was more acceptable than *not tall* (16% (20) versus 10% (13); see Fig. 3, left panel), with $p = .0003$.

3.4 Discussion

Our results show that speakers clearly distinguish between the meanings of *not late* and *not very late*, and provide strong support for the hypothesis that the latter is interpreted with an implicature to the negation of the former. The signature of this implicature is visible in the results in three ways.

First, in the range of (strictly) late arrivals, participants treat *not very late* as more acceptable than *not late*, with the largest difference in acceptability—as well as the highest rate of agree-% for *not very late*—occurring in the region just after 9am. This shows that, in this range, *not very late* is strictly less informative than *not late*. This is a pre-condition for the existence of the implicature from *not very late* to *late* (see step (*ii*) of the derivation proposed in §2.3).

Second, for early arrivals, *not very late* is instead degraded relative to *not late*—crucially in spite of the fact that early arrival should be perfectly compatible with the literal meaning of *not very late* (given the clear unacceptability of *very late*). This is strong evidence for the inference in (1): the fact that *not very late* is degraded precisely when *not late* is most acceptable suggests that the two expressions are in competition, as expected if *late* is an implicature of *not very late*.

Third, note that mean agreement rating for *not very late* is only at 58% prior to 9am—the region in which the literal reading is true but the strengthened reading is false. Variance is also higher for *not very late* than for all other constructions in this region ($sd = 33$, compared to the average *sd* of 18% for all others; difference visible from relative width of error ribbons). This pattern demonstrates uncertainty about whether being early counts as 'not very late', which is expected since the literal meaning is compatible with being early but the strengthened meaning is not. Again, this provides evidence for the *late* implicature of *not very late*.

A different picture emerges from the results of the "tall"-predicates. The constructions *not tall* and *not very tall* are treated almost completely alike by participants, with no indication of an implicature *tall* from *not very tall*. The close similarity of the curves for *not tall* and *not very tall* confirms our intuition that the implicature to the positive form is not drawn from *not very tall* in the way it is drawn from *not very late*. Unlike the results for *late*, we found no evidence for competition between these two expressions—there is no interval during which acceptance of one spikes while

acceptance of the other drops—and thus no evidence that an implicature would be derived from an assertion of one expression via reasoning about the non-asserted one. Moreover, the fact that the meaning of *not very tall* is almost identical to the meaning of *not tall* provides evidence not only for the lack of an implicature, but also for a meaning stronger than the mere negation of *very tall*.

Going back to the two puzzles outlined in §1, our results provide strong evidence that (*i*) unlike *not very late*, *not very tall* does not give rise to the positive inference 'tall'; and (*ii*) *not very tall* seems to be further "strengthened" to have a meaning very close to *not tall*, even though the positive forms *tall* and *very tall* were clearly distinguished.

Two final points relating to this aspect of the results. First, while Horn's (1989) proposed interpretation for *not very tall* ("rather short") was not precise enough for a predictive theory, our data precisely characterizes the relationship between the interpretations of *not very tall*, *not tall*, and *short*: as can be seen from visual inspection of Fig. 2, we found that *short* is consistently interpreted as stronger than *not very tall*, with mean agreement of 26% for the former versus 38% for the latter (collapsed over heights).

Second, we found that the response profiles for *not tall* and *not very tall* coincide for nearly the entire range of heights evaluated in the present study. Nonetheless, there is one exception: in the interval between 6ft 2in and 6ft 5in, *not very tall* is judged slightly more acceptable than *not tall* (16% versus 10% mean agree, respectively). In some sense this is unsurprising, as these heights intuitively correspond to someone who would be tall, but less than very tall. Albeit small in magnitude and range, this difference could be an indication that *not very tall* is still marginally weaker than *not tall*—and hence that the "strengthening" of *not very tall* to *not tall* is more of a gradient phenomenon than has been assumed in (limited) past discussions of the *not very* ADJ construction.

An anonymous reviewer pointed out that the experiment was very long, and that judgments on vague items may be affected by this (as participants may accommodate a metalinguistic QUD, such as "does 6ft 1in count as *tall*?"). If this had an effect, it would most likely be stronger for later trials, so we decided to test order effects. We observed very limited order effects, and crucially, they did not affect judgments on vague items.[15] In Experiment 2, order effects were completely obviated by showing each combination of construction and scale point to each participant only once.

In the next section, we propose a theoretical account of our results, which will be supported by a post-hoc analysis of the results of this first experiment. We follow with a second experiment which further tests the proposal and addresses some shortcomings of the first experiment.

[15] Specifically: for each adjective, we fit a model at each scale point with construction (7 levels, sum-coded), order (2 levels corresponding to first and second occurrence of a given construction at a given scale-point, sum-coded) and their interaction as fixed effects. We included random slopes for construction and order but not for their interaction (this would have saturated the model). The random effect structure was then simplified to avoid over-fitting using the procedure proposed in Bates et al. (2015). We observed no significant effect of trial order in any of the constructions from the 'tall' version (highest $\chi^2(7) = 13, p = .07$), but we found an effect in the 'late' version at 8:54am ($\chi^2(7) = 21, p = .004$, $p = .048$ after Bonferroni correction for 13 comparisons). Further inspection with a model in which construction was treatment-coded showed that *early*, *late*, *not early*, *very late*, and *not very late* were unaffected by order (all $t < 1$), but *on time* was judged slightly higher on its second occurrence ($t = 1.6$) while *not late* was slightly degraded ($t = -1.4$). This effect is therefore unrelated to vagueness (*not late* is not vague), and is most likely due to competition between these two expressions (both are true of someone who arrived just a few minute before 9:00am, but *on time* is more specific than *not late*).

4 A proposed model of vagueness-implicature interactions

In the results of Experiment 1, we found that a sentence like (17a) is interpreted with an inference to the negation of (17b). By contrast, a sentence like (18a) is interpreted as approximately synonymous with (18b)—with the caveat discussed in the previous section.

(17) a. John was not very late.
b. John was not late.

(18) a. John is not very tall.
b. John is not tall.

In §2.3 we argued that the first case can be explained as a standard structural (Manner) implicature triggered by competition between (17a) and (17b), and our experimental results confirmed this view. We have also seen that euphemism may partially explain how it is that (18a) can be used to communicate something that approximates (18b), instead of simply "less than very tall." But the explanation is inherently incomplete: euphemism cannot not explain why (minimum standard) absolute adjectives do not follow the same pattern, nor does it explain why reasoning about euphemism dictates meaning *instead of* the kind of Gricean reasoning that explains the interpretation of (17a). The remaining puzzle, then, essentially boils down to the question of why (18a) does not give rise to a *tall* implicature parallel to the *late* implicature of (17a).

Since the difference between adjectives that do implicate the positive form in *not very* ADJ and those that don't systematically aligns with the distinction between minimum standard and relative standard gradable adjectives, we believe that the explanation must be sought in the interaction between vagueness—the semantic dimension along which the two adjective classes vary—and the derivation of structural implicatures—the pragmatic mechanism that accounts for the positive form inference from *not very* ADJ, when present.

4.1 Deriving the vague/non-vague asymmetry

We propose that the asymmetry in implicature to the positive form between *not very tall* and *not very late* follows directly from basic facts about relative versus absolute scale structure, combined with a novel constraint on the derivation of implicatures.

Ripley (2011) defines a borderline contradiction as a conjunction of the form $(Fa \wedge \neg Fa)$, where F is a vague predicate and a is a borderline case for F (meaning it is not clear whether a is F or non-F). Such constructions have been shown to be more acceptable than ordinary contradictions, but are still clearly degraded (Alxatib & Pelletier 2011; Serchuk et al. 2011; Egré et al. 2013; Zehr 2014).

We propose to generalize the notion of a borderline contradiction in two different directions. First, we want to say that the sentence *John is tall and not tall* is a borderline contradiction *independently of John's height*. The reason that such a sentence is deemed a borderline contradiction is that in any context, it is either "clearly" false (since if one conjunct is true, then the other and hence the sentence is false), or else has a "borderline" status (when John's height is deemed sufficiently close to θ_{tall}). In short: no matter what height John is, *John is tall and not tall* will either be false or a borderline case. This suggests that it can be useful to think of "borderline contradictoriness" as a property of sentences, instead of (or in addition to) sentence-context pairs, as in Ripley's (2011) original formulation.

An advantage of this conception is that it enables us to ask whether a conjunction $Fa \wedge \neg Ga$ is a borderline contradiction, even when $F \neq G$. And this is the second way in which we propose to refine the notion: borderline contradiction can arise in cases of conjunction of two distinct predicates.

As a motivating example, if *The dog is big and not big* is a borderline contradiction, then so too should be *The dog is big and not large*.

These two desiderata can be captured by a formulation that counts a sentence as a borderline contradiction if it simply cannot ever be clearly true—this is general enough to capture cases with distinct predicates, and also counts ordinary contradictions as extreme cases. To define the notion more precisely, we adopt a trivalent system of truth-values where 1 is the value of a clearly true sentence, 0 is the value of a clearly false sentence, and a third value—often written '#'—represents borderline cases (sometimes called "indeterminacy"). We can then give the definition in (19) to capture our intended sense of "borderline contradiction."

(19) **Definition** A sentence S is a *borderline contradiction* if in every world w, $[\![S]\!]^w \neq 1$ (i.e., it's either false or borderline).

Under Ripley (2011), Tye (1994), and others' assumptions regarding the projection of the third value—that a conjunction is # when either (or both) of its conjuncts is—we correctly count *John is tall and not tall* as a borderline contradiction. If F and G are gradable predicates on the same scale, and if the parts of the scale that make F and G true do not overlap, then any conjunction of F and G applied to the same element is a borderline contradiction. This includes—but is not limited to—cases where G is the negation of F, as in *John is tall and not tall*.

A few comments are in order. First, phrasing this definition in a trivalent system for vagueness sweeps issues related to *second-order vagueness* under the rug (Dummett 1975). In practice, there is no definite threshold between truth and "borderline-ness," so the notion of a borderline case itself is vague (e.g. where is the cutoff point between being tall and a borderline case for tall?). This makes it difficult to operationalize the definition in (19).

Second, our definition is very general and covers ordinary contradictions as extreme cases of borderline contradictions.[16] This allows us to build upon previous accounts of implicatures with minimal change. As an example, Fox (2007) proposes a contradiction-free mechanism to derive implicatures based on the notion of innocent exlusion. We thus propose the simpler constraint in (20), which could be more explicitly formalized in terms of Fox 2007 or other alternatives-based theories of implicature.

(20) **Constraint on vague implicatures:** If a sentence S and its alternative S' are such that $S \wedge \neg S'$ would be a borderline contradiction, then the potential implicature resulting from the negation of S' is not derived.

The intuition behind (20) is that borderline contradictions are generally infelicitous and therefore if there is an option to not interpret a sentence in such a way that its implicature-strengthened meaning is a borderline contradiction, then such a strategy should be preferred.

Our explanation of Experiment 1's results in terms of (19) and (20) runs as follows. Consider first the case of *not very late*. The meaning conveyed by the assertion with its implicature can be paraphrased as "late but not very late." We predict that the implicature is derived only if there are arrival times which clearly satisfy both predicates *late* and *not very late*. Because *late* is minimum standard, any time shortly after 9am counts as clearly late, and by getting sufficiently close to 9am, there will be some times which also count as clearly not very late. Since there can be times which clearly count as both late and not very late, there is no borderline contradiction, and hence

[16] Though is not intended to apply to other uses of '#', such as presupposition failure (see e.g. Spector 2016).

Manner-based reasoning can apply and the implicature can be drawn.

By contrast, the proposal could block the derivation of the implicature for *not very tall* in the following way. For the implicature to be derived, we would need to find individuals which are simultaneously clearly tall and clearly less than very tall. However, in this case both predicates are vague. We argue that in general there are simply no heights which clearly satisfy both predicates together, because *very* does not increase the threshold enough to make *not very tall* clearly compatible with *tall*. Thus a borderline contradiction arises and the implicature is blocked.

In the following subsection, we provide empirical support for the mechanism proposed here, by showing that strengthening *not very tall* with *tall* does indeed lead to a meaning that is more contradictory than the corresponding strengthened meaning of *not very late*.

4.2 Evidence from further data analysis

Our theory of *not very* ADJ relies upon the crucial assumption that the conjunction of "tall" with "not very tall" is a borderline contradiction in a way that the conjunction of "late" with "not very late" is not. However, we have not justified this claim beyond intuition thus far. We will now do exactly this. Establishing the claim essentially reduces to showing that the phrase *tall but not very tall* is significantly less acceptable ("more contradictory") than *late but not very late*, which has a determinate lower-bound and hence should not give rise to borderline contradictions.

The most direct way to evaluate this assumption would be to elicit judgments about the acceptability of ADJ *and not very* ADJ. This strategy, however, is clearly problematic: the phrase *late and not very late*, for instance, seems pragmatically deviant, presumably because of lexical competition between *and* and *but* (cf. also *some but/and not all*). One could then consider using *but* to conjoin ADJ with *not very* ADJ, since *but* has a similar (enough) truth-functional meaning to *and* but is a more natural connective in this context. However, doing so would be equally problematic for the following reason: using *but* to conjoin ADJ and *not very* ADJ introduces an implication[17] that the two conjuncts "contrast" with one another, and thus—since their main predicates are identical—would attract focus to the modifier *very*. As we noted earlier, focal stress on *very* has the effect of forcing an inference to the positive form. Therefore, using *and* could make the conjunction of ADJ with *not very* ADJ infelicitous for independent pragmatic reasons, while using *but* would affect focus structure and hence provide judgments about a meaning that is probably distinct from the logical conjunction of the relevant predicates (the problem is even worse if one accepts Bach's (1999) and Potts's (2005) arguments that *and* and *but* are not even truth-conditionally equivalent in the first place).

Instead of collecting explicit judgments, we employed an indirect strategy that allowed us to estimate the interpretation of the abstract construction "ADJ $\wedge \neg$ *very* ADJ," while avoiding the two confounds of focus-attraction and pragmatic competition between *and* and *but*. In the following, we describe a post-hoc analysis on the data from Experiment 1 establishing that *tall*$\wedge\neg$*very tall* does indeed have a degraded interpretation when compared to *late* $\wedge\neg$*very late*, as hypothesized in §4.1 above.

By combining data about ADJ with data about *very* ADJ, we were able to "reconstruct" estimated

[17] There is no current consensus among researchers about the exact nature of the contrastive implication from *but*. Traditionally it was viewed as a "conventional implicature" (Grice 1975). But influential arguments by Bach (1999) and Potts (2005) conclude that the contrastive inference from *but* is actually a kind of "ancillary entailment," so that the truth-conditional contributions of *but* and *and* are not even identical. See also Karttunen (2016) and Blakemore & Carston (2005).

agree-%'s for complex expressions of the form ADJ $\wedge \neg$ *very* ADJ, and assess whether they behave like borderline contradictions. We used definitions of negation and conjunction from fuzzy logic (Zadeh 1965), a standard logical foundation for the analysis of vague language. Where $v(A) \in [0,1]$ represents the fuzzy truth-value of proposition A:

(21) **Definitions of (fuzzy) logical operators**

a. $v(\neg A) \quad =_{\text{def}} \; 1 - v(A)$ (negation of proposition A)

b. $v(A \wedge B) \; =_{\text{def}} \; \min(v(A), v(B))$ (conjunction of propositions A, B)

In other words, if a participant's mean rating for *Bill is tall* when Bill is 6ft is 70%, then we can infer their mean rating of $\neg$(*Bill is tall*) to be around $1 - 70\% = 30\%$. And indeed, this "artificial" meaning almost completely coincides with participants' actual judgments about the sentence *Bill is not tall*. The conjunction $A \wedge B$ is defined as the proposition with value equal to the minimum of $v(A)$ and $v(B)$.[18]

Using these definitions, we can estimate interpretations of "pseudo-predicates" of the form ADJ $\wedge \neg$ *very* ADJ for each participant and scale point (e.g. someone's judgment about the degree to which a 6ft-tall man counts as 'tall and not very tall') as follows: we first take the actual judgments for ADJ and for *very* ADJ, scaling them to the interval $[0,1]$ by dividing each by 100. We then take, for each participant and each height/time, whichever of the following two quantities is smaller: the mean value of their judgments about ADJ, or 1 minus the mean value of their judgments about *very* ADJ.[19] This allows us to compare estimated meanings of logically complex predicates in order to evaluate our assumption about the role of borderline contradiction in the present phenomenon. We also calculated estimates for 'ADJ $\wedge$ *not* ADJ' (using the actual judgments for *not* ADJ) as a reference point, since they are paradigm cases of borderline contradictions.

Estimated mean agree-%'s for each of the four reconstructed predicates are plotted in Fig. 4. Visual inspection suggests that *tall* $\wedge$ $\neg$*very tall* is clearly degraded compared to *late* $\wedge$ $\neg$*very late*, and is qualitatively closer to explicit borderline contradictions. To assess this quantitatively, we looked at the highest value a reconstructed predicate took across the whole scale.[20] We fit a linear mixed-effects model on the peak values of the constructions 'ADJ $\wedge$ $\neg$ADJ' and 'ADJ $\wedge\neg$*very* ADJ' with construction, adjective (*tall* or *late*) and their interaction as fixed effects, and a

[18] While the operation for negation is uncontroversial, several definitions for conjunction are possible in fuzzy logic. The min operator (also called Zadeh-conjunction or Gödel t-norm) is the most common, especially in the literature on vagueness, but one might imagine that there could be better options. We validated the choice of min by comparing ratings for *neither tall nor short* to 'min$(1 - \text{tall}, 1 - \text{short})$'. Not only was it a very good approximation, but if anything it slightly *under*estimated the actual agreement with *neither tall nor short*. Since the min operator is the pointwise largest t-norm, no other choice of a conjunction function would have yielded better results (in particular not the product of $v(A)$ and $v(B)$, nor the Łukasiewicz t-norm, which are salient alternative options). For relevant technical background on triangular-norms ("t-norms")—particular kinds of binary operations on the unit interval $[0,1]$—see especially §2 of Klement et al. (2004).

[19] Crucially, we used the judgments for 1−*very* ADJ and not those for *not very* ADJ—the data we aim to explain—as the latter strategy would have been circular.

[20] A borderline contradiction as defined in §4.1 should maintain low acceptability across the scale, whereas a non-contradictory proposition should have at least some range in which it has a reasonably high acceptability rate. Note that a perfectly acceptable predicate with a very narrow meaning would have a very small range of high acceptability ratings and close-to-zero ratings elsewhere, while a borderline contradictions is expected never to reach high agreement, but should in principle be rated higher than a clear contradiction when averaged across the entire scale. This shows that peak value is a more appropriate measure of borderline contradictoriness than area under the curve, which would not necessarily distinguish between these two hypothetical cases.

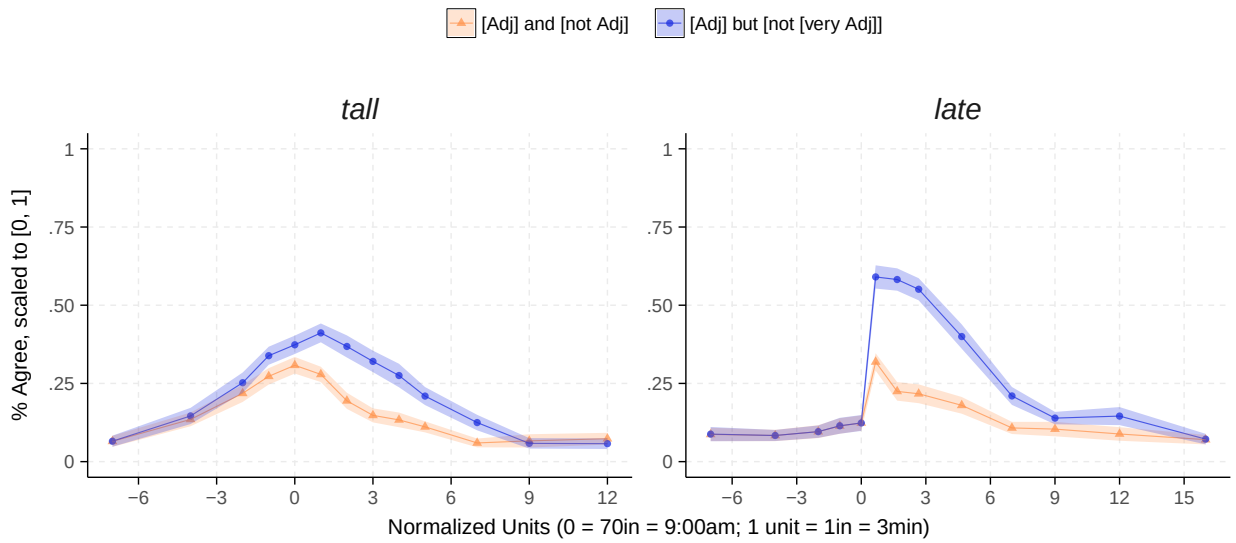

Figure 4: Estimated interpretations of complex expressions which would be problematic to elicit explicit judgments about. ADJ *and not* ADJ is defined as $\min(\text{ADJ}, \textit{not}\ \text{ADJ})$, ADJ *and not very* ADJ as $\min(\text{ADJ}, 1 - \textit{very}\ \text{ADJ})$. The units are normalized so that 0 is 9am for *late*-predicates, and 5ft 10in for *tall*-predicates (the average adult American male height).

random intercept for participant (with only two data points per participant, we could not include a slope for construction). This showed a significant interaction between adjective and construction ($\chi^2(1) = 9.6, p = .002$), establishing that *late* $\wedge$ $\neg$*very late* was significantly more acceptable than *tall* $\wedge$ $\neg$*very tall*.

Furthermore, while 17 of the participants agreed that the position of the peak for *late* $\wedge$ $\neg$*very late* was 9:02am, no more than 8 participants agreed on a single peak value for *tall* $\wedge$ $\neg$*very tall* (most frequent was 5ft 11in, 180cm).

In sum, the interpretive difference between *not very tall* and *not very late* follows from the fact that deriving the inference from *not very tall* would result in too much uncertainty and lack of consensus among speakers, whereas the absolute nature of *late* ensures agreement on at least a part of the resulting strengthened interpretation: few (if any) heights seem to count as 'tall but not very tall', and there is variation in what exactly those heights are; but it is clear across speakers that 2min late qualifies as 'late but not very late'. In §5.4 we show that the response data from Experiment 2 yield parallel patterns across the board when subjected to this same reconstructed predicate analysis.

Finally, we should note that while the proposal above explains why inference to the positive form is not derived for relative adjectives, it does not fully explain why the interpretation of *not very tall* is not fully equivalent to the mere negation of *very tall*. This missing piece of the puzzle could potentially be where evaluativity and understatement become relevant (Krifka 2007; Horn 1989), though assessing this possibility would require investigation of a much larger set of lexical items.

5 Experiment 2

5.1 Goals

Experiment 1 showed that the adjectives *tall* and *late* give rise to different inferences in the *not very* ADJ construction. We have argued that the difference between the two is driven by the underlying scale structure and vagueness associated with these adjectives. In this second experiment, we address

some limitations of Experiment 1 and confirm the central role of vagueness in the asymmetry.

Besides vagueness, there are a number of independent differences between *tall* and *late*, and thus it is possible that the judgment patterns observed in Experiment 1 were idiosyncratic to the adjectives we tested. In order to establish a more general role of vagueness in implicature derivation, we conducted a second experiment with *tall* and two additional relative adjectives: *fast* and *hot*. Instead of comparing relative and minimum standard adjectives, we tested the same relative adjectives in the positive form and in a comparative construction (*taller than the average American man*, *faster than the average mid-sized sedan*, and *hotter than the average summer day in Citytown*).[21] This enabled a minimal comparison between a vague and non-vague condition using the same measurement scale. That is, while the positive form of relative adjectives involves borderline cases, there is a fixed minimum reference point in the comparative construction. For example, the comparative *taller than the average American man* is true of any individual whose height exceeds the degree denoted by the average (assuming the average is known). This is analogous to the fact that any degree exceeding the threshold counts as ADJ in the case of minimum standard adjectives.[22] By using comparatives, we were able to test a continuous scale with scale points below the threshold, which is the region where the implicature can be observed (evidenced by the curve for *not very* ADJ dropping to a low agreement when that of *not* ADJ reaches full acceptability).[23] Comparing positive and comparative forms of an adjectives also obviated any remaining issue regarding evaluativity and euphemism, as discussed with example (16) in §2.4.

In Experiment 2, we provided an explicit reference in all conditions (as opposed to Experiment 1, where a 9:00am threshold was supplied for *late*, but none was supplied for *tall*). Specifically, in each trial the average measure of ADJ was provided as background information (relative to a specific scenario, which was held constant within each adjective). Moreover, we provided an explicit comparison class in the critical sentence. For comparatives, the average provided in the context sentence was used in a *than*-phrase; for adjectives in the positive form, we used *for*-phrases (e.g., *tall for an American man*). Crucially, relative adjectives are gradable in both their positive and comparative forms, allowing intensification in all constructions (the main difference is that the natural intensifier for comparatives is not *very* but *much*). By analogy to the previous experiment where we compared *not (very) tall* to *not (very) late*, here we are comparing '*not (very)* ADJ *for an X*' to '*not (much)* ADJ*-er than the average X*'.

To recap: we pursued three main goals in Experiment 2: (*i*) replicating the finding for *tall* in a context that provides a precise and explicit reference point and a comparison class; (*ii*) establishing

[21] We are grateful to Stephanie Solt for suggesting this approach.

[22] See especially Kennedy & McNally (2005b) for a unified analysis of comparative and minimum standard adjectives.

[23] This worked well with *late* since its "physical scale" can extend below the threshold of its corresponding "conceptual scale:" the ordered set of degrees on which θ_{late} lies is lower-closed, mapping all "early" times to a single value on the abstract "lateness" scale (usually $\mathbf{min}_{\text{late}} = \theta_{\text{late}}$). But lateness can also be measured on the fully-open time scale, on which arbitrary points can be constructed (even if times before $\mathbf{min}_{\text{late}}$ are not distinguished in the semantics). It is difficult to find other minimum standard adjectives which behave the same in this respect. Our impression is that minimum standard adjectives usually have lower-closed physical scales as well. As an example, *rainy* and *snowy* have lower-closed conceptual/degree scales (*slightly snowy/rainy*), and they also have clearly lower-closed physical scales—the idea of negative precipitation feels nonsensical. Because of this, a potential implicature from *not very rainy* to *rainy* would hardly be visible if measured on the precipitation scale (which would probably be the only choice for *rainy* within the present paradigm). Similar considerations contribute to the difficulty of studying maximum standard adjectives in this way. For relevant and useful discussion of measurement and different notions of "scale" in degree semantics, see Sassoon 2010, Solt & Gotzner 2012, and Lassiter 2011:Ch2.

the same pattern for two additional relative adjectives; and (*iii*) establishing that the vague/non-vague distinction plays a more general role in implicature derivation (rather than being a particular property of positive-form adjectives) by looking at comparatives.

5.2 Design

Experiment 2 used the same task and continua of degrees as Experiment 1. We simplified the design by dropping antonyms and constructions such as *neither tall nor short*. Participants judged four different constructions for a single adjective in both its positive (ADJ, *not* ADJ, *very* ADJ, *not very* ADJ) and comparative (ADJ-*er*, *not* ADJ-*er*, *much* ADJ-*er*, *not much* ADJ-*er*) forms. We created three analogous experiment versions for the adjectives *tall*, *fast* and *hot*. These manipulations form a 3 (adjectives, between-subjects) $\times$ 2 (positive or comparative form, within-subject) $\times$ 4 (constructions, within-subject) design.

5.3 Methods

5.3.1 Participants

Participants were recruited via Amazon Mechanical Turk in three survey versions. In total, 135 participants took part in the Experiment (45 in each survey version, age range: 21–64). Two participants in the *tall*-version were excluded from analyses because they performed below chance on the performance measure defined in Experiment 1 (based on answers to extreme scale points).

5.3.2 Materials and procedure

In each trial, participants were presented with a single display consisting of: a context sentence providing an explicit reference point, which was constant throughout the experiment (e.g., the average height of American men); a fact placing an individual at one of 13 possible scale points; and a statement involving this individual as the subject of a particular construction, in reference to a comparison class. For example, in the *tall*-experiment, participants judged the statement *John is tall for an American male* after reading that John was 6ft 2in. The task was the same as in Experiment 1: participants indicated whether they agreed with the statement on a continuous scale.

The eight sentences resulting from combining two forms (positive versus comparative) and four constructions followed the patterns: "*x* is ADJ/very ADJ/not ADJ/not very ADJ for an *X*" and "*x* is ADJ-er/much ADJ-er/not ADJ-er/not much ADJ-er than the average *X*." Examples with the positive form of *hot* and the comparative form of *fast* are provided in Figure 5.

In the case of *tall*, the comparison class and scale points were identical to Experiment 1 (American males, ranging from 5ft 3in to 6ft 10in). In the *fast*-version, statements were about cars (mid-sized sedans), the top speed of which ranged from 85mph to 161mph (average top speed: 113mph). In the *hot*-version, utterances were about summer days in the made-up city of Citytown, which could range from 66°F to 113°F (average high temperature: 83°F). Each participant saw all possible combinations of scale points and constructions once for a single adjective, totaling in 104 trials per experiment version (to reduce the experiment's duration, we did not present each combination twice as in Experiment 1). Trials were presented in a different random order for each participant.

5.4 Results

The results for the ADJ and *very* ADJ constructions for all three adjectives in both positive and comparative forms are presented in Figure 6. Cursory visual inspection confirms that unlike the

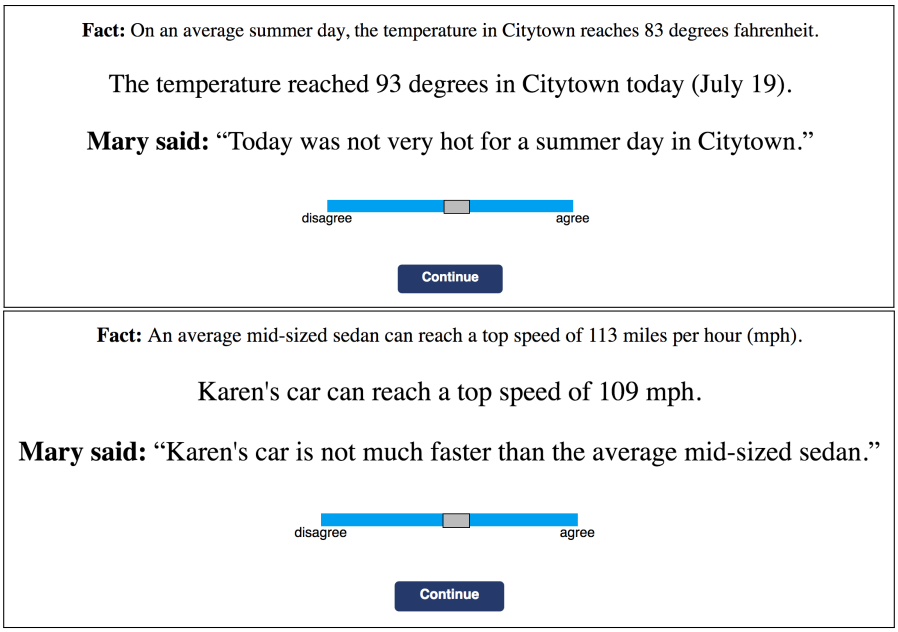

Figure 5: Sample displays from Experiment 2: positive *hot* item (top), comparative *fast* item (bottom).

positive forms of relative gradable adjectives, the comparative forms have clear thresholds.

Figure 7 presents the results for the critical *not* ADJ and *not very* ADJ constructions. As predicted, we observe that the vague positive forms show no trace of implicature, while the non-vague comparative forms show the same two effects as 'late' in Experiment 1: *not much* ADJ-*er* does not drop as fast as *not* ADJ-*er* after the threshold, but is degraded in the region just prior to it.

We submitted the results shown in Figure 7 to a cluster analysis similar to that of Experiment 1. The only difference resided in the way we obtained t-values: with only one item per participant and condition, we replaced the mixed-effects models with simple linear models. The analysis confirmed the first observation: none of the positive forms showed any significant clusters (*fast*: $p = .88$; *hot*: $p = .11$; *tall*: $p = .43$), while all comparative forms had a cluster of negative t-values before the threshold (*faster*: 85-109 mph; *hotter*: 66-83°F; *taller*: 5ft 3in-5ft 10in; all $p < .0001$) and a cluster of positive t-values after the threshold (*faster*: 117-129 mph; *hotter*: 85–90°F; *taller*: 5ft 11in-6ft 3in; all $p < .0001$).

We followed up with the analysis proposed in §4.2, indirectly measuring the contradictoriness of the potential implicature for each construction. The predicates ADJ *and not* ADJ and ADJ*(er) and not very/much* ADJ*(er)* were reconstructed following the same procedure as in §4.2 (see Figure 8). Peak values for each construction and each form (positive or comparative) were extracted for each participant (resulting in 4 data points per participant). We then fit a linear mixed-effects model with construction, form, adjective (sum-coded), and all their interactions as fixed effects, with random intercepts and slopes for construction and form, by-participant. The resulting random effects structure was simplified following the procedure of Bates et al. (2015) to avoid over-fitting. We observed an interaction between form and construction ($\chi^2(1) = 15, p = .0001$), which indicates that ADJ-*er and not much* ADJ-*er* was less contradictory than ADJ *and not very* ADJ (w.r.t. the baseline construction ADJ*(er) and not* ADJ*(er)*). We also observed a three-way interaction ($\chi^2(4) = 11, p = .025$), indicating that there was some variability between the three adjectives. Note however that all positive forms were clearly more contradictory than any of the comparative

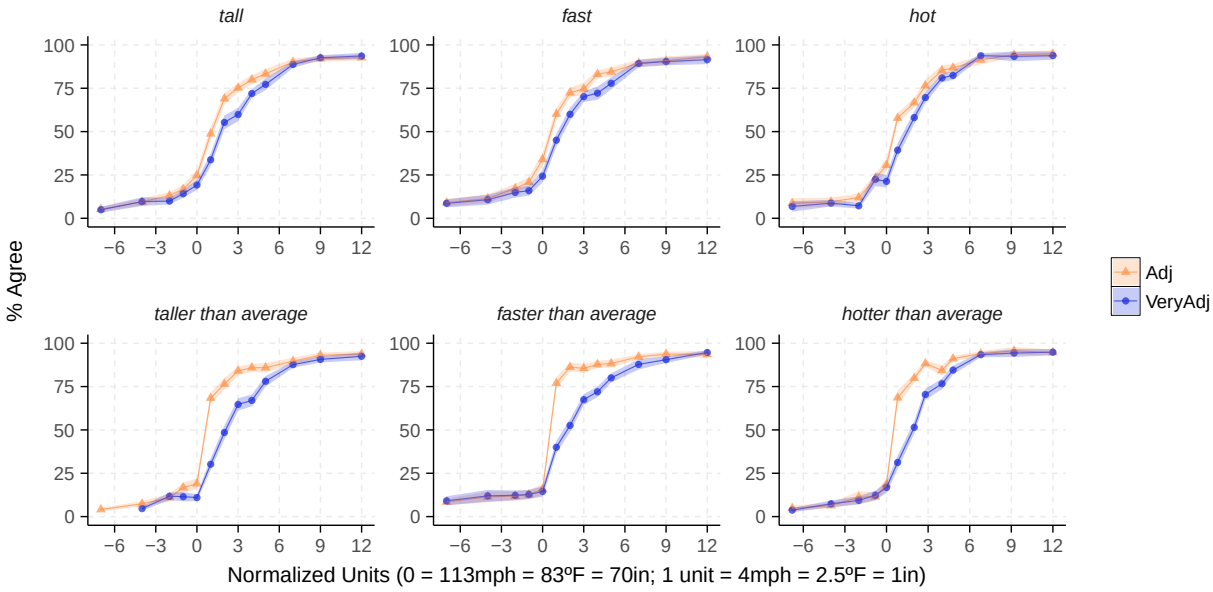

Figure 6: Mean % agree by construction across degrees. Top: *tall for an American man*, *fast for a mid-sized sedan*, and *hot for a summer day in Citytown*, with and without *very*. Bottom: *taller than the average American man*, *faster than the average mid-sized sedan*, and *hotter than the average summer day in Citytown*, with and without the intensifier *much* (in place of *very* for syntactic reasons).

forms.

As in the previous case, there was less agreement between participants regarding the peak for the vague ADJ *and not very* ADJ constructions (at most 13, 17, and 13 participants agreeing for *fast*, *hot*, and *tall* respectively) than with the non-vague ADJ-*er and not much* ADJ-*er* constructions (up to 20, 22, and 20 participants agreeing for *faster*, *hotter*, and *taller* respectively).

5.5 Discussion

In this experiment, we replicated and extended the results of the first experiment. First, we replicated the results with *tall*, and obtained very similar results with two other (vague) relative gradable adjectives, *fast* and *hot*. We showed that even when providing the comparison class explicitly with a *for*-phrase and mentioning an explicit reference point (the average value in the comparison class), these predicates remained vague and still showed no trace of an implicature from *not very* ADJ to ADJ. Second, we generalized the results with 'late' to other non-vague constructions. Using the comparative form of relative adjectives, we showed that the implicature resurfaced as soon as vagueness disappeared. Crucially, the use of comparative forms allowed us to address a number of potential worries regarding Experiment 1. In particular, the same participants could be tested on the positive and comparative form of a given adjective, and the scale points were exactly the same.

Since the positive and comparative forms of a given adjective are likely to share the same evaluative content, any proposal which would explain the effect as purely euphemistic cannot account for the results of this experiment. It is indeed unclear how such a theory would explain the difference between *John is not very tall for an American man* and *John is not much taller than the average American man*.

We also replicated the results of the post-hoc analysis we ran on the data from Experiment 1: we confirmed that the missing implicatures from positive forms (ADJ *and not very* ADJ) would

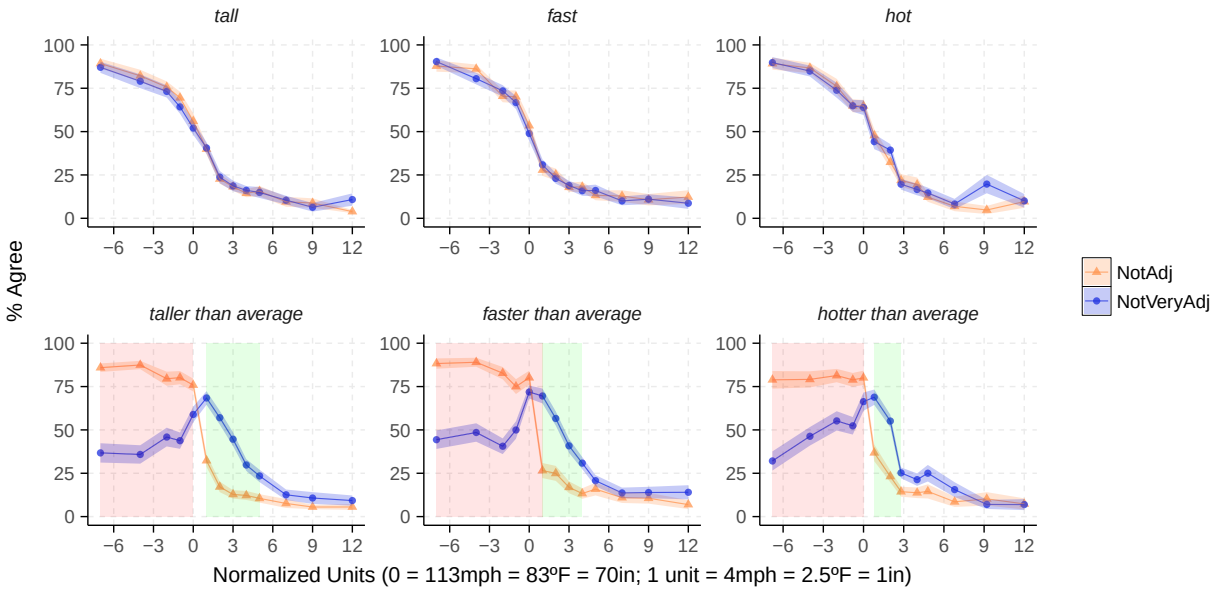

Figure 7: Mean % agree by construction across degrees for the target *not very* ADJ*/not much* ADJ*-er*, compared to its alternative *not* ADJ*/not* ADJ*-er*. Significant clusters of positive differences are indicated in green (*not very* ADJ above *not* ADJ); clusters of negative differences in red.

have been more contradictory than the attested implicatures from comparative forms (ADJ*-er and not much* ADJ*-er*).

6 General Discussion

Experiments 1 and 2 established that a predicate's interpretation in the '*not very* ADJ' construction depends upon its scale structure: non-vague predicates are associated with implicatures to the positive form (e.g. *not very late* $\rightsquigarrow$ *late*), whereas no such inference is drawn for vague predicates (e.g. *not very tall* $\not\rightsquigarrow$ *tall*).

Of course, the English language has a large number of gradable expressions with varying idiosyncratic properties. Future studies involving larger sets of predicates will clarify the underlying mechanisms, and probably reveal subtle patterns and differences not detected here. A few additional factors that affect interpretation were identified and mentioned in earlier sections—among them are euphemism, evaluativity, focus-structure, and the alignment (or lack thereof) of a predicate's "conceptual" and "physical" scales. All of these are potentially important issues for building a more comprehensive theory of vagueness-implicature interactions in gradable predicates.

Here we provide tentative proposals about how two of these factors might be explained by the theory we advanced in this paper: narrow focus on intensifiers (cf. *not very tall*, *not* $[very]_F$ *tall*), and intensifiers with stronger meanings than *very*/*much* (e.g. *extremely*).

In §2, we noted an intuition that stronger intensifiers like *extremely* are more likely than *very* to generate positive implicatures for vague predicates (cf. *not very/extremely tall*). Under the assumption that it is easy to find heights which are clearly tall without being anything close to *extremely* tall, our proposal in §4.1 predicts that *John is not extremely tall* should implicate that John is tall to a greater degree than does *John is not very tall*. The idea is that because there are more heights in $[\theta_{\text{tall}}, \theta_{\text{extremely\_tall}}]$ than there are in $[\theta_{\text{tall}}, \theta_{\text{very\_tall}}]$, the strengthened meaning of *not extremely tall* should be easier to satisfy ("less contradictory") than that of *not very tall*. This idea

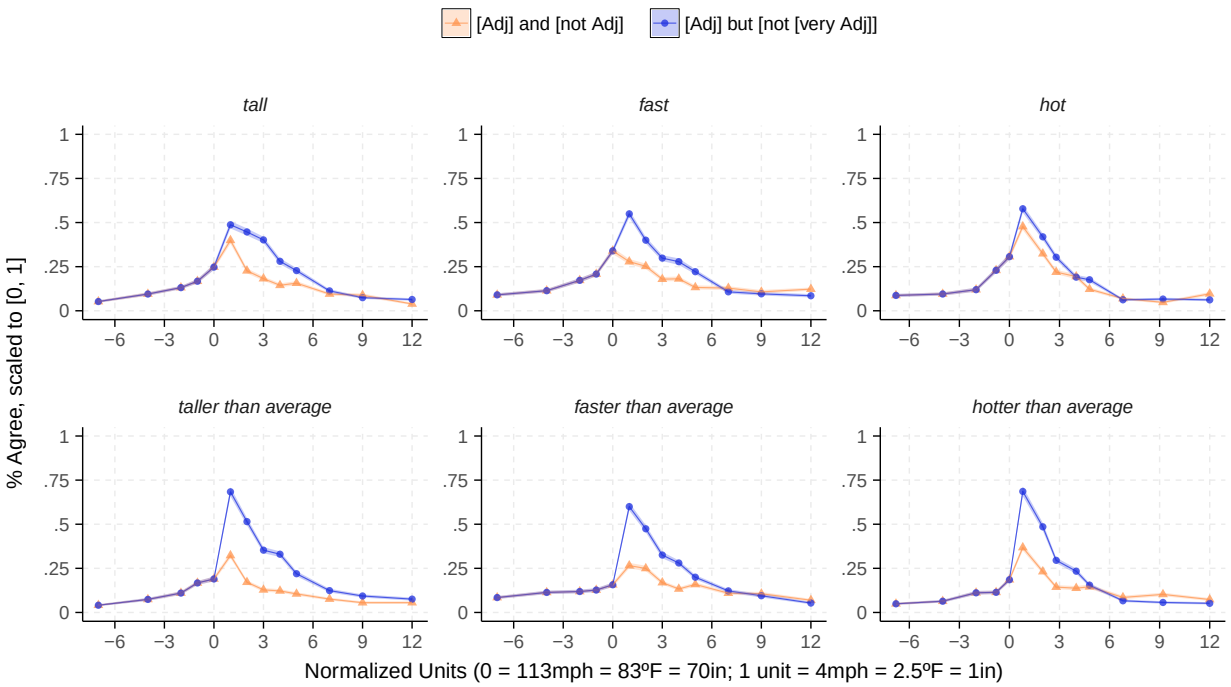

Figure 8: Estimated interpretations of complex expressions we could not explicitly ask for judgments about. ADJ *and not* ADJ is defined as $\min(\text{ADJ}, \text{not ADJ})$, ADJ *and not very* ADJ as $\min(\text{ADJ}, 1 - \text{very ADJ})$.

makes sense in the context of Bennett & Goodman's (2018) quantitative study on a wide range of intensifiers, including *very* and *extremely*. The results of their second experiment show that *very* is one of the lowest ranked intensifiers, while *extremely* is among the highest.

The second case of interest is narrow focus on *very*: the sentence *John is not* $[very]_F$ *tall* seems to imply that John is tall moreso than the sentence without focus on *very*. We see two ways in which the role of focus could be explained.

The first possibility builds upon the observation that narrow focus seems to require a very specific context: it sounds odd to utter *John is not* $[very]_F$ *tall* "out of the blue." This utterance seems to require previous discussion of—or perhaps a QUD directly relevant to—John's height (or just people's heights). This suggests that the comparison class for heights would already have been restricted by the point that the sentence is (felicitously) uttered. By focusing *very* but not *tall*, the speaker may signal that she takes it for granted that John is 'tall', and she is discussing John's position among the class of tall people. In effect, this would mean that "John is not $[\text{very}]_F$ tall" *presupposes*—rather than implicates—that John is tall.

A second possibility is that narrow focus on *very* leads to a more costly sentence (marked and requiring more effort to produce). On Bennett & Goodman's (2018) theory, intensifiers have no semantic contribution but they make the utterance more costly. But by increasing the cost of her utterance, the speaker signals that she is trying to convey more information; hence the intensifying effect. They show that the magnitude of the contribution of an intensifier correlates with its cost (which they model from frequency and number of syllables), as predicted by an RSA model. If this is on the right track, then narrow focus could increase an intensifier's effect, possibly to a point where the implicature could be drawn without leading to a borderline contradiction. Put another way, focusing *very* could make its contribution comparable to that of *extremely*, in which case the

existence of an implicature would be explained exactly as it is for *extremely*.

In both the case of strong intensifiers and narrow focus on *very*, our account predicts that the wider the "gap" introduced by the intensifier is, the more likely an implicature is to arise.

7 Conclusion

In two experiments, we showed that the inference to ADJ from *not very* ADJ is sensitive to scale structure: *not very* ADJ $\rightsquigarrow$ ADJ is a more attractive inference when ADJ has a determinate threshold than when ADJ is vague. Furthermore, we showed that when present, the positive inference has the hallmarks of a structural implicature derived from Manner-based Gricean reasoning. Finally, we showed that this pattern seems to hold not just for relative versus (minimum standard) absolute gradable adjectives, but also for other sets of predicates that differ minimally in the availability of a concrete, determinate threshold of application (here, morphologically unmarked relative adjectives versus their comparative forms).

Because of the systematic nature of the difference between vague gradable predicates and non-vague (or "precise") ones, we argued that the pattern of implicatures exhibited in the '*not very* ADJ' construction should be accounted for as an interaction between the mechanisms at the source of implicature and the semantic property of vagueness. We proposed in §4 that implicatures are not drawn if they lead to meanings that are borderline contradictions. This constraint explains the core patterns in our experimental data: that intensified gradable predicates under negation implicate the unmodified predicate when they have a precise, minimum-standard semantics, but not when they have a vague, relative semantics. Perhaps more importantly, the constraint we advanced is a theoretically well-motivated hypothesis from which clear empirical predictions can be derived in the future (provided clear definitions of the notions involved).

The present results illustrate that the way a gradable predicate applies to an object can affect whether an assertion should be strengthened via implicature. This situation can be viewed as an interaction between two sources of interpretive uncertainty, and might be fruitfully approached as such in future research. Theories that quantify semantic information in terms of entropy reduction, for instance, might shed light on the conditions under which vague implicatures are and are not likely to be drawn (see e.g. van Rooy 2004). Another potentially promising direction for future research in this domain is the integration of behavioral data like those collected here with linguistically-oriented computational modeling frameworks like RSA (Frank & Goodman 2012; Lassiter & Goodman 2014; Lassiter & Goodman 2017; a.o.) and evolutionary game-theoretic approaches to meaning in language (Qing & Franke 2014a; Qing & Franke 2014b; a.o.).

This study only scratches the surface of phenomena at the intersection of vagueness, scale structure, and conversational implicature. We suspect that there are other interactions between vagueness and implicature from which further insights can be extracted and more comprehensive theories developed. We showed that empirical data support the existence of a constraint along the lines of (20): that implicatures are not drawn if they lead to borderline contradictions. However, we ultimately suspect that (20) could itself be a consequence of some more general principle or principles regulating the derivation of implicatures in the presence of vague language. In future research, we will aim to identify and elucidate potential such principles.

References

Alxatib, Sam & Jeff Pelletier. 2011. On the psychology of truth-gaps. In *Vagueness in Communication*, 13–36. Springer.

Aparicio, Helena, Ming Xiang & Christopher Kennedy. 2016. Processing gradable adjectives in context: A visual world study. In *Semantics and Linguistic Theory*, vol. 25, 413–432.

Bach, Kent. 1999. The myth of conventional implicature. *Linguistics and Philosophy* 22(4). 327–366.

Bartsch, Renate & Theo Vennemann. 1972. The grammar of relative adjectives and comparison. *Linguistische Berichte* 20. 19–32.

Bates, Douglas, Reinhold Kliegl, Shravan Vasishth & Harald Baayen. 2015. Parsimonious mixed models. *arXiv Preprint* `arXiv:1506.04967v1`.

Bennett, Erin & Noah Goodman. 2018. Extremely costly intensifiers are stronger than quite costly ones. *Cognition* 178. 147–161.

Bierwisch, Manfred. 1989. The semantics of gradation. In *Dimensional adjectives: Grammatical Structure and Conceptual Interpretation*, 71–261. Springer-Verlag.

Blakemore, Diane & Robyn Carston. 2005. The pragmatics of sentential coordination with *and*. *Lingua* 115(4). 569–589.

Bolinger, Dwight. 1972. *Degree Words*, vol. 53. Walter de Gruyter.

Breheny, Richard, Nathan Klinedinst, Jacopo Romoli & Yasutada Sudo. 2016. The symmetry problem: Current theories and prospects. *Natural Language Semantics* 26. 85–110.

Burnett, Heather. 2014. A delineation solution to the puzzles of absolute adjectives. *Linguistics and Philosophy* 37(1). 1–39.

Chemla, Emmanuel. 2009. Similarity: Towards a unified account of scalar implicatures, free choice permission, and presupposition projection. Under Revision for *Semantics and Pragmatics*.

Cresswell, M. J. 1976. The semantics of degree. In Barbara Partee (ed.), *Montague Grammar*, 261–292. New York: Academic Press.

Dummett, Michael. 1975. Wang's paradox. *Synthese* 30(3-4). 301–324.

Egré, Paul, Vincent de Gardelle & David Ripley. 2013. Vagueness and order effects in color categorization. *Journal of Logic, Language and Information* 22(4). 391–420.

Ernst, Michael. 2004. Permutation methods: A basis for exact inference. *Statistical Science* 19(4). 676–685.

Fox, Danny. 2007. Free choice disjunction and the Theory of scalar implicature. In Uli Sauerland & Penka Stateva (eds.), *Presupposition and Implicature in Compositional Semantics*, 71–120. New York, NY: Palgrave Macmillan.

Fox, Danny & Martin Hackl. 2006. The universal density of measurement. *Linguistics and Philosophy* 29. 537–586.

Fox, Danny & Roni Katzir. 2011. On the characterization of alternatives. *Natural Language Semantics* 19(1). 87–107.

Frank, Michael & Noah Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084). 998.

Grice, Paul. 1975. Logic and conversation. In *The Logic of Grammar*, 64–75. Dickenson.

Heim, Irene. 1985. Notes on comparatives and related matters. Ms., University of Texas.

Heim, Irene. 2000. Degree operators and scope. In *Proceedings of Semantics and Linguistic Theory*, vol. 10, 40–64.

Hersh, Harry M & Alfonso Caramazza. 1976. A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General* 105(3). 254.

Horn, Laurence. 1989. *A Natural History of Negation*. Chicago: University of Chicago Press.

Karttunen, Lauri. 2016. Presupposition: What went wrong? In *Semantics and Linguistic Theory*, vol. 26, 705–731.

Katzir, R. 2007. Structurally-defined alternatives. *Linguistics and Philosophy* 30(6). 669–690.

Katzir, Roni. 2014. On the roles of markedness and contradiction in the use of alternatives. In S.P. Reda (ed.), *Pragmatics, Semantics and the Case of Scalar Implicatures*, 40–71. Springer.

Kennedy, Christopher. 1999. *Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison*. Garland.

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

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

Kennedy, Christopher & Louise McNally. 2005b. The syntax and semantics of multiple degree modification in English. In Stephen Müller (ed.), *Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar*, 178–191. CSLI Publications.

Klement, Erich Peter, Radko Mesiar & Endre Pap. 2004. Triangular norms. Position paper I: Basic analytical and algebraic properties. *Fuzzy Sets and Systems* 143(1). 5–26.

Krifka, Manfred. 2007. Negated antonyms: Creating and filling the gap. In *Presupposition and Implicature in Compositional Semantics*, 163–177. Springer.

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

Lassiter, Daniel. 2011. *Measurement and Modality: The Scalar Basis of Modal Semantics*: New York University dissertation.

Lassiter, Daniel & Noah Goodman. 2014. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Proceedings of Semantics and Linguistic Theory*, vol. 24, 587–610.

Lassiter, Daniel & Noah D Goodman. 2017. Adjectival vagueness in a bayesian model of interpretation. *Synthese* 194(10). 3801–3836.

Leffel, Timothy, Ming Xiang & Christopher Kennedy. 2016. Imprecision is pragmatic: Evidence from referential processing. In *Proceedings of Semantics and Linguistic Theory*, vol. 26, 836–854.

Maris, Eric & Robert Oostenveld. 2007. Non-parametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods* 164(1). 177–190.

Matsumoto, Yo. 1995. The conversational condition on horn scales. *Linguistics and Philosophy* 18(1). 21–60.

Potts, Christopher. 2005. *The Logic of Conventional Implicature*. Oxford University Press.

Potts, Christopher. 2008. Interpretive economy, Schelling points, and evolutionary stability. Ms., UMass Amherst.

Qing, Ciyang & Michael Franke. 2014a. Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In *Proceedings of Semantics and Linguistic Theory*, vol. 24, 23–41.

Qing, Ciyang & Michael Franke. 2014b. Meaning and use of gradable adjectives: Formal modeling meets empirical data. In *Proceedings of CogSci*, vol. 36, 1204–1209.

Ripley, David. 2011. Contradictions at the borders. In Rick Nouwen, Robert van Rooij, Uli Sauerland & Hans-Christian Schmitz (eds.), *Vagueness in Communication: International Workshop (revised selected papers)*, 169–188. Berlin, Heidelberg: Springer.

van Rooy, Robert. 2004. Utility, informativity and protocols. *Journal of Philosophical Logic* 33(4). 389–419.

Sassoon, Galit. 2010. Measurement theory in linguistics. *Synthese* 174(1). 151–180.

Sauerland, Uli. 2004. Scalar implicatures in complex sentences. *Linguistics and Philosophy* 27(3). 367–391.

Schlenker, Philippe. 2008. Be articulate: A pragmatic theory of presupposition projection. *Theoretical Linguistics* 34(3). 157–212.

Schwarzschild, Roger. 2008. The semantics of comparatives and other degree constructions. *Language and Linguistics Compass* 2(2). 308–331.

Serchuk, Phil, Ian Hargreaves & Richard Zach. 2011. Vagueness, logic and use: Four experimental studies on vagueness. *Mind & Language* 26(5). 540–573.

Simons, Mandy. 2001/2013. On the conversational basis of some presuppositions. In *Perspectives on Linguistic Pragmatics*, 329–348. Springer.

Solt, Stephanie. 2015. Measurement scales in natural language. *Language and Linguistics Compass* 9(1). 14–32.

Solt, Stephanie & Nicole Gotzner. 2012. Experimenting with degree. In *Proceedings of Semantics and Linguistic Theory*, vol. 22, 166–187.

Spector, Benjamin. 2016. Multivalent semantics for vagueness and presupposition. *Topoi* 35(1). 45–55.

von Stechow, Arnim. 1984. Comparing semantic theories of comparison. *Journal of Semantics* 3. 1–77.

Trinh, Tue & Andreas Haida. 2015. Constraining the derivation of alternatives. *Natural Language Semantics* 23(4). 249–270.

Tye, Michael. 1994. Sorites paradoxes and the semantics of vagueness. *Philosophical Perspectives* 8. 189–206.

Zadeh, Lotfi A. 1965. Fuzzy sets. *Information and Control* 8(3). 338–353.

Zehr, Jérémy. 2014. *Vagueness, Presupposition and Truth-Value Judgments*: École Normale Supérieure de Paris dissertation.