Rick Nouwen
Robert van Rooij
Uli Sauerland
Hans-Christian Schmitz (Eds.)

LNAI 6517

Vagueness in Communication

International Workshop, ViC 2009
held as part of ESSLLI 2009
Bordeaux, France, July 2009
Revised Selected Papers

Springer

Lecture Notes in Artificial Intelligence 6517

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

FoLLI Publications on Logic, Language and Information

Rick Nouwen Robert van Rooij
Uli Sauerland Hans-Christian Schmitz (Eds.)

Vagueness in Communication

International Workshop, ViC 2009
held as part of ESSLLI 2009
Bordeaux, France, July 20-24, 2009
Revised Selected Papers

Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors
Rick Nouwen
Utrecht University, Utrecht Institute for Linguistics OTS
Janskerkhof 13a, 3512 BL Utrecht, The Netherlands
E-mail: r.w.f.nouwen@uu.nl

Robert van Rooij
University of Amsterdam, Institute for Logic, Language and Computation (ILLC)
Oude Turfmarkt 141-147, 1012 GC Amsterdam, The Netherlands
E-mail: r.a.m.vanrooij@uva.nl

Uli Sauerland
Zentrum für allgemeine Sprachwissenschaft
Schützenstraße 18, 10117 Berlin, Germany
E-mail: uli@alum.mit.edu

Hans-Christian Schmitz
Fraunhofer FIT, Schloss Birlinghoven
53754 Sankt Augustin, Germany
E-mail: hans-christian.schmitz@fit.fraunhofer.de

ISSN 0302-9743 e-ISSN 1611-3349

ISBN 978-3-642-18445-1 e-ISBN 978-3-642-18446-8

DOI 10.1007/978-3-642-18446-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010942868

CR Subject Classification (1998): I.2.7, I.2, H.3, H.4, F.4.1, F.4.3

LNCS Sublibrary: SL 7– Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Most of the papers in this volume originate from the workshop Vagueness in Communication that was held during the 2009 European Summer School in Logic, Language and Information in Bordeaux, France. Although vagueness has long since been an important topic in philosophy, logic and linguistics, some recent advances have made the functions of vagueness in natural language communication an exciting and timely research area. This renewed interest has a distinct cross-disciplinary character and has spawned many new research questions. The workshop brought together researchers whose work contributes to the cross-disciplinary line of inquiry, in particular by broadening the empirical base for the study of vagueness, by offering a synthesis of theories from different disciplines, and by addressing the pragmatics of vagueness. It thereby provided a forum for lively discussions on recent and on-going work.

The workshop was organized by the four editors of the present volume and Manfred Krifka, who unfortunately could not participate in the editorship of this volume because of other commitments. We would like to thank all the workshop participants for their contributions to the success of the workshop and the quality of the papers in this volume. Specifically, we thank the Program Committee, Graeme Forbes, Peter Gärdenfors, Hans Kamp, Stefan Kaufmann, Chris Kennedy, Ewan Klein, Manfred Krifka, Manfred Kupffer, Louise McNally, Christian Plunze, Marieke Schouwstra, Markus Schrenk, Yoad Winter and Thomas Ede Zimmermann for their help in selecting suitable contributions. We would also like to acknowledge the additional referees that were willing to advise us and the authors on the papers for this volume: Marta Abrusan, Anton Benz, Chris Kennedy, Sveta Krasikova, Chris Potts, Diania Raffman, David Schlangen, Kristen Syrett and Frank Veltman.

We would like to thank Matthias Daubitz for TEXnical and editorial support during the preparation of the final version of this volume.

Rick Nouwen acknowledges support from the Netherlands Organisation for Scientic Research (NWO) for the Degrees Under Discussion project. Robert van Rooij is supported by the NWO for the 'On Vagueness – And How to Be Precise' project as well as the ESF VAAG project. Uli Sauerland thanks the German Research Foundation DFG for its financial support through grants SA/925-1 and SA/925-4, the latter within the Eurocores LogiCCC program as part of the research project VAAG. The completion of this volume was furthermore aided by a grant from the ESF within the Eurocores LogiCCC program for editorial support.

November 2010

Rick Nouwen
Robert van Rooij
Uli Sauerland
Hans-Christian Schmitz

Table of Contents

Introduction

Rick Nouwen, Robert van Rooij, Uli Sauerland, and Hans-Christian Schmitz

1 Vagueness

One could define vagueness as the existence of *borderline cases* and characterise the philosophical debate on vagueness as being about the nature of these. The prevalent theories of vagueness can be divided into three categories, paralleling three logical interpretations of borderline cases: (i) a borderline case is a case of a truth-value gap; it is neither true nor false; (ii) a borderline case is a case of a truth-value glut; it is both true and false; and (iii) a borderline case is a case where the truth-value is non-classical. The third of these is proposed in the fuzzy logic approach to vagueness. Three-valued approaches have only $\frac{1}{2}$ as a value in addition to the standard values 1 and 0. These approaches can be interpreted either as allowing for gaps or gluts, depending on how the notion of satisfaction or truth is defined. If a sentence is taken to be true only if its value is 1, it allows for gaps, but if it is taken to be true already if its value is at least $\frac{1}{2}$ it allows for gluts. The most popular theories advertising gluts and gaps, however, are supervaluationism and subvaluationism, both of which make use of the notion of precisifications, that is, ways of making things precise. Truth-value gaps in supervaluationism are due to the way truth simpliciter, or supertruth, is defined: A proposition is supertrue (superfalse) if it is true (false) at all precisifications. This means that a proposition can be neither true nor false in case there exist two precisifications, one of which make it true and one of which makes it false. Conversely, in subvaluation theory, the same scenario would lead to a truth-value glut. That is, the proposition would be both true and false. This is because subvaluationism defines truth simpliciter as being true at some precisifcation.

The vagueness debate is a lively one since there are quite a few additional aspects to vagueness that need to be accounted for. One is higher order vagueness: the fact that the very boundary between definite cases of a predicate T and borderline cases is itself vague. Traditionally, however, probably the most crucial burden of any theory of vagueness is to account for the Sorites Paradox: Why is it that we accept (1), but not the apparent consequence that thereby everyone should count as tall?

(1) If person x is tall, then someone ever so slightly shorter than x is tall too.

In section 2 of this introduction we will present an overview of the philosophical theoretical debate on vagueness, focusing in particular on this paradox.

The topic of vagueness, however, far extends the essentially logical issue of how to treat borderlineness and the sorites. In linguistics, the tight connection between vagueness and the grammatical notion of gradability has sparked a lively

R. Nouwen et al. (Eds.): ViC 2009, LNAI 6517, pp. 1–12, 2011.

line of research into the relation between the meaning of degree expressions and vagueness. The linguistics of vagueness is the topic of section 3 of this introduction. It is becoming apparent that for a true understanding of vagueness, however, one needs to look beyond just linguistics and philosophy proper. The psychology behind the sorites, and the use of vague terms in general, was until recently pretty much unexplored territory. Section 4 provides the backdrop for new directions the study of vagueness is taking.

2 Logic and Philosophy: The Sorites

The inductive premise of the sorites paradox (henceforth referred to as P), represents a crucial ingredient of vagueness, namely *tolerance*. Vagueness entails indifference with respect to small changes in the degree to which some quality holds. It is precisely this aspect of vagueness that is centre stage in the theoretical debate.

Formally, let us write $x \sim_T y$ for x and y are *neglectably different* with respect to the degree of T-ness. The general scheme for the Sorites Paradox is then the following, where given the possibility of a series $x_0 \sim_T \ldots \sim_T x_n$ from one extreme of T-ness to another, it would appear that C follows from P.

$(P) \quad \forall x \forall y[(x \sim_T y \ \wedge \ T(x)) \ \rightarrow \ T(y)]$

$(C) \quad \forall x[T(x)]$

There are two main theoretical options to account for the paradox. The first stance is to deny P, in which case the paradox simply disappears, but a more difficult problem surfaces of why it *seems* to us that P. The alternative is to tackle P head-on, by trying to understand how it follows semantically and, crucially, how it does not entail C.

Examples of accounts within the first tradition, where P is argued not to hold, include fuzzy logic approaches, which contend that our tendency to accept P is because it is *almost* true (i.e. it has a truth-value close to 1). Fuzzy logic gives rise to some unwelcome properties (see for instance the critiques in [9,14,40]). In particular, it predicts truth-values for complex propositions that are in many cases not entirely intuitive. A further often cited criticism is that the degrees of truth in fuzzy logic are unsuitable as a basis for a semantics of the comparative. It appears that fuzzy approaches would naturally interpret *John is taller than Bill* as *John is tall* having a higher truth-value than *Bill is tall*. However, this entails that all comparison takes place on the same scale, namely that of degrees of truth. This is problematic since comparative forms are restricted to certain dimensions. For instance, *The temperature is higher than John is tall* is uninterpretable. Fuzzy logic, however, suggests that this sentence should express the statement that it is more true that the temperature is high than it is true that John is tall.

A much more popular alternative is supervaluation theory. The core proposal in supervaluationist accounts is that vagueness is the result of many possible

ways in which things could be precise. A proposition is supertrue, if it is true irrespective of how we resolve our semantic indecision. The selling point of supervaluation theory is that it preserves all classical validities. Thus, or so it is claimed, logically speaking there is no difference between classical logic and supervaluation theory. But the non-standard way of accounting for these validities still comes with its logical prize. To see why, consider the following. Proponents of supervaluation theory hold that although there is a cutoff-point – i.e. the formula $\exists x \exists y[T(x) \wedge x \sim_T y \wedge \neg T(y)]$ is supertrue –, still, no one of its instantiations itself is supertrue. This is a remarkable logical feature: in classical logic it holds that $\varphi \vee \psi \models \varphi, \psi$ (meaning that at least one of φ and ψ must be true in each model that verifies $\varphi \vee \psi$). In supervaluation theory this doesn't hold anymore; $\exists x[T(x)] \not\models_{\text{supv}} T(x_1), \ldots, T(x_n)$. The relation between supervaluation theories of vagueness and classical logic is the topic of the contribution by Cobreros in this volume. His starting point is the observation that supervaluationist logic no longer has a classical notion of logical consequence once a "definite" operator is taken into account. Cobreros shows, however, that there exist deduction systems that come very close to being classical, thus showing new light on the alleged non-classicality (and its consequences) of supervaluationism.

Beyond the classicality debate surrounding supervaluationism, a problem of a more conceptual nature has been noted. Supervaluation theory makes use of complete refinements, and supervaluation theory assumes that we can always make sharp cutoff-points: vagueness exists only because in daily life we are too lazy to make them. But this assumption seems to be wrong: vagueness exists, according to Dummet [7], because we cannot make such sharp cutoff-points even if we wanted to.

An early variation on supervaluationism originates in [24]. According to Lewis, vagueness arises as a consequence of there being many possible precise language that can be used in communication. The contribution of Lassiter takes this idea as a starting point. He explores a theory of vagueness which locates vagueness not in semantics, but rather in the probabilistic representation of linguistic knowledge. In Lassiter's approach this uncertainty is probabilistically represented. That is, the context contains a probability distribution over the set of possible languages, where these possible languages differ in the threshold for what counts as *tall*, *smart*, *a heap* etc.

An alternative to supervaluationism especially popular in the 1980s among linguists (and later by philosophers as well) was the so-called 'contextualist' approach. This approach was initiated by Kamp [15]. To solve the Sorites paradox, he (i) makes use of a sophisticated mechanism of *context change* and (ii) adopts a non-truth conditional analysis of conditional sentences, and proposes a weak, but non-standard notion of entailment. The idea of context change is that once it is explicitly accepted within the discourse that x has property P, for any vague predicate, the initial contextually given valuation function V changes into (possibly) new valuation function V' such that indistinguishable, or at least sufficiently similar individuals to x must be counted as having property P as well according

to new valuation function (and context) V'. In other words, what Kamp proposes is that each of the inductive premises is true in case its antecedent is verified, because of context change.

In this volume, Forbes critically discusses contextualist accounts of the Sorites, most particularly that of Soames [37]. Contextualists typically tackle the individual steps in the inductive premise by assuming they involve a context-shift. He argues that certain versions of the Sorites are left unexplained by such accounts.

While most contextualists (e.g. Pinkal, Raffman, Graff) follow Kamp making use of context change, they normally seek to improve on (ii) above by making the resulting logic more classical.[1] In this volume, Egré's contribution follows the lead of Raffman [31] of comparing the role of context in Sorites series to comparable phenomena in perception. Egré explores an account of the sorites in which bordeline cases are ambiguous. He sketches the paradox as a combination of two plausible constraints: on the one hand the conservation of categorisation between adjacent items and on the other hand the existence of a category switch somewhere in a sorites series. He argues that these constraints are compatible if the switch occurs among items that are ambiguous between the two contrasted categorisations. Egré compares category switches to percept switches such as those in Fisher-type series [10]. He also discusses the consequences of an ambiguity approach to the principle of tolerance.

3 Vagueness and Linguistics

Although vagueness occurs in a variety of categories, such as nouns (*heap*), prepositions (*near*) and verbs (*enjoy*), in linguistics, it is naturally associated with adjectives like *tall*. This is because the linguistic study of vagueness is deeply connected to notion of *gradability*: the possibility to use modifiers to express the degree to which a predicate, typically an adjective, holds. Although the exact relation between gradability and vagueness is an interesting issue in itself (see below), there are several obvious reasons to connect the two phenomena. First of all, the inductive premise of a sorites paradox is based on a comparison with respect to *degree*, witness the comparative form of *short* in the inductive premise *if John is tall, then someone who is ever so slightly shorter is tall as well*. Second, degree modifiers interact with vagueness. That is, some introduce vagueness, while yet other degree modifiers remove it. For instance, the bare use of *straight* in (2-a) is hardly vague at all, but modification with *almost* introduces (more) vagueness. Conversely, the positive form of *tall* in (3-a) is vague, while its comparative form (modification by the comparative morpheme *-er*) is not.

(2) a. The rod is straight.
b. The rod is almost straight.

(3) a. John is tall.
b. John is taller than Bill.

[1] For a somewhat different contextual solution, see [11,29,32].

A final example of the connection between gradability and vagueness is a minimal variation on (3):

(4) a. Compared to Bill, John is tall.
b. John is taller than Bill.

These two sentences do not have an equivalent meaning. The crucial case is when John is taller than Bill to a degree that is just barely observable. In that case, (4-b) is true, but (4-a) is not. In other words, tolerance is an aspect of what Kennedy [20] calls *implicit comparison*, comparison using the positive form of an adjective. It is not an aspect of *explicit comparison* (the morphosyntactic comparative form). Kennedy proposes that the difference between (4-a) and (4-b) is compositional. The comparative morpheme *-er* imposes a strict comparison of degrees. In the case of (4) this amounts to comparing John's height to Bill's height. The positive form in (4-a) is the result of combining an adjective with the (silent) modifier *POS*. Kennedy proposes that the type of comparison encoded by *POS* is different from the one encoded by *-er*. It expresses that a degree *significantly exceeds* a contextual standard of comparison. One possible implementation of this involves Fara's notion of interest-relativity ([8]), which explains the non-crisp judgement for (4-a) as follows: if, given my interests, John's height exceeds the standard of comparison in a way that is significant, then it could not be that the slightly different height of Bill does not. Independent of the specific implementation, crucial is the understanding that the positive form encodes a fundamentally different mode of comparison from the morphosyntactic comparative form. This was stressed too by van Rooij [33], who presents an alternative approach to Kennedy's within a framework based on Klein's comparison class-based delineation approach. Van Rooij stresses that explicit comparison involves a weak order, while implicit comparison involves a semi order. The difference is best explained in measure-theoretic terms. Let $f(x)$ be some measure of x (say, height), and e be some fixed value which acts as a margin of error, then the following is a definition of a semi order $\succ_T$.

(5) $x \succ_T y$ iff $f(x) > f(y) + e$

If e is 0, then $\succ_T$ is a weak order. Clearly, the difference between a weak and a semi order is closely related to Kennedy's proposal for the difference between the positive and the comparative: weak orders represent a strict mode of comparison, while semi orders represent comparison based on significant differences in measurement. To account for the contrast, van Rooij proposes that (4) involves two different kinds of uses of comparison classes. While (4-b) involves existential quantification over comparison classes, (4-a) is based on comparing just John and Bill. Crucially, not all comparison classes are admissible. A comparison class is only pragmatically appropriate if the gap between individuals that have the relevant property and those that do not is significant. In other words, for the case in which John and Bill hardly differ in height, {John, Bill}

is not an admissible comparison class, hence the unacceptability of (5-a) in such a context.

The two theories of Kennedy and van Rooij are representative of the two main contenders among linguistic semantic approaches to degree phenomena: the degree approach, which maintains that the semantics of gradable predicates necessitates the use of some notion of degrees, versus what is often called the delineation approach, where gradable predicates lack degree arguments. There is considerable variation among degree approaches. For instance, Kennedy [17] takes an adjective to be a measure function, a mapping from entities to degrees. A popular alternative is to treat adjectives as relations between entities and degrees [36,38,13]. Opposing the degree approaches are proposals inspired by supervaluationist or contextualist theories of vagueness. Most prominent is the comparison class approach of Klein [22], and recent reincarnations of that theory (for instance, [6,33]). According to these approaches, a predicate like *tall* is always evaluated with respect to a comparison class. So, $\text{tall}_c(x)$ is true if x is tall with respect to class c. To a large extent, these theories are equivalent to a (certain kind of) degree semantics for gradable predicates. For instance, a glance at the semantics of comparatives shows that degrees and comparison classes are not entirely dissimilar. The following two forms represent the interpretation of *John is taller than Bill* in the two frameworks.

(6) $\exists d[\text{tall}(j, d) \;\&\; \neg\text{tall}(b, d)]$ degree semantics

(7) $\exists c[\text{tall}_c(j) \;\&\; \neg\text{tall}_c(b)]$ delineation semantics

What is different, however, is the interpretation of the positive form. Kleinian analyses offer a direct interpretation of the positive form: *John is tall* is true iff John is tall in the relevant comparison class. In degree approaches, however, the positive needs to be interpreted indirectly, since the semantics of the adjective yields not the interpretation of the positive, but rather a degree relation or function. In degree approaches, *John is tall* is therefore interpreted by first quantifying the degree argument of the adjective. An example of this is Kennedy's approach discussed above, where the positive is the result of applying a silent modifier *POS* expressing that the relevant degree significantly exceeds a contextual standard c:

(8) [John is [*POS* tall]] is true
$\Leftrightarrow$
$\exists d[\text{tall}(x, d) \;\&\; d \text{ significantly exceeds } c]$

Part of the debate is based on which approach is somehow more natural. Klein [22] argued that the comparison class proposal is more in line with the principle of compositionality than its degree counterpart is, for it predicts that the comparative form is derived from the positive form, as is the case in (almost) all natural languages. Von Stechow [38] and others have argued, however, that

this argument is not entirely conclusive, and hold that comparison is cognitively primary.[2]

Comparison classes are not just relevant to the approach of Klein [22] and its offspring. A comparison class can be made explicit using *for* phrases, as in *John is tall for a basketball player*. When it is not made explicit, it is often assumed to be part of the interpretation of the positive form. It is not trivial, however, what contribution a comparison class makes to the standard of comparison of a positive form. Kennedy [19] points out that theories along the lines of Cresswell [4], where the standard of comparison is the average measure of the individuals in the comparison class, is untenable. If such analyses were on the right track, then examples like (9) would be expected to be contradictory:

(9) John is taller than the average height of a basketball player, but he is still not tall for a basketball player.

Solt, this volume, addresses this issue further and argues that a crucial ingredient of the semantics of the positive depends on the distribution of measures of the individuals in the comparison class. In her analysis the comparison class is an argument of the positive operator. She furthermore builds on von Stechow [39] in assuming that positive forms make use of a neutral region on the relevant scale (as opposed to the single value on the scale represented by the standard of comparison). This so-called standard range, Solt argues, depends on the distribution of the individuals in the comparison class (with respect to the relevant quality dimension). She furthermore discusses other semantic aspects of *for* phrase comparison classes, such as their alleged presuppositionality. (If John is tall for a basketball player, then he has to be a basketball player. Cf. [19]).

So far, we have pointed out the relevance of the linguistic study of gradability to vagueness mostly by discussing the role of comparatives and positive forms in the Sorites and in regulating vagueness. However, a look at a typology of gradable adjectives (and arguably other gradable expressions) yields a more fundamental look at vagueness. Here we enter the question of which concepts give rise to vague natural language expressions, and what exactly is the relation between vagueness and gradability.

Standardly, it is assumed that there are two kinds of gradable expressions: (i) those that are context-dependent or *relative*, like *tall* and (ii) those that are context-independent or *absolute*, like *straight*. That is, while our understanding of (10) depends on what counts as tall in the given context, (11) expresses the same in any context, namely that the rod is not bent.

(10) John is tall.

(11) This rod is straight.

[2] The degree/delineation debate goes beyond this foundational issue of compositionality, however. A number of empirical phenomena have been used to argue either in favour of or against the use of degrees in semantics. Such considerations are well beyond the scope of this introduction, however. See [38, 17, 25, 32, 6, 5].

The difference between expressions like *tall* and *straight* is often connected to a notion of scale structure ([30,21,34,19]). Following the line of reasoning of Kennedy [19] (in part based on [21]), scale structure influences the (likely) standard of comparison an adjective is evaluated against. Degrees of height are positioned on a principally open-ended scale, which yields no salient reference point as to what counts as tall in any context.[3] In contrast, *straight* is associated to a scale of bendedness, which contains a zero point.[4] This scalar end-point is used as a context-independent standard of comparison: *straight* entails *no* bendedness, *bent* entails *some* bendedness. A similar construal of *tall* would simply be meaningless.

The absolute/relative distinction is particularly relevant to the relation between vagueness and gradability. It shows that not all gradable predicates are vague. This is particularly interesting in the light of the tight relation that certain theories predict between the two notions. In delineation theories, vagueness often entails gradability. That is, the existence of different delineations for a predicate (via, for instance, comparison classes) is exactly what drives comparison. It has been argued ([19]) that this means that such theories will not be able to account for the absolute / relative distinction. This is conclusion is countered, however, by van Rooij [33] and McNally (this volume), among others.

An immediate problem for the above absolute analysis of terms like *straight* is that as a result it cannot be truthfully applied to any observable object: there is no object that is absolutely straight according to an ultimate high *standard of precision*. Thus, if we want to explain our use of absolute terms, we still have to make their meaning context-dependent, although this context dependence now involves standards of precision rather than standards of comparison (cf. Lewis, 1979). Standards of precision are also relevant for the interpretation of measure phrases. Intuitively, if you truly (enough) say that John is 2 meters tall, he can actually be taller than Mary, of whom it is truly (enough) said that she is 2.01 meters tall. One way to account for this observation is to assume that the underlying structure of measurement in the former case is *courser grained* than the measurement structure in the latter case. The *point* denoted by '2 meters' on the coarse scale corresponds with a *set of points* on the finer-grained scale, and might include, for instance, 2.02 meters. But why do we associate the different expressions with the different measurement structures? Krifka [23] argues that this can be derived by Horn's division of pragmatic labor: '2.02 meters' is a more complex expression than '2 meters' and its use thereby signals that a more complex, i.e. fine-grained, measurement structure is involved. In this volume, Bastiaanse provides a game theoretical account of round number interpretation, elaborating on Krifka's account. He also suggest applying a similar model to other vague expressions.

[3] Though you might wonder why according to Kennedy and associates *tall*'s antonym does not have such a salient reference point.

[4] This difference in scale structure arguably also explains why some modifiers better pair with some adjectives than others: one can say 'This bar is absolutely straight' without the modifier having an epistemic reading, while this doesn't seem to be possible with 'John is absolutely tall' (cf. [35]).

Kennedy has argued that the interpretation of absolute adjectives depends on pragmatics. Kennedy's interpretative economy principle states that the contribution of the conventional meanings of elements in a sentence should be maximised. Since scale structure is part of the conventional meaning of an adjective, this should be used as a basis for the standard of comparison, which would be an end-point for closed scale adjectives (rendering them absolute). McNally's contribution to this book challenges Kennedy's approach by identifying empirical problems for interpretative economy. She argues that the absolute / relative distinction should be compared to the distinct classification strategies of classification by rule (absolute) and classification by similarity (relative) [12].

The distinction between vague and crisp terms is especially enigmatic when one turn to non-adjectival predicates. For instance, a noun like *chair* allows for borderline cases, but is not gradable (cf. [14]).[5] Your typical four-legged wooden dinner table chair is not *more a chair* than some oddly shaped plastic 1-legged designer chair, even though the former is definitely more prototypical of the chair concept than the latter. The upshot seems to be that despite their obvious kinship, vagueness and gradability are distinct notions. This should maybe not be that surprising given the essentially grammatical nature of gradability. Gradability is the possibility of being degree modified and is thus, in contrast to vagueness, subject to a wealth of grammatical constraints. Only recently have linguists begun to unravel the full empirical scope of gradability.[6] Gradability is naturally associated with adjectives, since this category is involved in the bulk of degree phenomena. However, at least since Bolinger [2], it is known that gradability is not limited to adjectives. For instance, some (but far from all) nouns are gradable. The examples in (12) express that John is an idiot to a relatively high degree [26,28]. On the other hand, there is no option to interpret the adjectives in (13) as degree modifiers.

(12) a. John is a huge idiot.
b. John is an unbelievable idiot.

(13) a. That is a huge chair.
b. That is an unbelievable chair.

4 New Directions

The use of experimentally gathered data has proven very useful in linguistics and philosophy in recent years, for instance in investigating the nature of pragmatic components of meaning such as implicatures or presuppositions. It is obvious that the study of vagueness could benefit from experimental research too. So

[5] The influential [16] makes the more general point that issue of the relation between gradability, vagueness and prototype similarity is a highly complex one, where many questions remain open.

[6] See, for instance, [27] for a study that includes several non-adjectival degree phenomena. There is moreover a recent interest in cross-linguistic differences in degree phenomena [1,18].

far, however, very few have ventured in this direction. The notable exceptions show that there is some promising ground to be made. One particular issue that lends itself to experimentation is the question what borderline cases look like; are they gaps, gluts or something else? Bonini et al. [3] split a large group of subjects in two, asking one half when a certain property holds, and the other half when it is false to say that a certain property holds. The difference between the two groups sheds light on the nature of borderline cases. Bonini et al. concluded from their data that a gap-like theory should be preferred and they proceeded to argue for a specific form of epistemicism on the basis of their results. In the contribution to this volume by Alxatib and Pelletier we find a closely related example of experimental work on vagueness. On the basis of new experiments, they argue against the conclusions drawn in [3]. Alxatib and Pelletier instead use their experimental results to argue for a novel approach which combines sub- and supervaluationism.

A related study is Ripley's contribution. Ripley investigates sentences that express the logical form $Tx \wedge \neg Tx$ or $\neg(Tx \vee \neg Tx)$ where T is a vague predicate. Since such sentences are contradictions in classical logic, Ripley calls them "Borderline Contradictions". Such borderline contradictions have played an important role in the discussion of vagueness in language by Kamp [14] and Fine [9]. Ripley's shows that a majority of subjects find such sentences quite acceptable and discusses the consequences of this finding.

Klein's contribution to this volume exemplifies a different kind of experimentation altogether. He uses computational simulation experiments to explore the communicative functions of vagueness. Based on Parikh's idea that, given sufficient overlap in how agents interpret a vague term, vagueness is useful, Klein measures the success of communicating with a vague expression by simulating a task of two agents. He focuses on the vagueness inherent in temporal expressions like *morning* and computes how successful two agents are in actually meeting up on the basis of agreeing to meet at a vaguely indicated time.

The increasing use of experimental methods and of insights from experimental psychology illustrates that despite the fact that vagueness has long since been an important topic in philosophy, logic and linguistics, the function of vagueness in natural language communication is very much an exciting and timely research area. As the diversity of contributions in this volume shows, the renewed interest into vagueness has a distinct cross-disciplinary character and has spawned many new research questions.

References

1. Beck, S., Oda, T., Sugisaki, K.: Parametric variation in the semantics of comparison: Japanese vs. english. Journal of East Asian Linguistics 13, 289–344 (2004)
2. Bolinger, D.: Degree Words. Mouton, Den Haag (1972)
3. Bonini, N., Osherson, D., Viale, R., Williamson, T.: On the psychology of vague predicates. Mind and Language 14(4), 377–393 (1999)
4. Cresswell, M.: The semantics of degree. In: Partee, B. (ed.) Montague Grammar, pp. 261–292. Academic Press, London (1976)

5. Doetjes, J.: Incommensurability. In: Schulz, K., Aloni, M. (eds.) Proceedings of the 17th Amsterdam Colloquium, ILLC, AUP (2010)
6. Doetjes, J., Constantinescu, C., Součková, K.: A neo-kleinian approach to comparatives. In: Ito, S., Cormany, E. (eds.): Proceedings of Semantics and Linguistic Theory XIX, New York, Ithaca (2008)
7. Dummett, M.: Wang's paradox. Synthese 30, 301–324 (1975)
8. Fara, D.G.: Shifting sands: An interest-relative theory of vagueness. Philosophical Topics 20, 45–81 (2000)
9. Fine, K.: Vagueness, truth and logic. Synthese 30, 265–300 (1975)
10. Fisher, G.: Measuring ambiguity. The American Journal of Psychology 80(4), 541–557 (1967)
11. Gaifman, H.: Vagueness, Tolerance and Contextual Logic. Manuscript. Columbia University, New York (January 2002)
12. Hahn, U., Chater, N.: Similarity and rules: distinct? exhaustive? empirically distinguishable? Cognition 65, 197–230 (1998)
13. Heim, I.: Degree operators and scope. In: Proceedings of SALT, vol. 10. CLC Publications, Ithaca (2000)
14. Kamp, H.: Two theories of adjectives. In: Keenan, E. (ed.) Formal Semantics of Natural Language, pp. 123–155. Cambridge University Press, Cambridge (1975)
15. Kamp, H.: The paradox of the heap. In: Mönnich, U. (ed.) Aspects of Philosophical Logic, pp. 225–277. D. Reidel, Dordrecht (1981)
16. Kamp, H., Partee, B.: Prototype theory and compositionality. Cognition 75, 129–191 (1995)
17. Kennedy, C.: Projecting the adjective: the syntax and semantics of gradability and comparison. PhD. Thesis, UCSD (1997)
18. Kennedy, C.: Modes of comparison. In: Elliott, M., Kirby, J., Sawada, O., Staraki, E., Yoon, S. (eds.) Proceedings of Chicago Linguistic Society, vol. 43 (2007)
19. Kennedy, C.: Vagueness and grammar: The semantics of relative and absolute gradable predicates. Linguistics and Philosophy 30(1), 1–45 (2007)
20. Kennedy, C.: Vagueness and comparison. In: Egré, P., Klinedinst, N. (eds.) Vagueness and Language Use. Palgrave McMillan, Oxford (2010)
21. Kennedy, C., McNally, L.: Scale structure, degree modification and the semantics of gradable predicates. Language 81(2), 345–381 (2005)
22. Klein, E.: A semantics for positive and comparative adjectives. Linguistics and Philosophy 4, 1–45 (1980)
23. Krifka, M.: Approximate interpretation of number words: A case for strategic communication. In: Vogel, I., Zwarts, J. (eds.) Cognitive Foundations of Communication. Koninklijke Nederlandse Akademie van Wetenschapen, Amsterdam (2007)
24. Lewis, D.: General semantics. Synthese 22, 18–67 (1970)
25. Moltmann, F.: Degree structure as trope structure: A trope- based analysis of positive and comparative adjectives. Linguistics and Philosophy 32(1), 51–94 (2009)
26. Morzycki, M.: Degree modification of gradable nouns: size adjectives and adnominal degree morphemes. Natural Language Semantics 17(2), 175–203 (2009)
27. Neeleman, A., Koot, H.v.d., Doetjes, J.: Degree expressions. The Linguistic Review 21(1), 1–66 (2004)
28. Nouwen, R.: Degree modifiers and monotonicity. In: Egré, P., Klinedinst, N. (eds.) Vagueness and Language Use. Palgrave McMillan, Oxford (2010)
29. Pagin, P.: Vagueness and domain restriction. In: Klinedinst, N., Egré, P. (eds.) Vagueness and Language Use. Palgrave MacMillan, Oxford (2010)
30. Paradis, C.: Adjectives and boundedness. Cognitive Linguistics 12(1), 47–65 (2001)

31. Raffman, D.: Vagueness without paradox. Philosophical Review 103 (1), 41–74 (1994)
32. van Rooij, R.: Vagueness and linguistics. In: Ronzitti, G. (ed.) The Vagueness Handbook. Springer, Heidelberg (2010)
33. van Rooij, R.: Implicit versus explicit comparatives. In: Egré, P., Klinedinst, N. (eds.) Vagueness and Language Use. Palgrave McMillan, Oxford (2010)
34. Rotstein, C., Winter, Y.: Total adjectives versus partial adjectives: scale structure and higher-order modifiers. Natural Language Semantics 12, 259–288 (2004)
35. Sauerland, U., Stateva, P.: Two types of vagueness. In: Egré, P., Klinedienst, N. (eds.) Vagueness and Language Use. Palgrave McMillan, Oxford (2010)
36. Seuren, P.A.M.: The comparative. In: Kiefer, F., Ruwet, N. (eds.) Generative Grammar in Europe, pp. 528–564. Reidel, Dordrechtz (1973)
37. Soames, S.: Understanding Truth. Oxford University Press, Oxford (1999)
38. von Stechow, A.: Comparing semantic theories of comparison. Journal of Semantics 3, 1–77 (1984)
39. von Stechow, A.: Times as degrees: früh(er) 'early(er)', spät(er) 'late(r)', and phrase adverbs. Unpublished manuscript, Tübingen (2006)
40. Williamson, T.: Vagueness. Routledge, New York (1994)

On the Psychology of Truth-Gaps*

Sam Alxatib[1] and Jeff Pelletier[2]

[1] Massachusetts Institute of Technology
[2] University of Alberta

Abstract. Bonini et al. [2] present psychological data that they take to support an 'epistemic' account of how vague predicates are used in natural language. We argue that their data more strongly supports a 'gap' theory of vagueness, and that their arguments against gap theories are flawed. Additionally, we present more experimental evidence that supports gap theories, and argue for a semantic/pragmatic alternative that unifies super- and subvaluationary approaches to vagueness.

1 Introduction

A fundamental rule in any conservative system of deduction is the rule of $\wedge$-Elimination. The rule, as is known, authorizes a proof of a proposition p from a premise in which p is conjoined with some other proposition q, including the case $p \wedge \neg p$, where p is conjoined with its negation. In this case, i.e. when the conjunction of interest is contradictory, $\wedge$-elimination provides the first of a series of steps that ultimately lead to the inference of q, for any arbitrary proposition q. In the logical literature, this is often referred to as the Principle of Explosion:

(1) $p \wedge \neg p$ (Assumption)
(2) p (1, $\wedge$-Elimination)
(3) $\neg p$ (1, $\wedge$-Elimination)
(4) $p \vee q$ (2, $\vee$-Introduction)
(5) q (3, 4, Disjunctive Syllogism)

Proponents of dialetheism view the 'explosive' property of these deductive systems as a deficiency, arguing that logics ought instead to be formulated in a way that preserves contradictory statements without leading to arbitrary conclusions. One such formulation is Jaśkowski's DL [8], an axiomatic system that is adopted as a logic for vagueness by Hyde [7]. Hyde's reformulation provides a semantics for DL that relies on a system of precisifications[1]: a predicate P is

* We thank Paul Egré, James Hampton, David Ripley, Robert van Rooij, Phil Serchuk, the organizers and audience at the ESSLLI 2009 Vagueness and Communication worshop and the ENS Vagueness and Similarity workshop, and the two anonymous reviewers for their help and insightful input. Many further issues that have not been fully discussed in this paper are detailed in Alxatib and Pelletier (forthcoming) [1]. The research for this project was partially funded by F. J. Pelletier's NSERC grant #5525.

[1] Precisifications were first used in van Fraassen's work on presuppositions (cf. [5]).

R. Nouwen et al. (Eds.): ViC 2009, LNAI 6517, pp. 13–36, 2011.

associated with a set of classically-constructed 'sharpenings' (precisifications), each of which delineates a precise boundary between P's extension and its anti-extension. The semantics of Hyde's system is then set up so that a predicate P is considered to hold of an individual a iff a belongs to P's extension in at least one precisification. This creates a system that preserves what may in other logics be seen as inconsistencies, for it now becomes possible for P to simultaneously hold *and* not hold of a single individual, as would happen when the individual, say a, belongs to P's extension in one precisification, and to its anti-extension in another. In this case, $P(a)$ is said to fall into a truth-value 'glut'; since truth/falsity in this logic requires truth/falsity in at least one precisification, $P(a)$ would be true *and* false when a belongs to P in some precisifications but not in all.

Logics like Hyde's are called 'subvaluationary' logics. The truth-value gluts that are characteristic of these systems stand in contrast with truth-value *gaps*, which emerge in the *super*valuationary systems of Fine [4] and Kamp [9]. In these logics, which are also intended as logics of vagueness, truth/falsity is defined as *super*-truth/falsity, where super-truth is truth in *every* precisification, and super-falsity is falsity in every precisification. So, if an individual a belongs to P's extension in some but not all precisifications, the statement $P(a)$ will not be assigned any truth-value, for it is neither true in every way of sharpening P, nor false in every way of sharpening P.

Both families of logics are built on top of a system of precisifications, and in both logics the individual precisifications are respectful of classical predicate logic: every predicate within a precisification has an extension, and the complement of this extension is precisely the predicate's anti-extension. No individual is left behind[2]. It is only when truth is defined as super/sub-truth that borderline cases show non-classical properties, namely having two truth-values in subvaluations, and no truth-value in supervaluations. Both frameworks, however, share with classical logic the rule of $\wedge$-elimination: if $p \wedge q$ is true in some sharpening, then $p \wedge q$ is sub-true, and since the sharpenings are classically constructed, the sharpening in which $p \wedge q$ holds is a sharpening in which p holds and q holds. It follows, then, that there is a sharpening in which p is true, and there is sharpening in which q is true. This makes both p and q sub-true, and therefore true. The same can be said of a supervaluationary system: if $p \wedge q$ is super-true, then every sharpening is such that $p \wedge q$ holds in it, and because every sharpening is classical, every sharpening will be such that p holds in it and q holds in it. So p and q will be super-true, and therefore true.

Our goal in this paper is to show experimental evidence for a pattern that violates $\wedge$-elimination, and to show further that this pattern can be accounted for if both the sub- and the super-valuationary approaches are used together. In

[2] Actually, in many formulations of supervaluations (like Fine's for instance) there is mention of 'incomplete' precisifications. A precisification of a predicate is incomplete if its extension together with its anti-extension do *not* exhaust the domain of individuals in the model. But in these formulations, it is usually added that only complete precisifications are considered when evaluating whether or not a proposition holds, and this has the effect of making the system maximally faithful to classical logic.

the course of establishing our argument, we intend to show that the observations which we think reconcile the two approaches pose a considerable challenge to the epistemic hypothesis proposed in Bonini et al. (BOVW). In Sect. 2 we lay out the relevant theoretical foundations and provide a very brief description of the Sorites paradox, and of the solution claimed by supervaluationists, subvaluationists, and epistemicists. In Sect. 3 we describe BOVW's experiment and their epistemic interpretation of the data, and we argue against their criticism of gap-theories. In Sect. 4, we describe the experiment conducted for this study and show how the results pose problems for BOVW's view, and discuss in detail our interpretation of the data. Finally, in Sect. 5 we show what we think is evidence against $\wedge$-elimination, and propose a unification of sub- and super-valuations to account for it.

2 Background

Vagueness is most famously characterized as a logical problem in Eubulides's Sorites Paradox. In contemporary literature, the paradox is often formulated as an inductive proof of a false statement like (1c) from two unobjectionable premises like those in (1a) and (1b).

(1) a. A man standing 190 cm is tall.
b. A man who is just a millimeter shorter than a tall man is also tall.
c. A man standing 100 cm is tall.

Most of those who tackle the paradox concern themselves with the inductive step (1b). Fuzzy logicians like Machina [12], for example, observe that when one is afforded with an infinite number of truth-values, one can choose to assign the inductive step a truth-value just short of complete truth (hence its near-acceptability). To see how this resolves the sorites, consider a rewording of the inductive step as a conditional: if n is tall then $n-\delta$ is tall (for some small change δ). In many fuzzy logics, the truth value of a conditional is 1 iff the consequent is at least as true as the antecedent, and otherwise,

$$V(p \rightarrow q) = (1 - V(p)) + V(q) \tag{1}$$

Returning now to the sorites conditional, if we assign to its antecedent the truth-value n and to its consequent the value $n - \delta$, the conditional will turn out to be $(1 - \delta)$-true – just under completely true. The reason is that $(1 - n) + n - \delta$ will be $1 - \delta$, regardless of the value of n. If one were to apply modus ponens to this conditional together with the basic (completely true) premise (a) of the sorites, modus ponens will license a conclusion that is $1 - \delta$ true. But if we repeat the process, the next application of modus ponens will produce a conclusion that is slightly less true ($1 - 2\delta$ true), and as we advance down the height spectrum the conclusions will gradually become less true, so that by the time we get to 100 cm the truth of the conclusion will be much closer to falsity than to truth.

Like the fuzzy logician, the sub-/super-valuationist takes issue with the inductive step of the sorites. But on her account, the inductive step turns out false. Recall that sub-/super-valuationary semantics refer to precisifications, which are classical constructions. The conditions on truth, whether subtruth or supertruth, make the inductive step of the paradox false, for in *no* precisification is it true that small changes in degree go unnoticed; in every precisification, every predicate has a precisely defined extension, so in every precisification there is an n for which $P(n)$, but for which $\neg P(n-1)$. Since the inductive step is false in every precisification, it is sub-/super-false, and since it is sub-/super-false, it is false. Note, furthermore, that the inductive step is not true in *any* precisification, so it can never be true even under the subvaluationist's 'weaker' requirements.

The inductive step of the sorites is considered false in another view of vagueness: the epistemic view. Advocates of epistemicism, like Williamson [19] and Sorensen [16,17], dismiss the need for non-classical logics in the treatment of vagueness. They insist, instead, that *in reality* there is for each vague predicate a precise boundary that divides its extension from its anti-extension, but that the location of this boundary is unknown. In its defence of classicality, the view is similar to the supervaluationary approach, but it differs in that it claims a single, albeit unidentifiable, precise boundary for every vague predicate. This is the view that Bonini et al. claim to find experimental support for.

3 BOVW's Experiment

3.1 Method

BOVW administered in-class questionnaires (in Italian) to 652 students in Italian universities in two between-subject experimental conditions: True and False. We will follow BOVW and refer to the True group as the 'truth-judgers' and the False group as the 'falsity-judgers'. The two conditions had approximately the same number of students. The objective behind the questionnaires was to find, numerically, the boundaries that their subjects thought appropriate for attributing a vague predicate to a given entity/event. Participants were presented with scenario-question pairs such as the following example, using the vague predicate *tall* and the dimension of height. The difference between the conditions is highlighted by the italicized text. (English translations are taken from BOVW's paper).

A. Condition: TRUE

> When is it *true* to say that a man is 'tall'? Of course, the adjective 'tall' is true of very big men and false of very small men. We're interested in your view of the matter. Please indicate the smallest height that in your opinion makes it *true* to say that a man is 'tall'.
>
> It is *true* to say that a man is 'tall' if his height is greater than or equal to ___ centimeters.

B. Condition: FALSE

> When is it *false* to say that a man is 'tall'? Of course, the adjective 'tall' is false of very small men and true of very big men. We're interested in your view of the matter. Please indicate the greatest height that in your opinion makes it *false* to say that a man is 'tall'
>
> It is *false* to say that a man is 'tall' if his height is less than or equal to ___ centimeters.

Other items included the following: *mountain* (in terms of elevation), *old* (in terms of a person's age), *long* (in terms of a film's length), *inflation* (in terms of percentage), *far apart* (as between two cities, in kilometers), *tardy* (for an appointment, in minutes), *poor* (in terms of income), *dangerous* (cities, in terms of crimes per year), *expensive* (for 1300cc sedan cars), *high unemployment* (in percentage with respect to a country), and *populous* (for an Italian city, in population). In our study, we focus only on the adjective *tall*.

The data were collected through a series of studies, each differing (sometimes only slightly) in choice of predicate. BOVW also ran a set of studies in which the words 'true' and 'false' were removed from the query. In these questionnaires the instructions were modified as in (C) and (D), and were given to different participant groups than (A) and (B) above.

C. Condition: TRUE

> When is a man tall? Of course, very big men are tall and very small men are not tall. We're interested in your view of the matter. Please indicate the smallest height that in your opinion makes a man tall.
>
> A man is tall if his height is greater than or equal to ___ centimeters.

D. Condition: FALSE

> When is a man not tall? Of course, very small men are not tall and very big men are tall. We're interested in your view of the matter. Please indicate the greatest height that in your opinion makes a man not tall.
>
> A man is not tall if his height is less than or equal to ___ centimeters.

Following BOVW, we will refer to queries like (C) and (D) as the non-metalinguistic queries, in contrast with the metalinguistic queries seen in (A) and (B).

Theoretical Predictions

Before we show BOVW's results, we take a moment to outline what might be predicted by advocates of the three approaches we are focusing on: subvaluationism, supervaluationism, and the epistemic view.

To the subvaluationist, borderline cases are cases that fall in truth-gluts. So in a subvaluationary world, when one asks about the minimal value n that makes n tall (or makes it true to say that n is tall), one is asking for the n above which it is *subtrue* to say that n is tall. By definition, borderline cases qualify, because it is subtrue to say that a borderline case is tall. Similarly, when one asks about the m below which it is *false* to say that m is tall, one is asking about the m below which it is *subfalse* to say that m is tall, and again, this will include borderline cases. The subvaluationist therefore predicts n to be lower than m, that is, the responses of BOVW's truth-judgers should come out lower than those of the falsity judgers.

The supervaluationist predicts the opposite. Truth in this framework is super-truth, and falsity is super-falsity. So the lowest n that makes 'n is tall' true is the lowest n that makes it supertrue, i.e. makes n tall in every precisification. This will place n just above the borderline range because borderline cases will be excluded, for they are not tall in every way of making tall precise. The highest m that makes it false (i.e. superfalse) to say 'm is tall' will, for the same reasons, also exclude the borderline cases, and will land just below the borderline range. The prediction, then, is that the responses of the truth-judgers take a greater value than the responses of the falsity-judgers.

It is not entirely clear what the epistemicist might predict here, at least if he does not augment his view with one or more auxiliary assumptions. Indeed, BOVW add to their epistemic hypothesis an assumption that makes their predictions converge with those of the gap-theorist. We return to this after we show their findings.

3.2 Results

For almost every predicate they tested, BOVW find the average of the values provided by the truth-judgers to be significantly higher than that of the values provided by falsity-judgers. In the case of *tall*, for example, they find that the minimum height that makes a man tall – or makes it *true to say* that a man is tall – is higher than the maximum height that makes him not tall – or *false to say* that he is tall. The results from four of their six studies are shown in Table 1.[3]

While these findings contradict the predictions of glut-theories of vagueness, they seem to stand in support of gap-theories. Surprisingly, however, BOVW reject the gap account and instead promote the following epistemic hypothesis:

> Vagueness as Ignorance: S mentally represents vague predicates in the same way as other predicates with sharp true/false boundaries of whose location S is uncertain.

The reason that gaps appear, according to BOVW, is that speakers are in general more willing to commit errors of omission than commit errors of commission. In

[3] The predicate 'tall' was not used in their Study 3. Study 6, which did include 'tall', made explicit reference to the middle range, and is therefore excluded from the present discussion.

Table 1. Truth- and Falsity-judgments for 'n is tall' (from BOVW)

	Study 1	Study 2	Study 4	Study 5
Truth-judgers	178.30 cm	179.55 cm	181.49 cm	170.28 cm
Falsity-judgers	167.22 cm	164.13 cm	160.48 cm	163.40 cm

other words, speakers would rather withhold the application of a predicate to an individual with an uncertain degree of membership than incorrectly ascribe the predicate to an individual of whom the predicate might not hold.[4] As a result, truth-judgers will provide the lowest value that they *confidently* think the predicate in question applies to, and falsity-judgers, likewise, will provide the greatest value that they *confidently* think the predicate does not apply to. The former value will of course turn out greater than the latter, and thus gaps emerge with *all* predicates, not just the ones that are usually seen to be vague.

The grounds on which BOVW reject the gap hypothesis, which otherwise seems a natural consequence of their empirical results, are predominantly theoretical. Their main points of criticism of gap theories are (1) that gap-theories do not offer an elegant account of higher-order vagueness, and (2) that, when examined in light of their data, gap theories lead to contradictory statements. We evaluate each of these grounds in turn.

Higher-Order Vagueness. Higher-order vagueness is the phenomenon that seems inevitable whenever one proposes that there is a 'gap' between the extension and the anti-extension of a predicate. For example, if one wishes to propose that, because there is no sharp cutoff line between the bald and the not-bald men, there must be a gap between the bald men and the not-bald men, filled by borderline-bald men, it seems impossible to then try to justify a sharp cutoff line between the bald men and the borderline-bald men either. Nor, on the other side of the gap, between the borderline-bald men and the not-bald men. So, there should be borderline cases of borderline cases: a 'second order vagueness'. But once a theorist starts down this path, it seems not possible to stop at all: there will be all levels of higher-order vagueness. Any rationale that could be given to stop at some particular high-order could have been used to not admit of the original first-order gap.

It does not seem like an easy task for the supervaluationist to provide an account of higher-order vagueness, since the framework, as we described it at least, allows three *sharp* possibilities: true, false, and neither. But Keefe [10] proposes the following maneuver: suppose borderlineness were to apply not only to the predicate itself, e.g. *tall*, but also to the admissibility of the way the predicate is made precise. When the admissibility of the precisifications is subject to borderlineness, one can imagine some individual a who is tall in every admissible precisification, but who is not tall in some precisification s of borderline

[4] Based on studies by Ritov and Baron [15] and Spranca et al. [18].

admissibility. In this case, we cannot say that a is super-tall, for he can only be super-tall if we ignore s, and we can only ignore s if it was *in*admissible. But we cannot say that a is borderline either (gappy that is), for that requires that a be not tall in some admissible precisification, and s is not quite admissible. This makes a a borderline-borderline case (2nd-order vagueness). If further conditions are imposed in higher metalanguage(s) on, say, the admissibility of admissibility, then finer gradations become more visible in the system, for that makes room for borderline-borderline-bordline cases, etc. We refer the reader to Keefe for more details.

BOVW's problem with this approach, and one of their reasons for rejecting gap theories, is that 'the mental representation of all these vague boundaries seems psychologically implausible' (p. 388). They add, furthermore, that if the ascent to higher orders of vagueness is stopped, the blur surrounding the gappy region will be replaced with a sharp line, and 'there is no introspective evidence for such a line' (also p. 388).

We officially suspend judgement on the issue of psychological plausibility. But we object to the way BOVW use introspection as a test of acceptability of a semantic theory. We note, as they do also, that there is no introspective evidence for the sharp but unknown divider that is presumed by their epistemic theory, a charge that BOVW address by saying that 'other semantic/conceptual principles have been plausibly ascribed to people who do not reliably acknowledge them' (pg. 387). So in considering the very same feature that their theory shares with an opposing theory, they happily cite this principle to defend theirs but will not consider it as a possible defense of the opposing theory. We think, therefore, that these 'psychological arguments' they use to favor their hypothesis and reject gap-theories are inconsistent.

The Absurdity of Denying Bivalence. BOVW begin their second argument against gap-theories by claiming that no difference was detected between the size of the metalinguistic gaps and the size of the non-metalinguistic gaps. Recall that BOVW used two survey styles, in one inquiring about the n for which it was *true* to say that predicate P holds of an individual (the metalinguistic questionnaire), and in the other inquiring about the n that makes an individual P (no mention of truth – the non-metalinguistic questionnaire). The comparisons for *tall* are shown in Table (2).[5]

If BOVW are right, the gap-theorist has to admit that the truth-conditions for 'n is tall' and ' "n is tall" is true' are the same, and similarly for 'n is not

[5] Indeed there seems to be no significant difference between the metalinguistic truth-judgements and the non-metalinguistic ones, but it is questionable whether the same holds of falsity-judgements; the average of the n for the metalinguistic falsity-judgers – taken as the average of Studies 1 and 2 – is 165.68 cm. For the non-metalinguistic studies, 4 and 5, the average comes to 161.94 cm. The difference between the two is 3.74 cm, which is almost 30% of what subjects, on average, claim to be the difference between '"x is tall" is true' and '"x is tall" is false'. It thus seems quite likely that there is a significant difference between metalanguage falsity and object language negation. We continue our reply, however, as if this difference was insignificant.

Table 2. Comparison of BOVW's metalinguistic and non-metalinguistic judgements

	Truth-judgements	Falsity-judgements
Metalinguistic (Studies 1, 2)	178.30 cm; 179.55 cm	167.22 cm; 164.13 cm
Non-metalinguistic (Studies 4, 5)	181.49 cm; 178.28 cm	160.48 cm; 163.40 cm

tall' and ' "n is tall" is false'. But BOVW argue against the viability of this position for gap theorists, as follows. Suppose height n is borderline tall. On a supervaluationary account, the statement 'n is tall' will have no truth value, that is, 'n is tall' is not true and 'n is tall' is not false. They give the following argument (pp. 388–389) to show that this cannot be correct (they wish the $\equiv$ to be read 'has the same truth conditions as'):

(1) 'n is tall' is not true (assuming n to be borderline)
(2) 'n is tall' is not false (assuming n to be borderline)
(3) n is tall $\equiv$ 'n is tall' is true (as shown by their experimental results)
(4) n is not tall $\equiv$ 'n is tall' is false (as shown by their experimental results)
(5) n is not tall $\equiv$ 'n is tall' is not true (from equivalence (3))
(6) n is not not tall $\equiv$ 'n is tall' is not false (from equivalence (4))
(7) n is tall $\equiv$ 'n is tall' is not false (double-negation in (6))
(8) n is tall (from assumption (2) and equivalence (7))
(9) n is not tall (from assumption (1) and equivalence (5))
(10) n is tall and n is not tall (conjunction of (8) and (9))

Since (10) is contradictory, and furthermore goes against the anti-glut findings of BOVW's experiments, the assumptions (1) and (2) must be revised. But these assumptions are the very ones that define the supervaluation position! So, unless there has been a mistake in the reasoning that got us from these two assumptions and the experimental results, it appears that supervaluation theory has been refuted.

We think that a supervaluationist could legitimately complain about the inferences involving negation in BOVW's proof. Before we discuss this, we point out that the proof need not be explained in full in order to understand how the alleged absurdity arises; one need only look at (4) and (5) to see the problem: (4) and (5) have the same proposition to the left of the '$\equiv$' symbol, but they each describe a different state of affairs on the right side of '$\equiv$'. In (4), 'n is not tall' is claimed to have the truth-conditions that make 'n is tall' false, but in (5), 'n is not tall' is claimed to have the conditions that make 'n is tall' not true. This is trouble for the gap-theorist because in her theory the conditions that make 'n is tall' false are different from those that make it not true; 'n is tall' is false whenever it is superfalse, but it is not true whenever it is either false or neither true nor false. The two scenarios cannot *both* be said to have the same

truth-conditions as 'n is not tall', precisely because they describe different truth-conditions. If BOVW can show that the gap-theorist is forced to accept (4) and (5), their argument succeeds.

But the gap-theorist is not forced to accept (4) and (5) in the way intended by BOVW. (5) is derived from the equivalence in (3), which says that whenever n is tall, 'n is tall' is true. From this, it follows that whenever n is not tall, 'n is tall' is not true. (5), then, is to be understood as saying that whenever n is *anything but tall*, the sentence 'n is tall' is not true. Now if we turn our attention to (4), it simply says that, based on empirical evidence, the gap-theorist ought to say that the sentence 'n is tall' is false whenever n is not tall. In order for their argument to be convincing, BOVW must force the gap-theorist to say that 'not tall' in this context also means *anything but tall*, just like it does in (5). In other words, BOVW seem to be saying that, in order for the gap theorist to make her theory match the empirical findings, she must say that 'n is tall' is false iff n is *anything but tall*. But this is something that BOVW cannot do; the gap-theorist can respond by saying that 'n is not tall' in (4) means 'n is super-not-tall'. If the negation in the left-side of (4) is assigned a strong interpretation, the problem for the gap-theorist described in the previous paragraph disappears.[6]

Essentially, the gap-theorist's escape is to say that negation can have two interpretations: 'n is tall' is false whenever n is *strong*-not tall ('choice' negation), and 'n is tall' is not true whenever n is *weak*-not tall ('exclusion' negation). Ultimately, we will favor an account where negation is treated unambiguously in the semantics, but where its different interpretations arise from pragmatic principles (we refer the reader to Horn [6] and Levinson [11] for discussions of the inferences involving negation). For now, however, we use truth-tables (Table 3) merely to illustrate the difference between the strong/choice and the weak/exclusion interpretations of negation.

Table 3. Strong/Choice negation ($\sim$) and Weak/Exclusion negation ($\neg$)

φ	$\sim\varphi$	$\neg\varphi$
T	F	F
G	G	T
F	T	T

It can now be seen that, with the distinction between negations in place, the conclusion in (10) loses its contradictory reading; (10) becomes the proposition that (the borderline case) n is tall, in the sense of 'not *un*tall', as it were, and not tall, in the sense of weakly-not tall, or not definitely tall. In other words, the conditions under which (10) holds are the very conditions that make *tall* and *not tall* subtrue. The reader is invited to verify this claim.

[6] Further discussion against this and related 'logic of the argument' is given in more detail and against a wider group of similar arguments, in Pelletier and Stainton [13].

BOVW's concern, then, is that by virtue of its sub-tallness and sub-not-tallness, n can be said to be tall and not tall, which is contradictory. In Sect. 5, we reveal some empirical evidence not only that this 'contradictory' conclusion is often judged true, but that its conjuncts are often considered false at the same time. Before we get to that, however, we describe the experiment and discuss the findings that we think are problematic for the epistemicist.

4 Experiment

Participants. 76 undergraduates from Simon Fraser University participated in this study. 59 participants classified themselves as fluent English speakers, 10 as advanced, and 5 as intermediate (leaving 2 participants, who left the question unanswered).

Method. Participants were presented with an image of five suspects in a police line-up (Fig. 1). The suspects were shown with the following heights in pseudo-randomized order: 5′4″, 5′11″, 6′6″, 5′7″, and 6′2″.[7] The suspects were labeled with numbers on their faces, and were referred to by these numbers in the experimental material.

Participants were given a paper and pen questionnaire (on a separate page from the image) with five sets of four statements (one for each suspect). Each statement had three labeled checkboxes to the right. An example is included in (2) for suspect #1.

(2)

#1 is tall	True ❑	False ❑	Can't Tell ❑
#1 is not tall	True ❑	False ❑	Can't Tell ❑
#1 is tall and not tall	True ❑	False ❑	Can't Tell ❑
#1 is neither tall nor not tall	True ❑	False ❑	Can't Tell ❑

Before the survey was handed out, the participants were given the following instructions:

- You will be asked to describe the heights of the five suspects in the line-up shown below.
- Please use the height standards of adult males in present-day North America.
- This is not a test, and there are no correct answers. Upon reading the questions, simply check the first answer that pops in your head and seems to describe the situation as you see it.

In order to minimize the effect of order on the subjects' responses, each sheet was printed with the questions randomly ordered. This was done in every copy of the survey, so no two copies had the same order of questions.

[7] Both the metric measurement system and the imperial system are in common usage in western Canada.

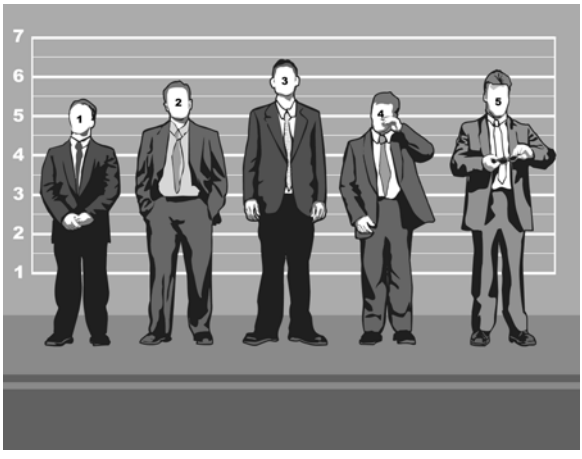


Fig. 1. Suspects of Different Heights in Police Lineup

Results and Discussion. Our reply to BOVW draws particularly on the responses to the first two statements. Later, in Sect. (5.1), we consider the other two sentences, in the course of presenting our own position. In Fig. 2, the percentages for *true* responses to *X is tall* are shown to increase with height, starting with 1.3% at 5′4″, reaching the median value of 46.1% at 5′11″, and peaking at 98.7% at 6′6″.

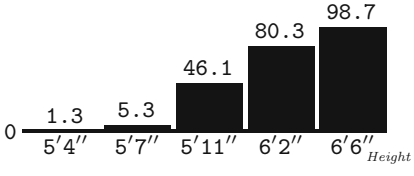


Fig. 2. % of 'True' responses to 'X is tall'

Conversely, the percentage of *false* responses, seen in Fig. 3, begins with a ceiling of 98.7% at 5′4″ and drops to 1.3% at 6′6″, passing the median at 5′11″ with a value of 44.7%.

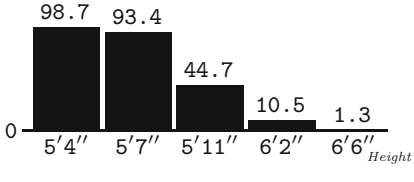


Fig. 3. % of 'False' responses to 'X is tall'

Figure 4 shows the percentage of *true* responses to *X is not tall*, which also reaches the median at 5′11″, this time at 25.0%, and peaks at 5′4″ at 94.7% and drops to 0.0% at 6′6″.

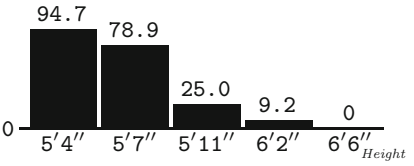


Fig. 4. % of 'True' responses to 'X is not tall'

The percentage of *false* responses to *X is not tall* is shown in Fig. 5: 3.9% at 5′4″, a median of 67.1% at 5′11″, and a maximum of 100.0% at 6′6″.

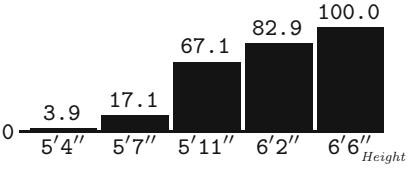


Fig. 5. % of 'False' responses to 'X is not tall'

It is the difference between these sets of answers that is problematic for the BOVW account. The numbers indicate a significant preference for rejecting a proposition over accepting its negation.[8] In classical logic, the statement 'a is tall' is true just in case its negation, 'a is not tall', is not true, and vice versa. But in a gap theory like supervaluations, the statement 'a is tall' is true if it is supertrue, and otherwise it is not true. The prediction, then, is that if a is borderline, the statement 'a is tall' is judged false more frequently than its negation 'a is not tall' is judged true, the reason being that the latter statement only holds if it is supertrue, which would not be the case if a was borderline. Similarly, a gap theory would predict more false responses to 'a is not tall' than true responses to 'a is tall'.

Here we see an immediate objection: falsity in supervaluations is *super*falsity, and this disqualifies the tallness of borderline individuals from being false. A supervaluationist should not expect a preference for 'False' responses any more than a preference for 'True' responses when it comes to a borderline case. Strictly speaking, this objection is accurate. But of course, this would not be an issue if the checkboxes in our questionnaires were instead labeled 'True', '*Not true*' (instead of 'False'), and 'Can't tell'. For 'Not true' would surely include the borderline range for the gap theorist. But now suppose that a participant was in disagreement with a statement, and the only three options (as in our questionnaire) were 'True', 'False', and 'Can't tell'. We find it quite reasonable to expect the participant in this case to check 'False', since among the available answers, 'False' is the only plausible substitute for 'Not true'. We feel, therefore, that it

[8] According to a χ^2 test for independence, the chance of the difference (between denial and assertion) in the case of #2 being drawn from the same distribution is less than 5%: $\chi^2(2) = 8.22; p < 0.05$.

is legitimate to interpret 'False' as a sign of rejection in our set-up, but we certainly allow that this unsupported claim needs to be bolstered by experimental evidence.

The data, as shown in Figs. 2–5, confirms the preference for 'False' responses. For suspect #2 ($5'11''$), our borderline poster-child, 46.1% thought it was *true* that he was tall, while 67.1% thought it was *false* that he was not tall. Similary, 25.0% thought it was true that he was not tall, whereas 44.7% thought it was false that he was tall. Both comparisons show that a significantly bigger sample of participants rejected the statement '#2 is tall' (or 'not tall') when compared to the sample of those who accepted the classical negation of each statement.

BOVW might claim that this could as easily be taken as support for their epistemic hypothesis. Recall that BOVW assume that errors of commission are considered by their participants to be graver than errors of omission. Thus the subjects prefer to withhold judgement regarding uncertain cases than incorrectly attribute the predicate to them. If our participants (as we claim) would rather reject a statement (by judging it false) than accept its negation (by judging the negation true), can we not interpret this preference also as a way of favoring errors of omission over errors of commission? If this interpretation of the data is available, then the evidence that we find supportive of gap theories can also be taken to support the epistemic hypothesis (together with BOVW's auxiliary assumption regarding error preferences). In response to this concern, we point out that our subjects were also given the option of checking 'Can't tell', but very few people chose to answer that way: for the statement 'x is tall', where x is $5'11''$, there were 44.7% false responses, and 9.2% 'Can't tell' responses; for 'x is not tall', at the same height, there were 67.1% false responses, and 7.9% 'Can't tell's.

Of course, one may also object that it is possible for the participant to have taken 'Can't tell' as meaning something like 'I give up', thus accounting for the low rate of 'Can't tell' responses (because we cannot expect our subjects to comfortably choose this way of answering). In this picture, the fact that there is a preference for falsity-judgement over truth-judgement may after all be due to a preference of omission errors over commission errors, and so this part of our argument against BOVW is not convincing. Evidence against this interpretation is available elsewhere in our data, however. For, if we maintain vagueness-as-ignorance and combine it with this error-preference pattern, we should expect these same falsity-judgers (who are choosing to answer safely, as it were) to also prefer answering 'False' for the apparently contradictory statement '#2 is tall and not tall'. After all, the epistemic theory is classical, so it should predict virtually *no* 'True' responses to this statement. But as we will show below in Sect. 5, subjects seem happy to claim that this statement is true.

In a last-ditch attempt to save epistemicism, such a theorist may say that our last considerations cannot be taken as a counterargument to the vagueness-as-ignorance hypothesis because, the theorist might say, speakers need not be *aware* of their ignorance. This reply is not relevant here. What is relevant is that if

errors of omission are indeed preferred to errors of commission, which is an assumption that the epistemicist requires, then we would expect a much larger number of 'Can't tell's, since this is the least committing answer with regards to borderline (or uncertain) cases.

We now return to the use of negation in this experiment and its role in the semantic/pragmatic account that we favor. Earlier we argued that BOVW were mistaken in assuming that only one type of negation could be understood in statements like 'a is not tall'. This assumption led them to conclude that ' "a is tall" is false' held under the same conditions as ' "a is tall" is not true', since both metalinguistic statements were 'equivalent' to 'a is not tall'. In response, we suggested that two interpretations are available for 'a is not tall': one in which the negation is identified with choice/strong negation (in which case 'a is not tall' holds if it 'super-holds'), and another in which the negation is identified with exclusion/weak negation (in which case the statement holds just in case 'a is tall' does *not* super-hold)[9]. A natural question that one can ask at this point is: which of these two types do we think arises when we present our participants with the statement 'X is not tall'? Surely, if the negation was interpreted as weak negation, then there should not be a significant difference between accepting the statement '#2 is not tall' and rejecting the statement '#2 is tall', since '#2 is not tall' (where 'not' is weak) would hold in the same set of circumstances that makes '#2 is tall' not hold. But since we do find a significant preference to deny the former, it would seem that the negation is interpreted as strong, and we should explain why.

We think that pragmatic factors contribute to the emergence of what resembles a strong/choice interpretation for 'not'. We begin our explanation of the pragmatic effects by inviting the reader to consider the following scenario: suppose John and Mary have a single friend named Lucy. Lucy is looking for a date, and John and Mary suggest that she meet their friend Bill. Suppose further that Bill is of average height. Now Lucy asks their friends about Bill's looks, and in response, Mary provides a few answers, one of which being 'he's not tall'. Here we find it felicitous of John to object to the way Mary described Bill's physical stature, and say in response: 'Well, he's not *not tall*. He's average.' The felicity of this interaction suggests that two different logical interpretations of

[9] We wish to note here that there is also room for interpreting negation as intuitionistic negation. The intuitionistic negation of p, $-p$, is true iff p is *false*, and is false when p is true or, on a gap-theoretic interpretation, when p is neither true nor false. In singly-negated statements, intuitionistic negation converges with choice/strong negation, since both assign the value True to $-p$ whenever p is false. However, intuitionistic negation differs from choice/strong negation in doubly-negated statements: $--p$ is true whenever $-p$ is false, and $-p$ is false iff p is true or truth-valueless. At first glance, this seems desirable, since we can derive the 'not untall' reading of 'not not tall' using only one definition of negation, and also derive the strong interpretation of negation in singly-negated expressions. But the consequence of having only intuitionistic negation in the language is that singly-negated expressions can *never* be given a weak interpretation. In Sect. 5.2 we show an example in which a weak interpretation of single negation is required (see Footnote 17).

'not' are involved, for otherwise John's comment would merely be equivalent to 'he's tall'.

If we assume that Bill is of average height, and if we are considering a gap-theoretic system, then the statement 'He is tall' will be neither true nor false. So, when Mary says 'He's not tall', if she is to be speaking truthfully[10] she must be intending a weak/exclusion interpretation of 'not', for that is the only negation that will convert an 'other-valued' statement into a truth. When John tries to correct Mary, or correct the impression left by Mary's statement, he takes what Mary said, with the truth value thus computed, and negates that. In this case, John is either negating a 'true' (if Mary was using weak/exclusion negation) or negating a 'other' (if Mary was using strong/choice negation). Note in these cases that if Mary were using weak/exclusion negation and were understood to be using weak/exclusion negation, then no matter what negation John is using to negate that, what he says is false, because all of the negations would take Mary's true into a false. We presume this is not right, since John is imagined to be speaking truthfully. From this it follows that, even if Mary were speaking truthfully (by using weak/exclusion negation), she could not have been understood that way. So it seems that John is taking Mary to be using strong/choice negation and he is denying the 'other' value to Mary's claim. This would mean that John was using weak/exclusion negation, since that is the only negation that maps an 'other' value to truth.[11]

So it seems that 'not' can be interpreted in a way akin to strong/choice negation in some cases, and to weak/exclusion negation in others. In order to provide a complete account of how negation is used, particularly with vague predicates, one must offer a description of the situations in which the strong/choice interpretation arises, and those in which the weak/exclusion interpretation arises. Here we follow Levinson (among others) and invoke familiar pragmatic principles: when we, as experimenters, present a group of participants wih questions or statements that contain negated (or even unnegated) vague expressions, we feel it reasonable to assume that these expressions are being interpreted by the participants with sufficient observance of the Gricean maxims, in particular the maxim of quantity. If it is also assumed by our participants that we intend for this principle to be observed by them, then we would expect that by '(not) tall' the participants will understand that we want them to pick up on the most informative reading possible, which to the participant must correspond to that definition of '(not) tall' which s/he thinks all (or most) people would agree upon and, also, that s/he assumes that we, the experimenters, think all (or most) people would agree upon (assuming, of course, a fixed context of use, comparison class, etc.). The closest match to this description is the super-interpretation, i.e.

[10] 'Speaking truthfully' here is to be understood as making a *semantically* true statement. The statement might not accord with various Gricean restrictions and therefore might not be a *pragmatically* felicitous statement.

[11] Here it is also possible to understand John as using his own negation twice, in which case it may be that John is in fact using intuitionistic negation. This possibility was brought up and discussed briefly in Footnote 9.

that 'is (not) tall' is read as 'is *super*-(not)-tall'. So, when the question arises as to whether a person standing 5′11″ is tall (or not tall) the addressee – who may reasonably be expected to comply with the Gricean principles – is very likely to say 'False'. In the next section we present more experimental findings (Sect. 5.1) and offer a more detailed theoretical account in which the maxims of quality and manner are involved (5.2).

5 Contradictions and Borderline Cases: Gaps vs. Gluts

In this section we turn to statements in our questionnaire that until now we have ignored: 'x is tall and not tall' and 'x is neither tall nor not tall'. The relevant data are by no means indicative of a knock-down argument in favor of any particular theory, but the implications they carry can be of great importance for the gap theorist as well as the glut theorist, and we intend to use them to further clarify and extend our account of the data we have already presented.

5.1 Data

Figures 6 and 8 show that the numbers of *true* responses to each of these statements, which we will call *both* and *neither*, increased when the suspect's height was closer to average, peaking at 44.7% and 53.9%, respectively, for the 5′11″ suspect.[12] The number of *false* responses followed a complementary pattern, *decreasing* as the heights approached 5′11″ and reaching a minimum of 40.8% and 42.1% at that midpoint, as shown in Figs. 7 and 9. Note that there are more subjects who say *true* to '#2 is tall and not tall' than say *false* to it. Note also that more subjects say *true* to '#2 is neither tall nor not tall' than say it is *false*.

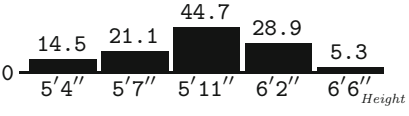


Fig. 6. % of 'True' responses to 'X is tall and not tall'

Particularly interesting, however, is how the two statements, *both* and *neither*, correspond with one another. The responses for the two questions are cross-tabulated in Table 4. Note that the response types are subscripted with the

[12] One may question the reliability of 'True' responses to the contradictory statement here. For example, it may well be that (relative) abundance of 'True' responses to *both* is due to a simple yes-bias. Regarding this concern, however, we find it unlikely for a yes-bias to increase the number of 'True' responses to the *both* statement and not to its individual conjuncts. This of course deserves further investigation. Experimentally, one could compare the frequency of truth-judgements to a statement like 'tall and not tall' to a stronger one like 'definitely tall and definitely not tall'. If we find significantly fewer truth-judgements to the latter, our findings will no doubt be more informative. We thank James Hampton and Lawrence Goldstein for bringing up this issue.

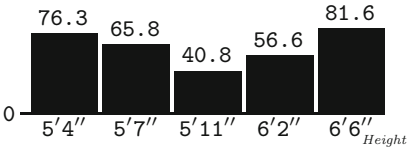


Fig. 7. % of 'False' responses to 'X is tall and not tall'

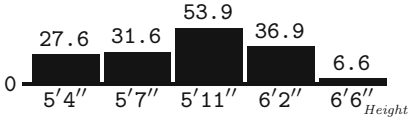


Fig. 8. % of 'True' responses to 'X is neither tall nor not tall'

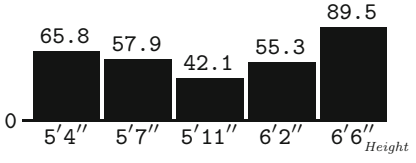


Fig. 9. % of 'False' responses to 'X is neither tall nor not tall'

relevant question: T_b, for example, is the number of truth-judgers for '#2 is tall and not tall'; F_n is the number of falsity-judgers for '#2 is neither tall nor not tall'.[13] What we want to highlight is that *neither*, whose truth can justify a truth-value gap, coincides in many cases (more than half!) of borderline-height with *both*, which, when true, suggests a truth-value glut.

Another interesting correlation is the one found between the questions 'x is tall' and 'x is not tall' on the one hand, and 'x is tall and not tall' on the other. Figure 10 shows that 32.4% of those who thought it was true that #2 was 'tall and not tall' also thought it was *false* that he was tall and *false* that he was not tall. Figure 11 illustrates the correlation in the other direction; it shows the percentage of *true* responses to 'x is tall and not tall' when the statements 'x is tall' and 'x is not tall' are judged false. The ratio is 68.8% at 5′11″, and 100% at 6′2″.[14]

There are other interesting findings that center around our borderline suspect #2. For example, we find the number of subjects who think '#2 is tall' and '#2 is not tall' are both true to be much higher than those who think they are both false (Table 5).

[13] A Bowker's test for symmetry gives an X^2 value of 8.04. With $df = 3$, the tail probability $p < 0.05$.

[14] We think the anomalous value of 100% for our 6′2″ subject is due to the fact that only four subjects thought that both 'tall' and 'not tall' were false for this subject. And all of these four thought #5 was 'tall and not tall'.

Table 4. Cross-tabulation of 'neither' and 'both' (Height = 5′11″): Response types for 'tall and not tall' are subscripted with b (for 'both'), and those for 'neither tall nor not tall' are subscripted with n. For example, the number of truth-judgers for both questions is in the cell where the T_b row intersects with the T_n column.

	T_n	F_n	C_n	
T_b	22	12	0	34
F_b	13	18	0	31
C_b	6	2	3	11
	41	32	3	76

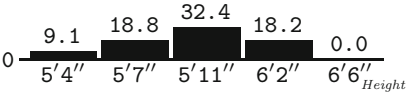


Fig. 10. % of Falsity of 'tall' and 'not tall' when *both* is true

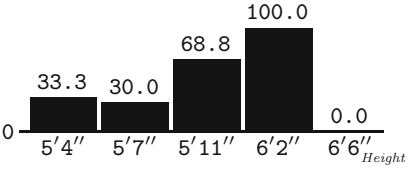


Fig. 11. % of Truth of *both* when 'tall' and 'not tall' are false

5.2 Analysis and Implications

Our goal in this section is to suggest a possible explanation for the pattern that we have just demonstrated: the pattern where 'is tall' and 'is not tall' are both considered false (when they are about a borderline individual), but where 'is tall and not tall' and 'is neither tall nor not tall' are considered true of that same individual.

Our idea, as we promised, relies crucially on the Gricean maxims of conversation. However, the solution also relies on an assumption that may seem somewhat controversial: that a given vague predicate has two possible interpretations, a *super*-interpretation and a *sub*-interpretation, in the same way that a vague expression containing negation can be interpreted strongly (i.e. super-interpreted), or weakly (i.e. sub-interpreted). Assuming this, together with the Gricean maxims, provides a way of accounting for the seemingly inconsistent patterns outlined above.

Our intuitive semantic theory is rather standard and classical, so far as our inclusion of vague and ambiguous predicates allows. Given a domain $\mathcal{D}$ of individuals,

the extension of a non-vague predicate[15] is interpreted normally, as some subset of $\mathcal{D}$ (the ones that manifest the property). The negation of a predicate is simply its complement, relative to $\mathcal{D}$ (note that we are assuming an unambiguous definition of negation here). A predicate that is vague but not ambiguous is represented as a set of ordered pairs, the first member of which is an (admissible, classical) precisification of the predicate and the second is the subset of $\mathcal{D}$ that satisfy the predicate in that precisification. The extension of that predicate is always relative to one or a group of the precisifications, and then becomes the subset of $\mathcal{D}$ that obeys that restriction on the precisifications. A (two-way) ambiguous predicate is interpreted as having two members, each one of which is an interpretation of the former types. One kind of meaning that a vague predicate can manifest is what we have intuitively called 'the super-interpretation': its extension is the subset of $\mathcal{D}$ of things that occur as values in *every* precisification. Another is what we called 'the sub-interpretation': the subset of $\mathcal{D}$ that appear as values of some precisification.

Table 5. Percent of Ss who gave same answer to both '#2 is tall' and '#2 is not tall'

#2 is tall	#2 is not tall	percent
T	T	3.9%
F	F	21.0%
C	C	5.3%

Our view here is that a vague predicate such as 'tall' can be ambiguous between the super- and sub-interpretations. A hearer is to find the more suitable interpretation of the predicate from these two possible meanings, or just to say that there is no way to choose and the sentence is simply ambiguous. The Gricean Maxim of Quantity can then be a condition on which one of these sets should be selected from the interpretation of the predicate. The condition is that the selected set may not be a superset of any other member of the set.[16] For 'tall', the result will be the set of the super-tall people, since it is the only set, from the two available options, for which the condition holds. For 'not tall', the two possibilities are the complements of the super-tall set and the sub-tall set, since 'not' unambiguously denotes the set-complement operation in this conception (complement with respect to $\mathcal{D}$).

So, the set of available interpretations will contain both the complement of the super-tall individuals, and the complement of the sub-tall individuals. The former set, the complement of the set of super-tall individuals, is the set of the individuals that are not super-tall, i.e., the borderline cases and the definitely not-tall cases. The latter set, the complement of the set of sub-tall individuals,

[15] We would extend this to n-place relations, but for the present paper we stick to monadic predicates.

[16] We see this as an application of the Strongest Meaning Hypothesis of Dalrymple et al. [3].

will contain individuals that are not sub-tall, that is, every individual *except* those that belong to the extension of tall in some sharpening. In other words, the set will contain the individuals for whom there are *no* sharpenings in which they belong to the extension of tall, and since each sharpening is classical, they are precisely the individuals that are super-not-tall. (Another name for this set might be the super-short individuals). So, the two possible interpretations for 'not tall' are the set of borderline-cases together with the super-not-tall cases (from the complement of the super-tall set of individuals), and the set of super-not-tall individuals (the complement of the sub-tall individuals, the super-short ones). And according to the condition of quantity, the latter set is selected since there are no subsets of itself that belong to the collection of interpretations. Formulated this way, the Maxim of Quantity will favor the super-interpretation both for 'tall' and for 'not tall', making it seem that negation is choice negation when it is really the effect of these pragmatic operations.

We now show how 'tall and not tall' might be made to mean 'borderline' on this approach. The set of interpretations will contain four elements, each of which resulting from the intersection of two sets (through the denotation of 'and'). The four elements are shown in (3). Note that, of the four options, only (3c) can be nonempty.

(3) a. $\{x : x \text{ is supertall}\} \cap \{x : x \text{ is supertall}\}^- = \emptyset$
b. $\{x : x \text{ is supertall}\} \cap \{x : x \text{ is subtall}\}^- = \emptyset$
c. $\{x : x \text{ is subtall}\} \cap \{x : x \text{ is supertall}\}^- \neq \emptyset$
d. $\{x : x \text{ is subtall}\} \cap \{x : x \text{ is subtall}\}^- = \emptyset$

(3a) and (3d) are empty because in both cases the intersection is applying to a set and its complement. (3b) is empty because the individuals it will contain are those that are tall in every precisification and at the same time not tall in any. Now, if in the formulation of the maxim of Quality one can block readings that are trivially false, then the maxim will allow only (3c) to emerge from the four options in (3). 'Tall and not tall' can therefore only denote the set of individuals who are sub-tall and sub-not-tall, i.e. the borderline individuals.[17]

Finally, in the case of 'not not tall', there are also two available interpretations: for 'tall' as super-tall we get the complement operation canceling itself, by applying twice, and yielding the set of super-tall individuals, and likewise, for 'tall' as sub-tall, we get the set of sub-tall individuals. Of the two options, it is the set of super-tall individuals that will qualify, and so we predict, incorrectly, that 'not not tall' means super-tall. But here we may add that the Gricean maxim of Manner, which penalizes prolixity, will block the super-interpretation, for if the super-interpretation was intended, the speaker would have had no reason to use 'not not' in his/her locution, but rather would say simply 'is tall'. It

[17] This is where intuitionistic negation fails to make the correct prediction: the negated conjunct is negated only once, so it can only be interpreted strongly. But in order for the conjunction to denote a non-empty set, the negation has to be interpreted weakly.

is difficult to precisely formulate a mechanism that blocks candidate interpretations on the basis of brevity, but as the maxim has generally proven useful in the theory of pragmatics, we feel it innocuous to invoke it for our purposes, hoping that however it can be made formal, it can be utilized to disqualify the super-interpretation from entering the set of denotations for 'not not tall', and instead interpreting the predicate as sub-tall.

We wish to emphasize that our theory does not suddenly use both super- and sub-valuationist interpretations in an ad hoc manner merely for the special case of apparently contradictory statements. Indeed, the assumption that they are both in play is not specific to borderline cases at all. Rather, we suggest that the interpretation of a vague predicate, regardless of whether or not the property is being predicated of a borderline individual, can be modeled using sets of precisifications, which is the architecture that both super- and sub-valuationists share. Where the two approaches diverge is in the use of the quantifier; in supervaluations, p is true when p holds in *all* precisifications, while in subvaluations, p is true when it holds in at least one precisification. What we suggest is that the use of the quantifier is pragmatically governed. Informativity (that is, Gricean quantity) demands the stronger of the two quantifiers, i.e., the supervaluationary interpretation, but in the case of contradictory statements like '#2 is tall and not tall', using the universal quantifier produces a trivially false statement; so the quantifier must be weakened in order to make the statement non-trivial, and we propose that it is weakened to an existential quantifier, thereby producing the subvaluationary interpretation.

There is a possible view – though not well-motivated, we hope to show – according to which our patterns are interpreted as support to the fuzzy approach to vague expressions. Recall that in fuzzy logic there is an infinite number of truth values, ranging from 0 (false) to 1 (true), and that the truth-value of $\neg p$ for any proposition p is $1 - V(p)$. Thus, for example, if $V(p) = 0.6$, the value of its negation $\neg p$ is $1 - 0.6 = 0.4$. Recall also that the truth value of a conjunction $p \wedge q$ is defined as the minimum of the truth values of the conjuncts p and q. If the truth-value of p were 0.6, for example, and the value of q were 0.3, then the value of $p \wedge q$ will be $min(p, q) = 0.3$. This makes it possible for contradictory expressions like $p \wedge \neg p$ to be more true than 0; for if the truth-value of p were 0.6, the value of $\neg p$ will be 0.4, and the value of the conjunction $p \wedge \neg p$ will be $min(0.6, 0.4) = 0.4$.

A fuzzy logician may point to Figs. 6 and 7 and claim that the findings they illustrate are in fact faithful to the predictions of fuzzy logic, specifically, the prediction that a contradictory proposition containing a vague predicate is false at the periphery, and gradually climbs to half-truth in borderline cases. The same could be said to hold with respect Figs. 8 and 9, if the disjunction of p and q is computed as $max(p, q)$. A defender of this view may add that the patterns in Figs. 2-5 lend further support, since the truth of relevant propositions seem to gradually climb from near-falsity on one end of the tallness spectrum, to near-truth on the other end.

The problem with this view is that it assumes a statistical notion of truth, that is, a definition of truth whereby a proposition is said to be true to a degree

determined by consensus. We think that proponents of this view argue in favor of the fuzzy approach without taking notice of how believers of contradictions – the truth-judgers of 'tall and not tall' – judge the truth of other related statements like 'x is tall' and 'x is not tall'. In other words, while the *percentages* of truth/falsity-judgements made by many different people can indeed be thought to resemble a fuzzy pattern, a closer look at how the same judgers, taken individually, responded to other queries reveals a recurrent pattern that the fuzzy approach cannot predict, namely, the pattern in which a borderline proposition, and its negation, are judged false, but in which their conjunction is simultaneously judged true.[18]

6 Conclusion

We have argued that the findings of BOVW were incorrectly interpreted as support for the VAGUENESS-AS-IGNORANCE hypothesis. In the course of our argument we suggested that BOVW's theoretical criticisms against the gap-theoretic account of higher-order vagueness are inconsistent with their defense of their own proposal. We also showed that BOVW question-beggingly presuppose a bivalent proof system in their claim that gap-theories lead to contradictory statements, and also that their experimental evidence for the logical equivalence of 'x is not tall' and '"x is tall" is false' was not convincing. Finally, we presented new experimental findings that contradict BOVW's explanation of gaps: the emergence of gaps, they claim, is due to a general preference for errors of omission. If this claim were valid, we would expect a much larger percentage of 'Can't tell' responses in borderline cases. This, however, was not the case.

We ended our discussion by shedding experimental light on a different view of vagueness, a view in which a predicate and its negation are each said to be false of a borderline individual, but in which their conjunction is said to be true. We acknowledge, however, that further experimental work (as well as further theoretical work) is called for to test the details of super- and sub-valuations and their interactions with the Gricean maxims. Our 'pragmatic story' is but a first step which needs further investigation.

References

1. Alxatib, S., Pelletier, F.J.: The psychology of vagueness: Borderline cases and contradictions. Mind and Language (forthcoming)
2. Bonini, N., Osherson, D., Viale, R., Williamson, T.: On the psychology of vague predicates. Mind & Language 14, 377–393 (1999)
3. Dalrymple, M., Kanazawa, M., Kim, Y., McHombo, S., Peters, S.: Reciprocal expressions and the concept of reciprocity. Linguistics and Philosophy 21, 159–210 (1998)

[18] For further criticism of the fuzzy account of vague predicates from an experimental point of view, see Ripley [14]. Note particularly his finding that subjects tend to *fully* agree with (allegedly) contradictory statements – choosing 7, 'Agree', on a scale of 1-7, rather than choosing a more moderate response, as the fuzzy logician would predict.

4. Fine, K.: Vagueness, truth and logic. Synthèse 30, 265–300 (1975)
5. van Fraassen, B.: Singular terms, truth-value gaps, and free logic. Journal of Philosophy 63, 481–495 (1966)
6. Horn, L.: A Natural History of Negation. University of Chicago Press, Chicago (1989)
7. Hyde, D.: From heaps and gaps to heaps of gluts. Mind 106, 641–660 (1997)
8. Jaśkowski, S.: A propositional calculus for inconsistent deductive systems. Studia Societatis Scientiarum Torunensis 1, 55–77 (1948); Article published in Polish with the title "Rachunek zdań dla systemów dedukcyjnych sprzecznych." English translation in Studia Logica 24, 143–157 (1969)
9. Kamp, H.: Two theories about adjectives. In: Keenan, E. (ed.) Formal Semantics of Natural Language. Cambridge University Press, Cambridge (1975)
10. Keefe, R.: Theories of Vagueness. Cambridge University Press, Cambridge (2000)
11. Levinson, S.: Presumptive Meanings: The Theory of Generalized Conversational Implicature. MIT Press, Cambridge (2000)
12. Machina, K.F.: Truth, Belief, and Vagueness. Journal of Philosophical Logic 5(47-78)
13. Pelletier, F.J., Stainton, R.: On 'the denial of bivalence is absurd'. Australasian Journal of Philosophy 81, 369–382 (2003)
14. Ripley, D.: Contradictions at the borders. Paper Presented at the 2008 Conference on Philosophy and Psychology of Vagueness at the Institut Jean-Nicod (2008)
15. Ritov, I., Baron, J.: Reluctance to vaccinate: Omission bias and ambiguity. Journal of Behavioral Decision Making 3, 263–277 (1990)
16. Sorensen, R.: Blindspots. Clarendon Press, Oxford (1988)
17. Sorensen, R.: Vagueness and Contradiction. Oxford UP, New York (2001)
18. Spranca, M., Minsk, E., Baron, J.: Omission and commission in judgment and choice. Journal of Experimental Social Psychology 27, 76–105 (1991)
19. Williamson, T.: Vagueness. Routledge, Oxford (1994)

The Rationality of Round Interpretation

Harald Bastiaanse*

Institute for Logic, Language and Computation
University of Amsterdam, Postbox 94242,
1090 GE Amsterdam, The Netherlands
H.A.Bastiaanse@uva.nl

Abstract. Expanding on a point made by Krifka [6, p.7-8], we show that the fact that a round number has been used significantly increases the posterior probability that that number was intended as an approximation. This increase should typically be enough to make assuming that an approximation was indeed intended a rational choice, and thereby helps explain why round numbers are often seen as simply having an approximate meaning. Generalization into non-number words is also discussed, resulting in a possible origin of (some) vagueness.

1 Introduction

This paper is about why round numbers are seen as round; that is, as an approximation that can be used to refer to other numbers close to them. Much has been said about round numbers already, but other work has mostly focused on explaining the distribution of round numbers (which I will not be getting into at all) and why a speaker would want to use round numbers.

Instead, we will look at things from the perspective of someone hearing a round number being used. The point will be to show that in addition to what other good reasons there may be, round meaning can also in large part be explained just by the mathematics of the situation and people making rational decisions when interpreting things. After that, we apply the analysis to vagueness.

Despite this difference in approach, I should mention that the idea for this analysis comes from the following remark in [6]:

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

> Let the a-priori probability on hearing *forty-five minutes* that one of the scales (17.c) or (17.d) be used be the same, say s. Then on hearing *forty-five minutes* the probability that the more fine-grained scale (17.d) is used is 5rs, and the probability that the more coarse-grained scale (17.c) is used is double the value of that, 10rs. Hence the hearer will assume the more coarse-grained scale.

* The research in this paper is supported by a grant from NWO as part of the *Vagueness – and how to be precise enough* project (project NWO 360-20-202).

R. Nouwen et al. (Eds.): ViC 2009, LNAI 6517, pp. 37–50, 2011.

This is almost a throwaway remark in the piece in question, but it suggests an underlying principle worth far more attention.

Now the central question I will look into in the next sections is: why is it rational for a hearer to interpret a round number as a rounding? I'll investigate this by looking into several questions and the mathematics behind them. The first question is a matter of conditional probability. Some game theory will follow later.

2 Conditional Probability

The first question is: *Given that a round number was used, what is the chance that it was meant roundly?* In Bayesian statistics there is a straightforward answer to this question: the probability of A given B is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

If A means it was meant roundly and B that a round number was used, then the formula is as above, so we are looking for the chance of both happening divided by the (prior) chance a round number gets used. Keeping in mind that our A is in B and therefore $P(A \cap B) = P(A)$, we obtain

$$P(\textit{meant roundly}|\textit{round number is used}) = \frac{P(\textit{meant roundly})}{P(\textit{round number is used})}$$

Let us look into these chances using an example.

The example we're going to use is as follows: First we take a round number, say, 30. Now there are a bunch of numbers close-by enough that you might round them to 30. We will use the simplifying assumption that only integers are relevant sufficiently close ones have a chance of being rounded to 30. (See Section 4.2 for notes on how to drop both of these assumptions.) Suppose these sufficiently close ones are 25-34, or 10 numbers in total.

Now one of these numbers is randomly selected (with equally distributed probability) and the speaker wants to talk about that number. Finally, the speaker may or may not decide to round that number. Since we are interested in the hearer's side of things, we are going to just assign a value x to the chance that the speaker will choose to round to 30. For this example let us suppose $x = 50\%$. (This is perhaps on the high side, but not much depends on this; the point is to show how much larger than x the final conditional probability is. Also, Section 3 will show that a much smaller x can in fact suffice.) Let us see what happens given this situation.

	30	25-34 but not 30
Speaker rounds	$0,5 \cdot \frac{1}{10}$	$0,5 \cdot \frac{9}{10}$
Speaker does not round	$0,5 \cdot \frac{1}{10}$	$0,5 \cdot \frac{9}{10}$

This table outlines the probabilities of the four (a priori) possible situations. In the left column are the situations where the randomly selected number was exactly 30, in the right the ones where it was close but not 30 itself. Similarly, in the top row are the situations where the speaker chooses to round, while in the bottom are the ones where he does not.[1]

Now to get from these numbers to the conditional probability we want, the main thing to do is to apply the condition we were using. That condition was *Given that a round number was used*. Of course, if the number is not actually 30 and the speaker does not round to 30, then he will not say 30. Thus the lower-right corner is irrelevant for us. That is a lot of the total chance we're throwing out, so we can already see where this is going. But let us take a look.

	30	25-34 but not 30
Speaker rounds	0,05	0,45
Speaker does not round	0,05	0,45

$$\begin{aligned} P(\textit{Speaker rounded}) &= P(\textit{rounded}; 30) + P(\textit{rounded}; \textit{not 30}) \\ &= 0,05 + 0,45 = 0,5 \\ P(\textit{“30” is used}) &= P(\textit{rounded}) + P(\textit{didn't round}; \textit{30}) \\ &= 0,5 + 0,05 = 0,55 \end{aligned}$$

The other steps are straightforward. To get the chance the speaker rounded, take the chance he rounded and it was 30 and the chance he rounded and it was not and add them together. These are the ones in the top row, and the result is 50% again. For the chance a round number was used, we add to that the chance that the number was 30 and he did not round it, so we get 0,55.

Now we simple divide these, as per the formula. This gives

$$P(\textit{Speaker rounded}|\textit{“30” is used}) = \frac{P(\textit{both})}{P(\textit{“30” is used})} = \frac{0,5}{0,55} = \frac{10}{11} > 90\%$$

Thus, while the chance of the speaker rounding was just 50%, the chance that 30 was meant as round and should be interpreted like that is over 90%.

For the general picture, we replace our 50% chance by x, use an arbitrary round number R, and let k be the number of numbers that could be rounded to it (i.e. 10 in the above example). As mentioned before, the exact values of x and k will prove not to be too important.[2]

[1] Keep in mind that when the actual number is exactly 30, “rounding” it still makes a difference: 30 meant sharply is not the same as 30 meant in a loose way that encompasses nearby numbers. Note also that the hearer cannot simply hear the difference between the two; indeed, figuring out how the hearer best deals with that is the point here.

[2] See Appendix 4.2 for a treatment on how to generalize away from the discrete scale and even probability distribution.

	Actually R	Merely close to R
Speaker rounded	$x\frac{1}{k}$	$x\frac{k-1}{k}$
Speaker didn't round	$(1-x)\frac{1}{k}$	$x\frac{k-1}{k}$

$$\begin{aligned} P(\textit{Speaker rounded}) &= P(\textit{rounded};30) + P(\textit{rounded};\textit{not 30}) \\ &= x\frac{1}{k} + x\frac{k-1}{k} = x \\ P(\text{“}R\text{” is used}) &= P(\textit{rounded}) + P(\textit{didn't round};\mathrm{R}) \\ &= x + (1-x)\frac{1}{k} = \frac{k-1}{k}x + \frac{1}{k} \end{aligned}$$

Given these probabilities, the chance the speaker meant the number R as round is as follows:

$$P(\textit{Speaker rounded}|\text{“}R\text{” is used}) = \frac{x}{\frac{k-1}{k}x + \frac{1}{k}} = \frac{kx}{(k-1)x+1} = \frac{k}{k-1+\frac{1}{x}}$$

With k on the large side, this is going to be close to 1. The only problem is if x is low, but for that to get problematic it has to get low enough to be inversely proportional to k.

Thus, just by the mathematics of it understanding numbers as round is the correct choice far more often than one might expect. It would seem to be the **rational** interpretation –and indeed we will be able able to say this with more confidence after section 3.

And it would be wrong to think that this will stay limited to hearers only. If round numbers are likely to be interpreted as such, a speaker is likely to anticipate and modify a round number if he actually means it non-roundly. But that makes round interpretation even more rational, since participants can expect this anticipation. This creates a self-reinforcing loop that makes round numbers get interpreted more and more as simply having a round meaning; in appropriate contexts, at any rate.

3 Game Theory

For the next part, we are going to look more closely into the rationality angle. The previous question was necessarily a bit indirect; but Game Theory is based on concepts like strategies and making the rational choice between them. Thus, it allows us to specifically ask *When is it rational to assume a round number was meant roundly?*, and to get an exact answer in the form of a value x has to exceed (where, as before, x is the chance of the speaker rounding). Furthermore, we will also be able to find out the exact importance of contextual factors.

To answer this question, Game Theory works by assigning so-called utility values to understanding and misunderstanding each other. Each outcome gets a value: the higher it is the better for everyone involved. These are just numbers, like the example values below. Each of the two hearer strategies then has an

expected utility depending on the other player, and round interpretation simply is rational if the expected utility is higher than for non-round interpretation.

For this example, suppose the speaker has asked the hearer to show up for an appointment at 2 o'clock. This could be meant sharply, or could be meant to allow about five minutes either way. Obviously it would be preferable for the hearer to correctly understand the speaker's intent, so these outcomes get a higher value than the rest. We also assume that a greater need for precision gives rise to some inconvenience for one or both parties, so the correctly interpreted strict appointment has a slightly lower score.

Furthermore, showing up sharply on a loosely meant appointment is obviously not as bad as taking a sharply meant appointment loosely, so the values are fixed accordingly.[3,4]

	Round interpretation	*Non-round int*
Round intention	3	1
Non-round intention	0	2

Now as before we are interested in the hearer's point of view and simply let x be the chance that the speaker will round a given number. The better strategy is picked by maximizing expected utility, so round interpretation is rational if and only if

$$\begin{aligned} &P(\textit{Round intention}) \cdot 3 + P(\textit{Non-round intention}) \cdot 0 \\ &> P(\textit{Round intention}) \cdot 1 + P(\textit{Non-round intention}) \cdot 2 \end{aligned}$$

Filling in x, this becomes

$$3x + 0(1-x) > 1x + 2(1-x)$$

which simplifies to $2x > 2(1-x)$ which is if and only if $x > \frac{1}{2}$. This result does not actually look all that good, but there is something very important being overlooked here.

The thing we are overlooking is not unlike the condition we posed earlier. Essentially, if the speaker uses a non-round number, there is no way it can be misinterpreted as round. So the real strategies the hearer chooses from are not round and non-round interpretation; they are to interpret roundly if a round number is used or to never interpret roundly. This changes the analysis considerably.

	Round int [if a round number]	*Non-round int*
Round intention	3	1
Non-round intention	1, 8	2

[3] There will also be some convenience in the fact that $3 - 1 = 2 - 0$, but this is not part of the story.

[4] Note that while the choice of payoffs here is convenient, it does *not* itself offer an advantage to round interpretation, as should become clear from the calculations as well as the generalized case later one.

In the lower-left corner instead of 0 we get 0-if-it's-round-and-two-if-it-isn't. That comes out to $0 \cdot \frac{1}{10} + 2 \cdot \frac{9}{10} = 1,8$.[5,6] This makes round interpretation look a lot better, yielding all the advantage and only a fraction of the disadvantage. As the calculation below shows, x need only be $\frac{1}{11}$ for round interpretation to be rational.

$$\begin{aligned} 3x + 1,8(1-x) &> x + 2(1-x) \\ 3x + 1,8 - 1,8x &> x + 2 - 2x \\ 1,2x + 1,8 &> 2 - x \\ 2,2x + 1,8 &> 2 \\ 2,2x &> 0,2 \end{aligned}$$

$$x > \frac{0,2}{2,2} = \frac{1}{11}$$

The general picture again is similar. In the general case we use not specific numbers but the following arbitrary game:

	Round interpretation	*Non-round int*
Round intention	a	b
Non-round intention	c	d

Any good example will of course have $a > b$ and $d > c$, but the numbers are otherwise open to be chosen freely. Of course, as before the factor k marginalizes

[5] Assuming we are being precise to the minute, resulting in what amount to a $k = 10$ as before.

[6] Readers trying to interpret in terms of signaling games should note that the type t has two independent parameters here: one is the preferred time (even distribution over ten options), the other is the importance of showing up on the minute, which also governs the payoffs. The latter has a probability of x of corresponding to the upper row and $(1 - x)$ of corresponding to the lower one.

Now formalize as follows:

t_{1i} : *preferred time is 14.00*
t_{2i} : *preferred time not 14.00*
S_1 : $t_{1i}, t_{2i} \rightarrow$ *"two o-clock"*
S_2 : $t_{1i} \rightarrow$ *"two o-clock"*
$t_{2i} \rightarrow$ *specific other time*
H_1 : *"two o-clock"* $\rightarrow$ *interpret as round*
specific other time $\rightarrow$ *interpret as precise*
H_2 : *any* $\rightarrow$ *interpret as precise*

That it is rational for the sender to pick S_1 iff showing up on the minute is unimportant is left to the reader. Given this relationship the second parameter and the sender's strategy are both governed by x, and the rest of the analysis follows.

the difference between c and d, so that this arbitrary game is transformed into the following actual game:

	Round int [if a round number]	*Non-round int*
Round intention	a	b
Non-round intention	$d - \frac{d-c}{k}$	d

The condition for round interpretation to be rational thus becomes

$$\begin{aligned} ax + \left(d - \frac{d-c}{k}\right)(1-x) &> bx + d(1-x) \\ (a-b)x &> \frac{d-c}{k}(1-x) \\ (a-b)kx &> (d-c)(1x) \\ ((a-b)k + (d-c))x &> d-c \\ x &> \frac{d-c}{(a-b)k + (d-c)} \end{aligned}$$

Thus because of the generally largish k at the bottom, x can safely be quite small. Usually the breaking point is where it gets inversely proportional to k. If $(d-c) = (a-b)$ (that is, if the cost for misunderstanding is the same either way) then x need only be as little as $\frac{1}{k+1}$ for round interpretation to be the rational choice.

Now context can matter a lot, and that will work its way into what a, b, c and d really are, but clearly the factor k strongly pushes things towards round interpretation.

4 Discussion

This paper shows that even a weak inclination to round can be enough to explain why rounding is [rationally] assumed: even if the chance the speaker chooses to round is low, round interpretation is still likely to be rational, and then people adapt and it gets more and more standard until it is a standard meaning. Roundness is a rational and natural outcome.

It does not purport to −and cannot− explain why speakers should have even a small inclination to round to begin with, but in this it should be favorably combinable with existing arguments focusing on the speaker side or on inherent benefits to rounding (eg arguments from irrelevance, high cost of precision, uncertainty on the part of the speaker, manipulation or mental restrictions). Such other arguments need no longer account for a preference for rounding, just for a sufficiently significant probability.

It also does not go into why such inclinations are limited to "round" numbers. In my opinion that matter is better dealt with through other methods of investigation, eg [2,4].

4.1 Generalization to Vagueness

Generalizing the results about round numbers to vagueness is often surprisingly straightforward. While vagueness doesn't have much to do with numbers as such, vague terms often do have an underlying scale that's numerical –or an underlying situation that is easily numerizable, so that the same arguments apply.

This is most clearly seen with absolute adjectives (using the term absolute adjective as used in [5]). Take for example the word "bald". Loose use of the strictest sense of the word could be interpreted as rounding the number of hairs to zero. But then, given the number of hairs on a normal person's head, the k −the number of hairs that can be rounded to zero− for this situation can easily be in the hundreds or even thousands. The required prior chance of rounding x is thus so low that it can be accounted for even with just the various kinds of uncertainty. In this analysis that's obviously not a stable situation, so the word will quickly get used more and more loosely.

Importantly, this process does not stop. As soon as the meaning has changed (and stabilized), it is again subject to the same analysis. There is a slight difference in that more than one case counts as strictly bald now, but this can be accommodated by replacing k with a factor dividing the number of cases of the looser meaning by that of the new 'strict' meaning. k will be smaller and x may or may not change as well, but even looser interpretation is likely to be rational several more times, and further and further loosening will occur so long as this is so.

So just how loosely will it get used and where does the repeated loosening stop? That question gets hard to answer. Even if we and the people involved are pursuing a rational answer, just how loosely people should use and interpret the word soon depends on all kinds of factors nobody really knows; matters like how loosely everyone else is, should be, has been and should have been using it. Given that people might not use words equally loosely there will be much uncertainty and legitimate disagreement about such things, and this becomes more and more relevant as the process of loosening goes on. Eventually, the word becomes vague.[7]

(Some people may prefer the following line of reasoning instead: if precise loose use is rational, there is also support for vague loose use, especially if people aren't actually capable of the former but can manage the latter. In this way we get a reduction of other vagueness to the vagueness inherent in loose use. When loosening stops, then, it is not so much because the term has become vague but because it has become vague enough/too vague, with further loosening making no difference: [current] vague terms are fixpoints of the loosening operator.)

What we have here then is a possible explanation for a lot of vagueness. Loose interpretation is often rational, this makes loose use become the norm over time, and therefore things eventually get vague.

[7] There is also another possible reason, which I will not expand on here. If the loosenings of two related words start to overlap, the extensions may stop expanding there, since it remains more rational to use the "closer" word. I hope to look into this matter in a later paper. Still, for the reasons above one would not expect the boundaries this results in to be sharp.

There are a number of reasons to hypothesize that this is indeed the origin of much vagueness. The context-dependence of most vague terms can be explained in terms of the context-dependence of loose use. It also correctly predicts that vagueness occurs mostly for cases where there is an associated measurable property on a continuous or extremely fine scale, as these are the cases the argument is most naturally and easily applied to.[8] A number of vague terms do indeed have an associated "literal" or "absolute" meaning, e.g. "bald", "flat", "full".[9]

Furthermore, if we think absolute adjectives like "flat" and "full" as having prototypes, then the suggestion in prototype theory that the prototypes are by and large clear and universal across while the boundaries between concepts are not is consistent with an account where modern concepts are the result of repeated loosening of concepts that originally coincided with these prototypes far more strictly. One example of such a suggestion is made in [13] and supported in [12, p. 58-78].

When we are investigating a word like "bald", one might object that even if it is commonly used to refer to more than just an endpoint, the endpoint still remains and can be referred to with modifiers like "completely" and "absolutely". There would seem to be a difference between the loose use of absolute adjectives and the vagueness of other adjectives such as "tall". However, the section below outlines a big problem with such a view, further suggesting that repeated loosening can in fact produce vagueness.

On *very*, and the Futility of Remaximizing. It is well-known that many kinds of expressions can be vague, including adjectives, nouns, quantifiers and modifiers. This also includes the word "very", which may in fact be an even better example of this theory than "bald". I suggested just now that modifiers like "completely" and "absolutely" can refer to the endpoint of words like "bald", but is this really the case? In modern times nobody associates the word "very" with any specific endpoint. It is simply a strengthener. But in earlier centuries, they did. There is a paragraph about this in Elena Tribushinina's work[12] which is worth quoting at length.

> It is also worth noting that *extremely* is probably undergoing a semantic change from a maximizer to a booster. A similar development has taken place for *quite* and *very*. In the times of Chaucer, *quite* was only used in the sense of 'entirely' (e.g. *quite right*). The weaker sense of 'fairly' (as in *quite tall*) is attested from mid 19^{th} century (Paradis 1997: 74).

[8] Loose use can involve situations where no clear measurable property is involved –e.g. "I need a Kleenex." (where in fact any tissue would suffice) [14]– but in such cases it cannot easily be argued that *repeated* loose use occurs often enough to achieve vagueness.

[9] In some cases, words that don't may have such a meaning at one point only for it to be evolved away or taken over by another word. See also the section on "very". Also, some vague terms may have evolved from other vague terms with the vagueness itself still coming about in the proposed way.

I wouldn't go so far as to propose that this process underlies *all* vagueness, though.

> Similarly, *very* originally meant 'true, genuine, really' (cf. Ger. *wahr*, Du. *waar*), and turned into a booster in the Middle English period (Cuzzolin & Lehmann 2004; Lorenz 2002; Mendez-Naya 2003; Peters 1994; Stoffel 1901).[10]

So as we can see here "very" originally meant something along the lines of "truly" or "completely", until it succumbed to the kind of pressures we have been talking about, which are also affecting "extremely", "totally", "completely" and pretty much every maximizer you can think of.[11] The phenomenon is well documented[12], and is entirely natural and perhaps because of these arguments also fairly predictable.

And of course, if even "very" can turn out to have come about in this way, so can any other word.

4.2 Schelling Points and Evolutionary Game Theory; a Problem?

During the course of writing this paper it has come to my attention that Christopher Potts has done a related game-theoretical analysis on a related phenomenon.[10] While his subject matter is different, one of his predictions contradicts an important one of my own. Before I mention how I account for this, a brief introduction of it is in order.

In [10], Potts seeks to derive Kennedy's Interpretive Economy principle [5], or rather, a substitute with the same practical consequences (in particular, solving Kennedy's puzzle) as that principle, from basic assumptions about cognitive prominence and evolutionary stability. This of course has little to do with general vagueness, much less round numbers, but his analysis would still be problematic for my own ideas discussed above.

Potts's argument rests on the notion that amongst the possible ways to interpret an adjective related to a scalar endpoint, the most strict one stands out as a so-called Schelling point, making it initially (at least marginally) more likely to be selected than other ways. The extent to which this is so is what he refers to as the strength of the "Schelling assumption". Insofar as the Schelling assumption is fairly weak, I will not argue against it here.

He then combines the Schelling assumption with evolutionary game theory, arguing that even a slight preference will result in strict interpretation becoming standard. This is a fairly straightforward application of evolutionary game theory, and I will mostly not argue against it either.

---

[10] The papers she cites are [1], [7], [8], [9] and [11], respectively.

[11] Indeed, many people have been annoyed at the way even "literally" gets (ab)used these days. From a discussion on the internet:

> A: I literally ROFL'd.
> B: You literally rolled over the floor laughing? Ouch.

People who understand both "literally" and "ROFL" can be hard to come by.

[12] See also [3].

However, it does go against my own notions: in Section 4.1 in particular I argued that the evolution is likely to go the other way around, with vague words possibly being a result of repeated loosening of previously much sharper words. So how do I account for this? Naturally, the answer lies in doing what I have been doing in this paper.

Potts's most important analysis starts from the following basic game:

	$[\![\textbf{full}]\!].$	$[\![\textbf{full}]\!]_d$
$[\![\textbf{full}]\!].$	10	9.9
$[\![\textbf{full}]\!]_d$	9.9	10

In this example, $[\![\textbf{full}]\!].$ represents the maximum (ie sharp) interpretation of "full" while $[\![\textbf{full}]\!]_d$ represents a looser interpretation. In order to let this conform more to the examples I have been using myself I will flip the table here, as follows:

	$[\![\textbf{full}]\!]_d$	$[\![\textbf{full}]\!].$
$[\![\textbf{full}]\!]_d$	10	9.9
$[\![\textbf{full}]\!].$	9.9	10

Now using evolutionary mechanics Potts shows that when a coordination game like this is repeated, even a very weak Schelling assumption will make the population evolve towards overwhelmingly favoring the Schelling point – in this case strict interpretation.

There is nothing inherently wrong with this analysis, except that it ignores the point I have been making in this paper. Stay with this example, a loose usage of the word "full" can be used in more situations than strict use. Following the analyses of this paper, we should assign a discrete scale or use the continuous analysis in Appendix 4.2 to find the appropriate number k for the amount/ratio of situations sufficiently close to be loosely referred to as "full".[13]

Assuming either an even distribution or one taken included as part of k as per Appendix 4.2, we should then follow Section 3 and replace the 9.9 in the lower-left by $9.9 \cdot \frac{1}{k} + 10 \cdot \frac{k-1}{k} = 10 - \frac{0.1}{k}$, thus replacing the basic game above by the following:[14]

	$[\![\textbf{full}]\!]_d$	$[\![\textbf{full}]\!].$
$[\![\textbf{full}]\!]_d$	10	9.9
$[\![\textbf{full}]\!].$	$10 - \frac{0.1}{k}$	10

[13] The value of k in this depends on what specific d is being used, but since the stricter reading consists of a single point it depends even more on how fine the scale is. Indeed, increasingly fine scales can render k arbitrarily high.

[14] or in Potts's notation,

	$[\![\textbf{full}]\!].$	$[\![\textbf{full}]\!]_d$
$[\![\textbf{full}]\!].$	10	$10 - \frac{0.1}{k}$
$[\![\textbf{full}]\!]_d$	9.9	10

By the math in the earlier Section 3, it follows that loose interpretation is rational if $x > \frac{1}{k+1}$. In this exaple the population distribution provides this x, and if loose interpretation is rational at the initial time t_0 it will only get more so, so the condition for loose interpreation to be the end result of evolution becomes $P^{t_0}([\![\mathbf{full}]\!]_d) > \frac{1}{k+1}$. Therefore a weak Schelling assumption (where it suffices for $P^{t_0}([\![\mathbf{full}]\!]_.)$ to be just barely higher than 50%) is nowhere near enough. To win, strict use would have to start out at more than $\frac{k}{k+1}$.

Given everything I've argued here, a factor benefiting strict use needs to be strong, not merely minimal, to be of much use against the k factor.

References

1. Cuzzolin, P., Lehmann, C.: Comparison and gradation. In: Booij, G., Mugdan, J., S.S., Lehmann, C. (eds.) Morphologie. Ein internationales Handbuch zur Flexion und Wortbiltung, vol. 2, pp. 1212–1220. Mouton de Gruyter (2004)
2. Dehaene, S., Mehler, J.: Cross-linguistic regularities in the frequency of number words. Cognition 43(1), 1–29 (1992)
3. Ito, R., Tagliamonte, S.: Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. Language in Society 32(02), 257–279 (2003)
4. Jansen, C.J.M., Pollmann, M.M.W.: On round numbers: Pragmatic aspects of numerical expressions. Journal of Quantitative Linguistics 8(3), 187–201 (2001)
5. Kennedy, C.: Vagueness and grammar: the semantics of relative and absolute gradable adjectives. Linguistics and Philosophy 30(1), 1–45 (2007)
6. Krifka, M.: Approximate interpretation of number words: A case for strategic communication. Cognitive Foundations of Interpretation, 111–126 (2007)
7. Lorenz, G.: Really worthwhile or not really significant? a corpus-based approach to the delexicalization and grammaticalization of intensifiers in Modern English. In: Wischer, I., Diewald, G. (eds.) New Reflections on Grammaticalization, pp. 143–161. John Benjamins, Amsterdam (2002)
8. Mendez-Naya, B.: On intensifiers and grammaticalization: The case of swathe. English Studies 4, 372–391 (2003)
9. Peters, H.: Degree adverbs in Early Modern English. In: Kastovsky, D. (ed.) Studies in Early Modern English, pp. 269–288. Mouton de Gruyter, Berlin (1994)
10. Potts, C.: Interpretive Economy, Schelling Points, and evolutionary stability. Draft version (March 2008), `http://www.stanford.edu/~cgpotts/manuscripts/potts-interpretive-economy-mar08.pdf`
11. Stoffel, C.: Intensives and Down-toners: A Study in English Adverbs. Carl Winter, Heidelberg (1901)
12. Tribushinina, E.: Cognitive Reference Points: semantics beyond the prototypes in adjectives of space and colour. PhD thesis, Leiden University (October 2008)
13. Wierzbicka, A.: The meaning of color terms: Semantics, culture, and cognition. Cognitive Linguistics 1(1), 99–150 (1990)
14. Wilson, D., Sperber, D.: Truthfulness and relevance. Mind 111(443), 583–632 (2002)

Appendix

Continuous Scale and k on Probability

It has been convenient to use the simplifying assumption of a discrete scale, but it is straightforward enough and interesting to drop this notion, especially in light of the discussion in section 4.1.

Starting from the general case scenario in Section 2, let R be some round number and as before let x be the prior chance that a sufficiently close number will be rounded to it. Let C be a set of real numbers sufficiently close to R to be rounded to it in this fashion. In order to avoid dividing by zero later on, we also let $A \subset C$ be a set of numbers so close to R as to be considered identical, or at least indistinguishable.[15]

Now let $B = C - A$, assume that the actual number is picked randomly with probability distributed evenly over C, and assume that $|\cdot|$ is an appropriate measure on $\mathbb{R}$.[16] Then we can "divide out"/ignore the probability part to obtain the following familiar-looking table:

	Actually R	Merely close to R
Speaker rounded	$x\|A\|$	$x\|B\|$
Speaker didn't round	$(1-x)\|A\|$	$(1-x)\|B\|$

I have not yet mentioned how k should be defined here, but by looking at the table it should surprise no one that the definition is simply $k = \frac{|C|}{|A|} = \frac{|A|+|B|}{|A|}$[17]. This leads to the following:

$$P(\textit{Speaker rounded}|\text{“}R\text{” is used}) = \frac{x|A| + x|B|}{x|A| + x|B| + (1-x)|A|} = \frac{|A| + |B|}{|B| + |A|/x}$$
$$= \frac{(|A| + |B|)/|A|}{\left(\frac{|A|+|B|}{|A|} - \frac{|A|}{|A|}\right) + 1/x} = \frac{k}{k - 1 + \frac{1}{x}}$$

which is of course the same result as in the discrete case.

Taking the probability distribution out in this way may seem suspect, and in any case it is interesting to consider the impact of non-even distributions. The resulting formula threatens to get convoluted, but this is easily avoided through cheating: redefine k as

$$k = \frac{P(A \cup B)}{P(A)}$$

[15] Of course in more general situations A may also simply be whatever R refers to sharply, so long as that has non-zero measure.

[16] In the more general case, pick an appropriate measure on at least C.

[17] In the general case, the equality obtains because A and B are disjoint and we picked an appropriate measure function.

Then it is clear that we can just combine area and distribution into probability to get the following table:

	Actually R	Merely close to R
Speaker rounded	$xP(A)$	$xP(B)$
Speaker didn't round	$(1-x)P(A)$	$(1-x)P(B)$

Thus the results are exactly as before[18],[19] except that now the effect of a change in probability distribution is a straightforward impact on k: for instance, the k in the above example could end up much fairly small if the distribution were a bell curve around R, with details depending on σ and the size of A.

[18] In this case the equality $P(A \cup B) = P(A)+P(B)$ follows from the laws of probability.

[19] Reobtaining the exact results from the sections involving game theory is not too difficult –and left to the reader.

Supervaluationism and Classical Logic

Pablo Cobreros*

Department of Philosophy
University of Navarra
31080 Pamplona, Spain
pcobreros@unav.es

Abstract. The supervaluationist theory of vagueness provides a notion of logical consequence that is akin to classical consequence. In the absence of a *definitely* operator, supervaluationist consequence coincides with classical consequence. In the presence of '*definitely*', however, supervaluationist logic gives raise to counterexamples to classically valid patterns of inference. Foes of supervaluationism emphasize the last result to argue against the supervaluationist theory. This paper shows a way in which we might obtain systems of deduction adequate for supervaluationist consequence based on systems of deduction adequate for classical consequence. Deductions on the systems obtained this way adopt a completely classical form with the exception of a single step. The paper reviews (at least part of) the discussion on the non-classicality of supervaluationist logic under the light of this result.

Keywords: Vagueness, Supervaluationism, Global Validity, Deductive Systems, Logical Consequence.

1 Introduction

1.1 Vagueness and Supervaluationism

My youngest daughter, Julia, is 3 months old (at the time I'm writing this paper). She is clearly a baby. Sofia and Carmen are 4 and 6 years old respectively; they are clearly not babies (you can ask them). Julia will probably cease to be a baby, and become, as her older sisters, clearly not a baby. But is there an exact time n such that Julia is a baby at n but Julia is not a baby at time n plus one second? The conclusion seems to be unavoidable if we want to keep to classical logic. Since from the fact that Julia is a baby at t_0 (today) and the supposition that for any time x if Julia is a baby at x then Julia is a baby at time $x+1$ it follows that Julia is a baby at time $t_0 + 10^8$ seconds (t_0 plus three years and a couple of months approx.) by a number 10^8 of applications of universal instantiation and modus ponens. Thus, if we grant that Julia is not a baby at time $t_0 + 10^8$, it

* I want to give thanks to Will Bynoe, Maria Cerezo, Paul Egre, Paloma Perez-Ilzarbe and the audience of the Workshop on Vagueness in Communication in the ESSLLI09 Bordeaux. Thanks also to an anonymous referee for some helpful observations. This paper is part of the research project 'Borderlineness and Tolerance' (FFI2010-16984) funded by the Ministerio de Ciencia e Innovación, Government of Spain.

R. Nouwen et al. (Eds.): ViC 2009, LNAI 6517, pp. 51–63, 2011.

classically follows that it is not the case that for any time x if Julia is a baby at x then Julia is a baby at time $x + 1$, classically in other words, there is a time x such that Julia is a baby at x but Julia is not a baby at $x + 1$ second.

Epistemicists in vagueness want to retain classical logic and they endorse the somewhat surprising claim that there's actually such an n (they claim we know the existential generalization 'there is an n that such and such' even if there is no particular n of which we know that such and such). Many philosophers, however, find this claim something too hard to swallow and take it as evidence that classical logic should be modified (at least when dealing with vague expressions). One standard way in which we might modify classical logic is by considering some extra value among truth and falsity; we then redefine logical connectives taking into account the new value. This strategy has motivated some philosophers to defend Kleene's strong three-valued logic for the case of vagueness.[1] Under this view, the conclusion that there is an n at which Julia is a baby and such that Julia is not a baby at n plus one second does not follow, since there are times at which the sentence 'Julia is a baby' has not a clearly defined truth-value. Thus, the strategy consists in a suitable weakening of classically valid principles like excluded middle along with other principles at work in the previous paradoxical result like the least number principle (see [4]).

In some sense, supervaluationists take a middle path among these two alternatives. Unlike epistemicists, supervaluationists hold that vague expressions lead to truth-value gaps and, thus, that at some time the sentence 'Julia is a baby' lacks a truth-value. Unlike philosophers endorsing Kleene's strong three-valued logic, however, supervaluationists endorse a non truth-functional semantics that allows them to endorse, broadly speaking, classical logic. How?

The basic thought underlying supervaluationism is that vagueness is a matter of underdetermination of meaning. This thought is captured with the idea that the use we make of an expression does not *decide* between a number of admissible candidates for making the expression precise. According to supervaluationism a vague expression like 'baby' can be made precise in several ways compatible with the actual use we make of the expression. For example, we can make it precise by saying that x is a baby just in case x is less than one year old; but the use of the expression will allow other ways of making precise like 'less than one year plus a second'. If Martin is one year old, the sentence 'Martin is a baby' will be true in some ways of making 'baby' precise and false in others. Since our use does not decide which of the ways of making precise is correct, the truth-value of the sentence 'Martin is a baby' is left unsettled. By supervaluationist standards, a sentence is true just in case it is true in every way of making precise the vague expressions contained in it (that is, 'truth is supertruth').

A *precisification* is a way of making precise all the expressions of the language so that every sentence gets a truth-value (true or false but not both) in each precisification. In this sense, a precisification is a classical truth-value assignment. However, precisifications should be *admissible* in the sense that some connections must be respected such as analytic relations between expressions. For example,

[1] For example, [7] and [8].

any precisification counting Martin as a baby should not count him as a child. Thus, the sentence 'If Martin is a baby then he is not a child' will be supertrue even if Martin is a borderline-baby. Also comparative relations must be respected by admissible precisifications. For example, any precisification making 'Nicolas is a baby' true (where Nicolas is one year and a month) should also make 'Martin is a baby' true. These restrictions on the admissibility of a precisification enables the supervaluationist theory to endorse Fine's so-called *penumbral connections*, that is, connections that might hold among sentences even if these have a borderline status [5, pp. 269-270]. Taking the previous example, if Nicolas is older than Martin but both are borderline cases of the predicate 'is a baby', the sentence 'If Nicolas is a baby then Martin is a baby' is true in every precisification (and, hence, true *simpliciter* for the supervaluationist) since every precisification in which the antecedent is true, the consequent is also true. At this point supervaluationists have some advantage over some truth-functional approaches such as those endorsing Kleene's strong three-valued logic, since for this semantics, the sentence 'If Nicolas is a baby, then Martin is a baby' comes out as indefinite.[2]

One consequence of supervaluationist semantics is that classical validities are preserved. A sentence φ is valid according to supervaluationist semantics just in case it is supertrue in every model. Since precisifications are classical truth-value assignments, classically valid sentences are true in each precisification and, thus, they are supertrue in every model. For example, though the sentence 'Martin is a baby' lacks a truth-value, the sentence 'Martin is a baby or Martin is not a baby' is supertrue since in each precisification some member of the disjunction is true (though not the same in every precisification). More generally, excluded middle is valid since, for every model, every precisification verifies $p \vee \neg p$. Furthermore, it can be shown that, as long as we stick to the classical language, supervaluationist consequence and classical consequence coincide.[3] At this point, supervaluationists seem to have again the upper hand over truth-functional approaches. In Kleene's strong three-valued logic φ entails φ even if $\varphi \rightarrow \varphi$ is not valid (thus, conditional proof is not a valid rule of inference).

The question now is how can supervaluationists explain the sorites paradox without committing themselves to an epistemic explanation of vagueness. If supervaluationist consequence coincides with classical consequence and the existence of an n such that Julia is a baby at n but Julia is not a baby at n plus one second follows by classical reasoning, the supervaluationist must be committed to that consequence as well. The supervaluationist explanation is that though they are committed to the truth of the claim 'there is an n such that Julia is a

[2] Fine claims that supervaluationism is the only view that can accommodate all penumbral connections [5, pp. 278-279]. For more on truth-functionality see [6, pp. 96-100].

[3] See [5, pp. 283-284] and [6, pp. 175-176]. Fine and Keefe identify supervaluationist consequence with what it is more precisely characterize as *global validity* below. The coincidence between supervaluationist and classical consequence is restricted to single conclusions; for example, the truth of a disjunction in a model classically guarantees the truth of some of its disjuncts but according to supervaluationist semantics a disjunction can be supertrue without either disjunct being supertrue. That is, $\{\varphi \vee \psi\} \vDash_{CL} \{\varphi, \psi\}$ but $\{\varphi \vee \psi\} \nvDash_{SpV} \{\varphi, \psi\}$.

baby at n but Julia is not a baby at n plus one second', they are not committed to the truth of any particular instance of that claim. The existential generalization is supertrue since every precisification of the language verifies it; but the n that makes the existential generalization true varies from one precisification to another so that there is no particular instance that is supertrue. (The case is analogous to the truth of the disjunction 'Martin is a baby or Martin is not a baby': the disjunction is verified in every precisification even if neither disjunct is verified in every precisification). The supervaluationist claims that the absence of a verifying instance suffice to show that the theory is not committed to a *sharp transition* from the times in which Julia is a baby to the times in which Julia is not a baby and, thus, to avoid an epistemicist explanation of vagueness while retaining classical logic.

The possibility of endorsing classical logic while avoiding an epistemicist account of vagueness is an appealing feature of the supervaluationist theory. In Fine's words, supervaluationism 'makes a difference to truth, but not to logic' [5, p. 284]. However, it is natural for a theory of vagueness to provide an explanation of the notion of *definiteness* and include a corresponding expression in the language in order to talk about borderline cases. Now supervaluationist logic is no longer classical when we introduce such an expression, and this fact is stressed by foes of supervaluationism to argue that the supposed advantage of supervaluationism over its truth-functional rivals is just an illusion.

1.2 Supervaluationism and Logical Consequence

Supervaluationist semantics for a propositional language containing a *definitely* operator ('$\mathcal{D}$' henceforth) might be modeled along the lines of a possible-worlds semantics for a propositional language with an operator for *necessity*. *Worlds* in a structure are informally read as *admissible precisifications* (admissible ways of making all the expressions of the language precise) and the accessibility between worlds is read as an *admissibility* relation between precisifications.[4] More explicitly, an interpretation for a propositional language with $\mathcal{D}$ is a triple $\langle W, R, \nu\rangle$ where W is a non-empty set of *precisifications*, R is an *admissibility* relation in W and ν is a truth-value assignment to sentences at precisifications. Classical operators have their standard meaning (relative to precisifications) and '$\mathcal{D}$' is defined as the modal operator for necessity:

$\varphi \rightarrow \psi$ takes value 1 at w if and only if at w: either φ takes value 0 or ψ takes value 1.

$\neg\varphi$ takes value 1 at w if and only if φ takes value 0 at w.

$\mathcal{D}\varphi$ takes value 1 at w if and only if φ takes value 1 at every precisification admitted by w.[5]

[4] This possible-worlds treatment of supervaluationist semantics is used, for example, in [10] and [11].

[5] When comparing local and global validity I shall talk about *points* instead of *precisifications* to remain neutral on the informal reading. In *Lemma 1* below we will write $\nu_w(\varphi) = 1$ to mean ν assigns value 1 to φ at w.

The question of which system gives the logic of *definiteness* for the supervaluationist reading of this notion depends on the informal reading of the semantics and on questions concerning higher-order vagueness. However, it is uncontroversial for any reading of $\mathcal{D}$ that the principle '$\mathcal{D}\varphi \rightarrow \varphi$' (if something is definite, then it is the case) should be valid and, consequently, that the *admissibility* relation should at least be reflexive.

So far supervaluationist and modal semantics coincide. The difference comes when we look at logical consequence. Logical consequence in modal semantics is standardly defined as *local consequence* [1, p. 31]:

Definition 1 (Local consequence). A sentence φ is a local consequence of a set of sentences Γ, written $\Gamma \vDash_l \varphi$, just in case for every interpretation and any point w in that interpretation: if every $\gamma \in \Gamma$ takes value 1 in w then φ takes value 1 in w.

In some sense, the notion of *local consequence* is a natural way of defining logical consequence in modal semantics. However, local consequence is not well defined in the supervaluationist reading of the semantics since for the supervaluationist that a sentence is true means that it is true *in every precisification* (that is, 'truth is supertruth'); and, thus, local consequence does not preserve the supervaluationist-relevant notion of truth. It is usually accepted in the literature that the supervaluationist is committed to something known as *global consequence*:[6]

Definition 2 (Global consequence). A sentence φ is a global consequence of a set of sentences Γ, written $\Gamma \vDash_g \varphi$, just in case for every interpretation: if every $\gamma \in \Gamma$ take value 1 at every point then φ takes value 1 at every point.

Global consequence preserves the notion of *truth-at-every-point* which is like the counterpart in this semantics of the notion of *supertruth*. In terms of modal semantics, we might see why supervaluationist consequence coincides with classical consequence for the classical language (this is Fine's and Keefe's previously mentioned result). If there are no modal expressions (operators whose truth-conditions depend on what's going on at points different of the evaluation-point) local validity will coincide with classical validity (since the truth conditions of classical expressions depend just on what's going on at the evaluation point which is a classical model). In turn, a language without this kind of operators will not be able to discriminate between global and local consequence. However, in the context of a theory of vagueness in which borderline cases play a key role (as it is the case of supervaluationism) it is natural to consider a '$\mathcal{D}$' operator; and in its presence, global and local validity no longer coincide. In particular, global validity is strictly stronger.

Every locally valid argument is globally valid. For if $\Gamma \nvDash_g \varphi$, then there is an interpretation such that every γ in Γ takes value 1 at every point and φ value 0

[6] I hold that global consequence is not fully adequate for the supervaluationist given the problem of higher-order vagueness (see [2] and [3]). In this paper, however, we will focus just on global validity.

at some. Now the point at which φ takes value 0 shows that $\Gamma \not\models_l \varphi$. The other direction is not true. In particular the inference from φ to $\mathcal{D}\varphi$ is globally valid (if φ takes value 1 everywhere, so does $\mathcal{D}\varphi$) but not locally valid (φ and $\neg\mathcal{D}\varphi$ might both take value 1 at the same point in an interpretation).

1.3 Counterexamples to Classically Valid Patterns of Inference

The characteristic inference of global validity, the inference from φ to $\mathcal{D}\varphi$, might be used to show that global validity leads to some counterexamples to classically valid patterns of inference as, for example, *conditional proof*:

Definition 3 (Conditional proof). $\Gamma \cup \{\psi\} \vdash \varphi \Longrightarrow \Gamma \vdash \psi \rightarrow \varphi$.

for $\varphi \models_g \mathcal{D}\varphi$, but it is not the case $\models_g \varphi \rightarrow \mathcal{D}\varphi$ (since the last would render the modality trivial, assuming reflexivity). In a similar manner, always making use of the inference from φ to $\mathcal{D}\varphi$, we might find counterexamples to other classically valid patterns of inference such as contraposition, argument by cases and *reductio ad absurdum* [10, pp. 151-152].

The next section reviews some discussion concerning these counterexamples to classically valid forms of reasoning. But before we proceed, there is a small remark concerning 'classical logic'. The '$\mathcal{D}$' operator is not a classical notion; in this sense any logic for a language containing the operator is not, strictly speaking, *classical logic*. However, it is assumed in the literature that the most standard logic of definiteness corresponds to some of the various normal modal systems, since standard rules like the ones mentioned above (conditional proof, contraposition etc.) are correct for this sort of systems. That's why in the following we will assume that, in the present context, 'classical logic' means local validity.

2 Problems with Global Validity

2.1 The Keefe-Varzi Debate

In her 2000 book on vagueness, Rosanna Keefe considers the issue of counterexamples to classically valid patterns of inference [6, pp. 178-181]. Keefe argues that the failure of those rules is a natural outcome of any non-epistemic reading of '$\mathcal{D}$' and suggests an alternative set of rules that are always *global-truth* preserving.[7] For example, instead of the standard rule of conditional proof, Keefe suggests the use of the following rule:

Definition 4 (Conditional proof*). $\Gamma \cup \{\psi\} \vdash \varphi \Longrightarrow \Gamma \vdash \mathcal{D}\psi \rightarrow \varphi$.

In an analogous way Keefe proposes other rules to deal with the other counterexamples [6, pp. 179-180].

[7] I've got some doubts, however, concerning the soundness of the proposed rule to substitute conditional proof, since it seems we might actually derive $\vdash \mathcal{D}\mathcal{D}\varphi \rightarrow \mathcal{D}\varphi$ which is not always globally true when R is not required to be transitive.

In a recent paper Achille Varzi discusses this suggestion of Keefe. In the first place, Varzi notes that the suggestion, as it has been presented, cannot be acceptable. The problem is that if we replace the old rules by Keefe's new rules the resulting system is doom to be incomplete. For example, the following consequence assertions are correct but not provable making use only of Keefe's rules:

(a) $\vDash_g p \to p$
(b) $p \vDash_g \neg\neg p$
(c) $p \vee q \vDash_g q \vee p$
(d) $p \to q \vDash_g (p \wedge r) \to q$ [9, p. 657].

Keefe's suggestion, however, can be understood in a broader sense. Keefe notes that the classical rules are perfectly sound when the $\mathcal{D}$ operator is not at play; her suggestion is, thus, that we should make use of both kind of rules depending on the presence or absence of the $\mathcal{D}$ operator in the premises: 'so when the $\mathcal{D}$ operator is involved, supervaluationism needs to modify some classical rules of inference, but the new rules are reasonable, and when no $\mathcal{D}$ operator is involved normal classical rules of inference remain intact.' [6, p. 180]. But Varzi does not find this strategy very convincing:

> I am not sure this would work, but even if it did, things would again begin to look ugly and one might as well think that the right thing to do is to bite the bullet and give up [global validity] altogether. [9, p. 657].

I understand that Varzi's objection to Keefe's strategy point out with certain pessimism to the difficulty of providing an adequate system of deduction for global validity based on classical rules in a simple and straightforward way. To some extent, this pessimism on Keefe's suggestion is reasonable since the suggestion is too general to provide any intuition on whether it really works. In order to avoid Varzi's pessimism, Keefe should provide precise constraints on the applicability of old rules; explaining when can we make use of the new ones and showing that the resulting system is adequate (correct and complete). We will consider this question later. Now we turn to a different objection based directly on the non-classicality of supervaluationist logic in the presence of $\mathcal{D}$.

2.2 Williamson's Objection

The cleanest exposition of the counterexamples to classically valid rules of inference is [10, pp. 150-152]. Based on this fact, Williamson argues against the supervaluationist theory. According to Williamson, patterns of inference such as conditional proof, contraposition, argument by cases and *reductio* play a central role in formal systems of deduction that are closer to our informal way of reasoning. Given that these rules of inference are not always correct under the global reading of logical consequence, Williamson draws the conclusion that 'supervaluations invalidates our natural mode of deductive thinking.' [10, p. 152].

It seems to me that this last claim is not completely fair. I concede to Williamson that the mentioned rules play a key role in those formal systems closer to our informal way of reasoning.[8] And, of course, there is a sense in which global validity invalidates these forms of deduction (since there are counterexamples to the corresponding patterns of inference). But there is still a sense in which the claim is unfair since we have not considered yet particular systems of deduction for global validity. My point is that, perhaps, these systems (or at least some of them) are relatively simple extensions of classical systems in which the applicability of the *controversial rules* have clearly defined restrictions and such that the form of deductions is, to certain extent, standard.

2.3 Two Questions

The foregoing discussion raise two related questions, one of technical character, the other more philosophical. The first question concerns the possibility of providing adequate systems of deduction for global validity. The second question concerns the aspect of these systems, whether we might include rules of inference such as conditional proof and whether the form of the corresponding deductions in those systems is relatively standard.

The following section aims to provide an answer to both questions. With respect to Varzi's pessimism, the section shows that there is a simple way to extend a deductive system for local validity to a deductive system for global validity. With respect to Williamson's claim that supervaluationism invalidates our natural modes of reasoning, it is shown a way to restrict the applicability of the relevant rules that provide to the deduction in these systems an *almost* classical form.

3 Deduction for Global Validity

This section presents a procedure to extend a given notion of deduction $\vdash_l$ for local consequence to a notion of deduction $\vdash_g$ for global consequence. We provide an argument to show that if $\vdash_l$ is complete with respect to local validity, $\vdash_g$ is complete with respect to global validity (section 3.1). Whether $\vdash_g$ is also correct with respect to global validity (that is, wether $\Gamma \vdash_g \varphi$ entails $\Gamma \vDash_g \varphi$) will depend on the original system defining $\vdash_l$. For the reasons given before, we are

[8] A qualification: Williamson's claim is not uncontroversial. For a start, conditional proof is not unrestrictedly valid in some presentations of first-order logic (in these accounts, $Px \vdash \forall x Px$ but $\nvdash Px \rightarrow \forall x Px$). But more generally, one might well doubt whether there is really anything like 'our natural mode' when we talk about deductive thinking. However, I concede to Williamson that claim in the text for the following reason. Classicality is one of the supposed advantages of supervaluationism over truth-functional approaches but the failure of those rules of inference in the presence of $\mathcal{D}$ calls this point into question. Thus, even if Williamson's claim is not uncontroversial, supervaluationists need to address the objection of non-classicality in the presence of $\mathcal{D}$ anyway.

interested in systems of deduction that make use of rules like conditional proof, contraposition, argument by cases and *reductio*. Section 3.2 shows a straightforward way to restrict the applicability of these rules to render $\vdash_g$ correct with respect to global validity (and that do not destroy the completeness argument in section 3.1). Section 3.3 evaluates in which way these results shed some light on the discussion in section 2.

3.1 Completeness

The completeness argument below will make use of the following connection between local and global validity:

Lemma 1 (Global-local connection). $\Gamma \vDash_g \varphi$ iff $\{\mathcal{D}^n\gamma \mid \gamma \in \Gamma, n \in \omega\} \vDash_l \varphi$.

The intuitive idea is that φ *globally* follows from Γ just in case φ *locally* follows from the *absolute definitization* of Γ, that is, the set containing: all the γ's plus all the $\mathcal{D}\gamma$'s plus all the $\mathcal{D}\mathcal{D}\gamma$'s etc.

Proof. (i) Right-to-left
Assume: $\Gamma \nvDash_g \varphi$. Then, there is an interpretation $\Im = \langle W, R, \nu\rangle$ where for all w and all $\gamma \in \Gamma$, γ takes value 1 at w and for some w, φ takes value 0 at w. Name w_0 the precisification at which φ is takes value 0. Since every γ in Γ is takes value 1 everywhere in the interpretation, every γ takes value 1 at w_0 for each iteration of $\mathcal{D}$. Thus, precisifcation w_0 shows that $\{\mathcal{D}^n\gamma \mid \gamma \in \Gamma, n \in \omega\} \nvDash_l \varphi$.

(ii) Left-to-right[9]
Assume: $\{\mathcal{D}^n\gamma \mid \gamma \in \Gamma, n \in \omega\} \nvDash_l \varphi$. Then there is an interpretation $\Im = \langle W, R, \nu\rangle$ and a precisification w_0 in it such that, for every γ in Γ and any iteration of $\mathcal{D}$, $\mathcal{D}^n\gamma$ takes value 1 at w_0 and φ takes value 0 at w_0. Let W' be $\{w|w_0R^nw\} \cup \{w_0\}$ (that is, w_0 plus the precisifications reachable from w_0 in any number of R-steps) and R', ν' the restrictions of R, ν to W'. We should demonstrate that the modified interpretation is a countermodel showing $\Gamma \nvDash_g \varphi$. We show first i) that both interpretations agree in the truth-values assigned to any formula in any w' in W'.

To show i), note first that if $w' \in W'$ then R' and R relate w' exactly to the same worlds, that is, if $w' \in W'$ then $w'R'w$ iff $w'Rw$. For if $w'R'w$ then both $w \in W'$ and $w'Rw$. On the other hand, if $w'Rw$, as $w' \in W'$, w_0R^mw' and thus $w_0R^{m+1}w$, that is, $w \in W'$. Thus, $w'R'w$.

i) is proved by induction over the set of wff. The case for propositional variables holds by definition. The case for non-modal operators is straightforward. For $\psi = \mathcal{D}\alpha$, suppose that $w' \in W'$:

$\nu'_{w'}(\mathcal{D}\alpha) = 1$ iff $\forall w*_{\in W'}$ such that $w'R'w*$, $\nu'_{w*}(\alpha) = 1$
iff $\forall w*_{\in W'}$ such that $w'R'w*$, $\nu_{w*}(\alpha) = 1$ (by IH)
iff $\forall w*_{\in W}$ such that $w'Rw*$, $\nu_{w*}(\alpha) = 1$ (by the fact noted above)

[9] The result is based on the fact that $\langle W', R', \nu'\rangle$ is a generated submodel of $\langle W, R, \nu\rangle$ [1, p. 56].

To show that the modified interpretation is a countermodel showing $\Gamma \nvDash_g \varphi$ note that w_0 has access to every world in W' (excluding, perhaps, w_0 itself) thorough some number of R-steps. Since for every γ in Γ and every $n \in \omega$, $\nu_{w_0}(\mathcal{D}^n\gamma) = 1$, every member of Γ takes value 1 at every world in W'. On the other hand, as $\nu_{w_0}(\varphi) = 0$, there is at least one world in W' in which φ takes the value 0. Thus, the modified interpretation shows that $\Gamma \nvDash_g \varphi$.

Since local consequence is standard we might assume that there are adequate systems of deduction for it. Let $\vdash_l$ be an adequate deductive relation for local consequence. Among other rules the following are locally valid (sometimes called *structural rules*):[10]

Definition 5 (Reflexivity). $\varphi \in \Gamma \Longrightarrow \Gamma \vdash \varphi$.

Definition 6 (Cut). $\Gamma \vdash \varphi$, $\Delta \vdash \gamma_1, \ldots \Delta \vdash \gamma_n \Longrightarrow \Delta \vdash \varphi$.

We consider an extra rule that is **not** locally valid:

Definition 7 ($\mathcal{D}$-introduction). $\Gamma \vdash \varphi \Longrightarrow \Gamma \vdash \mathcal{D}\varphi$[11].

The addition of this rule to $\vdash_l$ leads to a new notion of deductive consequence, $\vdash_g$. With the help of *Lemma 1* we can now show that $\vdash_g$ is complete with respect to global consequence:

Theorem 1 ($\vdash_g$-completeness). If $\Gamma \vDash_g \varphi$ then $\Gamma \vdash_g \varphi$.

Proof. (i) Assume: $\Gamma \nvDash_g \varphi$.

$\Downarrow$

(ii) $\{\mathcal{D}^n\gamma \mid \gamma \in \Gamma, n \in \omega\} \nvDash_g \varphi$.

$\Downarrow$

(iii) $\{\mathcal{D}^n\gamma \mid \gamma \in \Gamma, n \in \omega\} \nvDash_l \varphi$.

$\Downarrow$

(iv) $\{\mathcal{D}^n\gamma \mid \gamma \in \Gamma, n \in \omega\} \nvDash_l \varphi$

$\Downarrow$

(v) $\Gamma \nvDash_g \varphi$.

The step from (i) to (ii) is guaranteed by the rules of $\mathcal{D}$-introduction, Reflexivity and Cut. For assume that $\{\mathcal{D}^n\gamma \mid \gamma \in \Gamma, n \in \omega\} \vdash_g \varphi$; then, since formal proofs are finite, there's a finite $\Gamma^* \subseteq \{\mathcal{D}^n\gamma \mid \gamma \in \Gamma, n \in \omega\}$ such that $\Gamma^* \vdash_g \varphi$. Now,

[10] A third structural rule not used in the proof is *Monotonicity*: $\Gamma \vdash \varphi \Longrightarrow \Delta \vdash \varphi$ for all Δ such that $\Gamma \subseteq \Delta$.

[11] Given the Reflexivity rule, the inference from φ to $\mathcal{D}\varphi$ is a special case of this rule. This rule must not be confused with the *Necessitation* rule of standard modal logics that can be stated this way: $\Gamma \vdash \varphi \Longrightarrow \mathcal{D}(\Gamma) \vdash \mathcal{D}\varphi$, where $\mathcal{D}(\Gamma)$ is $\{\mathcal{D}\gamma \mid \gamma \in \Gamma\}$.

each $\gamma^* \in \Gamma^*$ is either an element of Γ or an element of Γ with a finite number of $\mathcal{D}$'s attached to it. Thus, making use of Reflexivity and $\mathcal{D}$-introduction, $\Gamma \vdash_g \gamma^*$ for any $\gamma^* \in \Gamma^*$ and thus, making use of Cut, $\Gamma \vdash_g \varphi$.

The step from (ii) to (iii) is guaranteed by the way we have defined $\vdash_g$: since this notion extends $\vdash_l$ (every local proof is a global proof), if we cannot provide a global proof, we cannot provide a local proof either. The step from (iii) to (iv) is based on the assumption that $\vdash_l$ is complete with respect to local consequence. The step from (iv) to (v) is based on the left-to-right direction of *Lemma 1*.

In order to prove the theorem we've had just to add the rule of $\mathcal{D}$-introduction. The intuitive explanation is (if any) as follows. *Lemma 1* shows that global consequence adds to local consequence the supposition that the premises are *absolutely definite*. For this reason we need to strengthen a system for local consequence with a rule reflecting the supposition that the premises are absolutely definite; and this is precisely what the $\mathcal{D}$-introduction rule does.

3.2 Correctness

The previous subsection shows that in order to obtain a complete notion of deduction for global validity all we need to do is to add the rule of $\mathcal{D}$-introduction to a complete system for local validity. But at this point we must be careful since a system obtained by the addition of $\mathcal{D}$-introduction might turn to be *too complete* as it is shown in the counterexamples to classically valid patterns of inference in section 1.3. If the system for local validity that we take to define the system for global validity contains rules that are not always globally valid (such as conditional proof), the addition of $\mathcal{D}$-introduction will render a system complete but not correct (we will be able to prove, for example, $\vdash_g \varphi \rightarrow \mathcal{D}\varphi$ which is not valid by supervaluationist standards).

At this point there are two possible alternatives. The first one would be focussing on rather succinct axiomatic systems in which the rules of deduction are always globally valid. Though this alternative might be logically satisfactory, it is not satisfactory from a more philosophical point of view. In particular, in order to address Williamson's objection, we should show how to incorporate deductive systems with rules like conditional proof etc. In order to incorporate such systems we should put some restriction on the applicability of *problematic* rules. Now, it should be noted that restrictions on the applicability of rules is a common place in formal logic (think, for example, on the rules of $\forall$-*introduction* and $\exists$-*elimination* in standard formulations for first-order logic) and so, it seems to me, that the fact that we should make use of restrictions does not constitute an objection *per se*.

The particular way in which we might formulate these restrictions would depend, partly, on the particular form of the deductive system in question. My proposal is to restrict the applicability of *problematic* rules to proofs that do not make use of the rule of $\mathcal{D}$-introduction. For example, if the proof showing that $\Gamma \cup \{\psi\} \vdash_g \varphi$ involves *any* application of $\mathcal{D}$-introduction, we are not allowed to use conditional proof to get $\Gamma \vdash_g \psi \rightarrow \varphi$. The restriction formulated this way might look a bit drastic and it is perhaps possible to formulate restrictions in

a more sensitive way, but the point is that this restriction guarantees the correctness of the system without destroying our previous completeness argument. When we consider restrictions on the applicability of rules of $\vdash_l$, the sensitive cases in the argument above are: the step from (ii) to (iii) and the step from (iii) to (iv). The step from (ii) to (iii) requires that every local proof is a global proof, but the previous restriction respects this fact since local proofs do not make use of $\mathcal{D}$-introduction (any local proof meets the restriction). The step from (iii) to (iv) is justified by analogous reasons: since local proofs do not make use of $\mathcal{D}$-introduction, the restriction on the applicability of rules do not restrict the number of local proofs ($\vdash_l$ is still complete after the restriction). This abstract consideration on the restriction of applicability of *problematic* rules might look a bit mysterious, so let us look to a particular example.

The inference from φ to $\psi \rightarrow \mathcal{D}\varphi$ is globally, but not locally valid. One might think that an appropriate way to provide a global proof would be something like this,

1 $\{\psi, \varphi\} \vdash_g \varphi$ [Reflexivity]
2 $\{\psi, \varphi\} \vdash_g \mathcal{D}\varphi$ [From 1, by $\mathcal{D}$-introduction]
3 $\{\varphi\} \vdash_g \psi \rightarrow \mathcal{D}\varphi$ [From 2, by conditional proof]

however, our restriction on the applicability of rules like conditional proof would render the step from 2 to 3 illegitimate. Now, if our previous remark on the restriction is correct (that is, if the restriction does not destroy the previous completeness argument), there must be a way to write the proof that respects the restriction. The *natural* way to do it (perhaps the only one) is this:

1 $\{\psi, \mathcal{D}\varphi\} \vdash_g \mathcal{D}\varphi$ [Reflexivity]
2 $\{\mathcal{D}\varphi\} \vdash_g \psi \rightarrow \mathcal{D}\varphi$ [From 1, by conditional proof]
3 $\{\varphi\} \vdash_g \mathcal{D}\varphi$ [Reflexivity and $\mathcal{D}$-introduction]
4 $\{\varphi\} \vdash_g \psi \rightarrow \mathcal{D}\varphi$ [From 2 and 3, by Cut]

Note that the restriction gives to any proof the same pattern as the one followed in the previous completeness argument (in particular step from (i) to (ii)). The general strategy to construct a global proof respecting the restriction is this: (i) assume the premises are as definite as you need for the proof and proceed classically (that is, here you might make use of any local rule, this corresponds to steps 1 and 2 in the example); (ii) *reduce* the $\mathcal{D}$'s attached to the premises making use of the rules of $\mathcal{D}$-introduction, Reflexivity and Cut (this corresponds to steps 3 and 4 in the example).

3.3 The Two Questions Revisited

The discussion in section 2 raised two questions concerning global validity. The first, more technical, whether we might provide adequate deductive systems for global validity. The second, concerning the form of these systems, whether it is possible to incorporate rules like conditional proof, contraposition, argument by cases and *reductio* and what is the aspect of formal proofs within these systems. In this third section we have found an answer to these two questions.

In the first place, we have provided a simple procedure to extend any adequate notion of deduction for local validity to a complete notion of deduction for global validity. In the second place, the section provides a positive answer to the second question by showing a way in which we might incorporate the aforementioned rules placing a suitable restriction on their applicability. It is worth noting that the aspect of global proofs respecting this restriction is completely classical, with the exception of the last step.

These answers to the two questions raised in subsection 2.3 contribute to the debate on the non-classicality of supervaluationist logic presented in subsections 2.1 and 2.2. Varzi's pessimism is overcome by providing a simple way to adapt classical systems of deduction for supervaluationist logic. *Problematic* rules are perfectly applicable with the exception of proofs in which the rule of $\mathcal{D}$-introduction has already been used. This last remark provides a precise sense to Keefe's quoting above according to which classical rules can be applied when the $\mathcal{D}$-operator is not involved. Thus, I think that section 3 shows a precise sense in which Keefe's original suggestion works perfectly fine. On the other hand, the result qualifies Williamson's claim according to which supervaluationism invalidates our natural form of deductive thinking. While there's a sense in which Williamson's claim is correct (since supervaluationist logic for a language with $\mathcal{D}$ gives raise to counterexamples to classically valid rules), there is another sense in which the claim must be qualified. Since we might employ systems of deduction correct and complete for global validity in which the problematic rules are present (with restrictions) and such that formal proofs in these systems are completely classical; with the exception of a single last step.

References

1. Blackburn, P., Rijke, M., Venema, Y.: Modal Logic. Cambridge University Press, Cambridge (2001) (Reprinted with corrections: 2004)
2. Cobreros, P.: Supervaluationism and logical consequence: a third way. Studia Logica 90, 291–312 (2008)
3. Cobreros, P.: Supervaluationism and Fara's paradox of higher-order vagueness. In: Egré, P., Klinedinst, N. (eds.) Vagueness and Language Use. Palgrave McMillan, Oxford (2010) (forthcoming)
4. Field, H.: This magic moment: Horwich on the boundaries of vague terms. In: Dietz, R., Moruzzi, S. (eds.) Cuts and Clouds: Vagueness, its Nature and its Logic. Oxford University Press, Oxford (2010)
5. Fine, K.: Vagueness, truth and logic. Synthese 30, 265–300 (1975)
6. Keefe, R.: Theories of Vagueness. Cambridge University Press, Cambridge (2000)
7. Tappenden, J.: The liar and sorites paradoxes; towards a unified treatment. Journal of Philosophy 90(11), 551–577 (1993)
8. Tye, M.: Sorites paradoxes and the semantics of vagueness. In: Tomberlin, J.E. (ed.) Philosophical Perspectives 8: Logic and Language. Ridgeview, Atascadero (1994)
9. Varzi, A.: Supervaluationism and its logic. Mind 116(463), 633–676 (2007)
10. Williamson, T.: Vagueness. Routledge, New York (1994)
11. Williamson, T.: On the structure of higher-order vagueness. Mind 108(429), 127–143 (1999)

Perceptual Ambiguity and the Sorites

Paul Égré*

Institut Jean-Nicod (CNRS)
DEC-ENS, 29 rue d'Ulm, 75005 Paris, France

Abstract. The sorites paradox results from two equally plausible constraints on categorization in sorites series: a constraint of category switch between the first and the last items, and a constraint of similarity or consistent judgment for adjacent items. Following the work of D. Raffman [27,28] this paper argues that both constraints can be met if we assume that borderline cases pattern as ambiguous items between opposing categories. I first review some empirical evidence in favor of this view. I then examine how it bears on the tolerance principle, from a descriptive and from a normative viewpoint. In particular, I discuss ways in which the account of tolerance outlined in [6] can be related to Smith's [37] fuzzy account, as well as to the similarity-based semantics for vague predicates proposed by van Rooij [34] and explored in recent work with Cobreros et al. [4].

1 Introduction

In previous work [6] I have argued that soritical series might be fruitfully conceived in relation to what I called Fisher series – after psychologist G. Fisher – namely sequences of ambiguous stimuli whose degree of ambiguity varies along the sequence (see [11]). Here I would like to give a more elaborate discussion of the relation I see between the two kinds of series. The aim is to propose an account of the sorites paradox based on the notion of ambiguity, and to say more about the tension there is between similarity and aspect change along a sorites series.

* I am indebted to D. Bonnay, S. Bromberger, P. Cobreros, L. Decock, R. Dietz, I. Douven, V. de Gardelle, L. Goldstein, J. Hampton, D. Lassiter, P. Pagin, D. Pressnitzer, D. Ripley, B. Spector and R. van Rooij for discussions and numerous intellectual exchanges on the issues addressed in this paper. Special thanks go to D. Raffman for the inspiration coming from her work and for earlier exchanges on the topic of this paper. Further special thanks go to R. van Rooij, D. Ripley and P. Cobreros for the work accomplished together on strict/tolerant semantics, as well as to I. Douven, R. Dietz and L. Decock for our work connecting vagueness and prototypicality. Each of these collaborations considerably widened and enriched my views about vagueness. I am also grateful to L. Goldstein, D. Ripley and to two ESSLLI reviewers for valuable comments that substantially improved the first version of this paper. I thank the ANR program 'Cognitive Origins of Vagueness' (ANR-07-JCJC-0070) for support, as well as the ANR program 'MultiStap' (ANR-08-BLAN-0167-01).

R. Nouwen et al. (Eds.): ViC 2009, LNAI 6517, pp. 64–90, 2011.

Rather than ambiguity, it may be more appropriate to talk of ambivalence. However, the notion of ambivalence, whether in its semantic or psychological version, is in some sense common to any account of vagueness, and it comes to mean different things depending on the theory. In supervaluationism, for example, borderline cases of a predicate P are presented as semantically indeterminate cases that can be precisified *one way or the other*. On the epistemic theory of vagueness, borderline cases of a predicate P are cases that give rise to uncertainty: though objectively they are exactly one of P or not P, from a subjective point of view *they might be* P *or they might be not* P. The view of borderline cases I favor is yet a third view, on which borderline cases are ambiguous items between two adjacent categories: they can legitimately be viewed in two distinct and even opposite ways, though to various degrees.[1]

The view differs from epistemicism in considering that even if our discriminative capacities were optimal, some cases along a sorites series would still leave us some choice in how best to categorize them. The view also differs from supervaluationism in that it tends to view ambiguity as more fundamental than indeterminacy. It may turn out that ambiguous cases between two categories will be resolved one way or the other, and it does not rule out that ambiguity could lead to semantic indeterminacy depending on the case. As I see it, however, it is their ambiguity that comes first, not their indeterminacy status.

On the view I favor, borderline cases are therefore best conceived as cases that are both P and not P, in agreement with fuzzy accounts of vagueness, dialetheist theories, and glut theories more generally. In presenting those cases as both P and not P, I really mean, however, that penumbral cases are cases to which both P and its negation can be legitimately applied (see [42,43,36]). This does not mean, however, that P and its negation are jointly applicable under exactly the same respects. Rather, I consider that borderline cases are cases that we perceive as unstable between P and its negation: on one occasion, we might resolve them as P, on another as not P. We may judge them to be both P and not P, but based on the evaluation we make of their very instability, namely based on some alternation between two distinct conceptualizations.[2]

The view I will sketch in this paper derives its main inspiration from the work of D. Raffman on the sorites [27,28]. Furthermore, it owes very much to ongoing work I have been pursuing with colleagues in several directions, both on the epistemology of vagueness, and on the semantics of vagueness. In [6], I outlined a view of borderline cases in sorites series as cases more or less *equipotential* between two competing percepts or categories, that is as cases coming with non-zero probabilities of eliciting either of the two categories alternatively.

[1] Schiffer [35] and MacFarlane [25] distinguish between vagueness as *ambivalence* (taking-to-be-partially-true) and vagueness as *uncertainty* (partially-taking-to-be-true). The connection between vagueness and ambiguity I countenance in this paper is much in the spirit of this technical notion of ambivalence, though it rests on a different foundation.

[2] See Cobreros et al. [4] for more on this, in particular in relation to the data obtained by Ripley (this volume) and Alxatib and Pelletier [1].

In subsequent work with Douven, Decock and Dietz, we built on a more general characterization of borderline cases essentially as cases *equidistant* between prototypes in conceptual space, given a suitable metric.[3] In the simplest setting, equidistant cases thereby correspond to ambiguous cases. In further work with Cobreros, Ripley and van Rooij [4], we investigated a semantics originally proposed by van Rooij (see [34]), on which borderline cases are cases *equisimilar* to P cases and to non-P cases, given a suitable similarity relation.

At bottom, I see all of these approaches as compatible with each other, and as converging on the idea that borderline cases between distinct categories are cases of overlap between concurrent representations. In the present paper, I will draw on these various approaches, but in an attempt to show this convergence and to further the earlier approach taken in [6]. In particular, taking a retrospective view, I would like to indicate how the degree-theoretic approach sketched in that paper can be related to the more general model-theoretic framework proposed by R. van Rooij and explored in our joint work with P. Cobreros and D. Ripley.

I shall proceed as follows. In section 2 of the paper, following Raffman, I argue that the sorites paradox essentially results from two opposing constraints, namely a static constraint of pairwise similarity and a dynamic constraint of category switch, which can be made compatible with each other if we assume that borderline cases pattern ambiguously in sorites series. In section 3, I review further evidence for Raffman's view of category switch as Gestalt switch in sorites series. More specifically, I argue that Fisher-type series of ambiguous figures give us further insight into the relation between the two constraints of similarity and category shift and into the graded structure of ambiguity. In section 4, I bring together these elements to refine the semantic account of the notion of tolerance outlined in [6]. I establish a correspondence in particular with the accounts of tolerance proposed independently by Smith [37] and by van Rooij [34]. In section 5, finally, several issues left open by the present account are discussed. In particular, the account of borderline cases as ambiguous cases allows us to weaken the tolerance principle so as to make it compatible with category switch. However, further work is needed to explain contextual effects and judgmental dynamics in sorites series, in particular to account for the phenomena of enhanced similarity between pairs and for the phenomenon of boundary displacement depending on the order of presentation.

2 Ambiguity in Sorites Series

2.1 Pairwise Similarity vs. Category Switch

The sorites paradox usually results from two intuitive constraints about the way categorization and discrimination ought to work together in sorites series.

[3] Equidistance is only the first component in the picture of vagueness we build there. The other main component concerns the multiplicity of prototypical points for a given concept.

The first constraint is a constraint of *similarity* or consistent judgment between adjacent items in the series. This constraint is articulated differently depending on the theory. Most of the time, it is presented as a principle of tolerance [40], namely as the idea that if x and y are highly similar and if x is P, then y too will be P. Occasionally, the constraint is presented as restricted to pairwise judgments (see [7,22]). It says that when we look together at two items x and y, such that these two items differ only very slightly in the relevant respects, we are reluctant or unwilling to judge the one P and the other not P.[4] For instance, if the difference in hue between two color shades is very small, then to the extent that we judge the first red, we will judge the second red and conversely.

This constraint is in tension with the intuition that as we are shown items of a sorites series consecutively, a jump or *category switch* should occur between the first and the last item, at least to the extent that those are stably or reliably assigned to distinct and exclusive categories. For instance, we expect that if we judge the first item stably as red, and the last stably as orange, then at some point in the series we will have to issue a differential judgment between at least two consecutive shades.

These two constraints of *similarity* and *category switch* appear to conflict with each other, because while the latter predicts that there will be two consecutive items x and y such that x will be judged red and y not red, the former seems to imply that x and y ought either both to be judged red, or both to be judged not red.

2.2 Similarity: Static vs. Dynamic

Importantly, however, there need not be a conflict between these two constraints, if we consider that the first essentially concerns *static* or simultaneous categorization, and that the second concerns *dynamic* or sequential categorization. A natural resolution of the conflict between those two constraints is indeed to consider that context plays a central role in the way we categorize (see [27,28,7]). In particular, whenever we see two very similar items presented simultaneously in pairs, the pair-context itself appears to enhance similarity between the two items. But it may happen that when we see the same two items each in isolation on different occasions, or one after the other but as part of a larger transition between other shades, we will issue a differential judgment.

[4] Fara's statement of the constraint is that "if two things are saliently similar, then it cannot be that one is in the extension of a vague predicate, or in its antiextension, while the other is not." Kennedy's statement is that "When x and y differ to only a very small degree in the property that a vague predicate g is used to express, we are unable or unwilling to judge the proposition that x is g true and that y is g false." Unlike the tolerance principle, the constraint of salient similarity makes an important restriction to pair contexts, and as such, it is not incompatible with the idea of category switch. I remain deliberately imprecise on the distinction at this point and refer to sections 4 and 5.2 below for a more precise comparison between the two principles.

Think, for instance, of two shades a and b such that a is only slightly redder than b, and b only slightly more orange than a. If we view them simultaneously next to each other on a fixed background, they might appear to present the same color. Now consider a case in which you are shown only shade a on one slide; then a is replaced by a redder shade c on a second slide; then the red shade c is replaced by shade b on a third slide. The dynamic interpolation of c may then enhance contrast rather than similarity between a and b in this case, and may have the effect of pushing b in the orange category (see [16] for evidence on such contrasts). In such a context we may be more likely to judge a red and b orange.

Similarly, imagine you present a and b consecutively, but in the context of a transition series from a very clear red to a very clear orange. In such a case, we may be able to perceive the overall direction of the transition from red to orange, even though we do not see differences locally: this very perception may help us to issue a differential judgment between a and b at the moment we go from a to b in the sequence.

2.3 The Ambiguity Hypothesis

How then can we integrate the two constraints of pairwise similarity between adjacent items and category shift within a sorites series? Following Raffman, the hypothesis we will make is that category shift will occur in a region of the series where shades are likely to be perceived ambiguously.

To take a toy example, suppose a series of three shades such that a_1 and a_3 are discriminable when viewed together side by side, in such a way that a_1 looks rather red, and a_3 rather orange. Now suppose an intermediate shade a_2 such that a_1 and a_2 are hard to discriminate and present the same color quality when viewed side by side, and similarly such that a_2 and a_3 are hard to discriminate and present the same quality when viewed side by side. Statically, the constraint of similarity predicts that to the extent that a_1 is viewed stably as red, a_1 and a_2 should be judged red together; on the other hand, it predicts that to the extent that a_3 appears stably as orange, a_2 and a_3 would be judged orange together. The upshot is that a_2 is a shade that has the potential of being judged either red or orange, depending on the context: it will look more red next to a_1 and more orange next to a_3.

a_1	a_2	a_3
Red	Red	Orange
Red	Orange	Orange

As a result, when shades are presented consecutively in a dynamic sequence, this view predicts that a category shift can occur either between a_1 and a_2, or between a_2 and a_3, without impugning the constraint of static similarity between adjacent items.

A further interesting aspect of the ambiguity view concerns the position of the shift depending on the direction of the transition from a_1 to a_3 or from

a_3 to a_1. It is known from the experimental literature that category switch in ordered series of gradually shifting stimuli tend to happen at different positions in the series, depending on the direction of change. Typically, sensitivity to the direction of change manifests itself as *hysteresis*, namely as the longer persistence of the category from which one is coming (see [28,30,15,6]). In our toy model, this would predict that to the extent that the end shades are assigned to the same respective categories across all conditions, the boundary between 'Red' and 'Orange' tends to be positioned between a_2 and a_3 when starting from a_1, and conversely that the boundary tends to be positioned between a_2 and a_1 when starting from a_3. Sometimes, however, the displacement of boundaries manifests itself in a dual form, that is the shift can happen earlier rather than later, though in a way that remains sensitive to the direction of the transition (see [20] for evidence on this fact).[5] In our example, this would predict that subjects would tend to switch from 'Red' to 'Orange' between a_1 and a_2 when starting from a_1, and between a_3 and a_2 when starting from a_3. Either way, the phenomenon of displacement of boundaries manifests that there is a range of intermediate shades that are categorized in opposite ways depending on the context. The static ambiguity of those shades is thus resolved in different ways depending on the dynamics of presentation.

Sorites series of only three elements are obviously too short for our example to be realistic.[6] For all its simplicity, however, the toy example we just discussed contains the main ingredients of the view of the sorites I favor. This view owes a great deal to the analysis of the sorites originally proposed by D. Raffman, who compared the phenomenon of category switch along a sorites series to a phenomenon of Gestalt switch [27,28]. However, while Raffman has made a number of central observations on the ambiguous patterning of intermediate items in sorites series, she did not quite propose to see ambiguity as the core of her account of vagueness. In particular, Raffman [29] considers that from a semantic point of view, if x is a borderline Red, then x is simply not Red.[7] As I will argue below, I am more inclined to the view that if x is a borderline case of Red, then x is indeed Red, though to a lesser degree than other Red candidates. By the ambiguity thesis, however, I am also committed to the idea that there is a semantically legitimate sense in which x is not Red. Before examining the ways in which this view can be semantically articulated in sections 4 and 5, in the next section I first review additional evidence in favor of the ambiguity view of borderline cases inspired by Raffman. Following [6], the idea of that section is

[5] I am indebted to an anonymous reviewer for bringing Kalmus's paper to my attention. See below for a more detailed discussion of hysteresis and of Kalmus' opposite finding.

[6] Indeed, as a reviewer points out, if a_1 and a_2 look the same pairwise, and a_2 and a_3 look the same pairwise, then presumably a_1 and a_3 look too similar for each of them to look stably red or stably orange singly. I consider more realistic examples in what follows, namely series including more shades.

[7] Raffman [29, p. 2] calls this the Incompatibilist view: "On the resulting view — I call it the Incompatibilist View— borderline cases for vague predicate 'Φ' are not Φ, the sentence 'x is not Φ' is true, and the sentence 'x is Φ' is false."

to show that the very phenomenon of Gestalt switch can be modulated in a way that accords with the graded structure of vague categories.

3 Percept Switch and Category Switch

Our picture of the structure of sorites series leaves two issues open. The first concerns the structure of the ambiguous region in more extended sorites series: are all ambiguous items ambiguous to the same extent? The second concerns the position where category switch can be expected to occur in the series, and also the question where it can legitimately occur. Our toy example obviously does not allow us to investigate these questions in full generality, given that only one shade, the middle shade, is predicted to be ambiguous. Suppose therefore that we are now dealing with a series of say 8, or 15, or 30 color patches making a gradual transition from a clear red to a clear orange. Is there a privileged position for the switch to occur along the series?

3.1 Fisher Series

Our answer to the first question is negative, based on the consideration that perceptual ambiguity is a gradable notion. An adequate illustration of this phenomenon can be found in the example of Fisher's series of ambiguous figures [11]. One particular example of Fisher's stimuli consists of a set of 15 ambiguous cards such that card 1 strongly favors the perception of a man's face ("the Gypsy"), while card 15 strongly favors the perception of a girl holding a mirror ("the Girl"). Consecutive cards in the series differ by small alterations. Overall, each card in the series supports the two concurrent percepts, but not to the same degree.

To test this phenomenon, Fisher presented cards in random order to a sample of 200 subjects and asked each of them individually to report which percept they saw first. What Fisher's data show is that while cards 1-3 strongly favor the identification of the 'Gypsy' as first percept (more than 80% of subjects reported the percept), cards 14-15 strongly favor the identification of the concurrent percept (less than 5% of subjects maintained the 'Gypsy'). The statistical distribution of answers gradually decreases from one end to the other and comes closest to a half toward the middle of the sequence (namely card 7).

A basic explanation for this phenomenon is that although the two percepts are coinstantiated in each card, they are prominent to different degrees, depending on the cues available in the stimulus. Arguably, the same ambiguity phenomenon can be expected for transition series of colors. We should expect that the closer a given patch of color stands to prototypical red in color space, the more likely it is to be categorized as red. The distance of a color shade to a prototypical value should thus constrain the probability that we perceive the shade as red rather than not red over repeated presentations. In what follows, we shall therefore assume that each item in a sorites series comes with a prior probability representing the degree to which it makes the relevant category prominent or salient.

3.2 Pairwise Similarity and Synchronization

The example of Fisher series is instructive in a second way, for although the likelihood of each percept varies from one card to the next, adjacent cards tend to elicit the same percept when considered pairwise. Thus, consider what happens if you screen off all cards but allow yourself to view only adjacent cards at a time. The effect is that to the extent that the percept 'Girl' comes to the fore in the right-hand card, it will usually come to the fore in the left-hand card. As for bistable stimuli more generally, one can draw one's attention to make the alternative percept salient, namely one can make the 'Gipsy' percept come to the fore, but in this case the 'Gipsy' will tend to come to the fore in the adjacent card.

Thus, although the percepts are not salient to the same degree in each card, a phenomenon of binding or synchronization appears to constrain one's perception of adjacent stimuli when they are presented pairwise together. This phenomenon is usually exemplified in the perception of multiple bistable stimuli: switches between percepts tend to be synchronized in this case (see [19] for a review; Flugel [12] originally investigated the perception of multiple Necker cubes). Interestingly, this constraint of synchronization is defeasible though. It means that even when we see the same ambiguous figures side by side, we may occasionally get non-synchronized switches between percepts. However, the important point is that there is a strong tendency for the percepts to be synchronized for identical bistable stimuli.

To my knowledge, the robustness of this synchronization phenomenon has not been tested for pairs of bistable figures that vary slightly, such as adjacent cards in Fisher series, or motion quartets with neighboring aspect ratios.[8] One may wonder, more generally, how the synchronization between switches varies as a function of the degree to which each stimulus makes each percept salient individually, and also as a function of the distance between the stimuli themselves. On the first issue in particular, one expectation might be that where the rivalry between percepts is the greatest in individual presentation (e.g. toward the middle cards), the binding between percepts, even for exactly identical stimuli, is more fragile; and by contrast, that where rivalry between percepts is minimal (e.g. for the end cards), the binding is more robust. A different expectation could be that synchronization is a more local phenomenon, which is maintained exactly to the same degree throughout. Thus we could imagine that while switches occur more often for motion quartets with 1:1 aspect ratio than for motion quartets with 2:1 aspect ratio, this higher frequency of switches does not affect synchronization between the switches when two quartets are presented together.

Experimental evidence is obviously needed to adjudicate this matter, but the evidence available suggests that while the stability of a percept is indeed a function of the stimulus position in the series and of its absolute distance to some focal value, synchronization is primarily a function of the relative distance

[8] Motion quartets (see [18, 17]) are dynamic stimuli consisting of dots that flicker alternatively at the opposite vertices of a virtual rectangle. They give rise to distinct percepts (horizontal or vertical motion) which can be biased depending on the aspect ratio of height to width of the rectangle.

or relative similarity between stimuli. The same distinction, arguably, is in play when we judge colors. Pairwise similarity appears to be primarily a function of the relative distance between items in the relevant perceptual space. On the other hand, category switch appears to be primarily a function of the absolute distance to some designated value, such as the perceptual distance to some prototypical red, or to some prototypical orange.[9]

3.3 Position of the Switch

Let us now turn to the question of where category shift can be expected to occur in a sorites series. This question can receive a descriptive as well as a normative interpretation. From a descriptive point of view, the question is where the shift typically occurs in a sorites series. From a normative perspective, the question is whether there are particular positions in the series where one ought to switch category or not.

Let us examine the descriptive question first. Where a category shift does occur in a sequence of items making a transition between two categories is known to be sensitive to the order of presentation of the stimuli. As already mentioned, the position of the switch can vary depending on whether the stimuli are presented in random order, in ascending order, or in descending order (we discuss this phenomenon in greater detail below). Our basic assumption, however, is that for all such conditions, the categorization of an item is primarily constrained by the degree of similarity that item has to a given prototype (see [27,5]). Suppose we are dealing with an ideal perceiver who would categorize solely according to the probabilities attached to these degrees of similarity, and who can only answer with "Red" or with "not Red". For instance, imagine a_1 is a prototypical red coming with probability 1 of eliciting the percept "Red", while a_2 is a slightly less typical red, more on the orange side, coming with a probability .9 of eliciting the percept "Red". In the random order case, we should expect that on most trials, a_1 and a_2 elicit the same percept. This will mean that we get a majority of (Red, Red,...) profiles, and only few (Red, not Red,...) profiles. In ordered sequence, a natural expectation is that the consecutive presentation of a_1 and a_2 would have the effect of enhancing their similarity. So we may expect to see even fewer (Red, not Red,...) profiles. The reasoning cannot be generalized, however. In some cases the ordered presentation of consecutive items in the sequence can displace the boundary up the sequence (hysteresis), in other cases it may displace it down the sequence (reverse hysteresis, or enhanced contrast). In some cases, finally, ordered presentation may have no main effect over random presentation (critical boundary).[10] The only robust expectation for all these cases, then, is that the shift will seldom if ever happen between the first and second items in a sorites series, at least for sufficiently fine-grained series between two categories.

[9] The distinction echoes Raffman's [27] distinction between *discriminatory* and *categorial* judgments: "Whereas the former correspond to comparisons of two presented patches, the latter presumably pertain to 'comparisons' of a presented patch with some sort of standard or prototype in memory" [27, p. 48].

[10] I am using the terminology of Kelso [21]. See below for details.

Consider now the normative question, the one we ultimately care about from a philosophical point of view. Descriptive considerations give us some indication on where switches are unlikely or likely to occur. But they do not tell us whether there is a privileged position where one ought to switch category in a sorites series. Is there such a position? The epistemicist postulates that there is one, namely that there is at some point a cut-off between say the last Red shade and the first non-Red shade. In principle, switching category at an earlier position or at a later position is a mistake. Suppose, however, that items in a sorites series only come with different potentials of eliciting this or that category, in the same way in which items in a Fisher series come with different degrees of ambiguity, namely with different probabilities of eliciting this or that percept. Then it would be very counter-intuitive to suppose that there is a unique and pre-determined cut-off along the series. Rather, if the vagueness of categories implies that some stimuli have an ambiguous status vis a vis at least two categories, and if this ambiguity itself is gradable, then the idea of a unique legitimate cut-off disappears.

This argument, in a nutshell, is the gist of the proposal made in [6]. In that paper, I suggested that it may be fruitful to think of soritical series in analogy with Fisher series. The point is that, in Fisher's original series, the two percepts ('Girl' vs 'Man') are co-present in each stimulus, though not to the same degree. Because of that, a percept switch appears to be legitimate at any point in the series, but also, it appears equally legitimate not to switch percept along the series. In other words, no stimulus in the series rationally mandates a shift, even though every stimulus rationally justifies such a switch. The general argument can be summarized as follows: if every sorites series contains a range of stimuli that can be compared to Fisher's stimuli in his series, then we ought to consider that from a normative point of view, a category switch is permissible for the whole range. No item of that range is such that one ought to switch. Only more extreme stimuli on each side of that range may be taken to mandate a shift one way or the other.[11]

[11] Quite remarkably, the same analogy between aspect-switching in series of ambiguous figures and category switching in sorites series was proposed independently by L. Goldstein [14, pp. 111-112], with no acquaintance with Fisher's work, but toward essentially the same point made here (I am grateful to L. Goldstein, p.c., who informed me about this after reading the first draft of this paper). There, Goldstein draws a series of five duck-rabbit figures with different biases towards one or the other interpretation. About those, he asks: "Is there a 'perfect' duck-rabbit, a maximally ambiguous figure such that each (non-aspect-blind) observer, gazing at the picture, switches between the two interpretations, seeing the rabbit aspect half of the time and the duck aspect the other half? No – different observers react differently to the same picture just as they do to the Necker Cube. There is no uniformity of switching-behavior. And perhaps for the same reason as there is no 'perfect' duck-rabbit, there is no 'perfect' red-orange marking the exact boundary in the series of colour patches between those that are really, objectively red, and those that are really, objectively orange." A slight divergence between our respective views is that I do not exclude the possibility of a maximally ambiguous figure, as Fisher sought to measure, namely a figure where both within-subject and between-subject ambivalence would come closest to a theoretical maximum.

On the present view, therefore, borderline cases of a category are cases for which it is rationally permissible to switch the category, or to maintain the category. The main benefit of this view is that it makes room for both within-subject and between-subject variability regarding category-switching. As the reader may feel, however, this view of borderline cases as ambiguous cases does not necessarily rid us of hidden boundaries. In particular, we need to consider the possibility that there is a last unambiguous case and a first ambiguous or borderline case along the series. In that case, this would imply that there is indeed a first position where one ought to switch, and a last position for which it is permissible not to switch category. The idea that there might be a first position where one ought to switch, and a last position where one can maintain the category may appear to beg the problem. Even under this assumption, however, I shall argue that what fundamentally matters is the idea of a sufficient gap between these two positions, namely the idea of an extended region between two cut-offs instead of a unique cut-off.

Smith [37], in his book on vagueness, criticizes epistemicism based on what he calls the 'jolt problem', namely the view that there would be two consecutive items x and y that are very similar in P-relevant respects, but such that $P(x)$ and $P(y)$ must have very different truth values. Much the same intuition is in play in the present argument, though articulated differently. Indeed, the core of our account is the idea that there should not be two consecutive and highly similar items x and y such that one ought to judge x P and one ought to judge y not P. There should be no 'normative' jolt in that sense. Arguably, there remains a jolt between the last item one can legitimately judge P and the first that one ought to judge not P. But arguably too, this jolt does not have the same status as the first one. In the next section I shall say more about how the view of borderline cases as ambiguous cases can be semantically articulated in relation to these normative intuitions.

4 The Tolerance Principle

The account of the sorites outlined in the previous section implies that some cases in a sorites series of color patches from red to orange can legitimately be categorized as red or as not red, depending on how their ambiguity is resolved along the series.

In [6], I argued that this view agrees with the conception of borderline cases between categories defended by Wright in particular [42,43], in which borderline cases are conceived as cases for which opposite verdicts are equally permissible. Based on this, I then proposed a normative version of the tolerance principle, intended to make explicit the relation between relative similarity and the deontic notions of obligatory and permissible judgments. I furthermore contrasted it with a descriptive version of tolerance, intended to make explicit the probabilistic relation between relative indiscriminability and sameness of categorization.

Since then it occurred to me that the normative version of tolerance I was stating in deontic terms can be seen as a particular case of the modal account of tolerance originally proposed by van Rooij [34] and elaborated in [4]. Similarly, I discovered that the descriptive version of tolerance I was stating in probabilistic terms parallels the concept of closeness defended by Smith [37] in terms of degrees of truth. In this section I revisit both of these formulations of tolerance. I show how the descriptive and normative understandings can be related, and discuss the correspondence with those accounts.

4.1 Closeness and Permissibility

The tolerance principle is stated in various forms in the literature on vagueness. Standardly, it is phrased in the following way:

(1) Whenever x and y are only very slightly different in respects relevant for the application of P, then if x is P, y is P.

To state the principle more formally, I will follow van Rooij [34] and will write $x \sim_P y$ to state that x and y are only very slightly different in respects relevant for the application of P, or equivalently that they are indiscriminable in the relevant respects:

(2) $x \sim_P y \rightarrow (P(x) \rightarrow P(y))$

Importantly, this principle relates a constraint on discrimination to a constraint on categorization. Under the assumption of symmetry of the similarity or indiscriminability relation $\sim_P$, (2) implies that when two objects are indiscriminable in the relevant respects, they are categorized alike, that is they are either both P, or both not P. As such, (2) comes very close to the similarity constraint explained above.[12] However, assuming classical logic this principle implies that the first and the last member of a sorites series should instantiate the same category, which is inconsistent with the idea that there ought to be a category switch along the series.

In [6], I argued that (2) is at best a rough paraphrase of the actual intent of (1). I proposed to distinguish between a descriptive and a normative articulation of (1). Both of those are considerably weaker than (2) from a logical point of view, and are therefore compatible with category switch in sorites series. In what follows I give a closer examination of those reformulations.

Tolerance as Closeness. As a descriptive approximation of the tolerance principle, I proposed that "if the probability for a given stimulus n to be seen as A is α, then the probability for a sufficiently similar stimulus $n+1$ to be seen as A should be sufficiently close to α" [4, p. 110]. That is, given a vague predicate

[12] However, I am not assuming $x \sim_P y$ to mean that x and y are *saliently* similar. As such, (2) fails to capture the notion of salient similarity. See below for an attempt to capture salient similarity proper.

P, a metric d and a probability distribution p, there must be relevantly small positive real values ε and δ such that for every x and y:[13]

(3) if $d(x, y) \leq \varepsilon$ then $|p(P(x)) - p(P(y))| \leq \delta$

Interestingly, this understanding of the notion of tolerance corresponds very tightly to what N. Smith independently describes as a principle of *closeness* in his degree-theoretic account of vagueness [37, p. 151]. For Smith, the standard principle of tolerance is false, but its adequate substitute is that if x and y are very close in P-relevant respects, then $P(x)$ and $P(y)$ will be very close in respect of truth, which Smith writes as follows (where $[P(x)]$ is the degree of truth of $P(x)$, $\approx_T$ is the relation of closeness between degrees of truth, and $\approx_P$ the relation of closeness in A-relevant respects):

(4) if $x \approx_P y$ then $[P(x)] \approx_T [P(y)]$

The main difference between Smith's account and the one explained here is that Smith's principle presupposes a notion of degree of truth, whereas the formulation I offered relies on the notion of probability.[14] Unlike Smith, I tend to consider that degrees of truth proper are not needed for a theory of vagueness to be adequate. However, other notions of degrees appear to me to play a functionally similar role. These include the degree to which an object is typical of a property, or the degree to which an object is similar to another. On that view, to say that some object is very red is equivalent to saying that it has a high degree of redness. But I do not consider that the degree to which "x is red" is true should be greater than the degree to which "y is red" is true when x is redder than y. Nevertheless, I believe that the redder a shade is, the more likely it is to be recognized as red.

In the case of (3), $p(P(x))$ should therefore be seen as denoting the degree to which x typically instantiates the property P, or the degree to which P is manifest in x. An assumption will be that such degrees are much like propensities, so that the degree to which x typically instantiates the property P probabilistically constrains the judgment that x is P. Thus, $p(P(x)) > p(P(y))$ may be taken to mean that the expected frequency of judgments of the form "x is P" is greater than the expected frequency of judgments of the form "y is P" in a categorization task in which ideal subjects have to respond by yes or no to each sentence. An idealization I will make is that $p(P(x)) = 1$ if and only if x prototypically and unambiguously instantiates property P. A simplification of the account will be that as we deal with ideal subjects, the answers they give to $P(x)$ are

[13] I follow [6] here. Strict inequalities could be used instead of large inequalities. As far as I can see, nothing of importance hinges on that. However ε and δ must be sufficiently small to make the constraint non-trivial. Note that, as in [37], the requirement is distinct from continuity, but basically asks for bounded variations in the stimulus to be matched by bounded variations in the judgment.

[14] See also Lassiter (this volume) for a probabilistic treatment of vagueness, leading to a similar diagnosis of the tolerance principle, though based on distinct premises. See [25] for an account combining degrees of truth with subjective probabilities.

directly constrained by $p(P(x))$, without interference from other personal factors or probabilistic limitations.[15]

The distance d, on the other hand, must really be thought of as associated with a metric for discrimination: in the case of colors we are dealing with, $d(x, y) \leq \varepsilon$ means that in a Same vs. Different discrimination task, the frequency of correct judgments lies in a certain interval (typically an interval of size ε centered around $1/2$, namely an interval around chance level). More generally, $d(x, y) \leq \varepsilon$ can be used as a definition of the relation $x \sim_P y$.[16] Thus, each side of the conditional expressed in (3) can be given probabilistic significance, but the point is that the antecedent corresponds to a metric for discrimination, while the consequent corresponds to a metric for categorization.

Tolerance as Permissibility. In agreement with this weakening of the notion of tolerance, I proposed to state a normative version of the notion of tolerance in terms of the deontic notions of obligation and permissibility. The principle I suggested is that if an individual x ought to be judged P, then it is not the case than an individual y that differs only slightly from y in the relevant respects ought to be judged not P. Letting $\mathcal{O}$ stand for the complex operator "it ought to be judged that", this principle corresponds to:[17]

(5) if $x \sim_P y$ then $\mathcal{O}P(x) \rightarrow \neg\mathcal{O}\neg P(y)$

If we define permissibility as the dual of obligation, and write $\mathcal{P}$ for this operator, then this principle corresponds to:

(6) if $x \sim_P y$ then $\mathcal{O}P(x) \rightarrow \mathcal{P}P(y)$

This says that whenever x and y are sufficiently similar in the relevant respects, to the extent that x ought to be judged P, it is permissible to judge y P (though not mandated).

In order to relate this principle to the notion of closeness, the following *ad hoc* semantics was offered. Consider a sorites series a_1 to a_n of objects and associate with each item a_j a prior probability $p(P(a_j))$ that it be judged as P, so that $p(P(a_0)) = 0$, $p(P(a_n)) = 1$, and for every individual x and property P, $p(\neg P(x)) = 1 - p(P(x))$. Assume that for any two consecutive items a_i and a_{i+1}, $p(P(a_i)) \leq p(P(a_{i+1}))$ and $|p(P(a_i)) - p(P(a_{i+1}))| \leq \delta$ for some $\delta < 1$. Declare ambiguous those individuals x in the series whose probability $p(P(x))$ lies strictly between 1 and 0, namely that have a non-zero potential of being seen as P as well as $\neg P$.

[15] I briefly return to this point below, see fn. 26.

[16] See [24, p. 34 sqq.] on the correspondence between algebraic and probabilistic accounts of indiscriminability. Van Rooij [33,34] handles indiscriminability relations algebraically on the basis of Luce's semi-order axioms; the metric approach of indiscriminability expressed in $d(x, y) \leq \varepsilon$ can be directly related to Luce's probabilistic definition of such a semi-order.

[17] This simplifies the more complex notation used in [6], in which "ought to" and "judge" were represented by distinct operators, though eventually treated as a semantic unit.

We can then stipulate that the only individuals that ought to be judged P are those whose probabilistic degree of P-ness is 1, namely those that unambiguously instantiate category P. Call P-similar all and only individuals that are pairwise adjacent in the series, namely $a_i \sim_P a_j$ iff $|i - j| \leq 1$. It is easy to check that any series satisfying the constraints laid out satisfies principle (5).

For illustration, consider a series of 8 individuals with the following stipulated P-potentials, assuming $\delta = .2$:

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
0	.1	.3	.5	.5	.7	.9	1

In this particular case, a_1 is the only shade that ought to be judged $\neg P$, while a_8 is the only individual that ought to be judged P. By duality, for all individuals distinct from a_1, it is permissible to judge them P, and all individuals distinct from a_8 it is permissible to judge them $\neg P$. A prediction is that the intermediate individuals a_2 to a_7 can be judged P as well as $\neg P$, without contradiction.

The idea of a connection between gradability and permissibility features in other recent accounts of vagueness based either on degrees of truth or on probabilities. In particular, Lassiter (this volume) gives a probabilistic interpretation of the main premise of the sorites exactly congruent with our model: no two consecutive items in a sorites series are such that the probability of the predicate applying abruptly switches from 0 to 1 from one individual to the next. Similarly, MacFarlane [25, p. 48] makes an explicit link between the degree of truth 1/2 and lack of error about opposite judgments when he writes that: "when it is true do degree .5 that Harry is bald, it will be just as correct to believe that Harry is bald as it is to believe that it is true that Harry is bald as it is to believe that it is false that Harry is bald. This, I think, admirably captures the 'ambivalence' felt in borderline cases." Our model is even more liberal in this respect, since in principle it is sufficient for an item to ambiguously instantiate a category to a non-zero degree to justify the correctness of the corresponding judgment.

Relation between the Principles. The descriptive rendering of tolerance as closeness means that when the probability of correctly discriminating between x and y is sufficiently low, the probabilities of categorizing them alike will be close. The principle of permissibility on the other hand means that the cases that one ought to judge P and those that one ought to judge not P should be sufficiently far apart. The exact relation between these two principles is not entirely obvious, however, and deserves further examination.

A first general remark is that both principles are obviously weaker than the standard tolerance principle. Both leave open the possibility that two consecutive items be categorized differently along a sorites sequence. Both principles are therefore compatible with the constraint of category switch. On the other hand, neither of these principles by itself gives us a direct handle on the constraint of simultaneous pairwise similarity, or on further contextual effects in categorization (such as hysteresis). We will say more about it in the next section.

A second remark concerns principle (5). In the semantics we stated, we assumed that cases x that ought to be judged P correspond to $p(P(x)) = 1$; cases that ought to be judged $\neg P$ correspond to $p(P(x)) = 0$; by duality, cases for which it is permissible to judge P are those such that $p(P(x)) > 0$ and cases for which it is permissible to judge $\neg P$ are those such that $p(P(x)) < 1$. In other words, it is assumed that obligation coincides with full absence of ambiguity. From a technical point of view, the choice of 1 and 0 is mostly a stipulation, however, and we could imagine that one ought to judge something P or not P provided it is *sufficiently* unambiguously P or *sufficiently* unambiguously not P. All we need for that is a sufficient gap between the thresholds relevant for $\mathcal{O}P$ and $\mathcal{O}\neg P$, not necessarily an extended gap such as between 0 and 1. This assumption may even be more realistic in some cases, since arguably, a very slight probabilistic degree for a property P, very near 0, may appear too thin to guarantee the legitimacy of the corresponding judgment.

More generally therefore, given a non-empty domain D of individuals, we may call a *p-model* (for potential or probabilistic model) a structure $M = (D, p, \alpha, \beta)$ where p is a function that associates to each individual d and property P what can be called the P-potential of d, that is the probability $p(P(d))$ that d will elicit the judgment that $P(d)$, or the degree to which P is manifest in d. By definition, $p(\neg P(d)) = 1 - p(P(d))$. Similarly, α and β are two functions that associate to each predicate P values α_P and β_P such that $0 \leq \alpha_P < \beta_P \leq 1$. These two functions fix the thresholds relevant for obligation and permissibility. In what follows we consider essentially one predicate, so we shall write α and β instead of α_P and β_P.

Let $M \models \mathcal{O}P(d)$ iff by definition, $p(P(d)) \geq \beta$ and $M \models \mathcal{O}\neg P(d)$ iff $p(P(d)) \leq \alpha$.[18] From this definition it follows that for every d, d' such that $|p(P(d)) - p(P(d'))| < \beta - \alpha$, we have that $M \models \mathcal{O}P(d) \rightarrow \mathcal{P}P(d')$. For instance, suppose $\alpha = .3$ and $\beta = .8$, then the permissibility principle (5) will be true iff $x \sim_P y$ implies that the difference between the P-potentials is less than .5. More generally, (5) will be true exactly on the condition that:

(7) if $x \sim_P y$ then $|p(P(x)) - p(P(y))| < \beta - \alpha$

The latter condition corresponds closely to the descriptive principle (3). In particular, (3) implies (7) provided $\delta < \beta - \alpha$. This means that the descriptive version of tolerance implies the normative version provided the probabilities of categorizing two adjacent individuals as P do not vary by more than the interval between what one ought to judge P and what one ought to judge not P.

4.2 Comparisons

Gap Principles. Our deontic principle (5) bears a relation with analogous weakenings of the tolerance principle discussed in the literature. Van Rooij [34] discusses the link between the tolerance principle and several modalized versions

[18] Alternatively, we could stipulate that $0 < \alpha \leq \beta < 1$, and use strict inequalities rather than large inequalities in the semantics of obligation.

that are weaker and that can be expressed in terms of operators of clarity or definiteness. Among those figures Williamson's *margin of error* principle, which says that if x is *clearly* P, then every y sufficiently similar to x in the relevant respects is P:

(8) if $x \sim_P y$ then $\Box P(x) \rightarrow P(y)$

An even weaker version of this principle is the principle that if x is *clearly* P, then no y sufficiently similar to x in the relevant respects is clearly not P:

(9) if $x \sim_P y$ then $\Box P(x) \rightarrow \neg\Box\neg P(y)$

This principle, first introduced by Wright, is called a *gap principle* by Fara [8], since it implies that a gap is mandated between cases that are clearly P and cases that are clearly $\neg P$.[19] As it turns out, the normative principle (5) that we put forward above is exactly a gap principle in this sense. The only difference is that we deal with obligation and not with clarity, but the difference is not substantial, since we basically propose that cases that ought to be judged P are those that are sufficiently clearly P, or sufficiently close to the prototype for P.[20]

Van Rooij [34] offers an in-depth discussion of the relation between these various modal weakenings of the notion of tolerance so I shall not discuss which of these principles might be the best substitute for tolerance. Rather, I would like to say more about the connection between the degree-theoretic semantics proposed to deal with (5) and the relational semantics proposed by van Rooij to internalize the modal version of tolerance expressed in (9) in a standard first-order language. I will not provide all the details of the semantics here, but refer to [34] and [4] for a systematic presentation.

Strict and Tolerant Semantics. Van Rooij's semantics is defined for first-order classical models M equipped for each predicate P with a binary relation $\sim_P$ that is reflexive and symmetric, but not necessarily transitive (such models are called T-models in [4]). The semantics rests on two dual notions of truth for vague predicates, a notion of strict truth (s-truth) and a notion of tolerant truth (t-truth). By definition $M \models^s P(a)$ iff every individual $d \sim_P a$ is such that M classically satisfies $P(d)$.[21] Furthermore $M \models^s \neg P(a)$ iff every individual $d \sim_P a$ is such that M classically satisfies $\neg P(d)$. Dually, $M \models^t P(x)$ iff it is not the case that $M \models^s \neg P(x)$, and similarly, $M \models^t \neg P(x)$ iff it is not the case that

[19] Gap principles were first discussed by Wright and Sainsbury in relation to higher-order vagueness. See for instance Wright [41] and Fara [8] for more on the genealogy of such principles.

[20] Interestingly, Fara [8] gives reasons not to take such gap principles on board, however her reasons are mostly based on the interaction of gap principles with a distinct rule of D-introduction present in supervaluationism. See [3] for a defense of gap principles, and the response in [9].

[21] For simplicity I make no distinction here between an object and its name. See [4] for a rigorous presentation.

$M \models^s P(x)$. The strict and tolerant truth of more complex sentences is defined recursively in the usual way.

Thus, strict truth for P internalizes the idea that P holds clearly, while tolerant truth for P internalizes the idea that the negation of P does not hold clearly. The semantics also allows us to distinguish between two modalizations of the standard notion of tolerance namely between:

(10) a. if $x \sim_P y$ then it is not the case that: ($M \models^s P(x)$ and not $M \models^s P(y)$)
b. if $x \sim_P y$ then it is not the case that: ($M \models^s P(x)$ and $M \models^s \neg P(y)$)

The two principles differ on the scope of negation in the second conjunct: negation takes scope over $\models^s$ in (10)-a (*qua* metalanguage negation), and it takes scope below $\models^s$ in (10)-b (*qua* object-language negation). The upshot of van Rooij's semantics is that (10)-a is not a valid principle of the induced logic, while (10)-b is a valid principle. Obviously, (10)-a is a very strong principle, since it mandates that if x is strictly P, then every individual sufficiently similar is strictly P as well. By contrast, (10)-b is in fact equivalent to the principle that if x is strictly P, then any y sufficiently similar is tolerantly P, namely:

(11) if $x \sim_P y$ then $M \models^s P(x)$ implies $M \models^t P(x)$

As explained in [34], we can view $M \models^s P(x)$ as shorthand for $M \models \Box P(x)$, and $M \models^t P(x)$ as shorthand for $M \models \neg\Box\neg P(x)$. Consequently, the version of tolerance we get in (10-b) can be seen as a reflection of Fara's gap principle or of our deontic version of tolerance in (5).

Starting from a potential model $M = (D, p, \alpha, \beta)$, we can easily define a correspondence with notions of strict and tolerant truth as follows:

(12) a. $M \models^s P(x)$ iff $p(P(x)) \geq \beta$
b. $M \models^t P(x)$ iff $p(P(x)) > \alpha$.

We can then define $M \models^s \neg\phi$ compositionally as in [34], that is as $M \not\models^t \phi$, and $M \models^t \neg\phi$ as $M \not\models^s \phi$. Similarly, let $M \models^s \phi \wedge \psi$ iff $M \models^s \phi$ and $M \models^s \psi$, and $M \models^s \forall x\phi$ iff for all d in D, $M \models^s \phi[d/x]$, and build up tolerant truth analogously for conjunction and universal quantification. Finally, assume that $x \sim_P y$ only if $|p(P(x)) - p(P(y)| < (\beta - \alpha)$. As in van Rooij's semantics, it will follow that if $x \sim_P y$ then $M \models^t P(x) \rightarrow P(y)$ (which is equivalent to (11), assuming a material interpretation of the conditional). By this correspondence, we thus equate tolerant truths with propositions that it is permissible to judge true, and strict truths with propositions that one ought to judge true.

A difference between van Rooij's original semantics and the present one is that van Rooij's approach starts from T-models, namely similarity structures in which we start from a reflexive, symmetric similarity relation $\sim_P$ for each predicate P, and we moreover fix classical extensions for each predicate P. Strict truth and tolerant truth, in the atomic case, are defined from classical truth and similarity.

Here, however, we do not have a notion of classical truth at hand, and we do not have an explicit definition of $a \sim_P b$, so we cannot write: $M \models^s P(d)$ iff for every d' such that $a \sim_P b$, $M \models^c P(d')$. However, starting from a p-model $M = (D, p, \alpha, \beta)$ it is possible to turn it into a T-model by letting:

(13) a. $a \sim_P b$ iff $|p(P(a)) - p(P(b))| < \varepsilon$ with $\varepsilon = (\beta - \alpha)/2$.
b. $M \models^c P(x)$ iff $p(P(x)) \geq \beta - \varepsilon$.

One can prove that for every p-model that is sufficiently rich, namely such that for every predicate P and p-value α there is an object x in the domain such that $p(P(x)) = \alpha$, assuming the conditions stated in (12) and (13), $M \models^s P(d)$ exactly if for every $d' \sim_P d$, $M \models^c P(d')$, and similarly $M \models^t P(d)$ exactly if there is some $d' \sim_P d$ such that $M \models^c P(d')$.[22]

Consider for instance such a p-model M in which for P, the threshold for strictness and tolerance are $\alpha = 0.2$ and $\beta = 0.7$ respectively. Then $\varepsilon = 0.25$, and $M \models^c P(d)$ iff $p(P(d)) \geq .45$. By definition, $M \models^s P(d)$ iff $p(P(d)) \geq 0.7$, and indeed, one can check that $M \models^s P(d)$ iff for every d' such that $d' \sim_P d$, $M \models^c P(d')$.

This correspondence between T-models and p-models is interesting, because it shows a bridge between a qualitative and a quantitative approach to the notion of ambivalence felt in borderline cases. A particularly nice feature of the strict/tolerant semantics is indeed that it provides a qualitative description of this ambivalence. On this account, a borderline case of P is a case that is similar to a classically P case and also that is similar to a classically non-P case. Classical extensions on that approach may be seen as playing the role of an underlying dividing line between competing representations. On the probabilistic approach we followed, by contrast, the idea of ambivalence is represented directly by the assignment of intermediate numerical degrees to particular sentences, as in fuzzy logics more generally. The possibility of defining strict and tolerant either numerically or relationally suggests that these two representations communicate in deeper ways.[23]

Central Gaps. The recipe we just gave to build a T-model from a p-model predicts that relative to the thresholds α and β, which set the boundaries for tolerant and strict extensions for a predicate P, the threshold for membership in the classical extension will be half-way between α and β. We may note that under those assumptions (13)-b allows us to strengthen the necessary condition stated above in (7) for similarity into a necessary and sufficient condition. In

[22] I am indebted to D. Ripley for pointing out the need for the richness assumption on p-models in order to secure those biconditionals.

[23] More is to be said about the correspondence between many-valued logics and similarity-based logics based on this, but this is left for further work. In [4], we discuss the connection between strict/tolerant and three-valued logic in particular. There we also discuss the psychological plausibility of the similarity-based approach to borderline cases and ambivalence in relation to psycholinguistic data obtained by D. Ripley (this volume) and by S. Alxatib and J. Pelletier [1].

effect, what this says is that the minimal difference in P-potential relevant for whether one ought to judge x and y either P or not P must be bigger than the minimal difference relevant to discriminate between x and y under the relevant respects, namely twice as big.

In this, the present proposal also bears a close affinity to the pragmatic treatment of vague predicates proposed by Pagin [26]. Basically, Pagin offers an account on which the use of vague predicates is constrained by a *central gap* whose size "must be at least equal to the tolerance level", where the latter designates the size of the step relevant to measure differences in the predicates application. The details of Pagin's proposal are different, but here the notion of a central gap corresponds to the difference relevant regarding what one ought to judge P and what one ought to judge not P: the correspondence we just laid out shows that this gap, as expressed by the difference $\beta - \alpha$, must indeed be greater than the difference in P-potential corresponding to a difference in discrimination.

5 Discussion and Problems

To close this paper, I propose to conclude with a brief discussion of some issues left open by the account here proposed of sorites series. The first issue concerns the determination of the parameters α and β in our p-models. The second issue concerns the account of contextual effects on judgment in the sorites.

5.1 Where Does Ambiguity Begin?

In section 2 we gave several reasons why appeal to ambiguity in sorites series could allow us to deal with the sorites paradox without inconsistency. First of all, by assuming an area of ambiguous stimuli between two categories, we can explain category switch along the series, without giving up on the idea that the same consecutive items that are judged differentially dynamically could be judged to match when considered pairwise and statically. Secondly, by assuming that ambiguity comes in degrees, and that ambiguity extends over several items along a series, we can explain that category switch can occur at different points without any error along the series. In this we surmised that there is an essential connection between ambiguity in category membership, and permissibility regarding opposite verdicts.

Formally, however, the normative view of tolerance that we presented is essentially faithful to the idea that we can make a tripartition of cases for a given predicate: there are the cases that one ought to judge P, the cases that one ought to judge not P, and an intermediate region of cases that one can judge P as well as not P. The same tripartite view is fundamentally in play in the distinction between strict cases of P, strict cases of not P and intermediate cases that are tolerantly P and not P. In [4], in particular, we show the existence of a natural correspondence between T-models for strict truth and tolerant truth and trivalent models underlying Kleene's strong logic and its dual, Priest's Logic of

Paradox, where the value 1 represents strict truth, 0 strict truth to the contrary, and the intermediate value $1/2$ the notion of borderlineness.[24] In the case of our p-models, similarly, this tripartition is determined by the parameters α and β relative to each predicate P.

As for any tripartite approach, the main objection we need to face concerns the actual vagueness of those three regions.[25] What is the status of the boundary between the last case we ought to judge P and the first case we can judge not P? More generally, where does ambiguity begin, and where does it end? In assuming the existence of two critical values α and β, we may seem to postulate the existence of arbitrary and overly precise boundaries regarding what counts as permissible. We owe an account of the way in which those values are set. More generally, we owe an account of how P-probabilities are bestowed upon individual items.

One solution to the latter problem, following [5], is to let P-potentials depend on the degree of similarity to some prototypical or focal values. In the case of a color predicate like "red", the value 1 should characterize cases that are prototypically red; for a predicate like "tall", it should characterize the tallest individual in the comparison class (viz. [33]), and so on, *mutatis mutandis*, depending on the predicate. By contrast, the value 0 should fall on cases that do not instantiate the relevant property in any respect (cases that instantiate a distinct prototype). In principle the value $1/2$ will be bestowed upon items that fall exactly between adjacent categories in the relevant conceptual space, so on items that are maximally ambiguous for that matter. More generally, we can conceive that for each predicate, given a context, we issue judgments by first locating the best instances of the target category and then build representations of counter-instances accordingly.

More realistically, however, we may conceive that there are as many p-models relative to a predicate and domain of objects as there are subjects who issue judgments. If we let P-potentials vary depending on the subject, much like personal probabilities, then we would predict that each subject starts with a possibly different appreciation of what counts as prototypical for a property. It seems a reasonable assumption, however, to maintain that what makes an item prototypical is the high degree to which it makes a property manifest across subjects. The main motivation to introduce the normative thresholds α and β in what precedes concerns precisely this fact. In principle, α and β may be seen as degrees that are sufficiently close to the values 1 and 0 to take account of the limited variability between subjects regarding what counts as prototypical for that property.

[24] See [32] for an in-depth discussion of such trivalent approaches to the sorites. See [4] for a correspondence between T-models and trivalent models.

[25] The same issue faces the metric account developed in [5], where borderline regions of a predicate are predicted to have sharp boundaries, as well as the similarity-based account of [4]. In the account developed by Douven et al., however, the size and shape of the boundary region is predicted to depend on the size and shape of the regions associated with prototypes in the relevant conceptual space.

The hypothesis I am making is that even if we assume the existence of such ideal values for each vague predicate, we nevertheless end up with a better account of borderline cases than if we assume a unique sharp boundary for each predicate. Epistemicism postulates that rules governing the use of vague predicates determine one essentially unknowable boundary for each predicate (see e.g. [39]). In my opinion, it might be less of an idealization to suppose the rules governing the use of vague predicates determine two such boundaries, an upper one and a lower one, rather than a single boundary.

The main motivation I see for this is that in sorites series we can recognize that some stimuli are ambiguous or equal candidates for opposite verdicts, and that we would not make any mistake in judging them either way. The values α and β may still fluctuate and be subject to uncertainty themselves, depending on how we appreciate typicality.[26] However, the basic idea is that the application we make of vague predicates is constrained for salient or typical values in a way in which it is not for intermediate values. If we declare a prototypical red yellow, for instance, we are indeed making an error. But as we move away from that prototypical red, we quickly enter an area where we become free to set the boundary in the best way we can. A fuller account of this problem would oblige us to engage with the issue of higher-order vagueness, but it would be beyond the scope of this paper to address it here.

5.2 Contextual Effects: Salient Similarity, Hysteresis, Monotonicity

In the previous section, we noted that both the descriptive version of tolerance expressed in (3) and the normative version expressed in (5) are logically weaker than the original tolerance principle (assuming classical logic as our background logic). Both of those formulations make room for the possibility that some item x is judged P, while the next item y, while not reliably discriminable from y, will be judged not P. Consequently, both of those principles make room for category switch along a sorites series. However, we can see that those principles do not

[26] A limitation of our account in this respect concerns the idealization on which P-potentials attached to an object x for a predicate P reflect both the likelihood of (ideal subjects) judging that object P and the degree to which P is manifest in x. Viewed as likelihoods, P-potentials were meant to account for how subjects behave in sorites series. Viewed as degrees attached to properties, and in relation to the parameters α and β, they are much like degrees of truth, namely objective values that subjects need to appreciate in order to make acceptable judgments. Ultimately, and as explained convincingly by MacFarlane [25], we need a more refined model distinguishing subjective uncertainties from the degrees attached to properties in relation to particular objects (whether as degrees of truth or as degrees of ambiguity, as on the present account). Note that if we maintain the identity of P-potentials with probabilities of P-judgments, we nevertheless predict the possibility of error for our ideal subjects, but in a rigid way. Suppose $\beta = .8$, and that $p(P(x)) = .8$. Then on average, in 20% of the cases in which the subject ought to judge $P(x)$, the subject will violate this normative requirement. Error cannot happen in that sense, however, when α and β are set at 0 and 1 respectively.

deliver any straightforward account of the idea of salient similarity between adjacent stimuli when viewed pairwise, nor of order effects in sorites series.

Salient Similarity. Let us consider first the phenomenon of salient similarity between pairs [7]. The constraint of similarity, as stated for instance in Kennedy [22] or van Rooij [33], says that if x and y vary only slightly, then we are unwilling to judge the one P and the other not P, when considering them together. The framework of tolerant/strict semantics suggests an elegant way of handling this constraint. Consider a model M in which $a \sim_P b \sim_P c$, and such that a and b both are classically P, while c is classically not P. It follows from the semantics that $M \models^s P(a)$, and that $M \not\models^s P(b)$, since b is in fact similar to a not-P individual. However, consider the submodel of M consisting of only a and b, which we shall write as $M, \{a, b\}$. Relative to it, we have that: $M, \{a, b\} \models^s P(a)$ iff $M, \{a, b\} \models^s P(b)$. This holds for any model consisting of just a pair of P-similar items, namely we have:

(14) a. if $a \sim_P b$ then $M, \{a, b\} \models^s P(a)$ iff $M, \{a, b\} \models^s P(b)$
b. if $a \sim_P b$ then $M, \{a, b\} \models^s \neg P(a)$ iff $M, \{a, b\} \models^s \neg P(b)$

If we suppose that strict truth is the norm of our judgment in this case, or indeed the norm of assertion,[27] then it will follow that when we consider items together pairwise, we either judge them both to be P, or judge them both to be not P, or refrain from making any judgment regarding whether they are P or not. This agrees with the idea that we are unwilling to judge one P and the other not P. Again, the important point is that this constraint does not hold in general, if we consider larger models. In this, van Rooij's semantics allows us to capture something like Fara's constraint of salient similarity between adjacent items.

Without further stipulations, however, or without the correspondence with the framework of strick/tolerant semantics, the basic machinery of p-models does not provide such a direct treatment of the notion of salient similarity in terms of model restriction. Indeed, p-models only purport to represent the probabilistic potentials of seeing items P or not P in isolation, and independently of context. When we view two shades of colors next to each other, or two cards side by side as in Fisher series, the phenomenon of enhanced similarity between colors, or the phenomenon of synchronization between figures, obviously invite us to consider additional parameters, namely some dependencies between probabilities.

Hysteresis. Regarding hysteresis, one possibility to model the judgmental dynamics in play in sorites would be to consider that depending on the direction in which a sorites is run, each of the prior probabilities $p(P(x))$ is increased or decreased uniformly. The effect would be that on average, the first switching point from P to not-P would automatically be displaced, to the right or to the left. We may note that hysteresis, namely the longer persistence of the category one is coming from, can be accommodated in the framework of tolerant/strict

[27] See [4] for a detailed discussion of this hypothesis.

semantics if we suppose that subjects judge tolerantly, as far as possible, in agreement with their initial judgment.[28]

As mentioned earlier, however, Kalmus [20] obtained data on color naming that indicate a phenomenon of 'reverse hysteresis'.[29] In a task in which subjects had to name color patches making a gradual transition from one color to an adjoining color, subjects were found to switch category closer to the category they were coming from, rather than later. In terms of the tolerant/strict distinction, this would suggest that the rule subjects then used was the opposite rule, namely to judge strictly, as far as possible, in agreement with one's initial judgment, and then to switch to the opposite category.

Prima facie, there may appear to be a conflict between Raffman's finding of hysteresis in a task of color judgment between Green and Blue and Kalmus' finding of reverse hysteresis in a task of color naming. However, the duality between these two findings suggests that such order effects do not merely depend on the direction of the transition, but on further parameters. Kalmus, for instance, points out that subjects in his task were given prior notification of the fact that they would make a transition from one color to an adjoining color. Kalmus conjectures that the prior notification may have reinforced the anticipation for change in most subjects. Furthermore, the time between the presentation of consecutive patches was fairly long (several seconds, as colors were changed manually). Raffman on the other hand used a different methodology, in which subjects after switching category were shown the immediately preceding shades, without explicit notification, and using a computer design. The rate of change in Kalmus's task, as well as the subjects' expectations regarding the switch, may thus have had an impact on people's resolution of the ambiguity of intermediate color patches. Kelso [21, p. 206] makes several observations on the dynamics of perception that appear to be compatible with this interpretation. In particular, he writes about motion quartets that: "Hock and colleagues found that the size of the hysteresis region was reduced when the rate of stimulus change was slowed. Dynamically, this reflects the presence of competing intrinsic tendencies: persistence under gradual parameter change favors the initially established perception, but slowing the rate of parameter change enhances spontaneous change."

Finally, I should add that in an unpublished pilot study conducted with Vincent de Gardelle and David Ripley in 2010, we were surprised not to find hysteresis in a task in which subjects had to judge true or false a sentence of the form "the shade is red" for each of twenty shades at the boundary between red and

[28] R. van Rooij (p.c.) mentioned such a possibility of accommodating hysteresis in the framework, based on a suggestion by M. Krifka.

[29] The phenomenon may be called *proteresis*, by parity of etymology: the term is coined and used by P. Girard and J-P. Boisset [13] in a pharmacology paper. The term used by Kelso [21, pp. 203-204] is that of *enhanced contrast*, which Kelso distinguishes from *hystereris* and from *critical boundary*. Critical boundary is the idea that the switch occurs at the same position, regardless of order. Hysteresis is when the switch happens at a larger position on the way up from a stimulus to another than on the way down. Enhanced contrast is when the switch happens at a smaller position on the way up than on the way down.

orange. We actually observed no significant difference depending on the order presentation, although we found hysteresis in an initial task of color matching run on the same stimuli. While more work is needed to prove the robustness of a possible contrast between matching and naming, we see that the variability in these various findings leaves open a number of questions about the source and exact manifestation of order effects in relation to vagueness.

Monotonicity. A last constraint that we cannot explain without further assumptions concerns monotonicity in one's judgment (see [16]; see [4] for discussion in the tolerant/strict framework). Generally, if subjects switch their judgment from 'P' to 'not P', for instance from 'red' to 'not red', they should rationally stick to their new judgment after the switch, if indeed they can perceive the direction of the transition (from more red to less red, or conversely). If we assume that subjects judge colors simply according to the probabilities $p(P(x))$, then we cannot rule out inconsistent profiles of answers, however. Again, a full account of this monotonicity constraint in judgmental dynamics calls for a more elaborate model.

6 Concluding Remarks

The idea that ambiguity might help to solve the sorites paradox is not new. As mentioned in the beginning of this paper, it underlies a number of different approaches to the sorites, most notably Raffman's approach, but also fuzzy theories and dialetheist accounts, depending on how the concept is articulated. The idea has also been criticized. Sorensen [38], for instance, considers that "the assimilation of vagueness to ambiguity makes the sorites paradox too easy to solve". I believe, however, that careful consideration of the way perceptual ambiguity works and gets resolved makes the hypothesis of ambiguity in sorites series more appealing and more solid than acknowledged by Sorensen. In agreement with Sorensen, however, we should not infer from that that vagueness is reducible to ambiguity. Lexical vagueness and lexical ambiguity in particular are two distinct phenomena (viz. [2]). Ambiguity is primarily a property of stimuli, words or expressions, depending on whether it is perceptual, lexical or syntactic. Vagueness is a property of concepts or representations. The principled distinction between the two phenomena does not imply, however, that some fruitful connections between them cannot be established.

References

1. Alxatib, S., Pelletier, J.: The psychology of vagueness: Borderline cases and contradictions. Mind and Language (2010) (forthcoming)
2. Bromberger, S.: Vagueness, Ambiguity and the "Sound" of Meaning. MIT, Cambridge (2008) (manuscript)
3. Cobreros, P.: Supervaluationism and Fara's argument concerning higher-order vagueness. In: Egré, P., Klinedinst, N. (eds.) Vagueness and Language Use. Palgrave Macmillan, Oxford (forthcoming)

4. Cobreros, P., Egré, P., Ripley, D., van Rooij, R.: Tolerant, classical, strict. The Journal of Philosophical Logic (2010) (forthcoming)
5. Douven, I., Decock, L., Dietz, R., Egré, P.: Vagueness: A conceptual spaces approach (2010) (manuscript, under review)
6. Egré, P.: Soritical series and Fisher series. In: Hieke, A., Leitgeb, H. (eds.) Reduction: Between the Mind and the Brain, pp. 91–115. Ontos Verlag (2009)
7. Fara, D.: Shifting sands: an interest-relative theory of vagueness. Philosophical Topics 28(1), 45–81 (2000)
8. Fara, D.: Gap principles, penumbral consequence, and infinitely higher-order vagueness. In: Beall, J. (ed.) Liars and Heaps: New Essays on Paradox, pp. 195–221. Oxford University Press, Oxford (2003)
9. Fara, D.: Truth in a region. In: Egré, P., Klinedinst, N. (eds.) Vagueness and Language Use. Palgrave Macmillan, Oxford (2010) (forthcoming)
10. Fine, K.: Vagueness, truth, and logic. Synthese 30, 265–300 (1975)
11. Fisher, G.: Measuring ambiguity. The American Journal of Psychology 80(4), 541–557 (1967)
12. Flugel, J.: The influence of attention in illusions of reversible perspective. British Journal of Psychology 5, 357–397 (1913)
13. Girard, P., Boisset, J.P.: Clockwise hysteresis or proteresis. Journal of Pharmacokinetics and Biopharmaceutics 17(3), 265–300 (1989)
14. Goldstein, L.: Clear and Queer Thinking: Wittgenstein's development and his relevance to modern thought. Rowman and Littlefield, Boston (1999)
15. Gregson, R.: Transition between two pictorial attractors. Nonlinear Dynamics, Psychology and Life Sciences 8(1), 41–63 (2004)
16. Hampton, J., Estes, Z., Simmons, C.L.: Comparison and contrast in perceptual categorization. Journal of Experimental Psychology: Learning, Memory and Cognition 31(6), 1459–1476 (2005)
17. Hock, H., Bukowski, L., Nichols, D., Huisman, A., Rivera, M.: Dynamical vs. judgmental comparison: hysteresis effects in motion perception. Spatial Vision 18(3), 317–335 (2004)
18. Hock, H., Kelso, J., Schöner, G.: Bistability and hysteresis in the organization of apparent motion patterns. Journal of Experimental Psychology 19(1), 63–80 (1993)
19. Hupé, J.M., Joffo, L.M., Pressnitzer, D.: Bistability for audiovisual stimuli: Perceptual decision is modality specific. Journal of Vision 8(7), 1–15 (2008)
20. Kalmus, H.: Dependence of colour naming and monochromator setting on the direction of preceding changes in wavelength. British Journal of Physiological Optics 32(2), 1–9 (1979)
21. Kelso, S.: Dynamic Patterns. MIT Press, Cambridge (1995)
22. Kennedy, C.: Vagueness and comparison. In: Egré, P., Klinedinst, N. (eds.) Vagueness and Language Use. Palgrave Macmillan, Oxford (2010) (forthcoming)
23. Lassiter, D.: Vagueness as probabilistic linguistic knowledge. In: Nouwen, R., et al. (eds.) ViC 2009. LNCS(LNAI), vol. 6517, pp. 127–150. Springer, Heidelberg (2011)
24. Luce, R.D.: Individual Choice Behavior. Dover, New York (1959) (Reedition Dover 2005)
25. MacFarlane, J.: Fuzzy epistemicism. In: Dietz, R., Moruzzi, S. (eds.) Cuts and Clouds: Vagueness, its Nature and its Logic, pp. 438–463. Oxford University Press, Oxford (2010)
26. Pagin, P.: Vagueness and domain restriction. In: Egré, P., Klinedinst, N. (eds.) Vagueness and Language Use. Palgrave Macmillan, Oxford (2010) (forthcoming)
27. Raffman, D.: Vagueness without paradox. Philosophical Review 103(1), 41–74 (1994)

28. Raffman, D.: Vagueness and context-relativity. Philosophical Studies 81, 175–192 (1996)
29. Raffman, D.: Borderline cases and bivalence. The Philosophical Review 114(1), 1–31 (2005)
30. Raffman, D.: Tolerance and the competent use of vague words. In: Unruly Words: A Study of Vague Language, ch. 5 (Book in Preparation) (2009)
31. Ripley, D.: Contradictions at the border. In: Nouwen, R., et al. (eds.) ViC 2009. LNCS (LNAI), vol. 6517, pp. 169–188. Springer, Heidelberg (2011)
32. Ripley, D.: Sorting out the sorites. In: Berto, F., Mares, E., Tanaka, K. (eds.): Paraconsistent Logic (tentative title) (2010) (forthcoming)
33. van Rooij, R.: Implicit vs. explicit comparatives. In: Egré, P., Klinedinst, N. (eds.) Vagueness and Language Use. Palgrave Macmillan, Oxford (2010) (forthcoming)
34. van Rooij, R.: Vagueness, tolerance, and non-transitive entailment (2010) (manuscript)
35. Schiffer, S.: The things we mean. Oxford University Press, New York (2003)
36. Shapiro, S.: Vagueness in Context. Oxford, New York (2006)
37. Smith, N.J.J.: Vagueness and Degrees of Truth. Oxford University Press, Oxford (2008)
38. Sorensen, R.: Ambiguity, discretion, and the sorites. The Monist 81(2), 217–235 (1998)
39. Williamson, T.: Vagueness. Routledge, London (1994)
40. Wright, C.: Language mastery and the sorites paradox. In: Evans, G., McDowell, J. (eds.) Truth and Meaning. Oxford (1976)
41. Wright, C.: Is higher order vagueness coherent? Analysis 52(3), 129–139 (1992)
42. Wright, C.: The epistemic conception of vagueness. In: Horgan, T. (ed.) Vagueness - Supplement of the Southern Journal of Philosophy. Oxford (1994)
43. Wright, C.: On the characterization of borderline cases. In: Ostertag, G. (ed.) Meanings and Other Things: Essays on Stephen Schiffer. MIT Press, Cambridge (2009) (forthcoming)

Context-Dependence and the Sorites

Graeme Forbes*

Department of Philosophy
University of Colorado
Boulder, CO 80309
USA
graeme.forbes@colorado.edu

Abstract. In Section 1 we describe the Sorites paradox and lay out options for a solution. In Section 2 we consider approaches which deny that all premises are true, and note that these solutions all seem open to a certain serious objection. In Section 3 we note a problem for the principle of transitivity of the conditional and present a *contextualist* resolution of the problem, according to which the "counterexamples" to transitivity involve the informal fallacy of shifting the context. In Section 4 we consider the possibility of applying the contextualist resolution of the general transitivity puzzle to Sorites arguments in particular, discussing the views of Kamp, Pinkal and Soames. Our negative conclusion, developed in Section 5, is that the paradoxes can be formulated in a way that does not commit the informal fallacy: context is held fixed. In the final section, we suggest a different defense against the objection used in Section 2.

1 Introduction

A Sorites paradox is an argument whose form is like this (the *Sorites scheme*):

$$\begin{array}{rcl} p_0 & & \\ p_0 & \rightarrow & p_1 \\ \vdots & & \vdots \\ p_{n-1} & \rightarrow & p_n \\ \therefore\ p_n & & \end{array}$$

There is also a *pure conditional* variant of the Sorites scheme, in which the minor premise is discarded and the conclusion is $\ulcorner p_0 \rightarrow p_n \urcorner$:

$$\begin{array}{rcl} p_0 & \rightarrow & p_1 \\ \vdots & & \vdots \\ p_{n-1} & \rightarrow & p_n \\ \therefore\ p_0 & \rightarrow & p_n \end{array}$$

* I thank David Barnett, Ewan Klein, Teresa Robertson, Ben Rohrs and two anonymous referees for discussion and comments which helped improve this paper.

R. Nouwen et al. (Eds.): ViC 2009, LNAI 6517, pp. 91–107, 2011.

What makes certain instances of either Sorites scheme paradoxical is that in them, p_0 is true and p_n is false, while because of the nature of the vocabulary in each p_i, and empirical facts about the objects mentioned in each p_i, each conditional premise of the form $\ulcorner p_{m-1} \rightarrow p_m \urcorner$ also seems true. Its seeming true is a result of its employing a *tolerant* concept, that is, a concept meeting the following condition: (i) the applicability of the concept to an object or in a situation depends on certain (perhaps only roughly) quantifiable features of the object or situation; (ii) significantly different quantities of these features can make an absolute difference between applicability and inapplicability of the concept; but (iii) there are differences in the quantities of these features of a magnitude *too small* to 'affect the justice with which [the concept] applies to a particular case' [20, pp. 156-157]. The problem is that insignificant differences can accumulate across cases into a significant difference.

Take, for example, the concept of being well-paid$_C$ by one's employer, where C is a reference-class and we are concerned only with individuals belonging to it; C might be, say, the class of Full Professors in humanities departments in the current *US News and World Report* top 50 research universities in the USA. It seems plausible that if any member of this group is well-paid, then any other member of the group who is paid less, but no more than \$100 less, than the given member, is also well-paid$_C$ (make it \$10 or \$1 if you have doubts about \$100). Suppose we agree as well that members of C are well-paid$_C$ if they are paid \$150,000 or more, and not well-paid$_C$ if they are paid \$75,000 or less. Then if $a_0, \ldots, a_{750}$ are 751 members of C such that a_0 is paid \$150,000 and a_i is paid \$100 more than a_{i+1}, we have an instance of the Sorites scheme that leads to a contradiction. For by our principles, 'a_0 is well-paid$_C$' is true, and each conditional 'a_i is well-paid$_C \rightarrow a_{i+1}$ is well-paid$_C$' is true, so 'a_{750} is well-paid$_C$' must be true as well; but a_{750} makes \$75,000, which means a_{750} is not well-paid$_C$. And the pure conditional version yields the conclusion 'a_0 is well-paid$_C \rightarrow a_{750}$ is well-paid$_C$', which is false. But chaining conditionals cannot lead from truths to falsehood.

There are only a few options for a resolution of this problem: (i) at least one conditional premise in the sequence is untrue[1], which, since there are only finitely many premises, means that some conditional premise is the *first* untrue premise; or (ii) there is something wrong with the logic; or (iii) there is some kind of semantic problem with the premises, such as an equivocation, which renders the logic inapplicable. In addition, any plausible resolution will have to include a good explanation of *why* the untrue conditionals seem true, or *why* the logic seems unassailable, or *why* the equivocation goes unnoticed. We take the options in turn.[2]

[1] I use 'untrue' as a catch-all, covering 'intermediate or false', 'neither true nor false, or false', 'not wholly true', 'not supertrue', and so on.

[2] Another constraint on a solution, which I will not directly address here, arises from an observation often made by Wright, namely, that the paradoxical arguments seem just as puzzling if the conditional premises are replaced by premises of the form $\neg(p \wedge \neg q)$; see, e.g., [21, *passim*]. Edgington [5] develops apparatus which addresses this; see further [7, p. 429, n. 11].

2 Untrue Premises

There are many attempts at solving the Sorites which claim to show that at least one conditional premise is untrue. The most conservative, *epistemicism*, says that there is a first premise in the list with a true antecedent and a *false* consequent, but it is impossible to know which premise it is, because it is impossible to know where the cut-off for being well-paid$_C$ lies.[3] The postulation of a sharp boundary at which a predicate like 'well-paid$_C$' abruptly ceases to apply, a boundary (dollar value, in this case) that is in principle unknowable, is the feature of epistemicism that generates the greatest resistance to it. But I will not attempt to evaluate the debunking explanation epistemicists give of the natural view that there is no such sharp boundary,[4] since the first thing I want to suggest here is that the problem for epistemicists is a problem for *everyone* who holds that some conditional premise in a Sorites is untrue.

It is hard to believe in the epistemicist's sharp boundaries because there appears to be nothing in virtue of which such a boundary could come to be. For there is no discontinuity in nature to mark the extension of a vague concept or predicate; nor is there some Board of Standards entitled to stipulate a boundary (except in cases like giving a legal definition of 'adult', which introduces a special sense of the word for special contexts). And there is no reason to think that the actual pattern of usage of a predicate like 'well-paid$_C$' in a community will somehow determine a sharp boundary.[5] Perhaps there is no reason at all why there is a sharp boundary, but there is one nevertheless. Perhaps, for example, a full professor in a humanities department in the current *USN & WR* top 50 national research universities is well-paid$_C$ iff he or she has a salary of at least $92,367.41, and that's just the way it is. That this is not really intelligible is the main objection to epistemicism.[6]

In the previous paragraph we have given an argument against sharp boundaries, not just cited an intuition: there must be an explanation why there is a sharp boundary, an explanation of how such a boundary comes to exist, but none of the possible grounds for such a boundary obtains. However, this argument applies equally well against other views about how there comes to be a first untrue premise in a Sorites. Such a premise must have a true antecedent, and for a conditional with a true antecedent to be untrue, on any view, is for its consequent to have a different semantic status than its antecedent.[7] So there

[3] For epistemicism, see especially [18] and [19].

[4] See [21, p. 87] for some negative comment with which I am in agreement.

[5] Williamson [19] argues that the meaning of a vague predicate supervenes on its use, which may be true. But it is a further question whether the meaning includes a sharp boundary and whether the pattern of usage fixes where that boundary is. There is no reason to think that it must, even if, by luck, it does.

[6] Two questions should be distinguished. One is why a sharp cut-off exists at all, another is why it falls where it does. I am arguing that there is no good explanation why a sharp cut-off exists. If there were, there might still be no explanation why it falls *here* rather than *there*. But that would be less objectionable.

[7] On supervaluationism, the consequent of the first non-supertrue premise is the first consequent for which there is an admissible sharpening making it false. See [8, pp. 253–254] for criticism of supervaluationism for introducing indefensible sharp distinctions.

will be an abrupt switch from the dollar values of salaries that put professors in the extension of 'well-paid$_C$' to dollar values that do not. We may be evaluating the conditional 'If Professor X is well-paid$_C$ then Professor Y is well-paid$_C$' in which the antecedent is true, or wholly true, or determinately true, while the consequent is not. So the specific \$100 (\$10, \$1) drop from the salary of Professor X to that of Professor Y marks a semantically significant transition. It's not a transition from true to false, certainly, but where it comes from is just as mysterious. For as before, there is no discontinuity in nature to mark the point of the transition; nor is there some Board of Standards entitled to stipulate the point at which 'well-paid$_C$' becomes, say, neither true nor false of a wage earner in C. And there is no reason to think that the actual pattern of usage of a predicate like 'well-paid$_C$' in a community will somehow determine a precise point of transition. That there is such a point appears to be a fiction.[8]

3 Faulty Logic

Perhaps a Sorites paradox can be taken as a refutation of the principles of classical logic on which the truth of its conclusion depends: there may be a problem either with *modus ponens*, or, for the pure conditional paradox, with transitivity of '$\rightarrow$'. Indeed, some semantics that render at least one premise untrue involve apparatus that can also be used to make trouble for these principles. For example, in fuzzy logic, the usual approach involves using the real interval $[0, 1]$ for degrees of truth from wholly false to wholly true, along with an account of '$\rightarrow$' on which the degree of truth of a conditional drops as the gap between the higher degree of truth of the antecedent and the lower degree of truth of the consequent widens, until we reach the limiting case when antecedent is 1 ($\top$) and consequent is 0 ($\bot$), which results in the lowest possible degree of truth, 0. The simplest clause with this effect is

[8] See [14, pp. 9–15], also in [5, pp. 257–263], for eloquent exposition of this theme. However, N. Smith [15, pp. 308–315] suggests that some precise points of transition can be determined *by vote*. In the present case, we simply survey language-users, and although they may disagree about what salaries make you well-paid$_C$, we can be sure that there is a *least* salary which they *unanimously* agree makes you well-paid$_C$. So we might say that only if you earn less than *that* salary does the claim that you are well-paid$_C$ fail to be unqualifiedly true. I have three doubts about this. First, while 100 % is a nice round number, 90 % is almost as nice: what makes it the case that unanimity is the correct requirement? Second, we need a criterion for excluding the judgements of certain voters, e.g., those who have had one drink too many [13, p. 331] and make judgements that diverge wildly from those of the vast majority (*that* can't be the criterion for Smith's purposes, because it is vague – 'wildly', 'vast'). Third, the envisaged polls are counterfactual: we are talking about what language-users *would* say if surveyed. But then the least salary which they would unanimously agree makes you well-paid$_C$ fluctuates from moment to moment (not just, as Smith allows, from today to tomorrow), since judgements about the cut-off may be sensitive to arbitrary factors, e.g., the current atmospheric pressure or wind strength. This makes it hard to see how the vote is reflecting some *fact* about how little you can be paid and still be well-paid$_C$.

(1) $v[p \to q] = 1 - (v[p] \dot{-} v[q])$.[9]

But (1) makes *modus ponens* and transitivity of '$\to$' *invalid* on an account of validity that generalizes classical validity in the way that (1) generalizes material implication (whose table is the special case when $[0,1]$ is replaced by $\{0,1\}$). Suppose we define $\models_g$ for finite premise sets by:

(2) $p_1, \ldots, p_n \models_g q$ iff $v[q] \geqslant \min\{v[p_1], \ldots, v[p_n]\}$.

In other words, a valid argument-form is one in which, for any interpretation v, the degree of truth of the conclusion on v is no lower than that of the least true premise on v. But then, if $v(\mathrm{A}) = .9$, $v(\mathrm{B}) = .8$, and $v(\mathrm{C}) = .7$, we have by (1) and (2) that $\mathrm{A}, \mathrm{A} \to \mathrm{B} \not\models_g \mathrm{B}$ and $\mathrm{A} \to \mathrm{B}, \mathrm{B} \to \mathrm{C} \not\models_g \mathrm{A} \to \mathrm{C}$. So the original Sorites schema is invalid, as is the equally paradoxical pure conditional variant.

However, classifying the Sorites scheme as invalid is not essential to the fuzzy logic solution of the paradox. *Modus ponens* and transitivity of '$\to$' are restored by a more orthodox account of validity, where we say $p_1, \ldots, p_n \models q$ iff q is wholly true on every interpretation on which $p_1, \ldots, p_n$ are all wholly true ($v[p_i] = 1$). But some of the apparently true conditional premises of a relevant instance of the scheme will still be slightly less than wholly true, according to this account. That they are slightly *less* than wholly true makes the argument unsound, and that they are *slightly* less than wholly true explains why we are inclined to take them to be true. So a complete solution of the paradox is available without challenging its logic. But, like other multivalued and supervaluationist solutions, this comes at the price of positing sharp boundaries, for example, the one marking the least salary $\$n$ such that if x earns $\$n$ it is wholly true that x is well-paid$_C$.[10]

A solution which *only* challenges the logic has the *prima facie* possibility of blocking the paradox without introducing sharp boundaries. But if we look to critiques of classical conditional logic which are not motivated by considerations about vagueness for the ingredients of an analogous critique of Sorites logic, we are liable to be disappointed. Suppose that Seb and Steve are two athletes who are about equally as good as each other at a certain event, say men's 1500 m track, and suppose also that they are far ahead of the rest of the competition. In this circumstance, the following conditional, concerning the Olympics in which they both compete at the peak of their powers, seems true:

(3) If Steve wins the gold, Seb will win the silver.

[9] '$\dot{-}$' is cut-off subtraction, $a \dot{-} b = a - b$ if $a \geqslant b$, and $a - b = 0$ if $a < b$.

[10] This price is also paid by the version of validity on which *modus ponens* and transitivity are invalid. And if we are generally explaining misjudgment (including misjudgments about validity) in terms of failure to notice very small differences, this version will seem unattractive on another ground, namely, that the conclusion of a *modus ponens* or transitivity inference can be much more false than its least true premise; e.g., two conditional premises for an application of transitivity that each have degree of truth .7 produce a conclusion with degree of truth .4. Surely we'd notice *that*? See further [19, p. 124].

Again, since they have no real rivals other than each other, it also seems true that

(4) If injury forces Seb to withdraw, Steve will win the gold.

But if we chain (4) and (3), the result is the surely false

(5) If injury forces Seb to withdraw, Seb will win the silver.

Of course, on the material reading, if (5) is false, i.e., if injury forces Seb to withdraw and he doesn't win the silver, then (3) and (4) are incompatible, and in view of (4), one might be tempted to revise the judgement that (3) is true. But if we make revisions for that sort of reason, the outcome will be that no conditional is true unless its antecedent *strictly* implies its consequent; for example, combine (4) with the evidently true 'if injury forces both Seb and Steve to withdraw, then injury forces Seb to withdraw'. So (4) is false too, and by the same token, most of the conditionals we ordinarily assert are false. But any philosophical analysis of some locution is highly suspect if it says that speakers who understand the locution and aim to speak the truth using it nevertheless typically produce falsehoods as a result of using it, no matter how expert they are about the subject-matter. It's much more likely that there is some error in the analysis.

A better account of what is going on in the above inference involves appeal to the *context in which* a conditional is evaluated. We may suppose that with each such context there is associated a set of *admissible* possible worlds, and the truth-condition for a conditional in a context T is that its consequent is true in every T-admissible world in which its antecedent is true. So we can say that in evaluating (3) as true we are in a context in whose admissible worlds both Seb and Steve run in the final (and where other background conditions are as close as possible to the actual world). But the antecedent of (4), for pragmatic reasons, puts us in a context with a wider class of admissible worlds, including some in which only Steve runs. Relative to this wider class, (3) is false, since if Steve gets the gold in a race without Seb, one of the less talented others will have picked up the silver. And in the narrower class of worlds, where both run, (5) has an impossible antecedent, which, at least arguably, suffices for its truth.

On this analysis, there is no real threat to transitivity in (3)–(5). For if the truthvalues of premises in an argument may vary with context, demonstrations of validity or invalidity require the context to be held fixed; while as we have just seen, we get (3) and (4) both true only when we let the context change from premise to premise. That the transitivity scheme turns out to be valid is an advantage for the hypothesis of context-dependency over other semantics which simply accept, in the light of cases like Seb-Steve, that it is invalid. For it is hard to see how it *could* be invalid, if a conditional asserts the sufficiency of the antecedent for the consequent.

How, exactly, are we to model the context-dependency we are attributing here? A very simple account takes the domain of worlds of the context as a domain for the interpretation of modal operators. The conditionals in (3)–(5)

are then analyzed as strict conditionals, that is, as the necessitations of material conditionals, formulae of the form $\ulcorner \Box_T(p \supset q) \urcorner$, in which $\Box_T$ expresses universal quantification over the domain of worlds of the context T. If context determines the relevant domain of discourse for $\Box_T$, the switch in moving from (3) to (4) is like the switch that occurs when two professors report on how the honors students did in their classes. If Professor X reports 'every honors students got an A' and Professor Y reports 'not every honors student got an A', there is no contradiction, since the domain of the restricted quantifier 'every honors student' changes from report to report. In the same way, the domain of $\Box_T$ changes from (3) to (4), and when (5) gets its intuitively correct evaluation as false, we are in a domain in which (3) is false.

4 Context and Sorites Conditionals

(3), (4) and (5) constitute a paradox only if we fail to notice switches in context. Granted that the example doesn't provide a reason to *reject* transitivity, might the grounds it provides for *rejecting the argument as an instance of transitivity* be generalizable to at least the pure conditional Sorites scheme, or both schemes? That is, can we make a similar analysis of puzzling instances of the schemes, to the effect that the premises are certainly all true, but each is true in a certain context, and the relevant context shifts at various points as we go through the premises? Approaches to the Sorites embodying this diagnosis are to be found in [8], [11], and, the treatment which will be our main focus here, [16].[11]

Soames's idea is that for a given vague predicate F, such as 'well-paid$_C$',[12] there is a *default* context T_0 (in [11] , the 'basic' or 'root' interpretation, in [8, pp. 253, 256] the 'minimal' context) in which the meaning of F provides F with a default extension, a default antiextension (Kamp uses 'positive extension' and 'negative extension'), and what Robertson [13, p. 332] calls an 'inextension'; in the case of 'well-paid$_C$' the extension in T_0 contains the professors in C which the

[11] We focus on Soames's approach because it is *prima facie* the most conservative of the three cited. Pinkal (who acknowledges Kamp [8]) is mainly concerned to develop a notion of 'practical consistency' which can be used to resist the Dummett-Wright charge of incoherence in language [4,20] and this turns out to involve non-transitive entailment [11, p. 338]. And Kamp's notion of (absolute, complete) truth in a context is such that it needn't be closed under *modus ponens* (see (a), (c) in [8, p. 260]). If Soames' version of contextualism were successful, it would show that Kamp's and Pinkal's approaches involve more complexity and revisionism than is needed to solve the problem.

[12] Soames' running example of a vague predicate is 'looks green', but it is clear that he intends his analysis to apply to 'bald', 'green', 'well-paid', and so on, not just subjectivized versions like 'looks green' and 'seems bald'. However, I am in agreement with Edgington [5, p. 309, n. 15] that the subjectivized predicates are a special case, for which an account in terms of context-dependency (perhaps in the style of Raffman [12]) may be appropriate in a way that it is not for the non-subjectivized predicates. Raffman's contextualism is rather different from the logical kinds under discussion here; see further [7, p. 424, n. 6].

rules of language combined with the empirical facts determine to be well-paid$_C$; the antiextension contains the professors in C which the rules of language plus the empirical facts determine to be non-well-paid$_C$; and the inextension contains the professors in C which the rules of language combined with the empirical facts are silent on – they are the elements of C for which 'well-paid$_C$' is undefined.

However, in the course of a conversation, it is permissible for the participants to extend the extension (or antiextension) of F in certain ways, for example by decreeing that such-and-such an object, hitherto in the inextension, is to be counted as F; for example, it might be stated that to be paid \$90,000 is to be well-paid$_C$. Barring objections, 'well-paid$_C$' will now have those members of C earning at least \$90,000 in its extension, but it will also have in its extension those members of C earning a sum less than \$90,000 but within the tolerance-range of 'well-paid$_C$'. Earlier, we suggested that \$100 is within this range (at the likely cost of running out of actual professors, \$10 or \$1 could be used instead). So, by stipulating that to be paid \$90,000 is to be well-paid$_C$, we have changed the extension of 'well-paid$_C$' to include those in C who earn at least \$89,900. Now, if tolerance for 'well-paid$_C$' implies that anyone earning \$100 less than someone who is well-paid$_C$ is also well-paid$_C$, then we could not stop at \$89,900: it would turn out, after sufficiently many steps, that someone who works *pro bono* is well-paid$_C$. But the contextualist conception is rather different: what makes a predicate F_C tolerant is that for any item $x \in C$ in the inextension of F_C, if x is *explicitly characterized* in the context as being F_C (Soames) or if its being F_C is *part of the background* of the context (Kamp), or if x is *focused under the aspect* of being F_C (Pinkal), then the extension of F_C expands to include everything that has at least the magnitude of F_C-making features that x has, and also those items which do not have that magnitude, but whose shortfall is within the tolerance range (*mutatis mutandis* for 'non-F_C' and the antiextension of F_C). However, until one of these falling-insignificantly-short items y is explicitly characterized as being F_C, or until the background is updated with the judgement that y is F_C, or until y is focused under the aspect of being F_C, there is no way of iterating the extension-expanding principle by applying it over again to y. So we do not, in our example, end up concluding that those who work for nothing are in the extension of 'well-paid$_C$'.

The change in extension of F consequent upon an accepted proposal that x be taken to be F (or an updating of the background with 'Fx', etc.) is a change in the standards for being F, or more generally, a change in context. Now suppose someone reasons through our running instance of the Sorites scheme in the following way (the *modus ponens walk-through*): Professor X_1 is well-paid$_C$; if Professor X_1 is well-paid$_C$ then Professor X_2 is well-paid$_C$; so, Professor X_2 is well-paid$_C$. But if Professor X_2 is well-paid$_C$ then Professor X_3 is well-paid$_C$, and Professor X_2 is well-paid$_C$; so, Professor X_3 is well-paid$_C$; so... so Professor X_n (who works *pro bono*) is well-paid$_C$. Each conditional $p \rightarrow q$ is true because, (i), according to the standards for being well-paid$_C$ *in force in the context* T *in which the conditional is asserted*, p is true, either by virtue of being default-true (true in T_0), or by virtue of an expansion of the extension of 'well-paid$_C$'

which occurred when at the previous step p was detached by *modus ponens* and asserted (the detachment and the assertion may be distinguished, if desired); and (ii) q is true, because each professor is paid at most \$100 (\$10, \$1) more than the next one, so the professor mentioned in q falls under 'well-paid$_C$', either by default, or as soon as standards are adjusted when the professor mentioned in p is asserted, at the conclusion of the previous step, to be well-paid$_C$ (see [11, p. 336] for a similar dymanic).

We therefore have a close parallel with the case of Seb and Steve: each conditional, in its own context, is true, but the argument, which seems to be a correct formal proof, is in fact fallacious because context sometimes changes from one step to another. In a genuinely correct formal proof, context is held fixed, or else the proof is carried out in a special formalism with a method of tracking change of context. In the present case, we *think* we have a genuine formal proof, because we fail to notice, or understand, how beyond a certain point each step in the walkthrough effects a context-change, a change in standards for the application of the relevant vague predicate.

There are, I think, some difficulties for this account. First, the diagnosis which Soames offers of the appearance of truth in each Sorites conditional is unpersuasive. According to the degree theorist, each conditional seems true either because it *is* true, or because it is so close to being true that it's entirely understandable that it is taken as true. So an understandable mistake is attributed. However, on Soames' view [16, p. 215], the mistake is not so understandable: for a vague empirical predicate F we are said to confuse a principle such as

(6) If x is F and y differs from x in respect of F-ness only to an empirically indiscriminable degree, then y is F

with a metalinguistic principle along the lines of

(7) Anyone who characterizes x as F is committed to a standard that counts y as being F too, when y differs from x in respect of F-ness only to an empirically indiscriminable degree.

(6) concerns when objects must agree on whether or not they are F, while (7) concerns the commitments of speakers consequent upon making certain judgements, so on the face of it they are rather dissimilar. But perhaps we wouldn't realize that it is (7) that is driving Sorites reasoning until it is pointed out to us. However, there is little reason to think that the judgements of sophisticated speakers about Sorites conditionals are really confused versions of (7), in which they assent to 'if x is F then y is F' when what they are really thinking is 'if I say x is F then I'm committed to a standard under which y is F', or that they fail to notice that their reason for thinking the former is only a reason for thinking the latter. Surely, if x is green and *there is no perceptible difference in color between* x *and* y, then y is green too; if $x \in C$ is well-paid$_C$, and though $y \in C$ is paid a little less *the difference wouldn't buy you anything at the local*

Five & Dime, then y is also well-paid$_C$. These are highly plausible claims in their own right, and produce hesitation only in those who foresee a Sorites paradox coming down the tracks at them.[13]

A second, more severe objection, from [13], is that Soames' theory suffers from a *spillover* problem. If Professor Y is in the inextension of 'well-paid$_C$', then it's a legitimate move to stipulate that Professor Y is well-paid$_C$. But it's conceivable that Professor Y is paid only a few dollars more than the highest-paid professor in the *default* antiextension of 'well-paid$_C$', a certain Professor Z. And when we stipulate that Professor Y is well-paid$_C$, we add to the extension of 'well-paid$_C$' not only those who earn exactly what Professor Y does, but also those who earn a few dollars less. So Professor Z gets added. But *no* member of the default antiextension of 'well-paid$_C$' can be added to the extension of 'well-paid$_C$' (the professors in the default antiextension are, if you like, the *definitely* not well-paid$_C$ ones). So Soames' apparatus generates a contradiction.[14]

Perhaps this problem arises because Soames takes there to be sharp boundaries that delineate three groups, default extension, default inextension, and default antiextension. But this doesn't seem to be the crux of the matter. For it is of no help to change three to five by adding default 'buffer' zones between extension and inextension and inextension and antiextension. Presumably buffer-zone professors can be stipulated to be well-paid$_C$ or not well-paid$_C$, and so can be absorbed into the extension or antiextension of 'well-paid$_C$'. Thus we quickly find ourselves back at the spillover point. It also does not help to blur the sharp boundaries that delineate default extension, inextension and antiextension, say by taking the rules of language that govern 'well-paid$_C$' themselves to involve vague terminology. We might agree that to be well-paid$_C$ it's enough to be paid an amount that is *close* to the salary of some person who is well-paid$_C$, or *about the same* as the salary of such a person. But so long as, for each salary, there are lesser salaries that are close to or about the same as the given one, and about the same amount less across cases, we can simply advance in the style of Robertson (*op. cit.*, pp. 332–333) to a professor, x, who is, in the current context, well-paid$_C$, such that there is a lesser-paid professor who is default not well-paid$_C$, but whose salary is close to or about the same as that of x.

The spillover problem therefore appears to be a serious one. Maybe it can be met by suitable principles, though it seems likely that the motivation for

[13] I would make a similar response to Pinkal's suggestion [11, p. 330] that observational indistinguishability of x and y only guarantees truth-functional equivalence of observational predications of x and y when one or other of x or y is focused under the aspect of the observational predicate in question. There is no reason to think that in our confusion, the qualification about focus is simply something we overlook the need for.

[14] Analogous problems appear to afflict the other contextualisms. For example, Kamp [8, p. 260] has a notion of *coherence* on which some contexts are coherent, some incoherent, and some neither. When we announce that Professor Y is well-paid, we are in an incoherent context, but it's unclear what's to stop us reasoning our way into it from coherent contexts by a *modus ponens* walkthrough.

such principles will *only* be that they avoid the problem.[15] However, I wish here to pursue another difficulty, which I think vitiates any kind of contextualist approach to the Sorites, that is, any approach that explains the force of a Sorites in terms of the truth of each conditional in its own context, with shifts in context rendering the arguments fallacious and the subtle nature of the shifts explaining away the impression that the premises are all true together. For if it is possible to *fix* the context in a way that retains the appearance that the premises are all true together, their appearing that way isn't explained away by shifts in context; and without shifts in context, the contextualist has no grounds to say that the arguments are fallacious.

[15] In his response to Robertson, Soames [17, pp. 443–444, n. 13] agrees that his principles will generate Robertson's contradiction from a case of objects y and z such that y is in the default inextension of 'looks green' and z is in the default antiextension, y and z are perceptually indistinguishable as regards color, and in the context c the speaker s asserts 'that looks green', demonstrating y. To repair the problem, Soames suggests that 'the two most promising alternatives' are as follows. (A1): s's assertion '*that* looks green' in c does not *semantically express* any proposition in c because 'looks green' does not semantically express any property in c, but s does succeed in *asserting* a proposition, namely, one which attributes to y a property we can call looking green*, which applies to anything perceptually indistinguishable from y as regards color. By this characterization, z looks green*. But that is unfortunate, since our intuition about the case is not merely that s seems to assert *something* in c, but that, being a normal speaker, what s asserts of y in c with 'looks green' is *not* true of z; for after all, z doesn't look green, it only looks green*. In other words, it's counterintuitive that one can use 'looks green' in c to correctly attribute a property possessed by something that doesn't look green by any acceptable standards. This makes (A1) rather unappealing. (A2): s's assertion '*that* looks green' semantically expresses a proposition in c, but the property expressed by 'looks green' in c isn't the result of adjustment consequent upon the assertion. Rather, it's the property P_x expressed by 'looks green' in '*that* looks green' in the context where 'that' denotes the object x immediately before y in the Sorites series (P_x has in its extension x and everything that looks no less green than x, including y). But this does not resolve the contradiction. If stipulating in c that x looks green is to put y but not z into the c-extension of 'looks green', x and z must be discriminable, while neither x and y nor y and z are. Therefore, there should be a perfectly coherent context c* in which x is stipulated to look green while z (which is default not-green-looking) is announced not to look green. By the stipulation, y looks green (it is indiscriminable from x), and by the announcement, y doesn't look green (it is indiscriminable from z), both in the same context. So if c* is not just to be ruled out of order by decree, it seems option A2 should be revised to say that 'looks green' in '*that* looks green', '*that*' denoting y, expresses P_w in c, where w is immediately before x in the Sorites series and is indistinguishable from x and distinguishable from y. So (i) the demonstrative utterance may well be untrue, even though, one wants to say, y really does look green to s. And (ii) there are no acceptable standards under which there is a property of looking green that y has. Consequently, if we say '*that* looks green' denoting x, we can't be semantically expressing P_x, but must instead be expressing P_w. It is not clear where this will end. (For discussion of what might be a related issue, see [16, pp. 222–223, n. 11].)

5 Conditionals and the Sufficiency Relation

For the purposes of this discussion, I focus on the pure conditional Sorites scheme. Each conditional in an instance of such a scheme is in fact derived from two other premises. For instance, the conditional 'if Professor X_1 is well-paid$_C$ then Professor X_2 is well-paid$_C$' in our running example is derived from two premises, (i) 'if Professor X_1 is paid at most \$100 more than Professor X_2, then if Professor X_1 is well-paid$_C$, Professor X_2 is well-paid$_C$', and (ii) 'Professor X_1 is paid at most \$100 more than Professor X_2'. In other words, the premises of the pure conditional scheme are derived by an inference of the form $p \to (q \to r), p \vdash q \to r$, where p is the *relational* premise that states that the next item differs only by such-and-such an amount, where the amount in question is within the tolerance range of the relevant predicate.

Orthodox accounts of the conditional are typically formulated in a metalanguage with material implication ($\supset$) and quantification over indices of some sort (e.g., the non-variably strict S5 conditional $p \to q$ is defined as 'for all w, p is true at $w \supset q$ is true at w'). These analyses are not indicative of realistic strategies for establishing conditionals. It is more realistic to suppose that when we consider the major premise of the argument *for* a specific Sorites conditional, for example,

(8) If Professor X is paid at most \$100 more than Professor Y, then if Professor X is well-paid$_C$ so is Professor Y

we *apprehend* a relation between antecedent ('Professor X is paid at most \$100 more than Professor Y') and (conditional) consequent that makes the main conditional in (8) true. The relation is that of *sufficiency*: the antecedent suffices for the consequent. The truth-condition of $p \to q$ is just that this relationship should hold between the propositions p and q, something which can be established by a derivation of q from p, but whose holding does not consist in such derivability.

There are competing accounts of sufficing, for instance, a material account, a strict account, and a relevant account. But there is a problem for contextualism independent of this choice. (8) and the other conditionals which, along with the relational premises, entail the premises of a Sorites, all seem equally good candidates for truth. In our current terms, this is to say that in each case it looks as if the relational premise is sufficient for the antecedent of the Sorites premise to suffice for its consequent: we perceive an *a posteriori* relation of sufficiency between propositions such as 'Professor X is paid at most \$100 more than Professor Y' and the proposition 'Professor Y is well-paid$_C$ *if* Professor X is well-paid$_C$'. So in the latter conditional, the antecedent also suffices for the consequent. What is important here is that apprehension of sufficiency of *its* antecedent ('Professor X is well-paid$_C$') for its consequent does not require any attitude of *endorsement* towards the antecedent. But this in turn means there is nothing in the apprehension of the truth of the premises (if they are all true) to trigger a change of context. So we have a *fixed* context in which all the premises of a Sorites seem equally true, because the relation of sufficiency

seems to hold between antecedent and consequent in each conditional (given the relational premise). Hence it is not the case that they only all seem true because in evaluating each we implicitly shift to a context in which the conditional in question *is* true. And insofar as we are apprehending a sufficiency relationship between antecedent and consequent of the actual premises, we are not stumbling into some conflation of these premises with metalinguistic principles. Once again, then, the only recourse the contextualist has to block a Sorites is to insist that some premises are untrue, because of the mysterious sharp divisions between extension, inextension and antiextension of the vague predicate in question.[16]

There are other accounts of conditionals on which evaluating a conditional *does* involve taking an endorsing attitude towards the antecedent, perhaps thereby changing context. One such account is the 'suppositional' one developed in [2]. On this account, 'if p then q' is said to be synonymous with 'supposing that p, then q', and for a conditional to be true is for its consequent to be true *under the supposition* of its antecedent; that is, a conditional is true iff, supposing its antecedent to be true, its consequent is true as well.[17] But even though this looks like it might lead to the conclusion that each premise is true in its own context, that is not so. For though supposing the antecedent changes the standards under which the consequent is evaluated, the supposition is *cancelled* when the truth-value of the consequent is transferred to the whole conditional; hence the resulting truth-value is the truth-value in the *default* context. This means the semantics of the conditionals *guarantees* that the context is held fixed. For example, if 'Professor X_{121} is well-paid$_C$' is default undefined, then

(9) If Professor X_{121} is well-paid$_C$ then Professor X_{122} is well-paid$_C$

is true in the default context, because when we evaluate 'Professor X_{122} is well-paid$_C$' under the supposition that Professor X_{121} is well-paid$_C$, we have changed the standards for being well-paid$_C$ to ones under which Professor X_{121} is in the extension of 'is well-paid$_C$'. This change in standards doesn't affect the relational premise, so Professor X_{122} is pied-piped into the extension of 'is well-paid$_C$' along with Professor X_{121}. Thus the consequent of (9) is true under the supposition of its antecedent, i.e., in the context created by supposing its antecedent. But this means that (9) *as a whole* is true in the default context, for its truth-condition

[16] Soames doesn't himself propose that context-shift explains why a standard Sorites (as opposed to a 'forced march') seems sound but is fallacious. He simply insists that one conditional must be untrue, and as described in §4, posits an error thesis according to which we deny this because we confuse 'if X is well-paid so is Y' with "if X is stipulated/assumed/agreed to be in the extension of 'well-paid' then Y is in it too". Pinkal's position [11, p. 338] is similar: we only advance towards the spillover point by mixing steps involving classical consequence with steps involving practical consequence, for if we restrict ourselves to classical consequence, the first conditional with a true antecedent and undefined consequent stops the reasoning.

[17] Barnett derives a non-classical logic for 'if. . . then' from this starting-point, but that doesn't appear to be intrinsic to the basic approach. He employs a very substantive notion of supposition, on which no moves can be made if a contradiction is supposed. On a more minimal notion of supposition, *ex falso quodlibet* could still be justified.

in the default context is just that if we hypothesize a context that verifies the antecedent, the consequent also holds in that context. The same is true for all the other premises, so they are all true in the original context (the intuitively correct result). Hence, assuming transitivity, the absurd conclusion that if Professor X_1 is well-paid$_C$ then so is Professor $X_{pro\ bono}$, is also true in the default context.

On Soames' view, we should not accept that every premise is true in the original context. This is not because two are untrue, one marking the crossing of the lines between default extension and default inextension, and the other the line between default inextension and default antiextension. For in each case, the effect of supposing the antecedent is to move the extension / inextension boundary beyond the point where it could cause trouble, if it is not already beyond that point. The spoiler is again the spillover case: if Professor X_m is the least well-paid member of the default inextension, then the premise 'if Professor X_m is well-paid$_C$ then Professor X_{m+1} is well-paid$_C$' won't be true even on the suppositional account, provided we have a principled reason why the pied-piping effect should fail for Professor X_{m+1}.

But this is not a satisfactory way of responding to the problem. For it relies on there *being* a principled reason why the pied-piping effect should fail for Professor X_{m+1} (see note 15 for my scepticism that such a reason exists). Secondly, even if such a reason were forthcoming, we would still have a valid argument whose premises are true in the default context but whose conclusion, by Soames' lights, is untrue. For although we won't be able to conclude from the full instance of the pure conditional scheme that if Professor X_1 is well-paid$_C$ then so is Professor $X_{pro\ bono}$, we *will* be able to conclude from a truncated instance that if Professor X_1 is well-paid$_C$ then so is Professor X_m (the least well-paid member of the default inextension).

The remaining move to make contextualism compatible with the suppositional account of conditionals is to formulate a semantics for conditionals on which transitivity fails.[18] But the following reasoning suggests that transitivity should hold, at least in a range of cases our current ones fall into. For suppose that q is the case on the supposition that p, and that r is the case on the supposition that q. Then we may suppose that p, allowing us to conclude that q, and next, take whatever reasoning showed that r is the case on the *supposition* that q, and use that reasoning to show that r is the case on the supposition that p, applying it to the q we *inferred* from the supposition that p.

The general form of argument here is unreliable, for it may be that while r can be concluded when q is *supposed*, it cannot be concluded when q is *inferred*. A familiar example of this phenomenon is the sequent $A \vdash_{S5} \Box\Diamond A$, where it seems that we should be able to assume A, infer $\Diamond A$ by $\Diamond I$, then $\Box\Diamond A$ by $\Box I$. But $\Box I$ requires that the formula it is applied to depends only on assumptions that are *fully modalized* (for a sentential language, p is fully modalized iff every sentence

[18] Barnett [2] endorses a probabilistic semantics on which transitivity does fail (pp. 549–559), but I don't think this is well-motivated: the 'counterexamples' to transitivity are like the Seb-Steve case, for which there is an independently plausible diagnosis that preserves transitivity.

letter in p is within the scope of a $\Box$ or $\Diamond$). Since A is not fully modalized, and $\Diamond$A depends on A, $\Box$I can't be used on $\Diamond$A. One solution is to combine two separate lines of reasoning: in the first we show that A $\vdash \Diamond$A by a use of $\Diamond$I, and in the second we show $\vdash \Diamond$A $\rightarrow \Box \Diamond$ A by $\Box$I and $\rightarrow$ I (since $\Diamond$A depends on itself and is fully modalized, $\Box$I can be applied to it). We get A $\vdash \Box \Diamond$ A by a final use of *modus ponens*.

Perhaps there is a comparable difficulty for our justification of transitivity:

(10) If Professor X is well-paid$_C$ then Professor Y is well-paid$_C$

says that Professor Y is well-paid$_C$ supposing a context in which Professor X is well-paid$_C$, while

(11) If Professor Y is well-paid$_C$ then Professor Z is well-paid$_C$

says that Professor Z is well-paid$_C$ supposing a context in which Professor Y is well-paid$_C$. If we try to apply the reasoning that establishes (11) within the scope of a supposition of the antecedent of (10), we have moved the reasoning from one context into another, and this may be thought to be dubious. But as with the modal case, there is a way round the problem. We have reasoning which establishes (10) and reasoning which establishes (11), so we may make the supposition that X is well-paid$_C$ and *within its scope* suppose that Y is well-paid$_C$. The reasoning that establishes (11) can now be applied, since the supposition-created context we are in is no different from the one created by the antecedent of (11). And once it's been established that Z is well-paid$_C$ we can use $\rightarrow$ I to get (11) *still within the scope of the supposition of the antecedent of* (10). We may then apply *modus ponens* (we already have that Y is well-paid$_C$) to conclude that Z is well-paid$_C$, and a final $\rightarrow$ I gets us

(12) If Professor X is well-paid$_C$ then Professor Z is well-paid$_C$

in the default context. So it looks as if transitivity is correct for this account of conditionals, meaning that if all the premises of a pure conditional Sorites scheme are true in the default context, so is the absurd conclusion. And the premises all seem equally true by the contextualist's own lights, given the suppositional analysis of what it takes to make them true. So it turns out that not even the suppositional account of conditionals eliminates the possibility of a single context in which all Sorites conditionals seem true. All the contextualist can say about this case is that the proofs are unsound, because there are untrue conditionals among their premises. The sharp boundary we hoped to avoid still confronts us.

6 Conclusion

So where does this leave us? The contextualist account is a version of the third option we distinguished at the end of §1, according to which Sorites reasoning involves a fallacy of equivocation. For if it were true that there is no single context in which all the premises hold, we could think of the supposed changes

in context (in standards) needed to evaluate all the premises as true, as changes in meaning. But if this third option leads to a dead end, we are thrown back on one of the other two.

I suggest that we ought to reconsider our argument in §2 that any approach which classifies some premises as untrue is in the same boat as epistemicism, positing a sharp division whose existence is not intelligible. David Kaplan has drawn attention to a distinction between 'those features of a model which represent features of that which we model' and 'those features which are intrinsic to the model and play no representational role... *artifacts* of the model' (Kaplan 1975:722).[19] For example, a scale model of HMS *Victory*, like Nelson's actual flagship, must have a physical length, but its value is an artifact of the model. It must also have a ratio of hull length to mainmast height, and this feature is representational, for the further it diverges from the ratio of the actual flagship's length to its mainmast height (roughly 7:5) the less accurate the model.[20]

In the same way, in a type-(i) formal model of a Sorites paradox (one that classifies it as valid but unsound) some specific premise must be the first untrue one. The question at issue is whether *that* premise's being the first untrue one is an artifact of the model, or rather, one of its representational features. For an epistemicist, its being the first untrue premise is a representational feature (identifying the wrong premise is like the model-builder who gets hull length to mainmast height wrong). The challenge for others, in particular, the degree theorist, is to make a case that, for each model and for whichever premise is the first untrue one on that model, it is only an artifact of the model that the premise in question is the first untrue one. Simultaneously, however, it has to be a *representational* feature of the model that *some* premise is the first untrue one. For on non-classical analyses of Sorites reasoning, though there is no fact of the matter which premise is the first untrue one, it is a fact that *some* premise is, since the conclusion is definitely false. Further investigation of the artifact / representation distinction may allow us to identify different types of sharp boundary, with epistemicism in sole possession of the least desirable.

References

1. Beall, J.C. (ed.): Liars and Heaps. Oxford University Press, Oxford (2003)
2. Barnett, D.: Zif is If. Mind 115, 519–565 (2006)
3. Cook, R.: Vagueness and Mathematical Precision. Mind 111, 225–247 (2002)
4. Dummett, M.: Wang's Paradox. Synthese 30, 301–324 (1975)
5. Edgington, D.: Vagueness by Degrees. In: Keefe, R., Smith, P. (eds.) Vagueness: A Reader, pp. 294–316. The MIT Press, Cambridge (1996)
6. Forbes, G.: Two solutions to Chisholm's paradox. Philosophical Studies 46, 171–187 (1984)

[19] In [6] I appealed to Kaplan's distinction to argue for the superiority of counterpart-theoretic over relativized possibility solutions to certain modal versions of Sorites paradoxes. I now think that this misapplies the distinction.

[20] A model ship illustration is suggested in [3, p. 235]. Cook's paper uses Kaplan's distinction to respond to some of the complaints about degree theory voiced in [14].

7. Forbes, G.: Identity and the Facts of the Matter. In: Dietz, R., Moruzzi, S. (eds.) Cuts and Clouds: Vaguenesss, Its Nature and Its Logic. Oxford University Press, Oxford (2010)
8. Kamp, H.: The Paradox of the Heap. In: Mönnich, U. (ed.) Aspects of Philosophical Logic, pp. 225–277. Reidel, Dordrecht (1981)
9. Kaplan, D.: How to Russell a Frege-Church. Journal of Philosophy 72, 716–729 (1975)
10. Keefe, R., Smith, P.: Vagueness: A Reader, pp. 133–136. The MIT Press, Cambridge (1999)
11. Pinkal, M.: Consistency and Context Change: The Sorites Paradox. In: Landman, F., Veltman, F. (eds.) Varieties of Formal Semantics. Foris Publications (1984)
12. Raffman, D.: Vagueness without Paradox. The Philosophical Review 103, 41–74 (1994)
13. Robertson, T.: On Soames's Solution to the Sorites Paradox. Analysis 60, 328–334 (2000)
14. Sainsbury, R.M.: Concepts without Boundaries, Stebbing Chair of Philosophy, King's College, London, pp. 251–264. Inaugural Lecture (1991) (reprinted in Keefe and Smith 1996)
15. Smith, N.J.J.: Vagueness and Degrees of Truth. Oxford University Press, Oxford (2008)
16. Soames, S.: Understanding Truth. Oxford University Press, Oxford (1999)
17. Soames, S.: Replies. Philosophy and Phenomenological Research 65, 429–452 (2002)
18. Sorensen, R.: Blindspots. Oxford University Press, Oxford (1988)
19. Williamson, T.: Vagueness. Routledge, New York (1994)
20. Wright, C.: Language-Mastery and the Sorites Paradox. In: Evans, G., McDowell, J. (eds.) Truth and Meaning, pp. 151–173. Oxford University Press, Oxford (1976) (reprinted in Keefe and Smith 1996) (page references to the 1996 reprinting)
21. Wright, C.: Further reflections on the Sorites Paradox. Philosophical Topics 15, 227–290 (1987)
22. Wright, C.: Vagueness: A Fifth Column Approach. In: Beall, J.C. (ed.) Liars and Heaps, pp. 84–105. Oxford University Press, Oxford (2003)

Temporal Vagueness, Coordination and Communication

Ewan Klein and Michael Rovatsos

School of Informatics, University of Edinburgh

1 Introduction

How is it that people manage to communicate even when they implicitly differ on the meaning of the terms they use? Take an innocent-sounding expression such as *tomorrow morning*. What counts as *morning*? There is a surprising amount of variation across different people.[1]

For Anna, morning starts 'when she gets up', and finishes 'when she has lunch'. For Bart (who verges on the pedantic), morning officially starts at 12:00 am and ends at 11:59 am. Yet another view, held by Cecile, is that morning starts sometime between 6:00 and 7:00 am, and ends sometime between 12:30 pm and 1:30 pm. Finally, Devendra (who regularly works into the small hours) believes that morning has barely started at 10:30 am and finishes around 3:30 pm. Nevertheless, if Anna says to Bart: *drop by my office tomorrow morning and we'll have a look at your proposal*, the chances are high that Anna and Bart will manage to meet (as long as they have no conflicting engagements).

In the kind of linguistic contexts we are concerned with in this paper, it seems plausible to treat *morning* as a grouping of time units at some level of granularity (e.g., seconds, minutes, quarter-hours), ordered in the usual way. According to this view, a sentence like *Let's meet tomorrow morning* is equivalent to *Let's meet at some point in tomorrow morning*. This allows us to claim, for example, that the moment 9:15 am belongs to the extension of *morning*, while 9:15 pm does not. It follows that *morning* is open to the Sorites paradox: if 9:15:00 am counts as morning, then so does 9:15:01 am (i.e., the moment that is one second later than 9:15:00 am). By tediously iterating through the process of adding one second at a time (or one millisecond, if preferred), we will ineluctably reach the unwanted conclusion that 9:15:00 pm counts as morning. If we take the Sorites paradox as criterial for vagueness, we can conclude that *morning* and its companion expressions, *afternoon*, *evening*, *day* and *night* are all vague. But what does this mean? On the face of it, some speakers (like Bart) assign crisp boundaries to the time unit *morning*, while others (like Cecile) assign indeterminate boundaries. We will return to this issue in Section 3.3, but for the time being, let us just assume that the concepts corresponding to familiar time units possess crisp boundaries. Instead, we want to explore how terms like *morning* might be used in communities of speakers.

[1] The variability in usage and interpretation of terms like *morning* and *evening* has been explored by Reiter [18] in the context of weather forecasts.

R. Nouwen et al. (Eds.): ViC 2009, LNAI 6517, pp. 108–126, 2011.

Vagueness and Utility

The approach we have adopted is inspired in large part by Parikh's [16] observation that even though two speakers differ in the way they interpret a vague term like *blue*, if there is sufficient overlap in their interpretations, there will be positive utility in using the vague term. In Parikh's example, Ann requests Bob to fetch "a blue book on topology" from the book shelves in her study. The descriptive term contains enough information that even though they disagree on what counts as blue, the set of 'blue-for-Bob' books reduces Bob's search space far enough to significantly increase his chances of finding the correct book relatively fast.

What we want to adopt from Parikh's scenario is the idea that the success of communication involving a vague term can be measured in terms of completing a task. In Parikh's case, the task is to identify a book; in our case, the task is for two agents to meet one another. Just as the term *blue* functions in Parikh's scenario to reduce the search space within which the required book is located, we will assume that a term like *morning* reduces the temporal period within which the meeting will take place. More specifically, we assume there are two agents, say A_1 and A_2, who wish to meet up. Suppose A_1 says to A_2: *Let's meet up tomorrow morning. Drop by my office.* A_2 accepts the proposal. Both A_1 and A_2 have their own interpretation of what is meant by the phrase *morning*. For each of them, the interpretation is modelled as an interval, but these intervals do not need to coincide. Not surprisingly, we can observe that if the intervals overlap sufficiently, then the two agents will tend to be successful in meeting.

Although we focus in this paper on temporal intervals, in principle we could generalize our approach to any linguistic term whose semantic extension is a set. We define **overlap** between sets as follows:

Definition 1. *The* ***degree of overlap between sets*** X *and* Y, $\circ(X, Y)$, *is the quotient*

$$\frac{|X \cap Y|}{|X \cup Y|}$$

i.e., the cardinality of elements in the intersection of X *and* Y *divided by the cardinality of elements in the union of* X *and* Y.

If $\circ(X, Y) = 1.0$ then we say that X and Y *completely overlap*. Given some error margin ϵ, we will say that X and Y *approximately overlap* iff $1 - \circ(X, Y) < \epsilon$. In this case, we can say that there is an indifference relation between X and Y: the difference between them is either indiscernable or has no practical impact for the agents.

Let us write $V_i(e)$ for the interpretation that agent A_i assigns to expression e; we will restrict our attention to cases where $V_i(e)$ is a set $X \subseteq \mathcal{D}$ for some domain $\mathcal{D}$. Two interpretations V, V' are *completely (resp. approximately) aligned on* e iff $V(e)$ and $V'(e)$ completely (resp. approximately) overlap.

We assume that holding a meeting always has higher utility than failing to meet, that is, $U(meet) > U(\overline{meet})$. $P(meet \mid V_i(e))$ is the probability that a

meeting will take place, given the interpretation that A_i assigns to expression e. More generally, let's assume that S is an event whose occurrence is conditioned by linguistic meanings. Then the *expected utility* of an interpretation V_i, relative to outcome S, is given as:

$$EU(V_i) = P(S \mid V_i(e))U(S) + P(\overline{S} \mid V_i(e))U(\overline{S})$$

If $EU(V_i') > EU(V_i)$, then a rational agent A_i should adopt V_i' in place of V_i in order to maximize her expected utility. Using these notions, we go beyond Parikh's scenario, and make the following claim:

> *If individual agents in a given community maximize the expected utility of their interpretations, then over the course of successive interactions these interpretations will become approximately aligned.*

Much of the remainder of this paper will attempt to flesh out and substantiate this claim. However, we should emphasize that the model that we develop does not attempt to directly compare the utility of all possible interpretations at a given point in the interaction. Rather, the principle of utility maximization is comparable to an abstract specification which admits various computational implementations.

Conceptual Structures

The standard assumption in formal semantics (and indeed in much computational semantics) holds that linguistic meaning is a mapping from language to the world (or a model): meanings have an objective existence independently of speakers. By making the interpretation function V relative to agents, we are implicitly subscribing to a cognitive view, where meanings are psychological entities in the heads of agents. From the perspective of building some kind of computational system of interacting agents (such as mobile robots), the cognitive approach has obvious attractions. Each agent has only partial knowledge of the world in which it finds itself, including both the physical environment and its fellow agents. It does not have direct access to 'external reality', but has to build representations of the world on the basis of input from its sensors (which may well be noisy). It could be argued that it is enough to equip the agents with a mental language, such as some flavour of first order logic, in order to reason and communicate. But this begs the question of how the agents can be sure that they are using the non-logical terms of the language in the same way as their dialogue partners.

Within Gärdenfors' framework of conceptual spaces [8,26], concepts (and hence linguistic meanings) are internal mental representations. However, the requirement of 'shareability' [7] places constraints on how far the concepts of one agent can diverge from those of the other agents it interacts with. Shared meanings of expressions develop during language games — communicative interaction between language users — and involve mappings between conceptual representations that are influenced by the need to act effectively in the world.

A so-called 'meeting of minds' occurs when the representations in the minds of the dialogue partners become sufficiently compatible. This is essentially the same as our notion of approximate alignment.[2]

Our approach to meaning is also influenced by work on ontology alignment [5] in the context of multi-agent systems [1]. Agents collaborating in a shared environment need to share an ontology (i.e., the conceptualization of a domain) in order to communicate with each other, but in an open system, different agents can in principle use quite heterogeneous ontologies. Wang and Gasser [24] present a model that, like ours, explicitly considers which instances fall within the extension of a concept, but do not provide a utility-based method for determining successful alignment. Somewhat closer to our approach in this respect is the work of McNeill et al. [14], where agents are involved in jointly planning a task; plan failure triggers an attempt to diagnose mismatches in ontology; the agents use heuristics to repair their ontologies (in the sense of modifying the ontology signature), and then re-engage in the planning task. This cycle — communicate / diagnose failure / repair the ontology — is similar to the kind of model that we are proposing. However, the type of mismatches considered by [14], and the mechanisms used to effect the repair, are very different.

One question which arises is whether it is plausible that agents are prepared to modify their interpretations in the way we have suggested. Although this point deserves closer consideration, it does seem to be a characteristic of vague terms (both adjectives and nouns) that their boundaries are somewhat flexible. Thus, we seem to be more willing to shift the boundaries of what counts as morning than, say, what counts as a dog (or other natural kind).[3] On the other hand, even if agents are prepared to 'negotiate' meaning, there are no doubt some aspects which are non-negotiable — Bob may be prepared to shift his interpretation of *blue* so that it encompasses a shade of violet, but will balk at shifting it to cover bright orange. This is an important constraint, but we will defer the topic to future work.[4]

Overview of Paper

We use a simple multi-agent simulation in order to provide an explicit model of task based communication. In general, we believe this has a number of attractive aspects. For one thing, the simulation allows us to explore the consequences of setting various parameters in different ways, and to consider the interaction of these parameters in a manner that would be hard to achieve using a pencil-and-paper analysis. The approach can be viewed as implementing a

[2] An elegant computational implementation of alignment of colour terms in Gärdenfors' framework is presented by Jäger and van Rooij [10].

[3] This is similar to Williamson's [25] proposal that the meanings of vague terms are *unstable*, in the sense that minor differences in use give rise to minor differences in the extension of the term.

[4] For more discussion of constraints on shifting meanings in a computational framework, see [13,4].

language game in the sense of Gärdenfors [8], where the representations of individual agents affect communication about shared activities and are modified as a result.

Section 2 describes the framework of the simulation in more detail. Section 3 and Section 4 present the two sets of experiments that we ran, while Section 5 gives some conclusions and suggestions for future work.

2 Approach

As we have already indicated, our treatment of temporal expressions is highly simplified. Most notably, we ignore the element of context dependence in the application of temporal terms. For example, people who work together in an office will probably adopt a different view of what counts as morning than people who are up before dawn to milk the cows. Another contextual factor is the day of the week: for most Westerners, the temporal location of *morning* during the weekend diverges considerably from its location during the working week. We will abstract away from these factors, and only consider the case where the population of speakers adopts a shared context of use.

A second simplification is in our treatment of the expression *morning*. Given a specific day (say Monday 9th November 2009) and a specific speaker, say Anna, *morning* will denote a closed interval of time units.[5] For our purposes, it does not matter too much what level of granularity is chosen, but we will think of the intervals used by our agents as containing quarter-hour units; in other words, an interval with 12 elements would correspond to a period whose duration is three hours.

We will describe two families of experiments (referred to as Experiment 1 and Experiment 2 respectively), using a multi-agent simulator that was implemented in the Python programming language.[6] The agents are modelled as processes in the SimPy Discrete Event Simulator.[7] Before discussing the specifics of the experiments, we will give more details of the agent coordination task.

Let $\mathcal{T}$ be a finite set of integers representing **time units**, and let $\mathcal{I}$ be a set of closed intervals over $\mathcal{T}$. Given a set Ag of agents, each $A_i \in Ag$ is associated with a **preferred interval** $\iota_i \in \mathcal{I}$. We will assume that $\iota_i = V_i(morning)$, i.e., A_i's interpretation of the temporal expression *morning*. $V_i(morning)$ is private in the sense that for any $j \neq i$, A_j has no direct access to V_i.

Note that although $V_i(morning)$ is unique for each agent A_i, the inverse need not hold — that is, we let the cardinality of $\mathcal{I}$ be less than that of Ag. In

[5] This approach is intended to be compatible with that proposed by Ohlbach [15], who points out that a temporal expression such as *February* can be used to refer to a particular February; or to denote the set of all Februaries in the history of mankind; or, more generally, to refer to a function which given some year y returns the particular February of y.

[6] `http://www.python.org/`

[7] `http://simpy.sourceforge.net/`

our simulations, $\mathcal{I}$ is fixed as the set of intervals $\{[1, 10], [6, 15], [11, 20]\}$.[8] It is assumed in our model that the agents share the common time frame given by $\mathcal{T}$. For example, we might think of the three intervals in $\mathcal{I}$ as corresponding roughly to the time periods 7:00–9:30 am, 8:15–10:30 am and 9:30–12.00 am, respectively, where the time units 7:00 am, 7:15 am, ... have the same interpretation for all agents in Ag.[9]

On each run of a simulation, two agents A_i and A_j are selected at random. One of the agents is assigned the role of **proposer**, while the other takes on the role of **responder**; we'll refer to these as P and R respectively. P takes the lead in sending a "let's meet in the morning" message to R and chooses an arrival time arr_{P} from its period ι_{P}, while R chooses an arrival time arr_{R} from ι_{R}. One important feature of the model (which could however be relaxed) is that agents tend to pick an arrival time that falls somewhere in the middle of their preferred interval. This seems plausible when the proposed meeting time is some kind of approximation or vague interval. This feature is implemented by selecting A_i's arrival time (coerced to an integer) at random from a Gaussian distribution whose mean is the midpoint of ι_i, with standard deviation 1. In Experiment 1, the departure time of an agent A_i, dep_i, is simply set to the endpoint of ι_i. (We will later discuss a modification of this scheme used in Experiment 2.) P and R are judged to **meet** if $[arr_{\mathsf{P}}, dep_{\mathsf{P}}] \cap [arr_{\mathsf{R}}, dep_{\mathsf{R}}] \neq \emptyset$. We assume that on each run, P knows the arrival and departure time of R, even if they fail to meet.

It may be helpful to enumerate the four cases which determine whether or not a meeting occurs. (Although we mention a 'waiting cost' here, this feature does not come into play until Experiment 2.)

1. R arrives and departs before arr_{P}; the meeting fails with no waiting cost for P.
2. R has already arrived but not yet departed when P arrives; the meeting is accomplished with no waiting cost for P.
3. R arrives after arr_{P} but before dep_{P}; the meeting is accomplished with a waiting cost for P.
4. R arrives after dep_{P}; the meeting fails with a waiting cost for P.

These four options are shown graphically in Figure 1.

As mentioned before, each agent is assigned a preferred interval, which is intended to be a cognitive representation of a vague temporal expression. Since there is only one such expression in use in the community, we do not need to explicitly label it. The preferred interval, therefore, is a key aspect of each agent's mental state. Agents have no access to the mental states of others, and only observe their behaviour in arriving and departing at particular times. Each agent keeps a record of the other's arrival behaviour. More precisely, each agent

[8] These integer bounds are chosen for simplicity of implementation, but it would be conceptually straightforward to replace them with time points in *hh:mm* notation, or indeed to use seconds in Unix time (`http://unixtime.info/`).

[9] See [19] for discussion of the cultural and cognitive construction of time based time interval systems.

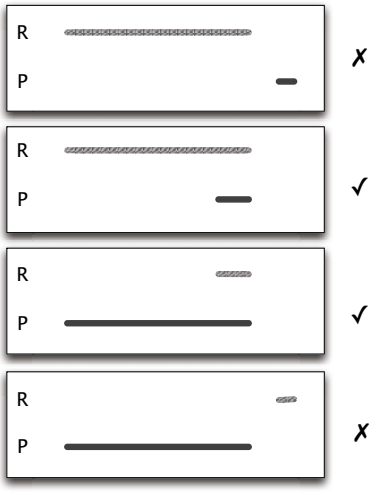


Fig. 1. Meeting Outcomes

A_i maintains a list $L(A_j)$ of observed arrival times for each other agent A_j, and the list is updated on any run in which A_i plays the P role and A_j plays the R role. Given $L(A_j)$, A_i can estimate the mean arrival time of A_j up to the current run in the simulation. We refer to this estimated mean as $\mu(\iota_j)$.

In order to provide a more concrete impression of the way the simulation works, in Figure 2 we have included a small extract from one simulation log file.

3 Experiment 1

3.1 Alignment

In the first set of experiments, we allow the proposer to update its preferred interval in the light of its experience so far. After each encounter, P attempts to **align** with R. It does so by adjusting ι_{P} so that the midpoint of ι_{P} approaches $\mu(\iota_{\mathsf{R}})$; that is, if t is the new target midpoint and *len* returns the length of an interval, then the adjusted interval is simply $[t - len(\iota_i)/2, t + len(\iota_i)/2]$. Let us refer to the midpoint of interval ι_{P} as $md(\iota_{\mathsf{P}})$ and let ι'_{P} be the new interval of P after alignment has taken place. Then we try to meet the following constraint after each run:

$$|\mu(\iota_{\mathsf{R}}) - md(\iota_{\mathsf{P}})'| < |\mu(\iota_{\mathsf{R}}) - md(\iota_{\mathsf{P}})| \tag{1}$$

In Experiment 1, we implemented the following update rule, where $\lambda \in [0, 1]$ is a scaling factor that we call the **learning rate**:

$$md(\iota_{\mathsf{P}})' = md(\iota_{\mathsf{P}}) + \lambda(\mu(\iota_{\mathsf{R}}) - md(\iota_{\mathsf{P}})) \tag{2}$$

```
activating agent-2 at 26
agent-2 proposed the following period:
[4, 5, 6, 7, 8, 9, 10, 11, 12, 13]
Proposer: agent-2, Responder: agent-4
Failed to meet!
agent-2 present: [11, 12, 13]
agent-4 present: [2, 3, 4, 5, 6, 7, 8]
agent-2 waited 2 mins
agent-2's cost: 0, net reward: 0
agent-2's cumulative reward over 8 proposals: -4
successes to date: 1.000
proposals to date: 8.000
success ratio: 0.125
reward ratio: -0.250
```

Fig. 2. Extract of a Simulation Log

3.2 Results

In analysing the results of Experiment 1, we focus on two dimensions for measuring the outcome: **interval overlap** and **proposal success ratio**. For convenience, we repeat a slightly modified version of Definition 1:[10]

Definition 2. *The **overlap between intervals** ι_P, ι_R is the quotient*

$$\frac{|\iota_P \cap \iota_R|}{|\iota_P \cup \iota_R|}$$

Definition 3. *The **success ratio** for an agent is the quotient*

$$\frac{\#\ \textit{of successful meetings}}{\#\ \textit{of proposals}}$$

In Fig. 3 we plot the average degree of interval overlap for a population of five agents over 250 runs.[11] We illustrate four cases, one where there is no learning, and three where the learning factor λ is set at increasingly high values. Fig. 3(b) shows that even a rather small value for λ is significantly better than no learning, and that the overlap between intervals ends up oscillating between 0.8 and 1.0. Fig. 3(c) shows a situation where complete alignment is achieved. By contrast, the setting of $\lambda = 0.5$ produces an oscillation similar to case Fig. 3(b), with the main difference being that this 'dynamic stability' is achieved more rapidly.

In Fig. 4, the outcomes for each agent are plotted separately, using the same four values for λ as in Fig. 3. In Fig. 4(c), it is striking that `agent-1` has much lower success than the other agents. This is due to the starting conditions in this particular run, where four of the agents started off with closely overlapping intervals and only `agent-1` happened to diverge sharply from this shared interval.

[10] This is known as the Jaccard index of similarity. We have also experimented with a related measure, Dice's coefficient, which yields comparable results.

[11] One average, each agent engages in 250/5 meeting proposals.

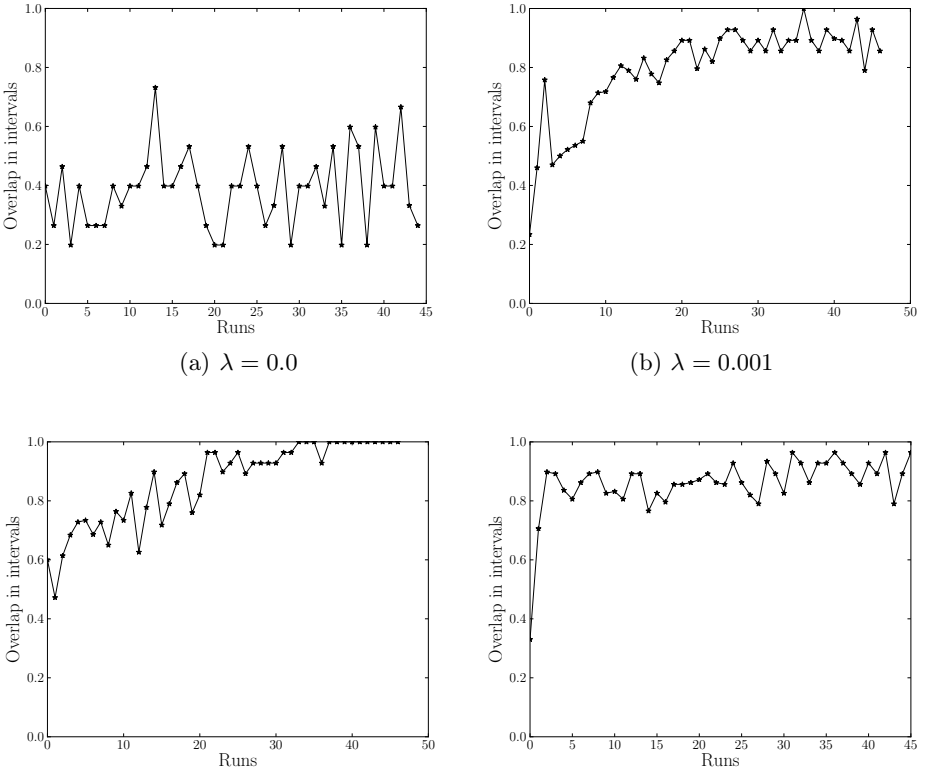
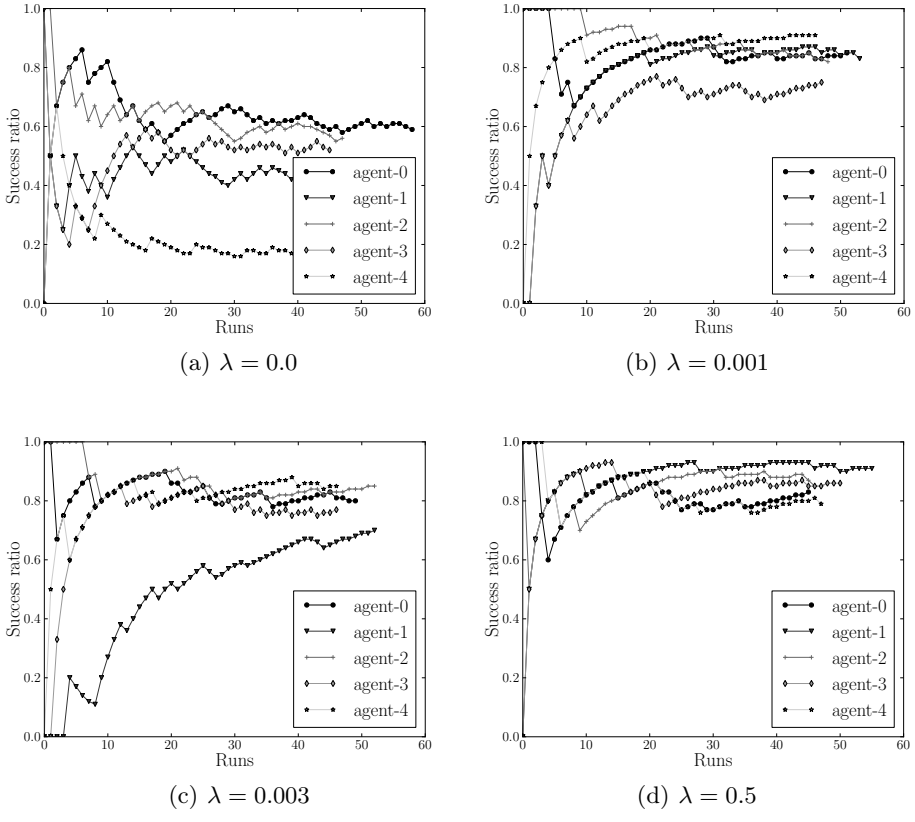


Fig. 3. Average Overlap in Preferred Intervals

Table 1 illustrates the intervals that are associated with each agent at the end of one complete simulation, after alignment has taken place. It can be observed that some of the intervals are left-shifted beyond the earliest point in $\mathcal{I}$. We will return to this issue later.

Table 1. Aligned Intervals after 350 runs, $\lambda = 0.5$

agent-0:	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
agent-1:	[-1, 0, 1, 2, 3, 4, 5, 6, 7, 8]
agent-2:	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]
agent-3:	[-1, 0, 1, 2, 3, 4, 5, 6, 7, 8]
agent-4:	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]



(a) $\lambda = 0.0$ (b) $\lambda = 0.001$

(c) $\lambda = 0.003$ (d) $\lambda = 0.5$

Fig. 4. Ratio of Successful Meetings to All Proposals

3.3 Discussion

As shown in Figure 4(a), when the discrepancy in preferred intervals is allowed to persist throughout the simulation, success in meeting tends to diminish over successive runs for all the agents. By contrast, Figure 4(b) shows gradual improvement to a mean success rate of around 0.8 when learning takes place. In addition, a positive value for λ enables the agent population to reach a relatively stable alignment of intervals. Despite this, complete alignment is not typically reached.

Fig. 5 gives an alternative visualization of how the preferred intervals of the five agents become increasingly aligned during the course of successive interactions (under the condition where $\lambda = 0.5$).[12] The greyscale intensity corresponds to the number of agents who share a given time period on a given run: the darker the

[12] In order to make this figure more legible, we have truncated the results to only show the first 100 runs.

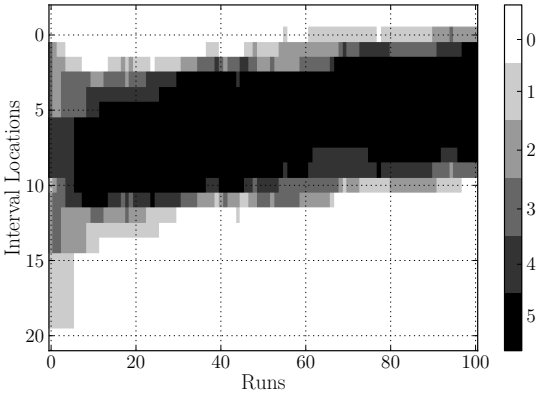


Fig. 5. Distribution of preferred intervals across agents over time

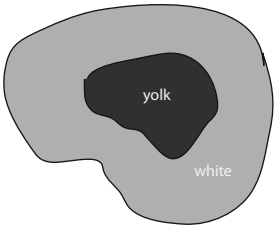


Fig. 6. Egg-Yolk Model of Vague Regions

shade, the greater the degree of overlap across the pool of agents. For example, it can be seen that only one agent has an interval which includes $[14, 19]$ during the first 5 runs, whereas four agents share the subinterval $[6, 9]$ One way of interpreting this result is to say that the chosen temporal unit is still vague, at the population level, but less so before alignment occurred: the concepts approximately overlap, in the terminology of Section 1, within the limit set by the error margin ϵ.

If we regard the shared concept as the union of the concepts of the constituent individuals, then we have a *range* of possible boundaries to the concept. This is reminiscent of the 'egg-yolk' theory [13,9] which represents a vague spatial region in terms of its maximal and minimal possible extensions. The maximal extension is called the **egg** and consists of two subregions, the **white** together with the **yolk** (or mininimal extension); cf. Figure 6.[13] If X_t is a snapshot at time t of the vague spatial interval depicted in Fig. 5, then using the Gotts and Cohn [9] predicates eggof and yolkof, we might try to identify the maximal and minimal regions of X_t in terms of the preferred intervals $\iota_{i,t}$ of agents A_i at time t:

$$\mathsf{eggof}(X_t) = \bigcup_{A_i \in Ag} \iota_{i,t}$$

[13] See [3] for a discussion of how the 'egg-yolk' model relates to supervaluation [6,11] approaches to vagueness.

$$\mathsf{yolkof}(X_t) = \bigcap_{A_i \in Ag} \iota_{i,t}$$

In effect, then, the vagueness is an emergent property of the interaction between the agents in the population, rather than inhering to the conceptual structure of any individual agent. At one level, this perspective has certain attractions, since the indeterminacy of a concept like 'morning' does appear to be related to the wide variability in the way that it is applied by individual speakers. Nevertheless, in order to do justice to the intuitions behind the egg-yolk model, we would need to enrich the representation of intervals within agents in order to accommodate something like the egg white region. This would then offer the possibility of agents conditioning their willingness to adapt according to the partition into yolk and white. For example, it might be plausible for a proposer P to only modify its preferred interval if $\mu(\iota_{\mathsf{R}})$ fell at least within the 'white' part of the interval.

One major disadvantage of the framework used in Experiment 1 is that we have, so to speak, 'hard wired' the goal of alignment into our agents. It could be argued that this has some plausibility; for example, there is considerable empirical evidence that human speakers do align to each other at numerous levels of cognitive representation in dialogue, ranging from phonetics up to the levels of semantic representation and the internal 'situation model' [17]. Nevertheless, the process captured in our simulation corresponds more closely to alignment *across* successive dialogues, rather than within a dialogue, which weakens the analogy. Is it possible instead to devise a more principled approach which allows agents to discover the advantages of alignment by themselves?

4 Experiment 2

4.1 Reinforcement Learning

In the second family of experiments, we adopt a simple form of reinforcement learning [23] to replace the alignment strategy of Experiment 1.

Before discussing the details, we need to briefly return to the way in which the proposer P selects a departure time. In Experiment 1, the departure time was set to be the end of the agent's preferred interval. We now modify this as follows:

$$dep_{\mathsf{P}} = \begin{cases} arr_{\mathsf{P}} + 1 & \text{if } arr_{\mathsf{R}} < arr_{\mathsf{P}}, \\ \text{end of } \iota_{\mathsf{P}} & \text{otherwise} \end{cases} \tag{3}$$

In other words, P departs at time $t+1$ if she knows that R has already arrived (and has either departed already or is still present at time t). Otherwise, P waits until the last point of ι_{P}. For simplicity, we do not consider the length of the meeting to be a factor in determining costs or utility.

In principle, P incurs a waiting cost which is proportional to the length of the interval $[arr_{\mathsf{P}}, dep_{\mathsf{P}}]$. However, for simplicity, we treat it as a fixed value, regardless of the length of the wait.[14]

[14] We assume that the cost is zero if $dep_{\mathsf{P}} = arr_{\mathsf{P}} + 1$.

Let us return to the learning scenario. To ease exposition, suppose that we have a pool of two agents, with a fixed assignment of roles. Each run t of the simulation contains a representation of a state s_t, on the basis of which P selects an action a_t. On the next run, P receives a numerical reward r_{t+1} and finds itself in state s_{t+1}; the reward is used to build a model of the long-term utility of performing action a in state s_t (taking into account sequences of state changes induced by the action, assuming utility-optimal behaviour thereafter). P maintains a mapping from states to probabilities of selecting each possible action. This mapping is called a *policy*, and is updated in the light of rewards received in states up to and including the current one.

We represent a state with the variable *alignment*. This takes as value one of five possible labels, each of which serves as a bin for a range of integers, corresponding to the difference σ between the median $md(\iota_{\mathsf{P}})$ of P's preferred interval and estimated mean $\mu(\iota_{\mathsf{R}})$ of R's arrival times. The correspondence between labels and the value of σ are shown in Table 2. For example, *alignment* would be assigned the value *other_ v_ early* just in case $md(\iota_{\mathsf{P}}) - \mu(\iota_{\mathsf{R}}) > 6$.

Table 2. Values of the *alignment* variable

bin labels	**range of** σ
other_ v_ early	$\sigma > 6$
other_ early	$6 \geq \sigma > 1$
aligned	$1 \geq \sigma > -2$
other_ late	$-2 \geq \sigma > -7$
other_ v_ late	$\sigma \leq -7$

The set $\mathcal{A}$ of possible actions for P are analogous to the set of possible alignments:

$$\mathcal{A} = \{\textit{shift\_far\_earlier}, \textit{shift\_earlier}, \textit{no\_op}, \textit{shift\_later}, \textit{shift\_far\_later}\}$$

Each action is a mapping from intervals to intervals. Thus the two actions *shift_far_ earlier* and *shift_far_ later* move their input five units earlier or later, respectively, while *shift_ earlier* and *shift_ later* only move their inputs one unit earlier or later. *no_ op* just returns its input unchanged. The actions are defined so that intervals cannot be shifted beyond a stipulated lower and upper boundary (taken to be 1 and 21 in the current model). This constraint is realistic to the extent that, for example, the start point of *morning* would not normally occur before 12.00 am. However, the way that we have implemented these constraints could definitely be improved (for example by defining a probability distribution over possible start times).

Note that despite the potential fit between actions and alignments, any association between the two has to be learned by the agents, rather than being stipulated in the model.

Table 3. Reward Matrix

	$wait = 0$	$wait > 0$
met = True	2	1
met = False	-2	-3

The reward received by an agent depends on the values of two variables *met* and *wait*. The first of these is boolean-valued, while *wait* takes a non-negative integer as value. Rewards are allocated according to the matrix in Table 3. In order to choose an action, the agent estimates the relative values of all members of $\mathcal{A}$. The estimated value of action a on the t^{th} run in state s is written $\mathcal{Q}_t(a, s)$, and we define this to be the average of the rewards received in s by the time the action was selected. That is, if a has been selected k times in s by the time of run t, giving rise to rewards $r_1, r_2, \ldots, r_k$, then its value is estimated to be the following:[15]

$$\mathcal{Q}_t(s, a) = \frac{1}{k} \sum_{i=1}^{k} r_i \tag{4}$$

When $k = 0$, we take $\mathcal{Q}_t(s, a) = 0$. $\mathcal{Q}_t(s, a)$ is re-computed on each run of the simulation.

The simplest strategy for action selection is the greedy method: choose an action which has the highest estimated value. However, it turns out to be advantageous to behave greedily most of the time while occasionally — with small probability ϵ — selecting at random some other action. We take $\epsilon = 0.1$ initially, and let it decrease over successive runs, so that the action space is sampled more broadly at the beginning of the simulation. This is termed an **ϵ-decreasing** strategy.

The purpose of this approach is to provide a decision-theoretic grounding for the usefulness of alignment. Instead of assuming a hardwired propensity to adjust towards the other agents' concepts, rewards received from the environment alone should be sufficient to cause the agent to behave in such a way, i.e., it would be rational for her to do so, purely on the basis of self-interest.

4.2 Results

In Experiment 1, we only required agents to update their preferred interval with respect to the observed behaviour of their most recent partner. For example, we might have agent A_1 moving ι_i earlier after interacting with A_2 and moving it later on a successive turn after interacting with A_3. However, in Experiment 2 we also consider the case where agents align not to the pattern of their individual partners, but rather to the mean behaviour of *all* their partners. We shall refer

[15] Although reinforcement learning typically involves learning the utility of *sequences* of actions, the more restricted version we have adopted here is sufficient to support our claim that alignment can be learned rather than being hard-wired.

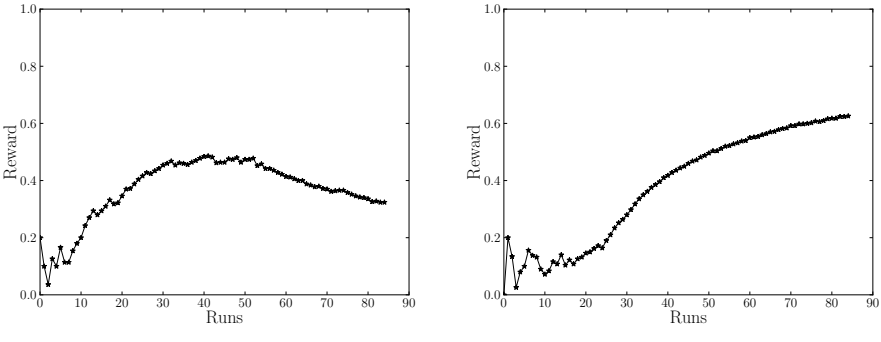
to these two conditions as ALIGN_TO_GROUP = *False* vs. ALIGN_TO_GROUP = *True*, respectively.

Figures 7(a), (b) show the **average reward ratios** achieved over 500 runs. For individual agents, the reward ratio is defined as follows, where K is set to be the maximum possible reward in a state, i.e. $K = 2$:

Definition 4. *The **reward ratio** for an agent is the quotient*

$$\frac{\textit{sum of rewards received}}{\#\ \textit{of proposals} \times K}$$

The average reward ratio is obtained as the mean of the reward ratio taken over the whole population. It remains low throughout the simulation under condition ALIGN_TO_GROUP = *False* (Figure 7(a)), but reaches a point above 0.6 when ALIGN_TO_GROUP = *True* (Figure 7(b)). Analysis of the behaviour of individual agents, illustrated in Figures 8(a), (b), shows that as in the case of Experiment 1, it is possible for one or more agents to persistently diverge from the rest of the group.



(a) ALIGN_TO_GROUP = *False* (b) ALIGN_TO_GROUP = *True*

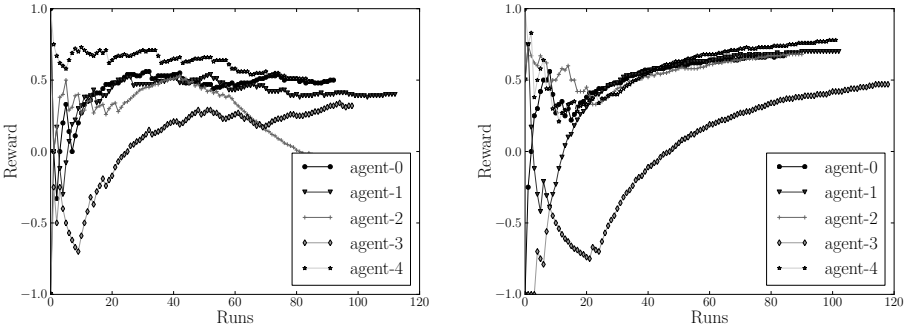
Fig. 7. Average Rewards over Time

Figures 9(a), (b) suggest that a reasonably stable alignment of intervals only emerges under condition ALIGN_TO_GROUP = *True*.

Finally, Figure 10 shows the extent to which proposals by agents lead to successful meetings.

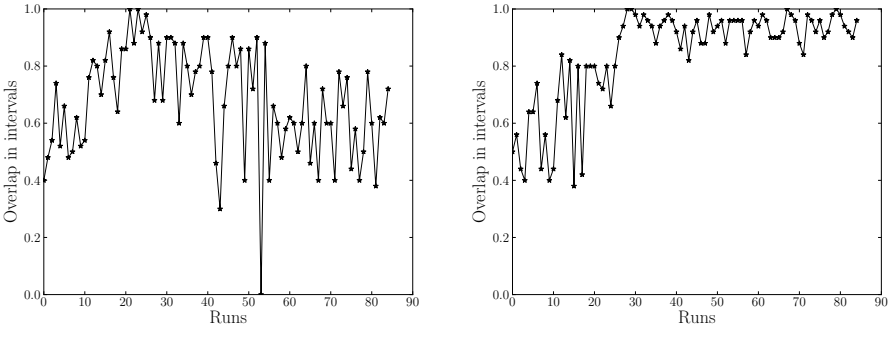
In general, the framework using reinforcement learning yields alignment results that are comparable with those achieved with the 'hard wired alignment' approach, but only when ALIGN_TO_GROUP = *True*.

These results are comparable to the work on word meaning and multi-agent language games carried out by Steels and colleagues [2,22,21]. However, unlike Steels, we are not concerned with how new terms emerge as bearers of meaning, but rather with how pre-existing 'unstable' meanings come to stabilize as a result of interaction. In addition, feedback about the interpretation of terms is



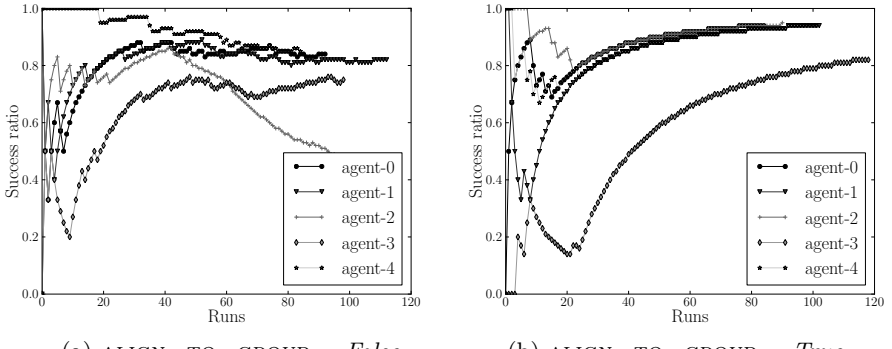
(a) ALIGN_TO_GROUP = *False* (b) ALIGN_TO_GROUP = *True*

Fig. 8. Rewards to Individual Agents over Time



(a) ALIGN_TO_GROUP = *False* (b) ALIGN_TO_GROUP = *True*

Fig. 9. Average Overlap in Preferred Intervals



(a) ALIGN_TO_GROUP = *False* (b) ALIGN_TO_GROUP = *True*

Fig. 10. Ratio of Successful Meetings to All Proposals

Table 4. Aligned Intervals after 500 runs, using Reinforcement Learning

agent-0:	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
agent-1:	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
agent-2:	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
agent-3:	[8, 9, 10, 11, 12, 13, 14, 15, 16, 17]
agent-4:	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

not acquired through explicit correction and deictic coordination, as for example in [2], but has to be inferred from whether proposals to meet are successfully consummated or not.

5 Conclusions and Future Work

This paper opened with the question *How is it that people manage to communicate even when they implicitly differ on the meaning of the terms they use?* We typically assume that the person we are talking understands our words in just the way in which we ourselves understand them; this is a crucial component of our shared 'common ground' [20] in the dialogue. Yet for many items of our core vocabulary, this assumption is probably too strong. Given differences in perceptual apparatus and in personal experience, meanings as mental entities surely differ somewhat from speaker to speaker. Despite these differences, communication usually succeeds, as far as we can tell. We have argued that a notion of **approximate semantic alignment** may be sufficient for communication in a task-oriented scenario. In order to support our claim, we have modelled the utility of a temporal expression for achieving coordinated action, specifically for pairs of agents to arrange meetings between themselves. We have shown that, given certain assumptions, the utility of the expression increases in line with interpretive alignment. That is, when the proposer's extension for the term overlaps more greatly with that of the responder, then the term is more effective in circumscribing the range of possible meeting times. This in turn increases the likelihood that two agents will successfully meet. If the agents adopt reinforcement learning, then over numerous interactions, they will tend to converge on more tightly aligned sets of interpretations, leading to a stable pattern of successful meeting proposals. However, as we pointed out earlier, our current model only achieves this convergence if agents align to the group average arrival time, rather than successively attempting to align to the average of their immediate partner.

Despite the fact that increased alignment correlates with increased utility, the way we have modelled multi-agent simulation rarely if ever leads to complete alignment. This adds support for the contention that vague terms provide robustness to communication — they work 'well enough' in the absence of complete agreement on boundaries. In order to explore this point more fully, let's return

to the details of Experiment 2. Figure 10(b) indicates that one of the five agents (namely `agent-3`) is less successful than all the others — achieving a score of 0.8 against an average of 0.95 for the other four. Inspection of the simulation log shows that `agent-3` has ended up with a preferred interval that diverges markedly from the rest of the members of the agent pool; see Table 4. Regardless of the reasons why `agent-3` has arrived at a sub-optimal policy, there is one striking fact: since there is sufficient overlap in preferred intervals, a 'good enough' policy can persist. In other words, the residual divergence between intervals across the population does not seriously impede the agents in achieving their goal of meeting.

From a methodological point of view, simulations of the kind presented in this paper do not allow strong conclusions to be drawn, and some kind of analytic model would be desirable. On the other hand, we would argue that simulations do allow us to be explicit about the assumptions we are making and to refine the kind of questions we want to ask. There are a number of issues which we plan to explore in future work, most notably the following:

1. representing temporal intervals using an 'egg-yolk' style representation;
2. allowing a responder agent to reject the proposer's suggestion, and to negotiate an alternative;
3. including in the community certain agents who refuse to adapt to other agents;
4. expanding the range of interactions by providing agents with a lexicon of complementary time expressions (such as *morning*, *midday*, *afternoon* and *evening*);
5. allowing agents to choose temporal expressions at different levels of granularity and to use approximations [12], and
6. applying the approach to the spatial domain.

References

1. Bailin, S., Truszkowski, W.: Ontology negotiation: Ontology negotiation: How agents can really get to know each other. In: Truszkowski, W., Hinchey, M., Rouff, C.A. (eds.) WRAC 2002. LNCS, vol. 2564. Springer, Heidelberg (2003)
2. Baillie, J.C., Steels, L.: Shared grounding of event descriptions by autonomous robots. Robotics and Autonomous Systems 43(2-3), 163–173 (2003)
3. Bennett, B.: Application of supervaluation semantics to vaguely defined spatial concepts. In: Montello, D.R. (ed.) COSIT 2001. LNCS, vol. 2205, pp. 108–123. Springer, Heidelberg (2001)
4. Burato, E., Cristani, M.: Learning as meaning negotiation: A model based on english auction. In: Håkansson, A. (ed.) KES-AMSTA 2009. LNCS (LNAI), vol. 5559, pp. 60–69. Springer, Heidelberg (2009)
5. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Berlin (2007)
6. Fine, K.: Vagueness, truth, and logic. Synthese 30, 265–300 (1975)
7. Freyd, J.: Shareability: The social psychology of epistemology. Cognitive Science 7(3), 191–210 (1983)
8. Gärdenfors, P.: Conceptual Spaces: The Geometry of Thought. A Bradford Book, MIT Press, Cambridge, Mass (2000)

9. Gotts, N., Cohn, A.: A mereological approach to representing spatial vagueness. In: Proceedings of the 5th International Conference on Principles of Knowledge Representation and Reasoning (KR 1996), pp. 230–241 (1996)
10. Jäger, G., van Rooij, R.: Language structure: psychological and social constraints. Synthese 159(1), 99–130 (2007)
11. Kamp, H.: Two theories of adjectives. In: Keenan, E. (ed.) Formal Semantics of Natural Languages. Cambridge University Press, Cambridge (1975)
12. Krifka, M.: Approximate interpretation of number words: A case for strategic communication. In: Bouma, I.K.G., Zwarts, J. (eds.) Cognitive Foundations of Communication, pp. 111–126. Koninklijke Nederlandse Akademie van Wetenschapen (2007)
13. Lehmann, F., Cohn, A.: The EGG/YOLK reliability hierarchy: Semantic data integration using sorts with prototypes. In: Proceedings of the Third International Conference on Information and Knowledge Management, p. 279. ACM, New York (1994)
14. McNeill, F., Bundy, A., Walton, C.: Planning from rich ontologies through translation between representations. In: Proceedings of ICAPS 2005 Workshop on The Role of Ontologies in Planning and Scheduling, Monterey, CA, USA (2005)
15. Ohlbach, H.J.: Calendar logic. In: Gabbay, D.M., Reynolds, M.A., Finger, M. (eds.) Temporal Logic: Mathematical Foundations and Computational Aspects, vol. 2, pp. 477–573. Oxford University Press, Oxford (2000)
16. Parikh, R.: Vagueness and utility: the semantics of common nouns. Linguistics and Philosophy 17, 521–535 (1994)
17. Pickering, M.J., Garrod, S.: Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences 27(02), 169–190 (2004)
18. Reiter, E., Sripada, S., Hunter, J., Yu, J., Davy, I.: Choosing words in computer-generated weather forecasts. Artificial Intelligence 167, 137–169 (2005)
19. da Silva Sinha, V., Sinha, C., Zinken, J., Sampaio, W.: When time is not space: The social and linguistic construction of time intervals in an Amazonian culture. Accepted for Publication in the Journal of Pragmatics (to appear)
20. Stalnaker, R.: Assertion'. P. Cole. Syntax and Semantics 9: Pragmatics, 315–332 (1979)
21. Steels, L.: The origins of ontologies and communication conventions in multi-agent systems. Autonomous Agents and Multi-Agent Systems (1998)
22. Steels, L., Vogt, P.: Grounding adaptive language games in robotic agents. In: Proceedings of the Fourth European Conference on Artificial Life, pp. 474–482 (1997)
23. Sutton, R.S., Barto, A.G.: Reinforcement Learning. A Bradford Book, MIT Press (1998)
24. Wang, J., Gasser, L.: Mutual online concept learning for multiple agents. In: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2002, pp. 362–369. ACM, New York (2002)
25. Williamson, T.: Vagueness. Routledge, London (1994)
26. Warglien, M., Gärdenfors, P.: Semantics, conceptual spaces and the meeting of minds (2007), http://logic.sysu.edu.cn/Article/UploadFiles/200810/20081022091135682.pdf

Vagueness as Probabilistic Linguistic Knowledge

Daniel Lassiter*

New York University

Abstract. Consideration of the *metalinguistic* effects of utterances involving vague terms has led Barker [1] to treat vagueness using a modified Stalnakerian model of assertion. I present a sorites-like puzzle for factual beliefs in the standard Stalnakerian model [28] and show that it can be resolved by enriching the model to make use of probabilistic belief spaces. An analogous problem arises for metalinguistic information in Barker's model, and I suggest that a similar enrichment is needed here as well. The result is a probabilistic theory of linguistic representation that retains a classical metalanguage but avoids the undesirable divorce between meaning and use inherent in the epistemic theory [34]. I also show that the probabilistic approach provides a plausible account of the sorites paradox and higher-order vagueness and that it fares well empirically and conceptually in comparison to leading competitors.

Keywords: Vagueness, probability, lexical representation, higher-order vagueness.

1 Introduction

One grain of sand is clearly not a heap. It seems plausible that, if you have something that is not a heap and you add one grain of sand to it, you still do not have a heap. But from these two premises it follows that no amount of sand can constitute a heap. That is, the following is a valid argument:

(1) Sorites Paradox
   a. One grain of sand is not a heap.
   b. If you add one grain of sand to something that is not a heap, then you still will not have a heap.
   c. ∴ No pile of sand, no matter how large, is a heap.

Slightly more formally, taking S_n to mean "an aggregation of n grains of sand":

(2) a. $\neg(heap(S_1))$
   b. $\forall n[\neg heap(S_n) \rightarrow \neg heap(S_{n+1})]$
   c. ∴ $\forall n[\neg heap(S_n)]$

* Thanks to Chris Barker, Philippe Schlenker, Gregory Guy, Chris Potts, Paul Egré, Stephanie Solt, and a reviewer for the current volume for helpful comments and discussion. Thanks also to Joey Frazee and David Beaver, whose very interesting paper [7] reached me too late to affect the content, but has important overlap with the current work.

R. Nouwen et al. (Eds.): ViC 2009, LNAI 6517, pp. 127–150, 2011.

In classical logic, the inductive premise (2-b) is equivalent to a denial that there is any number n such that n grains of sand do not constitute a heap, but $n+1$ grains do. That is,

(3) $\neg\exists n[\neg heap(S_n) \wedge heap(S_{n+1})]$

The problem of the sorites is that both variants of the inductive premise, (2-b) and (3), are intuitively plausible, but the conclusion is not. We know that the first premise is true (one grain is not a heap), and that the conclusion is false (heaps of sand do exist). It follows that the inductive premise is false. But it is very difficult to determine precisely where the inductive premise goes wrong.

In addition to the problem of the sorites, the existence of vagueness presents a serious challenge to the foundations of formal semantics. One way to identify this problem is in Grim's [8] claim that a precise theory of vague terms is impossible:

> Any successful account of vagueness will have to incorporate vagueness in one way or another; at the core of the Supervaluational approach, for example, lies the vagueness of 'acceptable precisifications'. Any hope for a fully precise account of vagueness is doomed.

A basic assumption of formal semantics is that natural language meaning and inference can be modeled in a mathematically precise fashion. This assumption is embodied directly in the principle of COMPOSITIONALITY: the meaning of an expression is built up from the meaning of its parts and the way that they are put together. But if if the basic objects of computation are not themselves precise, the idea that we can *compute* the meaning of a sentence (or even a discourse) from the meanings of words is problematic from the start. Since many, perhaps most, expressions of natural language are vague, formal semantics is in deep trouble if Grim is right.

The best-known theory of vagueness that is capable of avoiding Grim's problem is the epistemic theory [34]. According to this theory, meanings *are* precise, and the phenomena of vagueness are the result of humans' imperfect knowledge of the meanings of words. If this is right, formal semantics can carry along merrily. However, a common objection to the epistemic theory is that it requires an implausible divorce between meaning and humans' knowledge and use of language. In Williamson's own words, "[a]lthough meaning may supervene on use, there is no algorithm for calculating the former from the latter"; again, "meaning may supervene on use in an unsurveyably chaotic way" [35, pp. 206,209] . That is, there is no hint in the epistemic theory as to where meanings do come from. To those who view the study of language as part of (or at least closely connected to) the study of human psychology and sociology, this consequence of epistemicism tends to come across as a *reductio ad absurdum* of the theory.

Despite these criticisms, I think that the epistemic theory contains a crucial insight by treating vagueness as a result of IMPRECISE KNOWLEDGE OF LANGUAGE. That is, epistemicists locate vagueness not in the semantic theory proper, but in the relation between language users and the semantic theory.

Further, since (as Williamson [34] emphasizes) uncertainty is a very general fact of the human condition, the epistemic theory has the advantage of parsimony: it allows us to reduce vagueness in language to independently motivated features of human knowledge and belief.

The theory of vagueness presented in this paper incorporates this aspect of the epistemic theory while also allowing for a close connection between meaning and use. It can be seen as a development of David Lewis' suggestion that "languages themselves are free of vagueness but ... the linguistic conventions of a population, or the linguistic habits of a person, select not a point but a fuzzy region in the space of precise languages" [20]. On this approach, like the epistemic theory, vagueness resides in the relation between humans and precise languages. However, it is not necessary to endorse the epistemicist's claim that there is a single precise language that is being spoken – "English" or "Swahili", say, or some ultra-precise variant of these. Rather, there is always a range of languages that are contenders for being the language of conversation, and the epistemic problem is to use prior knowledge and context to select the most plausible candidates for being the language of conversation (see also [19,3]). This approach does not in itself rule out the epistemic theory; but, I will suggest, it makes it possible to claim the advantages of epistemicism without also taking on its less palatable consequences.

The theory developed here, like Lewis' and Williamson's theories, holds that vagueness is not a special semantic phenomenon, but a consequence of the nature of linguistic knowledge and general principles of language use. The precise development of this claim, however, will not be in terms of "fuzzy" regions but, like Edgington [4], in terms of PROBABILISTIC linguistic representations. There is independent reason to believe that human knowledge is represented probabilistically. I argue that this perspective leads naturally to a model of interpretation as an interpreter mapping words and other utterance-types to a probability distribution over precise resolutions. This approach allows us to extend existing models of probabilistic knowledge representation and reasoning to the interpretation and representation of vague terms. In this way, we can explain how formal semantics is possible, and show not only why the conclusion of the sorites paradox is false, but also why the premises seem so compelling.

The essay is structured as follows. Section 2 describes the set-theoretic model of assertion and belief update and arguments from Stalnaker [28] and Barker [1] that this model also encompasses beliefs and assertions about language, including vague language. Section 3 argues that a probabilistic enrichment of this model is needed to account for partial belief and to avoid a sorites-like paradox for factual beliefs involving continuous sample spaces. A similar argument suggests that linguistic representations are also probabilistic. I explicate this by offering a possible-languages model of linguistic knowledge akin to the familiar possible-worlds model, and show how this leads naturally to a picture of lexical representations as probability distributions over model-theoretic objects. Section 4 shows how the probabilistic approach resolves the sorites paradox, focusing on the differences between model-theoretic and probabilistic variants of

the inductive premise and why they come apart. Section 5 points out some important features of the model relating to assertibility, negation, borderline cases, and common ground. In section 6 I compare the theory to alternative accounts, showing that it is preferable in terms of empirical coverage and the intuitive correctness of its predictions, and I answer objections from Stephen Schiffer and Nicholas Smith, who have claimed that partial beliefs about the applicability of vague terms do not behave like subjective probabilities. Finally, section 7 is a brief treatment of higher-order vagueness, showing that this perspective can make sense the dual role of *definitely* as an epistemic modal and a metalinguistic modifier.

2 Metalinguistic Assertion and Linguistic Knowledge

2.1 Assertion and Metalinguistic Assertion

In Stalnaker's model, the role of an assertion is to eliminate certain possibilities from the common ground, which is construed as a set of worlds considered by the conversational participants as live possibilities for how the actual world might be [28]. Suppose you don't know whether it is raining outside. If a reliable source tells you, "It is raining", you will normally update your beliefs so that you no longer consider worlds in which it is not raining to be candidates for being the actual world. Taking propositions and common grounds to be sets of worlds, Stalnaker suggests treating the update operation as intersection: that is, to update the common ground with the information that p, simply intersect the current common ground with the p-worlds. Private belief update is treated similarly: if a person comes to believe that p, then their new belief state is simply the old belief state intersected with the proposition that p.

Stalnaker [28] also notes that assertions do not only convey information about the non-linguistic world (states of weather and the like). Rather, utterances may give information about how other utterances are to be interpreted. For example, suppose someone asks you, "What is an optometrist?" In this context, the reply "An optometrist is an eye doctor" serves to inform the linguistically uncertain interlocutor that, in the current language of conversation, the sequence of sounds "optometrist" is not an appropriate way to communicate any concept other than EYE DOCTOR. Importantly, "An optometrist is an eye doctor" gives no information about the non-linguistic world, but rather functions as an instruction to interpret a particular sequence of sounds in a certain way.

We might think that ignorance about technical vocabulary is a special case, though: perhaps, in the general case, there is a single clear-cut "current language of conversation". Barker [1] and Stalnaker [31] argue that there is not: rather, there are typically many languages which are viable candidates for being the language of conversation, just as there are typically many worlds that are viable candidates for being the actual world. Barker's discussion focuses on vague adjectives like "tall". In Barker's example, imagine that you arrive in a new town and you have no idea about the typical heights of local inhabitants. You ask a local: "What counts as 'tall' around here?", and the local responds: "See John over there? John is tall."

Even if we have quite precise information about John's height – say, we know that he is $5'11.4''$ – this utterance can be informative because it has the metalinguistic effect of narrowing down the range of possible interpretations of "tall". If this utterance is true, then the local meaning of "tall" must be such that John counts as tall; so we can eliminate languages where John does not count as "tall" as candidates for being the current language of conversation.

Barker's technical implementation of this effect is very close to the original Stalnakerian model (especially [30]). He assumes that in each world in the common ground there is a unique precise language that is the current language of conversation. Metalinguistic effects are modeled by the same update procedure as ordinary assertions (in Barker's implementation, using the apparatus of Dynamic Semantics). So, for example, before the local's utterance, there were worlds in the common ground where the standard for counting as "tall" in the current conversation was $6'0''$, so that people who are $5'11.4''$ do not count as "tall". After this utterance, these worlds are eliminated.

2.2 A Model for Metalinguistic Belief and Assertion

Barker's model of metalinguistic assertion has one feature worth flagging here. Since each world is associated with a unique precise current language of conversation, the model effectively assumes an epistemic theory of vagueness: if we were able to discern precisely what world we are in, we would also be able to fix the precise interpretation of all vague terms. I have already indicated my discomfort with this idea: I think it extremely implausible that human linguistic practices are somehow able to determine (and without speakers' knowledge) a single precise language that is being spoken in a given conversation. But it is not difficult to construct an alternative version of Barker's model that does not have this commitment.

Suppose that there is a domain of worlds W and a domain of possible languages L, where each $l \in L$ is a partial function from utterance-types to model-theoretic objects. Each conversational participant $x \in X$ has a belief-set $B_x \subseteq W \times L$ – that is, beliefs are sets of world-language pairs. For any x, x's *factual* belief-set is

$$\{w \mid \exists l \in L : (w, l) \in B_x\},$$

and her *metalinguistic* belief-set is given by

$$\{l \mid \exists w \in W : (w, l) \in B_x\}.$$

Define the *factual common ground*, more or less standardly, as the intersection of the conversational participants' factual beliefs (and so the union of their belief-sets):

$$\bigcup_{x \in X} \{w \mid \exists l \in L : (w, l) \in B_x\}$$

The *linguistic common ground* will then be the intersection of the conversational participants' metalinguistic beliefs.

$$\bigcup_{x \in X} \{l \mid \exists w \in W : (w, l) \in B_x\}\,.$$

The reason for separating the two components of B_x is that they serve very different roles in conversation. The goal of inquiry is to exchange information in an effort to narrow down the domain of possible worlds [29]. The purpose of gaining metalinguistic knowledge, I take it, is to enable people to conduct inquiry more efficiently: the more we know about each others' linguistic habits, the smaller the linguistic common ground, and the more effectively non-metalinguistic assertions will be able to narrow down the factual common ground. On this model there is no "fact of the matter" about which language is being spoken that goes beyond the attempt to coordinate on a set of languages as small as possible, in order for conversational participants to exchange information as efficiently as possible. In other worlds, the current languages of conversation are just the languages in the linguistic common ground. There is no need for reference to some extrinsically given set of facts about which language is being spoken: choosing a (set of) common language(s) in a given conversation is a coordination game in the sense of Schelling [25] and Lewis [19].

Within this model we can treat the various types of information gain as follows. Call the informational effect of an assertion (or other event) *purely factual* if the same languages are considered possible in the prior state S and the posterior state S': that is, if the factual common ground in S' is a proper subset of the factual common ground in S, but the linguistic common grounds at S and S' are equal. The effect of an assertion is *purely metalinguistic* if the linguistic common ground in S' is a proper subset of the linguistic common ground in S, but the factual common grounds are equal. The effect of an assertion is *mixed* if it is neither purely factual nor purely metalinguistic, i.e. if both worlds and languages are eliminated from the common ground.[1]

I should add that there is no reason in principle why the proposal I will develop could not be treated as an implementation of the epistemic theory of vagueness. To construct such a variant, simply ignore the contents of this subsection and treat the set-theoretic model as Barker gives it as the base of the probabilistic enrichment I will develop. However, it should be clear that the apparent connection with the epistemic theory is not a deep commitment of the model developed here, but simply one possible interpretation.

[1] An example of a mixed assertion is if someone says "John is tall" in a context in which we are unsure both about John's height and about the interpretation of "tall". Then, for any height h such that we are sure that John is at least as tall as h, we can eliminate languages from the common ground where the standard for "tall" is greater than h. Likewise, if we are sure that the standard is at least h', then we can eliminate worlds in which John's height is less than h'. In the probabilistic model to be developed, Bayesian update for mixed assertions is defined in similar fashion. This aspect is important because beliefs about the world and beliefs about language obviously do interact: we would not want a theory that separates them completely. See section 5.1 for the probabilistic version.

3 Probabilistic Semantic Representation

3.1 Toward a Probabilistic Model: Factual Information

The trouble with the model we have just outlined – both in the Stalnaker-Barker form and in the modified form just presented – is that it is not sufficiently rich to deal with cases where information change does not involve eliminating possibilities. You think you saw rain just now, but you consider it possible that you were mistaken. You may then increase your degree of belief in the proposition that it is raining without eliminating any worlds from the set of worlds considered possible. But if (factual) belief update is intersective, this option is excluded: you must either eliminate the worlds where it is not raining or leave them in. Intersection is not an appropriate model for this change in information, then.

Similarly, the effect of assertions is not always intersective. Consider the testimony of a witness of unknown reliability, who tells you "It is raining". (I am supposing that the meaning of this utterance is sufficiently clear that we do not need to worry about metalinguistic effects for the moment.) You might wish to increase your degree of belief in the proposition that it is raining, without eliminating the possibility that it is not – i.e., the possibility that the witness is misinformed or lying. Again, the simple intersective model of belief update is inadequate, both to model the update and to deal with degrees of belief.

An obvious way to deal with problems of this type is to represent partial belief using an enriched model such as subjective probability. On this approach, after looking out the window you should update your beliefs about the likelihood of rain using Bayes' rule, according to the estimated likelihood that the evidence of your senses really did indicate rain (and not, say, a sprinkler outside the window). Similarly, in the case of the unreliable witness, your belief in rain after the witness' assertion "It is raining" will depend on your estimate of the probability that the witness is a truth-teller.

A second problem with the set-theoretic model of belief is an analogue of the sorites paradox with beliefs involving continuous sample spaces. There is a real number r such that the top of the Eiffel Tower is r kilometers away from the top of Big Ben. I know for sure that r is not less than 100, and that it is not more than 1000, but I certainly do not know what r is with any precision. However, my knowledge is even more imprecise than this characterization may suggest. Intuitively, (4) holds:

(4) There is no real number r' such that my belief state allows for the possibility that Big Ben and the Eiffel Tower are r' kilometers apart, but excludes the possibility that they are $r' \pm \epsilon$ kilometers apart for sufficiently small ϵ.

Going through a forced march – "Are Big Ben and the Eiffel Tower 400 kilometers apart? Are they 400.01 kilometers apart?" etc. – there is no point at which I would be comfortable switching from "maybe" to "no" in a single increment of 0.01 kilometers.

In the set-theoretic model, (4) entails that there is no sharp cut-off in which worlds are considered possible: for any r', if there are r'-worlds in the belief-set,

then there are r''-worlds in the belief-set for any $r'' \in [r' - \epsilon, r' + \epsilon]$. It follows that my belief-set contains worlds where r is any number you like, including, e.g., 1 kilometer and 1,000,000 kilometers. This contradicts the assumption that I know that r is not less than 100 or more than 1000. The paradox, in effect, is that imprecise knowledge is no knowledge at all.

Again, an obvious approach to this problem – though not the only one, to be sure – is to treat factual beliefs as probability distributions over (appropriate subsets of) W. The reason that (4) holds, on this model, is that r' is a continuous random variable, so that there are no sharp cut-offs in the likelihood that $r = r'$. Suppose, for example, that $prob(r = r')$ is normally distributed with $\mu = 400, \sigma = 100$. Then it is much more likely that r is in $[390, 400]$ than that it is in $[190, 200]$, which is in turn much more likely than that r is in $[10, 20]$. Incidentally, since the latter has an extremely small but non-zero probability, this example technically requires a semantics for the adjective "possible" that does not equate possibility with non-zero probability. For relevant discussion see [32,37,38,18].

3.2 Toward a Probabilistic Model: Metalinguistic Information

Factual beliefs involving continuous sample spaces create problems for the set-theoretic model of belief that are uncannily similar to the sorites paradox. This is not, I will suggest, accidental: standard examples of vagueness involve predicates ranging over sample spaces that are either continuous (e.g. "tall") or involve very small increments (e.g. "heap"). I will suggest that the solution to the problem of vagueness is effectively the same as the solution to the problem in (4) suggested in the previous section.

For an example of metalinguistic belief involving a continuous sample space, we can stick with our stock example "tall". Consider Barker's scenario again. We know exactly how tall John is – he is $5'11.4''$ – and we know that he counts as "tall" in the local community. Barker's approach to metalinguistic update encounters a problem similar to (4), given in (5):

(5) In the context just described, there is no real number r' such that my belief state allows for the possibility that "tall" means "having height at least r' inches", but excludes the possibility that "tall" means "having height at least $r' - \epsilon$ inches" for sufficiently small ϵ.

(We use $r' - \epsilon$ rather than $r' \pm \epsilon$ here because there clearly is an r' in the linguistic common ground such that we can be sure that "tall" does not mean "having height greater than $r' + \epsilon$ inches": just set $r' = 5'11.4''$.)

(5) is of course a variant of the inductive premise of the sorites paradox, tailored to the set-theoretic model of metalinguistic belief. It strikes me as highly plausible: surely there is no point, as we move from $5'11.4''$ down to $1''$ at 0.01-inch increments, where I could comfortably agree that some precise interpretation of "tall" is possible, but the next interpretation is totally impossible. Add a few plausible premises and we have a full-blown sorites paradox:

(6) a. My belief state entails that someone who is $6'9''$ tall counts as "tall" in any language in the common ground.
b. My belief state entails that someone who is 1 foot tall does not count as "tall" in any language in the common ground.

(6-a) and (6-b) are mutually incompatible with (5). In other words, if (5) is correct, then my belief state must admit interpretations of "tall" where the standard is below $5'11.4''$, even ones where "tall" means "having a height of 1 inch or more".

This consequence is surely unpalatable: I am quite confident, for example, that someone who is one foot tall does not count as "tall" in any English-speaking community. Within this approach, our options are to reject (5) or to reject (6-b). Rejecting (6-b) seems to be out; but rejecting (5) would mean that there is a sharp cut-off in my metalinguistic beliefs regarding "tall" which I am unaware of (and presumably have no introspective access to). This response is no more plausible, I think, than it would be in the case of factual uncertainty.

The solution in this case should be, I think, just as in the factual case: enrich the belief space using probability measures or something with an equivalent effect. Let a *probabilistic belief space* be a triple $\langle W, L, \mu \rangle$, where W is a set of possible worlds, L a set of possible languages, and $\mu : (W \times L) \rightarrow [0, 1]$ a function from world-language pairs to probabilities obeying the usual axioms. If we like, we can define separate probability measures for languages and worlds:

$$\mu_W(w') \underset{def}{=} \sum_{l \in L} [\mu(w', l)]$$

– the total probability of a world w is just the sum probability of all world-language pairs in which w occurs – and likewise

$$\mu_L(l') \underset{def}{=} \sum_{w \in W} [\mu(w, l')]$$

– the probability simpliciter of a language l is the sum probability of all world-language pairs in which l occurs.

I will make use of these abbreviations in what follows, but it is important to keep in mind that languages and worlds do not, in general, vary independently: facts about the world impose serious constraints on what languages are plausible candidates for receiving significant probability mass. To pick an obvious case, if we are talking about cats, no world-language pair (w', l') will receive significant probability if l' does not yield a value for the sequence of sounds "cat", or if it assigns "cat" some value that is totally unrelated to cats. So the choice of language is obviously constrained by facts about the world, in this case facts about what noises are being made in the current conversation.

Less trivially, if someone says "Mary is tall" in a context where we are uncertain both about Mary's height and about the interpretation of "tall", the probabilities of worlds and languages will certainly not be affected independently by this utterance. In particular, if we are certain (i.e., probability 1) that Mary counts as "tall" in the local context and that Mary is at least h inches tall, then we can assign probability 0 to any world-language pair (w, l) in which the interpretation of "tall" in

l does not include people of height h. If the probability is small but non-zero that Mary is at least h' inches tall, then we should adjust our probability for languages where "tall" receives a value of at most h' inches to some appropriately small value. (I don't want to get involved in spelling out the complete update procedure here, but it is a straightforward application of Bayes' rule.) In any case, the idea is that the choice of world will constrain, but not fully determine, the probabilities that can appropriately be assigned to world-language pairs.

The explanation of the modified sorites paradox for metalinguistic belief (5) within this approach will be essentially the same as the explanation of (4) given above: the value of *tall* is a continuous random variable. This interpretation of (5) says that there is no point at which the probability that "tall" means "having height at least r inches" is substantially greater than the probability that "tall" means "having height at least $r - \epsilon$ inches" for small ϵ. Nevertheless, the probability that "tall" means "having height of at least r' inches" approaches zero as r' gets smaller. The probability that "tall" applies to people who are 1 foot tall, for example, is effectively zero.

3.3 Lexical Probabilities

The notion of a possible language, though useful for our purposes, suffers from much the same psychological implausibility as the notion of a possible world: just as people do not reason about factual issues using an infinite set of discrete possibilities, they surely do not reason metalinguistically using an infinite set of discrete languages. The solution, in the factual case, is to think of beliefs as probability distributions over propositions. Probabilities of factual events can of course be defined equivalently in either way. Similarly, we can simplify our agents' task by treating metalinguistic belief in terms of probability distributions over precise resolutions of individual words, and relate to an equivalent formulation in terms of the probability function μ (and its derivative μ_L).

An agent's belief-set determines the representation of a word or utterance-type in the following way. As usual, $\mathbf{D_e}$ is the set of possible objects whose members are $\mathbf{o_1}, \mathbf{o_2}, \ldots$. For simplicity's sake we will restrict attention to model-theoretic objects of type e and $\langle e, t\rangle$, although the definition could easily be expanded to account for objects of arbitrary type.[2]

[2] I am supposing that gradable adjectives have semantic type $\langle e, t\rangle$ (cf. [17,1]), though nothing hinges crucially on this assumption. If gradable adjectives turn out to be of type $\langle d, et\rangle$ [33,10] or type $\langle e, d\rangle$ [14,15,16], they will need special treatment. For example, von Stechow [33] and Kennedy [15] convert gradable adjectives in the positive form into properties of individuals using a silent degree morpheme. The value of this silent morpheme is given by a contextual parameter or something with equivalent effect, and is generally quite underdetermined. All of the arguments given here for the probabilistic model apply equally to the contextual determination of a value for the silent positive morpheme, I think, and we could easily extend the current approach to include probability distribution over contexts – something that is probably needed anyway. And, however gradable adjectives are best handled, the model described here is needed for vague terms like *heap* and *Mount Everest* that are not gradable adjectives, and do not show any sign of reference to degrees in their semantics.

Define the *lexical probability function* $LP_{u,A}$ of a word u according to an agent A as a function $f_A : \mathbf{D_e} \rightarrow [0,1]$, subject to the condition in (7):

(7) $$f_A(u(\mathbf{o_m})) = \sum_{l \in L} [\mu_L(l) : l(u)(\mathbf{o_m}) = \text{True}]$$

(μ_L should of course be understood as relativized to A's belief state here. Note also that (7) is only appropriate for finite L; the infinite case is not significantly different except that a bit of calculus is needed.)

In words, (7) says that the probability according to A that the word u applies to the object $\mathbf{o_m}$ is equal to the sum probability of all possible languages such that the value of u in that language, applied to $\mathbf{o_m}$, yields the value True. Since no ambiguity arises, I will use "*prob*" as a probability function for both languages and words/utterance-types from now on. The reader should understand "$prob_A(u(\mathbf{o_m})) = d$" as meaning

$$f_A(u(\mathbf{o_m})) = d$$

or, equivalently, as

$$\sum_{l \in L} [\mu_L(l) : l(u)(\mathbf{o_m}) = \text{True}] = d$$

As an example, suppose that the lexical representation of *tall*, $LR_{tall,A}$, for a particular speaker A has the following form. A considers possible these resolutions of *tall*: 5′6″, 5′7″, ... 6′5″ (spaced at 1″ to simplify the model; the extension to continuous probability spaces is straightforward). Letting italics represent utterance-types as above and boldface indicate model-theoretic objects, we'll call these $\mathbf{tall_1}, \mathbf{tall_2}, ...\mathbf{tall_{12}}$. Each resolution of *tall* denotes (the characteristic function of) a set of individuals who satisfy certain conditions. So, for example, $[\![\mathbf{tall_1}]\!] = \lambda x.x$'s height $\geq$ 5′6″; $[\![\mathbf{tall_2}]\!] = \lambda x.x$'s height $\geq$ 5′7″; and so forth. The probability distribution in the bottom row of Table 1 assigns a probability in the range $[0,1]$ to each resolution of *tall*. For example, the third column of Table 1 should be read: "The probability that the denotation of *tall* is '$\lambda x.x$'s height$\geq$ 5′8″' is 0.03".

As in (7), for any word u all of whose interpretations denote sets of individuals, the probability that x is u is the sum of the probabilities of all interpretations of u of which x is an element. In the case of scalar adjectives like *tall* it is a straightforward matter to calculate the probability that an individual x is u, because all of the available interpretations of *tall* in Table 1 are upward monotonic. That is, if an individual x is a member of the set denoted by $tall_n$, and the height of an individual y is greater than that of x, then y is also in the set denoted by $tall_n$.

Thus, for any height, the probability that an individual of that height will count as *tall* is simply the sum probability of all thresholds of height less than or equal to that height, i.e. the cumulative probability. To illustrate, consider 12 individuals in this height range, spaced at 1 inch, called $x_1, x_2, ...x_{12}$. Using

Table 1, we can calculate for each individual x_n the probability that x_n is *tall* as follows. For example,

$$prob_A(tall(x_5)) = \sum_{i=1}^{5} prob(tall = \mathbf{tall_i}) = 0 + .01 + .03 + .09 + .14 = .27.$$

Table 2 gives the values of $tall(x_n)$ according to (7) and the probability distribution in Table 1 for a representative sample of individuals of various heights.

Table 1. Sample lexical probability function for *tall*

Name	$\mathbf{tall}_1$	$\mathbf{tall}_2$	$\mathbf{tall}_3$	$\mathbf{tall}_4$	$\mathbf{tall}_5$	$\mathbf{tall}_6$	$\mathbf{tall}_7$	$\mathbf{tall}_8$	$\mathbf{tall}_9$	$\mathbf{tall}_{10}$	$\mathbf{tall}_{11}$	$\mathbf{tall}_{12}$
Threshold	5′6″	5′7″	5′8″	5′9″	5′10″	5′11″	6′0″	6′1″	6′2″	6′3″	6′4″	6′5″
$prob_A$ $(tall = \mathbf{tall_n})$	0	.01	.03	.09	.14	.23	.23	.14	.09	.03	.01	0

Table 2. Cumulative probability distribution corresponding to Table 1

Individual	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
Height	5′6″	5′7″	5′8″	5′9″	5′10″	5′11″	6′0″	6′1″	6′2″	6′3″	6′4″	6′5″
$prob_A(tall(x_n))$	0	.01	.04	.13	.27	.5	.73	.87	.96	.99	1	1

As Table 2 suggests, the upward monotonicity of all resolutions of *tall* in Table 1 explains the intuitive fact that the probability of an individual x_n being tall increases gradually the greater x_n's height is. Graphically, Table 2 yields the cumulative distribution in Figure 1.

Since we are looking at increments of 1″, this graph only approximates a smooth curve. We can increase the resolution of Tables 1 and 2 by considering

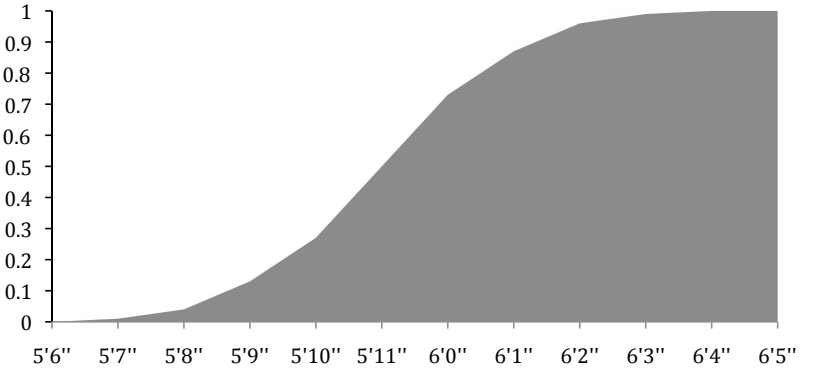


Fig. 1.

more intermediate cases while maintaining the shape of the curve. The representation of *tall* (for our agent A) is a function which yields the values considered here at intervals of $1''$. In the limiting case in which we consider a dense scale of heights, the curve will be smooth. Essentially, I am suggesting that the distinguishing characteristic of vague predicates like *tall* is that their lexical representations are described by CONTINUOUS PROBABILITY FUNCTIONS. Non-vague terms, then, will be those whose representations are (in the relevant regions) given by discontinuous probability functions.

4 The Sorites Paradox

The original motivation for our discussion was the sorites paradox, which is restated, substituting **tall** for **heap**, in (8). As above, x_1 is $5'6''$, and x_{12} is $6'5''$. Again, boldface indicates model-theoretic objects while italics indicate utterance-types. The use of boldface **tall** in (8) makes explicit the implicit assumption in the original statement of the sorites paradox that the words in question denote unique model-theoretic objects.

(8) a. $\neg\mathbf{tall}(x_1)$
b. $\forall n[\neg\mathbf{tall}(x_n) \rightarrow \neg\mathbf{tall}(x_{n+1})]$
c. $\therefore\ \forall n[\neg\mathbf{tall}(x_n)]$

Intuitively, premise (8-a) is clearly true – someone who is $5'6''$ is not tall. But the conclusion that no one is tall is clearly false. Our only hope is to deny (8-b); but this is strange, since this premise is intuitively very plausible.

Within the probabilistic theory of linguistic representation that I have sketched, the paradox in (8) could be restated in two ways. Suppose first that we consider the intended interpretation of **tall** as a model-theoretic object. **Tall** in itself does not represent any object in the theory proposed here: a subscript is needed to indicate which possible language $\mathbf{tall}_n$ is intended to belong to. So, if *tall* is assigned a value in l_{34}, then $l_{34}(tall)$ is written $\mathbf{tall}_{34}$. The paradox now appears as in (9):

(9) a. $\neg\mathbf{tall_{34}}(x_1)$
b. $\forall n[\neg\mathbf{tall_{34}}(x_n) \rightarrow \neg\mathbf{tall_{34}}(x_{n+1})]$
c. $\therefore\ \forall n[\neg\mathbf{tall_{34}}(x_n)]$

But since possible languages are perfectly precise, the inductive premise is plainly false: whatever l_{34} is, there must be some sharp boundary between the extension of $\mathbf{tall_{34}}$ and that of $\neg(\mathbf{tall_{34}})$, and so the conclusion does not follow.

Suppose now we rewrite the paradox using words in place of model-theoretic objects. *Tall* is a word, and it is interpreted by some possible language l_n as a semantic object $\mathbf{tall_n}$. Crucially, *tall* is not in itself a model-theoretic object. If we attempt to restate the paradox using *tall* in place of **tall**, we get (10).

(10) a. $\neg tall(x_1)$
b. $\forall n[\neg tall(x_n) \rightarrow \neg tall(x_{n+1})]$
c. $\therefore\ \forall n[\neg tall(x_n)]$

None of the clauses in (10) are well-formed within the theory I have introduced. I have occasionally spoken of "the probability that u applies to **o** (according to A)", but this was explicitly introduced as an abbreviatory convention. The bare claim that x_n is *tall* is meaningless: we can only say that *tall* applies to x_n with some probability d.

Suppose we rewrite the paradox using probabilities, as the present approach demands. It is plausible within a probabilistic theory (though not quite right, as we will see) to suppose that "x_n is not tall" is expressed as "$prob_A(tall(x_n)) = 0$". If we accept this, the restatement of the sorites paradox is:

(11) a. $prob_A(tall(x_1)) = 0$
b. $\forall n[prob_A(tall(x_n)) = 0 \rightarrow prob_A(tall(x_{n+1})) = 0]$
c. $\therefore$ $\forall n[prob_A(tall(x_n)) = 0]$

(11) is logically valid, but premise (11-b) is much less compelling than the original inductive premise (8-b) seemed to be. There is simply no reason to assume that, if the probability that something is *tall* is 0, the probability that an adjacent item is *tall* must also be 0 (rather than some small but non-zero amount).[3]

Much more plausible is the probabilistic version of the existential variant of the inductive premise, "$\neg\exists n[\neg tall(x_n) \wedge tall(x_{n+1})]$". A reasonable translation is:

(12) $\neg\exists n[prob_A(tall(x_n)) = 0 \wedge prob_A(tall(x_{n+1})) = 1]$

But (12) is not equivalent to (11-b) in the current theory: denying that there is a point at which the probability function jumps from 0 to 1 is not the same as denying that it ever increases from 0. (12) is true of *tall* and, suitably modified, any other vague predicate, but this creates no problem: it is, if anything, just a necessary condition for a predicate's being vague.

Delia Graff Fara has claimed that a convincing theory of the sorites, if it denies the validity of the inductive premise, must answer three separate questions (slightly modified from [5]).

[3] An influential objection to supervaluational treatments of vagueness is that, even though there is no sharp boundary between the positive extension and the negative extension of a vague predicate, the sentence "There is a sharp boundary between the positive extension and the negative extension of ϕ" is supertrue for all predicates ϕ, whether or not they contain vague terms. Eytan Zweig points out that one could construct an analogue to the supervaluationist's problem for my theory: the sentence "there is an n such that $prob_A(u(x_n)) = 0$ and $prob_A(u(x_{n+1})) \neq 0$" has probability 1 whenever, for all l_n such that $\mu_L(l_n) \neq 0$, $l_n(u)$ is defined. One possible approach is simply to say that people do not have reliable intuitions about infinitesimal differences in probability, so that the validity of this statement is not problematic. An alternative would be to deny that probabilities every really reach 0 or 1 except for logical contradictions and validities, respectively: so, for example, the probability that a 2-foot-tall person counts as tall is infinitesimal, but not zero, and the probability that a nine-foot-tall person is tall is extremely close, but not quite equal, to 1. See the following section for more discussion of the second approach.

1. **The Semantic Question:** If the inductive premise is not true, then must the classical equivalent of its negation, the "sharp boundaries" claim, be true?

   The "sharp boundaries" claim: $\exists n[\neg tall(x_n) \wedge tall(x_{n+1})]$

   (a) If the sharp boundaries claim *is* true, how is its truth compatible with the fact that vague predicates have borderline cases? For the sharp boundaries claim seems to deny just that.
   (b) If the sharp boundaries claim is *not* true, then given that a classical equivalent of its negation is not true either, what revision of classical logic and semantics must be made to accommodate that fact?
2. **The Epistemological Question:** If "$\forall n[\neg tall(x_n) \rightarrow \neg tall(x_{n+1})]$" is not true, why are we unable to say which one (or more) of its instances is not true – even when all the heights of the possible values of "x_n" are known?
3. **The Psychological Question:** If the universal variant of the inductive premise is not true, why were we so inclined to accept it in the first place? In other words, what is it about vague predicates that makes them seem tolerant, and hence boundaryless to us?

Let's address these questions in turn.

1. **The Semantic Question**
   (a) If we replace **tall** in Fara's formulation of the "sharp boundaries" claim by a model-theoretic object acceptable in our system such as $\mathbf{tall}_n$, the claim is true. This is not problematic because our original intuition that the sharp boundaries claim is false, and that the universal sorites premise is true, was not an intuition about some model-theoretic object $\mathbf{tall}_n$ but an intuition about the meaning of the word (utterance-type) *tall*.
   (b) If we replace **tall** in Fara's formulation of the "sharp boundaries" claim by a word such as *tall*, making appropriate adjustments, the "sharp boundaries" claim is false. However, no revision of classical logic and semantics is required to explain these facts; rather, the falsity of the sharp boundaries claim follows from the fact that the word *tall* does not denote a unique object, but denotes various objects with differing probabilities. The semantic metalanguage is nevertheless classical.
2. **The epistemological question.** "$\exists n[\neg \mathbf{tall}(x_n) \wedge \mathbf{tall}(x_{n+1})]$" is not well-formed in the present theory. If we substitute $\mathbf{tall_m}$, as in "$\exists n[\neg \mathbf{tall_m}(x_n) \wedge \mathbf{tall_m}(x_{n+1})]$" we can identify which n satisfies this formula given a complete specification of the language l_n or of the extension of $\mathbf{tall_m}$. If we consider the sharp boundaries claim substituting the word *tall*, our language does not permit us to ask which n makes "$\exists n[\neg tall(x_n) \wedge tall(x_{n+1})]$" true, because this sentence is not well-formed. But the probabilistic version of this formula – $\exists n[prob(tall(x_n)) = 0 \wedge prob(tall(x_{n+1})) = 1]$ – is plainly false, an intuitively correct result.
3. **The psychological question.** I suggest that we are inclined to accept the inductive premise because we interpret it as a claim about words/utterances

rather than about model-theoretic objects (which are probably not accessible to introspection anyway, like most grammatical objects). Speakers know that, given a pair of very similar objects, vague words like *tall* will not apply to one with probability 0 and to the other with probability 1. So it is almost always safe to assume that if an individual counts as *tall* in some context then an adjacent individual will also count as *tall* in the same context. However, informal deductions involving vague terms of natural language are not reliably truth-preserving because the terms are not associated with a unique model-theoretic interpretation.

5 Some Loose Ends

5.1 Assertibility, Joint Distributions, and Borderline Cases

When is the sentence "x is tall" assertible? In the previous subsection we supposed that it is when $prob$("x is tall") $= 1$. But this cannot be quite right: if the meaning of "tall" is a continuous random variable, then the cumulative probability approaches 1 in the limit as height goes to infinity, but never reaches 1 at finite height.

With respect to factual beliefs, it is commonly supposed that the norm of assertion is knowledge [36]. If this is correct, then we should expect, as a descriptive matter, that a cooperative speaker A will typically assert things that she *thinks* she knows. On standard assumptions, A's subjective probability for a proposition p will rarely if ever reach 1; but p may have high enough probability that A thinks that she may profitably make an assertion calculated to communicate the information that p. How high is judged "high enough" will depend on various features of the context, such as the conversational stakes and perhaps even aspects of the speaker's personality. I will use α as a placeholder for the threshold of assertibility, however this is determined in particular contexts. Cooperative speakers will assert a proposition p only if the probability of p is greater than α, and the information that p is deemed relevant, useful, etc.

The extension to vague terms is straightforward: in cases where the height of an individual x is known, and a speaker wants to decide whether to describe him as "tall", she can simply compute whether or not $prob(tall(x)) > \alpha$; if so, "x is tall" is assertible. We have already seen how this would be done in cases in which the individual's height is known with precision. In cases where there is both linguistic and factual uncertainty, the speaker can compute the joint probability distribution for the two random variables in question – x's height h, and the probability that someone of height h counts as "tall". In the finite case, this is just the average of the probability that x counts as "tall" in various worlds, weighted by the probability that these worlds are the actual world.

$$prob(\text{“}x \text{ is tall”}) = \sum_{l \in L} \sum_{w \in W} [[\mu_L(l) : l(tall)(x)(w) = \text{True}] \times [\mu_W(w)]]$$

That is, using the abbreviatory convention defined above,

$$prob(\text{“}x \text{ is tall”}) = \sum_{w \in W} [prob(tall(x)(w)) \times \mu_W(w)]]$$

(Again, the infinite case is an straightforward extension.) So “x is tall’ is assertible just in case its probability is greater than α, taking into account both linguistic and factual uncertainty.

Since our metalanguage is classical, $prob(\text{“}x \text{ is not tall”}) = 1 - prob(\text{“}x \text{ is tall”})$. Since “$x$ is not tall” is assertible just in case $prob(\text{“}x \text{ is not tall”}) > \alpha$, it follows that “$x$ is not tall’ is assertible just in case $prob(\text{“}x \text{ is tall”}) < 1 - \alpha$. A characterization of borderline cases – those for which neither “x is tall” nor “x is not tall” is assertible – follows immediately:

A *borderline case* of F is an individual x for which $1 - \alpha < prob(\text{“}x \text{ is F”}) < \alpha$.

5.2 Common Ground

Since introducing the probabilistic model, we have dealt exclusively with subjective probabilities, avoiding the issue of common ground. But of course a probabilistic theory of assertion, interpretation, and shared belief is needed at some point. To give a fully explicit model of this type would be a major undertaking, but a few preliminary comments may be useful.

The probabilistic model seems to require a weaker notion of common ground than Stalnaker’s: in particular, common knowledge seems unattainable. We can, however, use a metric of shared belief: x and y share factual beliefs to the extent that their probability distributions over W overlap. If the ultimate goal of inquiry is to find out how the world is [29], then the ultimate goal in our model ought to be to assign probability 1 to the actual world. Of course this goal is unattainable, but we can approach it by gaining more and more information about the world, reducing uncertainty as much as possible.

In the case of linguistic beliefs, there is no “way the world is” to be discovered; however, each participant in a conversation has a belief set including a probability distribution over a set of possible languages, and they wish to share linguistic forms so that they can exchange information. One goal of metalinguistic inquiry is to make your personal μ_L overlap as much as possible with your interlocutors’ personal μ_L. However, this is not enough – after all, one way to achieve total overlap would be for all interlocutors to have a uniform distribution over L, but this distribution would be useless for communication. Another goal of metalinguistic inquiry, then, must be to assign most of the probability mass to a relatively small set of languages – i.e., to minimize linguistic uncertainty so that the comprehension process produces a manageable set of candidate interpretations. A general theory of how language users balance these needs would be useful, but is beyond the scope of the present paper. Decision-theoretic and game-theoretic models of language use and evolution seem to me to provide a promising starting point, though [19,21,22,2,23,11].

6 Comparison with Alternatives

6.1 Supervaluationism, Interest-Relativity, and Their Kin

As discussed in section 1, the theory of vagueness outlined here treats vagueness in terms of the relation between language(s) and language users. Other theories that share this feature are, for example, Fara [5] and Barker [1]. These approaches share with the present theory the fact that the interpretation of vague terms is influenced by the discourse context, whether in Stalnakerian fashion (Barker) or by the interests of the conversational participants (Fara). However, such approaches to vagueness – and that of Kennedy [15] – have two related drawbacks. First, they treat vagueness as a special property either of particular predicates (Barker) or of the positive form (Fara and Kennedy), rather than deriving it from more general principles. Second, these theories only offer an account of vague scalar adjectives. But vagueness extends far beyond scalar adjectives; indeed, it is surprisingly hard to find examples of natural language expressions which are **not** vague. The present theory, unlike these alternatives, predicts that vagueness, not precision, should be the norm in natural language.

The theory outlined is here is to some extent an elaboration of Lewis' suggestion in [20] that vagueness is a question of language choice, rather than an issue of semantics proper. In her discussion of Lewis' idea and Burns' [3] defense of this approach, Keefe [13] argues that Lewis' "pragmatic" theory is, for all practical purposes, just a restatement of the supervaluationist approach advocated by Keefe herself and by Kamp [12] and Fine [6]. It is undeniable that Lewis' suggestion, and the elaboration I have offered, look similar in broad outline to supervaluationism. However, there are important conceptual and empirical differences. First, the theory makes no appeal to specialized semantic concepts such as "supertruth". (Perhaps a rational agent is obliged to assign probability 1 to certain sentences, but such sentences have no special semantic status.) Second, supervaluation theory (in the relevant form) is essentially an ad hoc theory designed specially to account for vague terms. The approach advocated here has a conceptual advantage over supervaluation theory in that it does not rely on any special semantic mechanisms, either enrichments of the semantic metalanguage or otherwise unmotivated stipulations about the lexical properties of particular words. Instead, the formal treatment of linguistic representations is maximally similar to the treatment of factual beliefs, and inherits independently motivated properties of this theory.

The current theory is also preferable to supervaluationism in at least one empirical respect: supervaluation theory founders on the issue of higher-order vagueness. That is, the theory predicts that *John is tall* is vague, but *John is definitely tall* is not vague. This is clearly incorrect: it is just as easy to construct a sorites series for *definitely tall* as it is for *tall*. Keefe [13] responds to this criticism by appealing to a vague metalanguage, so that the vagueness of *definitely tall* resides in the vagueness of the notion of an "acceptable precisification". But as Williamson [34] notes, pushing vagueness back from the object-language to the meta-language is not a satisfactory solution to the problem. I will show in

section 7 that the present theory can deal with higher-order vagueness without resorting to a vague metalanguage.

6.2 Fuzzy Logic

As noted above, a natural interpretation of the probabilistic apparatus argued for here is in terms of degrees of belief, usually treated in terms of subjective probabilities when factual beliefs are at stake. However, Schiffer [26] and Smith [27] have argued in various ways that, although we are correct to treat borderline status as involving an intermediate degree of belief that a term applies to an object, vagueness-related degrees of belief do not behave like subjective probabilities but like degrees of truth in fuzzy logic.

Schiffer's argument is this: if the probability of p is 0.5, the probability of q is 0.5, and p and q are independent, then the probability of $p \wedge q$ is necessarily 0.25. However, Schiffer [26, p. 225] claims that

> when Sally believes to degree .5 that Tom is bald, thereby making him a paradigm borderline case of baldness for her, she also believes to degree .5 that he is thin, making him also for her a paradigm borderline case of thinness ... Can we expect eminently rational Sally to believe to degree .25 that Tom is bald and thin? I submit not. I submit that she'll believe the conjunction to degree .5.

If Schiffer is correct, then the present theory is indeed on the wrong track, and fuzzy logic does better (since in fuzzy logic the degree of truth of a conjunction is the minimum of the degrees of truth of the conjuncts: $\mathbf{min}(0.5,0.5) = 0.5$).

Clashes of intuition probably will not take us far, but for what it is worth, I do not share Schiffer's intuition about this scenario. Whatever the status of Schiffer's claim, though, his proposal has a consequence that is much more counter-intuitive than the problem it is meant to solve. If we embrace fuzzy logic for vague terms we predict that the degree to which Tom is bald AND thin is $\mathbf{min}(0.5,0.5) = 0.5$, the same as the degree to which he is either bald OR thin ($\mathbf{max}(0.5,0.5) = 0.5$). This is strange: surely it is more likely that he is one or the other than that he is both. In contrast, the probabilistic theory predicts, I believe correctly, that *prob*(Tom is bald and thin) $\leq$ *prob*(Tom is bald) $\leq$ *prob*(Tom is bald or thin).

Smith [27] gives another argument against a probabilistic theory based on a story along the following lines. Suppose your long-lost brother is coming to visit, and you know that he is either very tall or very short, but you do not know which. This situation is qualitatively different from one in which you know that your brother is a borderline case of *tall*, and yet the probabilistic theory seems to collapse the two: in both cases your subjective probability that he is tall should be roughly 0.5. Thus, Smith concludes, a theory which treats linguistic and non-linguistic uncertainty using separate machinery is preferable.

The argument is interesting, but if it is interpreted as Smith wishes, it proves too much. Using similar reasoning we could show, without bringing in issues relating to vagueness, that many partial beliefs about factual issues do not behave

like subjective probabilities either. For example, suppose in a contest I will win $1 million if I pick the winning team in a sports contest between teams A and B. Consider two cases. In the first case, I know nothing about the teams and have no basis for choosing. In the second case I have seen these teams play each other hundreds of times and I know that they are evenly matched: each team has won precisely 50% of the games I have watched. In both cases, according to the standard theory, it is rational for me to assign probability 0.5 to the proposition that Team A will win and 0.5 to the proposition that Team B will win. However, it is clear that these two situations are qualitatively different. So, the argument would go, a theory which treats uncertainty due to ignorance and uncertainty due to statistical knowledge using completely different theoretical machinery is preferable. (Presumably fuzzy logic would not be a candidate in this case.)

I actually agree with Smith's claim that linguistic and factual uncertainty should be treated differently, at least in part: this is why we separated world and language components of belief-sets. But the problem that Smith brings up is a very general issue for the representation of uncertainty. There are numerous proposals for how to enrich probabilistic representations to deal with examples like the one just given, e.g. ranges of probabilities or upper and lower probabilities; see Halpern [9, ch.2] for an overview. However this issue is best dealt with technically, the undeniable fact that there is a qualitative difference between the two types of uncertainty in Smith's example does not show that metalinguistic uncertainty cannot be modeled using subjective probability. At most it shows that, in an enriched probabilistic model using probability ranges or the like, metalinguistic uncertainty about borderline cases will be more similar to factual uncertainty stemming from high-quality statistical information than to uncertainty stemming from having several distinct and very different options.

Another argument which favors the probabilistic approach over fuzzy logic is noted by Edgington [4, p. 305]. Imagining that d and e are two objects that are both borderline cases of being red (R),

> [L]et $val(Re) = 0.5$ and $val(Rd)$ be a little less than 0.5, say 0.4. What is $val(Rd \,\&\, Re)$? Here [fuzzy logic] gives a plausible answer: 0.4, the minimum of the two. But note: $val(\neg Re)$ is also 0.5 ... [and so] $val(Rd \,\&\, \neg Re)$ is also 0.4. This is immensely implausible. e is redder than d. How could it be other than completely wrong, in any circumstance, to say "d is red and e is not"? $val(Rd \,\&\, \neg Re)$ should be zero.

This is indeed implausible. The probabilistic theory gets this case right, however: because of the upward monotonicity of gradable adjectives, any resolution of red which makes "d is red" true will make "e is red" true as well, and so the probability that d is red and e is not is 0. This is, as Edgington also concludes, a strong argument for the probabilistic approach over a theory based on fuzzy logic.

Finally, we can note that, as section 7 will explore in more detail, English (like many other languages) uses the same modal adverbs to express a high degree of certainty in the truth of non-vague propositions and a high degree of certainty that a vague predicate applies to an object. We can explicate pairs like *John will clearly/definitely leave tomorrow* and *John is clearly/definitely tall*

without treating these operators as ambiguous if we simply assume that they place conditions on the probability of the expressions they modify. However, if linguistic and non-linguistic uncertainty are as fundamentally different as Schiffer and Smith argue, then we must treat these operators as ambiguous, and the fact that they occur in both contexts as a semantic accident.

7 Higher-Order Vagueness and Metalinguistic Modality

The brief discussion of higher-order vagueness above points to an important fact: *definitely* is vague, no less than *tall*. I suggest treating *definitely* as we did *tall* above, using a probability distribution over possible precise resolutions. However, *definitely* is different in that it is a modal operator. This is clear from its double life as an epistemic modal: *John is definitely coming* does not tell us about the meaning of "coming", but about the likelihood of John doing so. In epistemic logic *John is definitely coming* is usually taken to mean something like the following: In all of the worlds which the speaker considers live options for being the actual world, John is coming. Since the model we have adopted treats linguistic and non-linguistic beliefs similarly, we might think to define an epistemic metalinguistic modal in similar terms: *John is definitely tall* means that, in each of the languages that the speaker considers live options for being the language of conversation, John is in the extension of the interpretation of *tall* in that language.

However, since *definitely* is vague, we cannot treat it as a universal quantifier over accessible worlds. The universal quantifier is not vague, and indeed the assumption that *definitely* corresponds to a strong modal in Kripke semantics was the source of the problem of higher-order vagueness in the first place. For a standard use of epistemic *definitely* we can try instead: *John is definitely coming* is true iff the probability that John is coming exceeds some high threshold. I suggest a unified treatment of epistemic and metalinguistic *definitely*: *definitely* u establishes a minimum threshold for the probability of u, cashed out either as the probability of applicability of u to an individual (if u is resolved as a type $\langle e, t\rangle$ expression) or the utterance's truth (if u is resolved as an expression of type t).

Definitely takes a constituent as an argument and returns a function which quantifies over the meaning of that constituent in all relevant possible languages. Since *definitely* is vague, there will be a range of resolutions available with different thresholds, just as there were for *tall*. One possible resolution of *definitely tall* – call it $\mathbf{definitely}_6\ \mathbf{tall}_6$ – is given in (13):

(13) $[\mathbf{definitely_6 tall_6}](x)$ is true iff $prob_A(tall(x_n)) \geq 0.97$

Definitely tall as interpreted by a language l_7 is true of x iff the sum of the probabilities of all languages that make *tall* true of x is greater than some threshold, here 97%.

On this interpretation, *definitely* is effectively a special type of epistemic modal. (Note that, just as the actual world determines the truth-value of an

epistemic modal only indirectly, so the resolution of *tall* in l_n plays no privileged role in calculating the meaning of *definitely tall* in l_n.) We may suppose that the metalinguistic use of *definitely* is equivalent to the ordinary epistemic use, except that in the latter case the linguistic facts are held constant while the non-linguistic facts vary, while in the metalinguistic case the non-linguistic facts are held constant while the linguistic facts vary. Thus, the simplest extension of the theory proposed here predicts that *definitely* should be vague. It does seem that *definitely* is at least gradable, in that *John will very definitely come* is acceptable, while *very* is unacceptable with non-gradable adjectives and adverbs like *geological* [15]. A detailed comparison of metalinguistic and epistemic *definitely* will have to wait for future work; but see Sauerland & Stateva [24] for a detailed consideration of some closely related issues.

Imagine that the common ground contains seven resolutions of *definitely* with the probability distribution in Table 3. Applying (13) (with appropriate replacements) and the probabilities in Table 3 to the probability distribution for *tall* in Table 1 we get Table 4.

Table 3. Sample probability distribution for *definitely*

Name	$\mathbf{def}_1$	$\mathbf{def}_2$	$\mathbf{def}_3$	$\mathbf{def}_4$	$\mathbf{def}_5$	$\mathbf{def}_6$	$\mathbf{def}_7$
Threshold	≥ 0.82	≥ 0.85	≥ 0.88	≥ 0.91	≥ 0.94	≥ 0.97	$= 1$
$prob_A(definitely = \mathbf{definitely_n})$	0	.05	.1	.15	.2	.3	.2

Table 4. Cumulative probability distribution for *definitely tall*

Individual	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$
Height	5′6″	5′7″	5′8″	5′9″	5′10″	5′11″	6′0″	6′1″	6′2″	6′3″	6′4″	6′5″
$prob_A(tall(x_n))$	0	.01	.04	.13	.27	.5	.73	.87	.96	.99	.1	1
$prob_A([definitely\ tall](x_n))$	0	0	0	0	0	0	0	.05	.5	.8	1	1

In general, *definitely* u will always have probability less than or equal to the probability of u alone, which is the correct result. Note also that iterated *definitely* is also unproblematic, and simply lowers the probability further as compared to u alone.[4]

[4] A reviewer asks: Can this approach account for the possibility that x is neither definitely definitely tall nor definitely not definitely tall? It can. Let $\alpha = .8$ and *definitely* and *tall* be resolved as in the text. x_{11} and x_{12} count as *definitely tall* because *tall* applies to them with probability greater than α. By reapplying Table 3 to the third line of Table 4 we see that x_{12} counts as *definitely definitely tall*, but x_{11} does not. x_9 (and all shorter individuals) count as *not definitely tall*, since $prob(not\ definitely\ tall(x_9)) = 1 - prob(definitely\ tall(x_9)) = 1 - .05 = .95$. However, x_9 does not count as *definitely not definitely tall* ($prob = 0.5$); only x_8 and shorter enjoy this privilege. So x_9 is neither definitely definitely tall nor definitely not definitely tall.

To sum up, the theory advocated here offers the promise of an explanation of higher-order vagueness with only a slight modification of existing theories of modality. This is a considerable advantage over supervaluationist treatments, which must assume that the semantic metalanguage itself is vague (as in Keefe 2000). Note also that, if this approach is viable, it constitutes a counter-example to Grim's [8] claim, quoted above, that vague terms cannot not be modeled accurately in a precise metalanguage.

8 Conclusion

The theory of vagueness described here has an important advantage over many competitors: it stipulates no special semantic apparatus for vague terms, as does supervaluation theory. Nor does it rely crucially on the claim that the meanings of words float free of speakers' knowledge of language, as does epistemicism. Rather, the probabilistic account relies on general and independently motivated properties of language use and human cognition. This account also yields an account of the sorites paradox which is explanatory with respect to Fara's three questions. Empirical advantages of the present approach over competing theories include its ability to account for vagueness outside of the realm of gradable adjectives, intuitively correct results with conjunctions and disjunctions of sentences containing vague terms, and avoidance of problems with higher-order vagueness.

References

1. Barker, C.: The dynamics of vagueness. Linguistics and Philosophy 25(1), 1–36 (2002)
2. Benz, A., Jäger, G., Van Rooij, R.: Game theory and Pragmatics. Palgrave Macmillan, Oxford (2005)
3. Burns, L.: Vagueness: An Investigation into Natural Languages and the Sorites Paradox. Kluwer, Dordrecht (1991)
4. Edgington, D.: Vagueness by degrees. In: Keefe, R., Smith, P. (eds.) Vagueness: A Reader, pp. 294–316. MIT Press, Cambridge (1997)
5. Fara, D.G.: Shifting sands: An interest-relative theory of vagueness. Philosophical Topics 20, 45–81 (2000)
6. Fine, K.: Vagueness, truth and logic. Synthese 30(3), 265–300 (1975)
7. Frazee, J., Beaver, D.: Vagueness is rational under uncertainty. In: Aloni, M., Bastiaanse, H., de Jager, T., Schulz, K. (eds.) Logic, Language and Meaning. LNCS, vol. 6042, pp. 153–162. Springer, Heidelberg (2010)
8. Grim, P.: The buried quantifier: An account of vagueness and the sorites. Analysis 65(2), 95 (2005)
9. Halpern, J.: Reasoning about Uncertainty. MIT Press, Cambridge (2003)
10. Heim, I.: Degree operators and scope. In: Fery, Sternefeld (eds.) Audiatur Vox Sapientiae: A Festschrift for Arnim von Stechow. Akademie Verlag, Berlin (2001)
11. Jäger, G., van Rooij, R.: Language structure: psychological and social constraints. Synthese 159(1), 99–130 (2007)
12. Kamp, H.: Two theories about adjectives. In: Keenan, E. (ed.) Formal Semantics of Natural Language, pp. 123–155. Cambridge University Press, Cambridge (1975)

13. Keefe, R.: Theories of Vagueness. Cambridge University Press, Cambridge (2000)
14. Kennedy, C.: Projecting the adjective: The syntax and semantics of gradability and comparison. PhD thesis, U.C., Santa Cruz (1997)
15. Kennedy, C.: Vagueness and grammar: The semantics of relative and absolute gradable adjectives. Linguistics and Philosophy 30(1), 1–45 (2007)
16. Kennedy, C., McNally, L.: Scale structure, degree modification, and the semantics of gradable predicates. Language 81(2), 345–381 (2005)
17. Klein, E.: A semantics for positive and comparative adjectives. Linguistics and Philosophy 4(1), 1–45 (1980)
18. Lassiter, D.: Gradable Epistemic Modals, Probability, and Scale Structure. To appear in Proceedings of Semantics and Linguistic Theory XX (2010)
19. Lewis, D.: Convention: A Philosophical Study. Harvard University Press, Cambridge (1969)
20. Lewis, D.: General semantics. Synthese 22(1), 18–67 (1970)
21. Merin, A.: Information, relevance, and social decisionmaking: Some principles and results of decision-theoretic semantics. In: Moss, L., Ginzburg, J., de Rijke, M. (eds.) Logic, Language, and Computation, vol. 2, pp. 179–221. CSLI Publications, Stanford (1999)
22. Parikh, P.: The Use of Language. CSLI Publications, Stanford (2001)
23. Pietarinen, A.: Game theory and linguistic meaning. Elsevier, Amsterdam (2007)
24. Sauerland, U., Stateva, P.: Scalar vs. epistemic vagueness: Evidence from approximators. In: Gibson, M., Friedman, T. (eds.) Proceedings of Semantics and Linguistic Theory XVII. CLC Publications, Cornell University (2007)
25. Schelling, T.: The Strategy of Conflict. Harvard University Press, Stanford (1960)
26. Schiffer, S.: Vagueness and partial belief. Philosophical Issues 10, 220–257 (2000)
27. Smith, N.: Degree of belief is expected truth value. In: Dietz, R., Moruzzi, S. (eds.) Cuts and Clouds: Vaguenesss, Its Nature and Its Logic. Oxford University Press, Oxford (2010)
28. Stalnaker, R.: Assertion. In: Cole, P. (ed.) Syntax and Semantics: Pragmatics, vol. 9. Academic Press, London (1978)
29. Stalnaker, R.: Inquiry. MIT Press, Cambridge (1984)
30. Stalnaker, R.: On the representation of context. Journal of Logic, Language and Information 7(1), 3–19 (1998)
31. Stalnaker, R.: Assertion revisited: On the interpretation of two-dimensional modal semantics. Philosophical Studies 118(1), 299–322 (2004)
32. Swanson, E.: Interactions With Context. PhD thesis, MIT (2006)
33. von Stechow, A.: Comparing semantic theories of comparison. Journal of Semantics 3(1), 1–77 (1984)
34. Williamson, T.: Vagueness. Routledge, New York (1996)
35. Williamson, T.: Schiffer on the epistemic theory of vagueness. Noûs 33, 505–517 (1999)
36. Williamson, T.: Knowledge and its Limits. Oxford University Press, Oxford (2000)
37. Yalcin, S.: Epistemic modals. Mind 116(464), 983–1026 (2007)
38. Yalcin, S.: Probability Operators. To appear in Philosophy Compass (2010)

The Relative Role of Property Type and Scale Structure in Explaining the Behavior of Gradable Adjectives

Louise McNally

Universitat Pompeu Fabra

Abstract. Kennedy [9] proposes a semantics for positive form adjectives on which the standard for ascribing an adjective A makes the individuals that are A stand out from those that are not. To account for the differences between absolute and relative adjectives, Kennedy posits that the maximal and minimal degrees on closed scales naturally make individuals stand out in a way that degrees found away from the endpoints of a scale cannot. I argue that the ability of a degree to make individuals stand out is due less to scale structure than to the nature of the property the adjective describes. Thus, degrees that are not endpoints can behave like absolute standards as long as the application criteria for the property are clear. I relate the identifiability of such criteria to whether the property ascription can be modeled in terms of rule- vs. similarity-based classification (see e.g. [5]).

Keywords: semantics, adjectives, gradability, Sorites paradox, comparison class, classification, vagueness.

1 Introduction

There is a substantial literature on the semantics of gradability that distinguishes so-called RELATIVE adjectives such as *tall* from ABSOLUTE adjectives such as *closed* [14,8]:

(1) a. Marta is tall.
b. The door is closed.

Kennedy [9] identifies three important differences between these two classes (see [12] for additional relevant discussion). First, the truth of sentences such as (1-a) is context-dependent while that of (1-b) is not. This can be seen in the fact the addition of a modifier that makes a standard of comparison or a comparison class explicit can render the former either possibly true (as in (2-a-b)) or almost certainly false (as in (2-c-d)), while such an addition is not even felicitous with the latter (see (3)), indicating that a contextually-specified standard or comparison class is incompatible with the adjective.

R. Nouwen et al. (Eds.): ViC 2009, LNAI 6517, pp. 151–168, 2011.

(2) a. Compared to her friend Andrea, Marta is tall.
b. Marta is tall for an 11-year-old.
c. Compared to Michael Jordan, Marta is tall.
d. Marta is tall for a professional basketball player.

(3) a. ?? Compared to Door #1, Door #2 is closed.
b. ?? That box of cookies is closed for a box my daughter has gotten into.

Second, relative adjectives generally fail to yield crisp judgments about truth, while absolute adjectives do. As a result, the former easily give rise to the Sorites paradox, exemplified in (4). For example, if the difference in height between Marta and Andrea is very small, it is unlikely that we will judge one to be tall and the other not. More generally, assuming Premise 1 in (4-a) and given Premise 2 in (4-b), the general form of which is widely held to be valid for relative adjectives, the absurd conclusion in (4-c) will result.

(4) a. Premise 1: A 1.65-meter-tall 11-year-old is tall (for an 11-year-old).
b. Premise 2: If x is a tall 11-year-old and y is an 11-year-old who is 1 millimeter shorter than x, then y is a tall 11-year-old.
c. Conclusion: A 1.05-meter-tall 11-year old is a tall 11-year-old.

In contrast, if we open a door which is closed even the smallest amount, we can easily determine that the door will no longer be closed, and, correspondingly, if we try to recreate the paradox in (4) with *closed*, Premise 2 will fail, as Kennedy argues it fails in general for absolute adjectives.

Finally, with relative adjectives we can easily find borderline cases for which it is difficult or even impossible to decide whether the adjective truthfully holds or not. For example, if we are discussing the height of a group of children of the same age and we limit ourselves to consideration of this group, ignoring general knowledge we might have about the height of children of that age, we may have trouble judging whether a 1.50-meter-tall child is tall, if the height of the children in the group ranges from, say, 1.30 to 1.70 meters. However, no such difficulty arises with absolute adjectives like *closed*.

The contrasts between relative and absolute adjectives have two sorts of implications for a general account of the semantics of adjectives. First, as will be discussed below, the contrasting behavior of the two kinds of adjectives is unexpected on a degreeless semantics of the positive form such as that proposed in [10] (though see [15] for a proposed solution); in contrast, it seems less difficult to account for if we assume a semantics for adjectives that includes degrees. Second, if we assume a degree-based semantics of adjectives, these contrasts make it difficult to provide a unique, general characterization of the truth conditions for the positive form. Kennedy [9] addresses the latter question, arguing that a unified (degree-based) semantics for the positive form is possible if we take into account the basic difference in the kinds of standard values for truthful application that are associated with each kind of adjective and properly exploit a principle of Interpretive Economy that he proposes.

The goal of this paper it to argue that, while the basic intuition behind the semantics Kennedy proposes for the positive form seems correct, the details of the analysis assign too great a role to the abstract gradability properties of the adjectives in question. I will suggest that by focusing instead on the nature of the properties that adjectives contribute, the role that adjectives play in classifying individuals according to their manifestation of these properties, and the strategies for classification that may be involved, we can arrive at a better characterization of the relative/absolute distinction.

The structure of the paper is as follows. Section 2 presents Kennedy's account of the absolute/relative contrasts (hereafter, the Interpretive Economy account) and some challenges to it. Section 3 relates the data discussed in Section 2 to two different classification strategies: classification by similarity, and classification by rule (see e.g. [5]). Finally, Section 4 discusses the implications for the semantics of positive form adjectives.

2 The Interpretive Economy Account and Some Challenges

The Interpretive Economy account of relative/absolute contrasts builds on an analysis of adjectives as measure functions (see e.g. [7]) and on the typology of scale structure developed in [8]. On this analysis, all gradable adjectives will denote functions from entities to degrees; specific examples appear in (5).

(5) a. $\textbf{tall}(\textbf{Marta}) = 1.65$ meters
b. $\textbf{closed}(\iota x.\textbf{door}(x)) = 0$ degrees

Kennedy and McNally argue that various linguistic phenomena are sensitive to whether the scale associated with the adjective is CLOSED, i.e. whether there are maximal or minimal values in the codomain of the measure function, or OPEN, i.e. whether there are no such values. *Tall* is an example of an open scale adjective, as in principle there is no maximal or minimal value on the height scale.[1] In contrast, *closed* is a closed scale adjective.

The denotation of an adjective is converted from a measure function into a property that can be predicated of an individual via degree morphology, which introduces a STANDARD VALUE that determines whether the property truthfully applies to its argument or not. Kennedy and McNally argue that the standard value for the truthful applicability of a gradable predicate is, like the scale itself, also subject to linguistically relevant parameterization: Specifically, it can be relative, i.e. determined contextually (typically with respect to a comparison class), or absolute, i.e. fixed at a particular value. Relative and absolute adjectives are so called because they have relative and absolute standards, respectively. The fact that an adjective like *tall* admits the expression of a comparison class, as in *tall for an 11-year-old*, is evidence that its standard is relative. In the case

[1] See [8] for justification of the perhaps counterintuitive claim that there is no minimal value on the height scale.

of absolute adjectives, Kennedy and McNally assume that there are only two possible standards: either the maximal or the minimal non-zero value on the scale in question. The standard is maximal just in case truthful application of the adjectival predicate entails having the property in question to a maximal degree. For example, when we try to assert that a closed door has less than the highest degree of non-aperture, we derive a contradiction, and thus can conclude that the standard for closedness is maximal (see (6-a)). In contrast, a standard is minimal just in case the truthful application of the adjectival predicate only requires having the property in question to the smallest possible degree, and denying that the adjective applies, as in (6-b), is incompatible with having any degree of the property in question.

(6) a. #The door is closed, but it's slightly ajar.
b. #Montjüic is not visible from my rooftop, but I can see a tiny part of it.

Obviously, if the standard is either maximal or minimal, adding information about a comparison class or compared individual will have no effect on interpretation, and thus *for*- and *compared to*-phrases are infelicitous with absolute adjectives.

There is a strong correlation between scale type and standard value: If absolute standards must be either maximal or minimal values on a scale, it will be impossible for an open scale adjective to have an absolute standard, since by definition such scales lack maximal and minimal values. Closed scale adjectives could in principle have either absolute or relative standards, but there seems to be a strong tendency for them to have absolute standards. Just how strong this tendency is will prove to be a crucial question.

Kennedy [9, pp. 17-18], building on earlier work, posits a null degree morpheme *pos* to convert the adjective into the positive form of a gradable predicate of individuals and assigns it the semantics in (7), where g is a variable over measure functions (of type $\langle e, d\rangle$) and '$\mathbf{s}$ is a context-sensitive function that chooses a standard of comparison in such a way as to ensure that the objects that the positive form is true of "stand out" in the context of utterance, relative to the kind of measurement that the adjective encodes.[2]

(7) $pos : \lambda g \lambda x.g(x) \succeq \mathbf{s}(g)$

But what degree allows the objects that the positive form is true of to stand out? The answer to this question must be different for relative and absolute adjectives. In the former case, Kennedy argues that the value $\mathbf{s}$ returns will depend on the context. If, in a given context, the set of individuals under consideration is such that none of them stand out with respect to any of the others in the degree to which they possess the property in question, no appropriate standard value

[2] The semantics in (7) differs from that in e.g. [8] in that it is uniform for all gradable adjectives; earlier formulations distinguished between different types of *pos* for adjectives with relative vs. absolute standards.

can be chosen to differentiate among those individuals. This, Kennedy argues, is what underlies the intuitive validity of the second premise of the Sorites paradox and the failure of relative adjectives to yield crisp judgments. Specifically, when this premise is presented and evaluated, the context is typically restricted so as to involve the comparison of just two individuals. If the difference in the degree to which they manifest the property in question is very small, the reasoning goes, that difference will not be sufficient to make one stand out with respect to the other, and thus, in the absence of any additional information, if one of the individuals is considered to have that property, the other will be as well.

In the case of absolute adjectives, we have seen that the degree that stands out and thus constitutes the standard is not determined by context. Kennedy suggests that the difference between a property holding to no degree vs. to a minimal degree, and between one holding to a non-maximal vs. a maximal degree, is sufficiently salient to make maximal and minimal degrees stand out and constitute possible standard values, even though the difference in degree is just as small as in the cases that give rise to the Sorites paradox. The difference, he suggests [9, section 4.1.], is that the transition from the absence of a property to a minimal degree of its manifestation, or from a near-maximal to maximal degree of manifestation, constitutes a 'natural transition' (a term he borrow from [16]), whereas no such natural transition is obviously identifiable between degrees in the middle of a scale.

A final question that arises in this comparison of absolute and relative adjectives is why adjectives with closed scales should prefer standards that are endpoints over the sort of context-dependent standards that relative adjectives use, given that nothing about the nature of a closed scale forces the standard to be an endpoint. The answer that Kennedy proposes is that natural language follows a principle of Interpretive Economy:

(8) **Interpretive Economy:** Maximize the contribution of the conventional meanings of the elements of a sentence to the computation of its truth conditions. [9]

This principle is intended to guarantee that when an adjective's scale is closed, its standard will be maximal or minimal, since the scale is presumably part of the conventional meaning of the adjective, insofar as it is derivable from the possible values of the measure function that the adjective denotes. This standard will be preferred over a standard which is determined contextually. Since relative adjectives are not associated with closed scales, there will be no such conventionally provided degree that meets the requirement of making some individuals stand out with respect to others, and there will be no choice but to choose the standard contextually.

Kennedy (ibid.) tentatively suggests that Interpretive Economy is a constraint on semantic processing and comments, '[t]he intuition that Interpretive Economy is designed to capture is that although participants in a discourse may not be in full agreement about those properties of the context that play a role in the computation of context-dependent features of meaning, they are in agreement

about the conventional meanings of the words and complex expressions in the sentences they use to communicate (assuming they share the same lexicons and grammars).' Independently of whether this is demonstrably the case, Interpretive Economy can be taken to reflect the view that speakers and hearers tend to be fairly conservative in their use of language, relying on conventions that have been established in the use of language rather than innovating on a constant basis. In this latter sense the principle seems a plausible one to assume. What is less clear is the viability of the reasoning that leads from the conventionalized closed adjectival scale to the inevitable choice of an endpoint standard as an account for the contrasts underlying the relative/absolute distinction.

The crucial test cases for the Interpretive Economy account are of two kinds: 1) cases of adjectives which are (or can be) interpreted with non-endpoint standards despite having closed scales, and 2) cases of adjectives with standards that behave as absolute despite not being maximal or minimal. I begin with the first sort of case.

It is difficult to see how Interpretive Economy would allow a closed-scale adjective to have a standard which is not an endpoint, unless it is simply a default principle. But such adjectives do exist: a good example is *familiar*. In (9) we find two instances of this adjective. It is very difficult to see how these instances could be interpreted differently from each other, particularly since the second sentence seems intended to defeat the scalar implicature generated by the first that the familiarity property is not held to a higher degree. The only difference is that in one case the standard is established with respect to a comparison class (with *very* offering evidence for this; see [8]), while in the other, it is identified with respect to the number of things the student has to be familiar with. In this latter case, the standard is a minimal degree of familiarity; *completely* indicates that the scale associated with the adjective is closed [6].

(9) For a student who has just moved here, she is very familiar with the class routines and her teachers' expectations. In fact, she's completely familiar with them.

On the Interpretive Economy account, the adjective contributes the same measure function in the two cases: a function from entities to degrees of familiarity whose maximal and minimal values are determined by the volume of stuff that corresponds to the *with* argument. The unanswered question is why this conventional input can be ignored and a comparison class used to establish the standard on some occasions. Some kind of contextually licensed override of the economy principle would have to be possible, or perhaps one could argue that the meaning of *familiar* has been conventionalized in such a way that the standard is no longer a maximal value. Either way, we do not gain much insight into how and why such deviations are possible.

The Interpretive Economy account, as developed, also does not predict the existence of adjectives whose standards behave as if they were absolute without being endpoints on a scale. This is not due to the principle itself, but rather to the ancillary claim that maximal and minimal degrees are fundamentally

different from degrees which are not at the endpoints of scales in providing natural transitions. I now turn to a couple of examples that challenge this ancillary claim. These counterexamples do not call into question either Interpretive Economy as a principle, or the semantics for the positive form in (7), but they do shift the burden of explanation from scale structure to the specific nature of the adjectival properties themselves.

The first of these examples involves the adjective *full*. Kennedy and McNally [8] argue that *full* is an absolute adjective with a maximal standard, and Syrett [13] presents psycholinguistic evidence in support of that claim, based among other things on the fact that when presented with the two pictures in Figure 1, adult subjects consistently considered a request to indicate 'the full [jar]' infelicitous.

Fig. 1. Stimulus for the test item 'The full one' from [13], Appendix E

However, the facts are more subtle than these results indicate at first. Foppolo and Panzeri [3] present experimental data that indicates that subjects' intuitions about what counts as the standard for the equivalent of *full* and certain other absolute adjectives in Italian is sensitive to the type of object being ascribed the property. Indeed, Kennedy and McNally [8] already observed that speakers are in some cases willing to accept that a container is full even it is not full to the maximal degree. For example, when one is served a full glass of soda or beer, the beverage rarely reaches the top. Kennedy and McNally nonetheless maintained the claim that the standard is the maximal degree and offered two possible explanations for speakers' intuitions. One is that is that when a nearly-full glass is claimed to be full, the predication is, strictly speaking, false, but interlocutors are willing to speak loosely if the predication is sufficiently close to true for the purposes of the context (see [11]). The other is that the granularity of measurement might be made coarse enough in some cases so as to allow an almost full glass to count as full. However, neither of these explanations is

plausible when we consider cases such a glass of wine. Normally a wine glass is considered full if it is filled to about half of its capacity with wine. It is extremely difficult to imagine that in such cases we are taking the glass's maximal capacity as the standard but either speaking loosely or applying a very coarse granularity of measurement.

The defender of the maximal standard could argue as an alternative that we simply ignore part of the volume of the glass, thus conserving the maximality of the standard in the form of a degree that amounts to reaching the fill line for the glass, which might be below the glass's maximal capacity. This could be formalized by treating *full* as something like a function from container types to measure functions whose codomains have possibly different specific maximal values. Thus, a sentence like (10) could be accounted for by positing different choices of container type (e.g. the most generic form of glass in the first case; a glass in which wine is served in the second) for each of the uses of the adjective.

(10) The wine glass is half full; therefore, it is completely full as far as this restaurant is concerned.

This would preserve the Interpretive Economy account, but at the price of conceding that the measure function that *full* denotes depends on the specific sort of object to which it is applied as well as factors such as the function the container is fulfilling. In other words, the standard would depend indirectly on the context. Moreover, this analysis carries additional commitments. For instance, (11) should be true when the glass is filled to the fill line:

(11) This wine glass is completely full; it cannot be fuller.

But in fact it is difficult to deny that a functionally full wine glass can be made fuller, even under the interpretation that takes into account a 'fill-line' standard. It seems more promising to rethink the conditions on the use of the degree modifier *completely*, which might pose the biggest challenge to the claim that the standard is not a maximal value. Instead of requiring that *x is completely A* iff $A(x)$ is the maximal value on the volume scale, as proposed in [6], we could simply require it to be true just in case $A(x)$ is that degree which corresponds to the volume that would be occupied after a properly completed filling event involving the container in question. Such a degree might be unique for each kind of container, but it is far from obvious that it is maximal.[3]

Another challenge for the claim that the standard for *full* need not be a maximal value comes from the fact that, even when applied to the wine glass, it behaves like the standard of an absolute adjective. First, it is not compatible with *for*- or *compared to*-phrases:

(12) a. ?? Compared to the glass on the table, this glass is full.
b. ?? This glass is full for a wine glass.

[3] In fact, in some cases it might even exceed the volume capacity of the container by some amount that might be difficult to quantify precisely, as Ede Zimmermann (p.c.) observes might be the case with sake glasses.

Second it licenses crisp judgments and, correspondingly, reasoning about *full* does not fall into the pattern of the second premise of the Sorites paradox. Finally, it does not yield borderline cases within the limits of our ability to measure. If the standard is not a maximal degree, we are then left with the question as to what differentiates this non-maximal standard for *full* and the non-maximal standards for typical relative adjectives. I will return to this question in the next section.

An even clearer example of a standard that challenges the claim that all absolute standards are endpoints is that of color terms on a color-extension reading, where the gradable property makes reference to the amount of the object in question that has the color property. Consider, for example, the shirt in Figure 2:

Fig. 2. Boy's Gray Camo T-shirt (from http://www.teamcamogear.com)

Both (13-a) and (13-b) are arguably true statements about this shirt.

(13) a. This shirt is gray, but not completely gray.
b. This shirt is not white.

This suggests that in order for a color extension property to hold, the degree to which it holds must be more than minimal (otherwise, (13-b) would be false) but less than maximal (otherwise (13-a) would be false). Moreover, in contrast to the case of *full*, it also does not seem plausible to argue that we ignore those parts of the shirt that are not gray and that the standard is maximal with respect to the parts that are gray: (14) sounds like a contradiction.

(14) The shirt is completely gray, but it's not completely gray.

Nonetheless, as with *full*, this standard does not behave like a relative standard: The adjective does not allow overt mention of a comparison class (see (15-a)), nor does it admit degree modification by *very*, which Kennedy and McNally [8] argue is compatible with relative adjectives in general ((15-b)).[4]

[4] Of course, (15b) could be used to describe the intensity of the color of a shirt. The '??' indicates anomaly as a description of color extension.

(15) a. ?? This shirt is gray for a patterned t-shirt.
b. ?? This shirt is very gray.

Rather, the standard behaves like an absolute standard, with the difference that, instead of having to be maximal or minimal, it seems to be fixed to a degree such that the color gray predominates.

Though this standard behaves like an absolute standard in not being sensitive to a comparison class, we should ask whether it gives rise to crisp judgments and fails to yield borderline cases as well. In the case of crisp judgments, the answer is, I think, positive in principle. While it may be difficult to put a fixed value on the amount of color that counts as predominating (presumably it is above 50% but what exact degree it is might vary depending on e.g. where the color is distributed on the object in question), it is possible to imagine that a very small reduction in degree of color extension might make the difference between a color predominating or not, and thus the adjective applying truthfully or not. The fact that this might happen means that the second premise of the Sorites' paradox cannot hold as a rule for these adjectives. Borderline cases should also not arise, insofar as it should be possible for speakers to decide for any given case whether a color predominates in an object or does not. If it does not predominate, the color term will not apply truthfully.

Thus, the assumption that only endpoints on a scale can serve as natural transitions and, independently of reference to a comparison class, make some individuals stand out with respect to others, seems too strong. Some other explanation for the properties of absolute standards must be identified in order for the semantics in (7) to be maintainable.

Let us now turn very briefly to relative adjectives. Kennedy's claim is that the function **s** in (7) chooses a standard that makes the right individuals stand out as having the property in question. He eventually identifies standing out with being on the upper end of a natural transition. Since, on his view, no degree on an open scale provides a natural transition, a comparison class must be appealed to in order to identify the standard. However, nothing is said about how the standard is actually selected. Consider again *tall*. In what sense does the shortest member of a group of tall individuals stand out against the tallest member of the group of *non-tall* individuals, or lie on the upper end of a natural transition from non-tall to tall? This is the question that the Sorites paradox confronts us with. I want to suggest that we should look for the answer by attacking the problem from a different direction.

One fact that sometimes gets lost in discussions of the Sorites paradox is that one of the functions of adjectives is to group individuals according to the way they manifest a given quality (e.g. in the case of height, we have not only groups for *tall* and *short*, but also for some uses of e.g. *tiny* or *gigantic*). We might therefore consider the possibility that it is not the function **s** that 'chooses a standard...in such a way as to ensure that the objects that the positive form is true of "stand out",' but rather the possible groupings of objects manifesting a

given quality in different degrees that determine the standards for the adjectives that make reference to that quality.[5] In the following section, I suggest that by approaching the determination of the standard in this way, we can see the sense in which the standards eventually chosen for both absolute and relative adjectives can be considered natural transitions, even though the naturalness of those standards is not always evident by looking at the scale itself.

3 Classification and Standards

Let us take as our starting point the claim that the truthful application of an adjectival predicate requires the denotation of the adjective's argument be sortable into the category of objects for which the predicate holds, taking into account the range of relevant predicates that might characterize that individual along the quality or dimension of interest. We can then ask whether there are relevant differences in the way this sorting or classification task might proceed for relative vs. absolute adjectives, and whether these differences correlate with the contrasts presented at the beginning of this paper. The answer, I suggest, is positive: specifically, the truthful predication of relative adjectives can be insightfully modeled as CLASSIFICATION BY SIMILARITY, whereas the truthful predication of absolute adjectives can be modeled as CLASSIFICATION BY RULE.

Hahn and Chater [5] propose two criteria for distinguishing similarity- vs. rule-based reasoning, including in particular categorization or classification.[6] First, they claim that rule-based classification depends on a strict matching between the classification criterion/a and the relevant properties of the object being classified. In contrast, similarity-based classification requires only a partial match. Second, they maintain that rule-based classification involves comparing a representation associated with a specific individual (for instance, one concerning the degree of fullness of a specific glass) against a representation that is more abstract (e.g. a degree of fullness for glasses in general), whereas similarity-based classification involves comparing a representation of a specific individual or property of that individual against another representation of an equally specific individual or one or more of its properties.

Consider now the prototypical absolute adjectives. When the standard for such adjectives is a maximal or minimal degree, it is trivial to see how the decision about whether they apply to their arguments could be formulated as a simple rule which would not require comparison with any specific individuals. To know whether a (generic) container is full, we need only know how much of its volume is occupied. If all of it is, the container is full; if not, it is not. To

[5] Barker's analysis [2] of adjectives such as *stupid* might be considered to represent the spirit of this perspective on the way the standard is determined.

[6] Space limitations preclude a full discussion of the difficulties involved in distinguishing these two kinds of classification; the reader is referred to Hahn and Chater's article for details and a defense of the position that they *are* meaningfully distinguishable.

know whether a door is open, we need only check whether it has any aperture at all, and so on.

The view that the property contributed by an absolute adjective is ascribed via rule directly accounts for the behavior of these adjectives described above. First, since no comparison needs to be made to specific individuals, no comparison class is called for; thus, the examples in (3), repeated in (16), should be odd.[7]

(16) a. ?? Compared to Door #1, Door #2 is closed.
b. ?? That box of cookies is closed for a box my daughter has gotten into.

Second, if absolute adjectival properties are ascribed via rule, we can account for their failure to yield the Sorites paradox. Classification by rule, as described, requires an exact match of the classification condition associated with the property. If this condition is not precisely met, there is no reason to think that the adjectival predicate will truthfully apply. Of course, this does not exclude the possibility of a certain variability in the granularity of measurement or other criteria of precision that will be applied in order to decide whether there is a match and that the individual counts as having the property; however, the imprecision in these cases will turn out to be more circumscribed than in the case of classification on the basis of similarity.[8] Finally, in the case of absolute adjectives, indeterminacy or borderline cases will only arise to the extent that the application criteria for the rule are not fully defined or, in the case of a measurable property, the granularity of measurement is coarse.

Crucially, there is no reason to think that rule-based classification will be possible only for cases where the standard for a gradable adjective is either maximal or minimal. We only need a precise and principled way to identify the standard degree in question in order to be able to formulate the rule we need. For example, there is nothing problematic about a rule to the effect that a wine goblet is full if half of its volume is occupied, or that a beer glass is full if it contains 33 centiliters of liquid. Such rules can be applied without any consideration of a specific comparison class - they only require reference to the type of container involved insofar as that plays a role in the rule for determining fullness. Such criteria not only permit but require an exact match in the manifestation of the property in the individual in question. Thus, even though the standard is neither maximal nor minimal, we expect these adjectives to behave like other absolute adjectives. Similarly, a rule for determining color extension can be formulated, as suggested above, in terms of perceptual predominance of the color.

While classification by rule is thus clearly possible in ascribing properties denoted by absolute adjectives, we should ask whether anything in principle forces it. The answer is clearly negative. The use of *full* to characterize density of volume occupation, as in (17-a-b), is a good example.

[7] Note that direct comparison to a specific individual or class of individuals is not the same as reference to a specific *sort* of individual to clarify which variant of a rule for property ascription will apply.

[8] See [15] for a fuller discussion of the relation between granularity and vagueness.

(17) a. For a Friday, the dentist's schedule is very full.
b. Compared to the last box you packed, this one is very full.
c. The dentist's schedule/The box is completely full.

Clearly there is a maximal value for the volume occupation of a schedule or a box (see (17-c)), but in (17-a-b) the standard is determined by a comparison class or compared individual. Since most relative uses of *full* that I have identified involve characterizing density of volume occupation, we can perhaps look for an answer to the question of what factors lead to the maintenance of an absolute standard or the introduction of a relative one by asking specifically whether there is something about a property like density of volume occupation that lends itself to ascription via similarity-based reasoning.[9] Here, I think perceptual factors come into play: one thing we might expect from our unmodified, non-technical vocabulary is that it be usable without the help of measuring tools.

Density of volume occupation is at least partially independent of whether the contents of a container come into contact with the physical limits of that container. It thus may be difficult to measure precisely without some kind of measurement tool when the physical limits of a container are reached by the contents, though we can be sure that a container whose limits are not reached will not be fully occupied. Containers whose physical limits *are* reached by contents with a certain perceptible density will share important properties in common with a container whose volume is completely occupied, and they will be decidedly distinct from any container whose physical limits are not in contact with the contents. These similarities and differences may be sufficiently useful and salient for speakers to classify the former containers as full and the latter as not full, even though in such a situation it may nonetheless be clear that the similar containers are not identical. In such cases, the addition of information about comparison class or specific compared individuals may be added to improve the precision of the property ascription.

Now consider prototypical relative adjectives, again using *tall* as an example. Why should classification by rule with such adjectives be impossible? Here Kennedy's intuition about the relevance of the lack of endpoints on the scale associated with the adjective seems exactly on target. Rule-based classification involving non-maximal/minimal standards can work for adjectives interpreted with respect to a closed scale because such adjectives describe properties that can be held to proportional degrees. Since at least certain proportions are perceptually salient and can be easily estimated without knowing absolute values, it can be comparatively easy to know when a given property is held to a specific proportion. Nothing like this will be possible when applying an adjective that contributes a gradable property characterized by a single, unbounded dimension, uncorrelated with any other easily and consistently observable characteristic. In the case of height, the only plausible rule we could apply would be to stipulate

[9] Relative uses of *full* where density is arguably not at issue, like *a very full glass of water*, can also be found; I would give an explanation for them that is similar in spirit, if different in detail.

a specific height value as the standard. But such a value can only be expressed in an arbitrarily chosen measurement unit. Ascribing tallness to someone under such circumstances would depend on our ability to carry out this kind of measurement, and in the context of everyday language use this simply might not be feasible often enough.[10]

In contrast, if we have previously identified exemplars of tall and short individuals,[11] classifying any given third individual as tall or short on the basis of relative similarity to these exemplars is entirely feasible. In particular, it does not necessarily entail measuring in any precise way the heights of any of the individuals in question; to identify an individual as tall, it is sufficient to be able to judge that individual as more similar to the tall exemplar(s) than to the short one(s).[12] Nor should we worry about where the necessary exemplars to do this classification would come from: note, for example, that an entire sub-genre of children's literature is devoted to conveying certain qualities in terms of opposites, obviously facilitating familiarity with reliable pairs of exemplars on the basis of which children can learn to ascribe the relevant properties to new individuals.

If relative adjectives are ascribed on the basis of similarity rather than rule, we can immediately explain their behavior. First, the ascription of the property contributed by the adjective will depend crucially on a comparison class or compared individual because the classification involves ascribing the property on the basis of a comparison to representations of specific individuals; the job of the comparison class is to provide these individuals, most crucially, the exemplars that will serve as the basis for the classification. Second, we can explain the intuitive validity of Premise 2 of the Sorites paradox: This premise embodies a basic principle of similarity-based classification, which is that one classifies by maximizing within-class similarity and between-class distance. Finally, borderline cases can arise because some individuals may prove to be equally similar to the exemplars of the classes under consideration and thus difficult or impossible to classify in a non-arbitrary way. We can also now characterize the sense in which the standard for relative adjectives is a natural transition: it will be that degree which marks the boundary between the classes that result from grouping the individuals in question according to their similarity.

The reader might be concerned that, on this view, we would end up with two very different kinds of satisfaction conditions for gradable adjectives, which

[10] Of course, an explicit standard can be introduced via a measure phrase or a comparative expression (e.g. *taller than Max*). Unsurprisingly, comparatives do not allow modification by comparison clauses, do not give rise to the Sorites paradox, and do not have borderline cases.

[11] For simplicity, I will make this point assuming that with height there is just a binary classification into tall and short, but of course additional classes are arguably called for, such as neither-tall-nor-short, midget-sized, etc. These can be easily incorporated into the analysis.

[12] This is just the basic sort of algorithm that is used in the simplest forms of clustering, a standard similarity-based classification technique. See e.g. [4] for application of notions from clustering-based approaches to classification to a theory of conceptualization, specifically the theory of Conceptual Spaces.

might roughly be characterized as follows, setting aside degrees for the moment and assuming that we take adjectives to denote properties, and assuming that multiple properties (such as tallness or shortness) can be related to a given quality or dimension (such as height):

(18) $\mathbf{AdjP}_{absolute}(x)$ is true in context C iff x manifests the property contributed by **AdjP** as required in C.

(19) $\mathbf{AdjP}_{relative}(x)$ is true in context C and relative to a comparison class K iff x is more similar to the exemplar from K for the property contributed by **AdjP** than it is to the exemplar for any other property under consideration for classifying x with respect to the quality or dimension in question.

However, note that there is no need to assign the two sorts of adjectives to different logical types, and thus the fact that the satisfaction conditions are different should not be a cause for concern.

Summarizing, the position defended here is that abstract scale structure properties are only indirectly related to the contrasts in the behavior of absolute vs. relative adjectives. What is crucial is the possibility of establishing clear applicability conditions for the property. Various factors can facilitate or impede the establishment of such conditions. I have suggested that one important factor is the ease with which the degree that constitutes the standard can be perceived; maximal and minimal standards are a special case of such degrees, but not the only one.

I now turn briefly to the implications of the preceding discussion for the semantics of the positive form.

4 Implications for the Positive Form

Kennedy [9] takes the contrasting behavior of relative and absolute adjectives as an argument against a degreeless semantics for adjectives on which they denote simple properties of individuals, and specifically against the analysis defended in [10]. Klein's semantics treats adjectives as functions which assign individuals to positive extensions, negative extensions, and extension gaps, where extensions are defined relative to a domain D – effectively equivalent to a comparison class – of cardinality greater than or equal to 2. For any well-defined comparison class, Klein assumes that the positive and negative extensions of the adjective with respect to that class must not be empty; he defines the semantics in this way so that quantification over comparison classes can be used to induce the ordering needed to support a semantics for comparatives. Specifically, on this view *x is Adj-er than y* iff there is a comparison class for which x falls into the positive extension of the adjective and y does not. But as Kennedy notes:

> [it is not clear] how such an approach can account for the basic facts of the relative/absolute distinction in a non-stipulative way. Since vagueness (i.e., allowing for variable interpretations/precisifications) is a

> necessary condition for comparison, the expectation is that all gradable predicates should be vague. The challenge for a non-degree based analysis is to explain why only relative adjectives are vague in the positive form, while absolute adjectives have fixed positive and negative extensions, but remain fully gradable. [9, p. 41]

In other words, since comparison classes are irrelevant for, and indeed infelicitous with, the positive form of absolute adjectives, it should not be possible for an object to be in the positive extension of an absolute adjective with respect to one comparison class and in the negative extension with respect to another, and if this is not possible, then it will not be possible to correctly apply Klein's analysis of comparatives to absolute adjectives. However, this criticism might be overcome if there was some other way to induce the ordering needed to support comparatives involving absolute adjectives. [1] and [15] propose such alternatives.

The starting point of the discussion in the previous section was the degree-based semantics in (7), repeated in (20), so we should therefore reconsider it in light of that discussion:

(20) $$pos : \lambda g \lambda x.g(x) \succeq \mathbf{s}(g)$$

This sort of degree-based semantics has been criticized on the grounds that it is very abstract, that there is never or almost never any morphological manifestation of *pos*, and that it would appear that the semantics of the positive is effectively defined in terms of a comparison relation (see e.g. [10,1,15]). But let us set aside these concerns and focus on the key issue, which is whether the standard-fixing function **s** can be defined in a unified way for both absolute and relative adjectives.

We have, in a sense, found a way to characterize what it means for a standard to be a natural transition both for absolute and for relative adjectives. In the former case, it is that degree which the rule for truthful application of the adjective effectively makes reference to. In the latter, it is that degree which marks the boundary between the groupings that are derived from a similarity-based classification of a set of individuals according to the conventionalized labels we have for identifying the different manifestations of the quality or dimension in question. Nonetheless, at a deeper level, it is difficult to see how the value returned by **s** could be characterized in any truly unified terms across absolute and relative adjectives other than as 'the degree that makes the adjectival predicate truthfully hold.' Moreover, in the case of relative adjectives, I do not see any way to derive this standard degree except in an *a posteriori* fashion on the basis of the way the members of the comparison class are sorted in any given context. But if this is the case, it would seem that, as mentioned above, the identification of the standard presupposes that we are able to successfully use the adjective. This is of course not a problem for a strictly formal account of adjective semantics, but it lends support to the criticisms of such a semantics as a useful model of our semantic competence.

5 Conclusion

The discussion of gradable properties in the formal semantics literature has been heavily conditioned by a focus on orderings, scales, and standards, rather than on the nature of the properties themselves or the general role that adjectives play in categorization. While we must not lose sight of the need to derive orderings or precise measurements to support a semantics for comparative and related constructions, I have argued here on the basis of some new and overlooked data that a focus on the nature of the properties described by adjectives, as well as the classification strategies that might successfully model their ascription, can be particularly useful in providing a better understanding of the relative/absolute distinction.

Acknowledgments

I am very grateful to Christopher Kennedy, not only for many years of collaboration on the semantics of adjectives but also very specifically for giving me the opportunity to first develop the ideas presented in this paper, in the form of a commentary on his 'Vagueness and Grammar' paper at the Chicago Workshop on Scalar Meaning held in May 2006. I would also like to thank Carla Umbach, who first pointed out to me the work by Hahn and Chater, to Kristen Syrett both for comments and for permission to reproduce the material in Figure 1, and to the editors of this volume for very helpful criticism. Finally, thanks to the organizers of the 2007 Amsterdam Colloquium and the 2008 DGfS Workshop on Comparison and Similarity for giving me the chance to present this work there, to the audiences there and at CIMEC in Rovereto for their feedback, and to the organizers of the 2009 ESSLLI Vagueness and Communication workshop, who perhaps expected to hear a version of this paper at the workshop but didn't. The work reported here was supported by the Spanish Ministry of Science and Innovation under grant HUM2007-60599/FILO and by a Fundació ICREA Acadèmia award.

References

1. Bale, A.: The Universal Scale and the Semantics of Comparison. PhD thesis, McGill University (2006)
2. Barker, C.: The dynamics of vagueness. Linguistics and Philosophy 25, 1–36 (2002)
3. Foppolo, F., Panzeri, F.: Do children know when their room counts as clean? Ms., University of Milano-Bicocca (2010)
4. Gärdenfors, P.: Conceptual Spaces: The Geometry of Thought. MIT Press, Cambridge (2000)
5. Hahn, U., Chater, N.: Similarity and rules: distinct? exhaustive? empirically distinguishable? Cognition 65, 197–230 (1998)
6. Hay, J., Kennedy, C., Levin, B.: Scale structure underlies telicity in degree achievements. In: Matthews, T., Strolovitch, D. (eds.) Proceedings of SALT IX, Ithaca, NY, pp. 127–144. CLC Publications (1999)

7. Kennedy, C.: Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison. Garland, New York (1999) (1997 UCSC Ph.D thesis)
8. Kennedy, C., McNally, L.: Scale structure and the semantic typology of gradable predicates. Language 81(2), 345–381 (2005)
9. Kennedy, C.: Vagueness and grammar: The semantics of relative and absolute gradable adjectives. Linguistics and Philosophy 30, 1–45 (2007)
10. Klein, E.: A semantics for positive and comparative adjectives. Linguistics and Philosophy 4, 1–45 (1980)
11. Lasersohn, P.: Pragmatic halos. Language 75, 522–551 (1999)
12. Rotstein, C., Winter, Y.: Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. Natural Language Semantics 12, 259–288 (2004)
13. Syrett, K.: Learning about the Structure of Scales: Adverbial Modification and the Acquisition of the Semantics of Gradable Adjectives. PhD thesis, Northwestern University (2007)
14. Unger, P.: Ignorance. Clarendon Press, Oxford (1975)
15. van Rooij, R.: Vagueness and linguistics. In: Ronzitti, G. (ed.) The Vagueness Handbook. Springer, Heidelberg (to appear)
16. Williamson, T.: Vagueness and ignorance. Proceedings of the Aristotelian Society 66, 145–162 (1992)

Contradictions at the Borders

David Ripley*

Institut Jean Nicod
DEC-ENS
davewripley@gmail.com

1 The Issue

The purpose of this essay is to shed some light on a certain type of sentence, which I call a *borderline contradiction*. A borderline contradiction is a sentence of the form $Fa \wedge \neg Fa$, for some vague predicate F and some borderline case a of F, or a sentence equivalent to such a sentence. For example, if Jackie is a borderline case of 'rich', then 'Jackie is rich and Jackie isn't rich' is a borderline contradiction. Many theories of vague language have entailments about borderline contradictions; correctly describing the behavior of borderline contradictions is one of the many tasks facing anyone offering a theory of vague language.

Here, I first briefly review claims made by various theorists about these borderline contradictions, attempting to draw out some predictions about the behavior of ordinary speakers. Second, I present an experiment intended to gather relevant data about the behavior of ordinary speakers. Finally, I discuss the experimental results in light of several different theories of vagueness, to see what explanations are available. My conclusions are necessarily tentative; I do not attempt to use the present experiment to demonstrate that any single theory is incontrovertibly true. Rather, I try to sketch the auxiliary hypotheses that would need to be conjoined to several extant theories of vague language to predict the present result, and offer some considerations regarding the plausibility of these various hypotheses. In the end, I conclude that two of the theories I consider are better-positioned to account for the observed data than are the others. But the field of logically-informed research on people's actual responses to vague predicates is young; surely as more data come in we will learn a great deal more about which (if any) of these theories best accounts for the behavior of ordinary speakers.

1.1 Contradictions and Borderline Cases

In [19], I defend a theory of vague language based on the paraconsistent logic LP.[1] LP can be thought of as a three-valued logic; it is dual to Strong Kleene

* This paper has been drastically helped by discussions with Paul Egré, Patrick Greenough, Graham Priest, Diana Raffman, Greg Restall, Robert van Rooij, Nick Smith, comments from two anonymous referees, and *especially* Joshua Knobe. This research was partially supported by the Agence Nationale de la Recherche, Project "Cognitive Origins of Vagueness", grant ANR-07-JCJC0070.

[1] LP is so christened in [16].

R. Nouwen et al. (Eds.): ViC 2009, LNAI 6517, pp. 169–188, 2011.

logic, which has been recommended as a logic for vague language by eg [23] and [27]. If we use the numbers 1, .5 and 0 as the three values, then we can assign each atomic sentence A a value $\nu(A)$, and calculate the values of compound sentences as follows:

- $\nu(\neg A) = 1 - \nu(A)$
- $\nu(A \wedge B) = \min(\nu(A), \nu(B))$
- $\nu(A \vee B) = \max(\nu(A), \nu(B))$

It follows from these clauses that when A takes value .5, so too will $A \wedge \neg A$. But what do the values *mean*? We can, as usual, take the value 1 to represent truth and 0 to represent falsity. When it comes to the value .5, LP and Strong Kleene logic differ from each other. The Strong Kleene theorist reads .5 as a *gappy* value – one taken by sentences that are neither true nor false. Since such sentences aren't true, they aren't to be asserted, and they aren't part of the Strong Kleene theorist's theory. On the other hand, the LP theorist reads .5 as a *glutty* value – one taken by sentences that are *both* true and false. Since such sentences are true, they are to be asserted, and they are part of the LP theorist's theory.

An LP-based theory of vagueness uses this middle value for borderline cases. That is, where Egbert is a borderline case of 'old', the sentence 'Egbert is old' receives value .5. As above, this ensures that the sentence 'Egbert is old and Egbert isn't old' also receives the value .5. Since sentences with the value .5 are true, this theory predicts borderline contradictions to be true (it predicts them to be false as well). For similar reasons, whenever a is a borderline case of a vague predicate F, I claim that 'a is F and a is not F' is true. Similarly, I claim that 'a is neither F nor not F' is true as well, since this follows from the former by a single De Morgan law plus an application of a double-negation rule, both of which are valid in LP. This is a dialetheist theory, since it takes some contradictions to be true.

Other theorists, of various stripes, have not been so sanguine about the truth of borderline contradictions. A few quick examples: Fine [7] dismisses the idea in a single sentence – "Surely $P \wedge \neg P$ is false even though P is indefinite".[2] Williamson's [28] much-discussed argument against denials of bivalence works by arguing the denier to a contradiction; assuming the denial of bivalence was initially made about a borderline case, this contradiction will itself be a borderline contradiction. If Williamson thinks this is a dialectically strong argument, as he gives every indication of, borderline contradictions had better not be true. Keefe [11] offers: "many philosophers would soon discount the paraconsistent option (almost) regardless of how well it treats vagueness, on the grounds of ... the absurdity of p and $\neg p$ both being true for many instances of p". And Shapiro [21] claims, "That is, even if one can competently assert Bh and one can competently assert its negation, one cannot competently contradict

[2] Notation changed slightly; note that Fine is here treating borderline cases as "indefinite".

oneself (dialetheism notwithstanding)."[3] None of these rejections of borderline contradictions offers much in the way of argument; it's simply taken to be obvious that borderline contradictions are never true, presumably since no contradictions are ever true.[4]

Not all theorists – not even all non-dialetheist theorists – have been so quick with borderline contradictions, though. For example, fuzzy theorists allow for borderline contradictions to be partially (up to half) true.[5] Let's see how. The usual way of doing things assigns each sentence A a real-number truth value $\nu(A)$ from 0 to 1, inclusive. Then, the values of compound sentences are determined truth-functionally from the values of their components, according to the same clauses given above for LP. It follows from this that a contradiction (conjunction of a sentence with its own negation) can take a value as high as .5. It takes this maximum value when its conjuncts themselves each take value .5 – right in the middle of a vague predicate's borderline. A fuzzy theorist interprets the value .5 as a degree of partial truth, in particular as half truth, so a fuzzy theorist predicts borderline contradictions to be at least partially true, as much as half true. This prediction is often held up as a liability of fuzzy theories; see for example [28].

1.2 Predictions about Ordinary Speakers

Smith [22, pp. 252-253] lists ten sorts of sentence for which we don't as yet have clear empirical data about speakers' intuitions; he resists making many predictions about speakers' intuitions pending the data. At least three of his categories are borderline contradictions, in my sense, and he's right: there isn't much data on speakers' responses to them.

Some experimenters have taken brief looks at ordinary speakers' intuitions surrounding vague predicates (for example [3]), but these have primarily looked at atomic sentences, whereas the crucial action for theories of borderline contradictions is clearly in compound sentences; empirical work here is still in its infancy.[6]

[3] Since, for Shapiro, the relevant cases in which one might competently assert Bh and competently assert its negation are all cases where h is a borderline case of B, this is a rejection of borderline contradictions.

[4] Williamson might claim to have an argument for his rejection of borderline contradictions: his defense of classical logic on grounds of its simplicity. Note, though, that that defense is dialectically out of line in the midst of the argument Williamson gives against denials of bivalence; why bother arguing the bivalence-denier to a contradiction, and then appeal to the truth of classical logic to reject the contradiction, when you could simply appeal to the truth of classical logic directly to counter a denial of bivalence? Presumably, Williamson thinks the rejection of borderline contradictions is dialectically more secure than his defense of the full apparatus of classical logic.

[5] At least the usual sort of fuzzy theorists do. See for example [22].

[6] Since this paper was prepared, a few other studies have appeared that explore compound sentences like those considered here: [2] and [20].

Few logically-minded theorists of vagueness, then, have bothered being very explicit about what their theories predict about ordinary speakers. This does not mean, however, that there is no relation between these logical theories and experimental data. We have supervaluationist and contextualist and fuzzy theories of vagueness, and we can take these theories to be formal semantic theories, answerable to speaker intuitions in just the same way that other semantic theories – about gradable adjectives, or quantifier inferences, say – are.

It may well be, of course, that some theorists don't intend for their logical theories to be interpreted in this way. They might be offering hypotheses, for example, about the structure of reality itself, independent of how we talk about it; or they might be offering hypotheses about how we *ought* to use our language, rather than about how we *do*. These are worthy questions in their own right, but I won't explore them here. Rather, I'll present an experiment and weigh various possible explanations for the result; as such, the goal here is to consider various hypotheses about how speakers actually use vague language.

These hypotheses are best understood, I think, as theories of speakers' linguistic competence, and there is of course much more to participant responses than simply their competence; any number of performance factors may intervene. While there is no direct inference to be made from data about participants' responses to conclusions about their competence, the two are still related. The connection is provided by theories of the intervening performance factors. Given data x, we can compare theories of competence y and z by seeing what theories of performance would need to be conjoined to them, respectively, to explain x. If we find that y needs an odd story about performance factors to explain x, while z can explain x when conjoined with a natural (ideally, an independently-motivated) theory of performance, then this gives us some reason to favor y over z.

As we've seen above, different logical theories accord different status to borderline contradictions – some predict them to be fully true, some predict them to be at best half-true, and some predict them to never be true at all. I'll present and consider some evidence about which of these predictions seems to accord best with speakers' intuitions. Where predictions seem to come apart from participants' intuitions, I'll consider various performance-based explanations that might be offered.

2 The Experiment

To explore intuitions about contradictions in borderline cases of vague predicates, I conducted an experiment. Participants were 149 undergraduate students at the University of North Carolina.[7] They saw a slide (projected onto a screen) with seven circle/square pairs on it, labeled 'Pair A' to 'Pair G'. In Pair A, at the very top of the slide, the circle was as far from the square as it could be, while

[7] No demographic information was collected. All students were within the first month of introductory-level non-logic philosophy courses; it would be odd but possible for some of them to have taken other philosophy courses (including logic) in the past.

in Pair G, at the very bottom of the slide, the circle was touching the square. In between, the remaining five pairs moved the circle bit-by-bit closer to the square. (See Figure 1.)

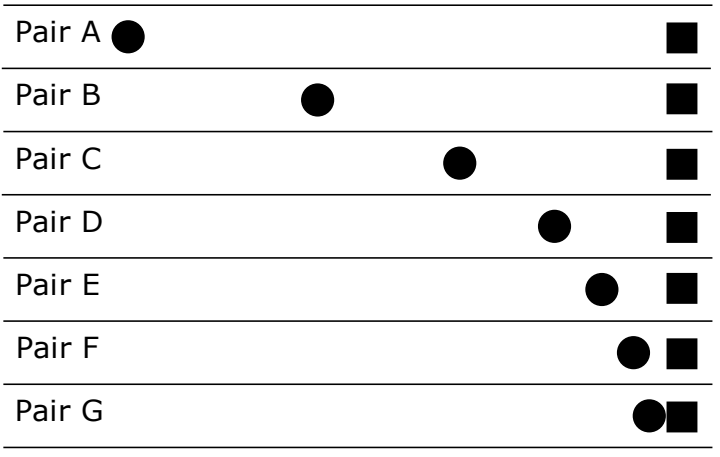


Fig. 1. Experimental Stimulus

It's difficult to tell exactly what's a borderline case of 'near the square'; as many authors have pointed out, the extension of vague predicates like 'near' is quite context-dependent, and it can be difficult to tell where the borderline is. For example, if we're discussing distances between cities, this provides a context in which the circle is near the square in *every* pair; the distance in the farthest pair is never more than the size of the screen being used, which is surely smaller than the distance between even the closest cities. Nevertheless, I take it that the context provided by this experiment is one in which: in Pair A, the circle is a clear countercase of 'near the square' (that is, it is clearly not near the square – after all, it's as far away from the square as can be projected on the screen), and in Pair G, the circle is a clear case of 'near the square'. Somewhere in between are the borderline cases.

Participants were randomly assigned to one of four conditions. In each condition, participants were asked to indicate their amount of agreement with a particular sentence as applied to each of the seven circle/square pairs. The four conditions involved four different sentences; each participant, then, saw only one sentence and rated it seven times, once for each pair. Ratings were on a scale from 1 to 7, with 1 labeled 'Disagree' and 7 labeled 'Agree'.[8] The four sentences were:

[8] As will emerge in §3.4, offering participants a range of responses is crucial to evaluate how well fuzzy theories describe participants' responses.

Conjunction, Non-elided: The circle is near the square and it isn't near the square.
Conjunction, Elided: The circle both is and isn't near the square.
Disjunction, Non-elided: The circle neither is near the square nor isn't near the square.
Disjunction, Elided: The circle neither is nor isn't near the square.

I'll discuss the difference between the elided and non-elided cases later. For now, note that each of these sentences has the form of a contradiction. The conjunctions wear their contradictoriness on their faces, while the disjunctions are a bit disguised; but one application of a De Morgan law reveals them to be contradictions as well.

2.1 Agreement to Contradictions

The mean responses to each pair formed a hump pattern: higher in the middle than at the ends. This is true overall, and it's also true of each of the four conditions (see Figure 2 on page 175). The highest overall mean occurred in response to Pair C; there the mean response was 4.1, slightly above the midpoint of the 1 to 7 scale. In other words, participants exhibit higher levels of agreement to these apparent contradictions when they are about borderline cases; they do not reject what appear to be borderline contradictions. They seem to make it to at least ambivalence. In fact, they go considerably further. The means are as low as they are because the participants do not agree amongst themselves as to which stimulus should receive the highest response. If we forget about *where* the highest responses occur, and look only at *how high* each participant's highest response is (see Figure 3 on page 175), we see that the modal maximum response is 7 – full agreement – and that the majority of participants offer a maximum response of either 6 or 7.

Similar results are reported in [2]; they do not measure degree of agreement, but also record agreement with apparent contradictions in borderline cases.

2.2 Response Types

Over 90% of participants gave responses that fall into one of four groups. I'll call these groups *flat*, *hump*, *slope up*, and *slope down*. Here are the defining characteristics of these groups (see Figure 4 on page 176 for frequencies):

Flat: A *flat* response gives the same number for every question. (24 participants)
Hump: A *hump* response is not a flat response or a slope response, and it has a peak somewhere between the first and last question; before the peak, responses never go down from question to question (although they may go up or remain the same), and after the peak, responses never go up from question to question (although they may go down or stay the same). (76 participants)

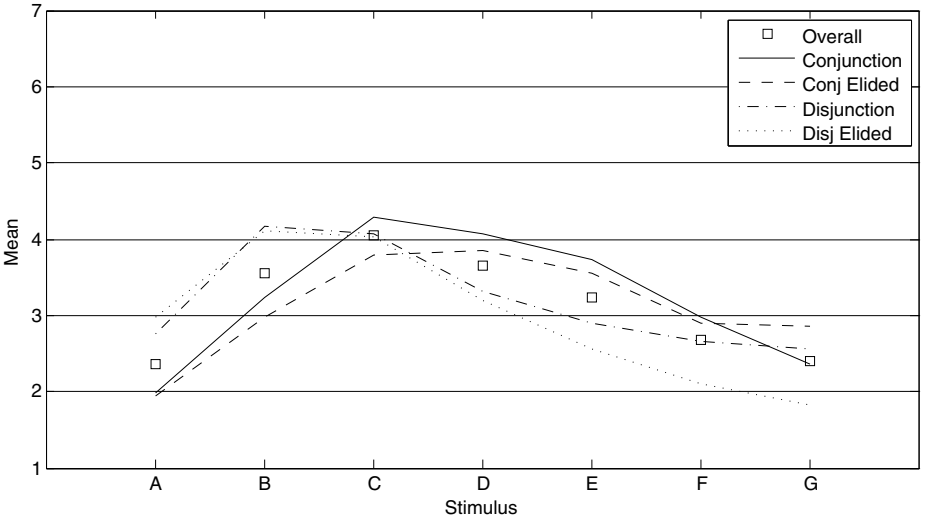


Fig. 2. Mean responses

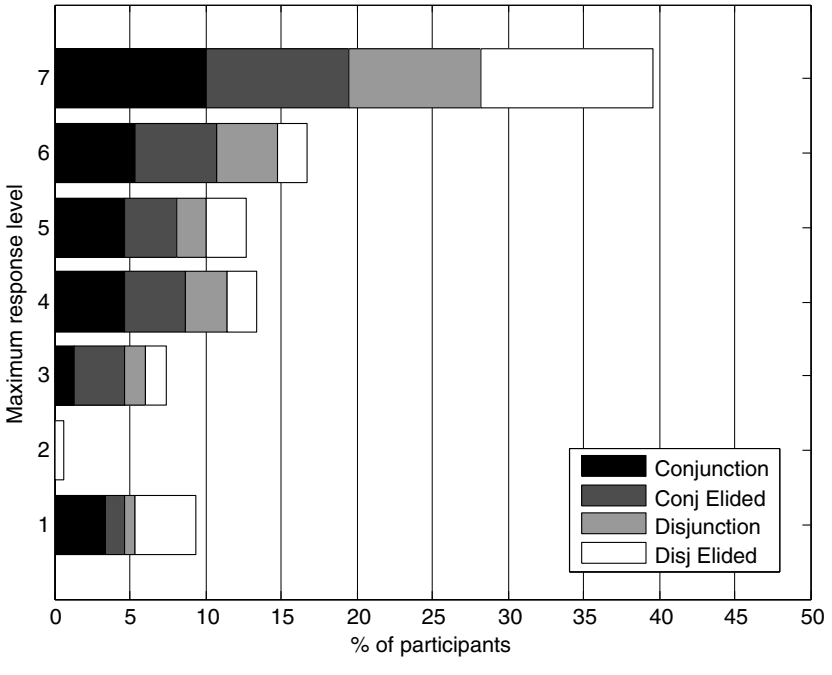


Fig. 3. Maximum responses

Slope up: A slope up response is not a flat response, and it never goes down from question to question (although it may go up or stay the same). (22 participants)

Slope down: A slope down response is not a flat response, and it never goes up from question to question (although it may go down or stay the same). (18 participants)

Other: There were a few responses that didn't fit any of these patterns. (9 participants)

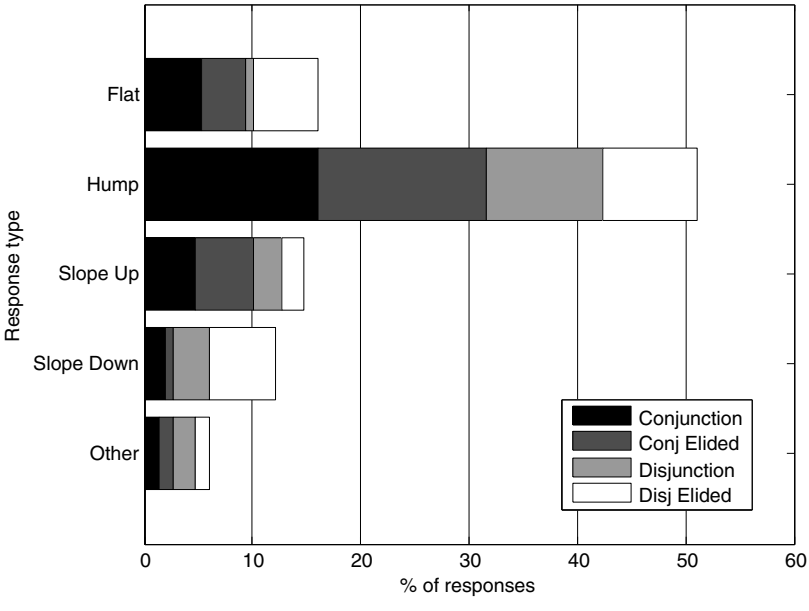


Fig. 4. Response type frequencies

Flat responses, in particular flat 1s (14 participants), look like the sort of response that would be predicted by all those theorists who hold that no contradiction is ever true, even a bit, even in borderline cases. But the majority of responses (76/149 participants) were hump responses.

The discussion that follows in §3 will consider various explanations for participants' agreement, partial or full, with these sentences. I'll focus discussion on the relatively large number of hump responses; a fuller discussion would consider potential explanations for the flat and slope groups as well.[9]

[9] Question type (conjunction vs. disjunction) had a significant effect on response type ($\chi^2(4, N = 149) = 11.27, p < .05$). However, this effect disappeared when the two slope response types were lumped together ($\chi^2(3, N = 149) = 2.76, p = .43$). Slope up responses occurred more in response to conjunctions, and slope down responses in response to disjunctions. This makes it seem as though the slope responders tended to ignore the second conjunct in each case, treating 'both near and not' as 'near' and 'neither near nor not' as 'not near'. More study would be needed to definitively interpret the slope responses.

3 Interpretations

It seems at first blush that we have substantial numbers of participants agreeing, at least somewhat, with borderline contradictions of various sorts. As in §1.1, however, theorists of varying stripes have not only claimed, but taken it to be obvious, that borderline contradictions can never be true. If those theorists are right, then participants in the present study either 1) were not really agreeing with contradictions, but rather with something else, or 2) were really agreeing with contradictions, but were mistaken in doing so. In this section, I'll consider a variety of potential explanations along these lines for the observed results. I'll also consider potential explanations of a third sort: those that hold that participants were really agreeing with contradictions, and that they agree because those contradictions are (partially or wholly) true. In the end, I'll argue that two potential explanations – one of the first type and one of the third type – are better positioned to explain the data than are the others.

3.1 Contextual Factors

This explanation falls into the "not really a contradiction" category.

Here's one way to explain the relatively high levels of assent to sentences like 'The circle is near the square and it isn't near the square' and 'The circle neither is near the square nor isn't near the square': participants take the phrase 'near the square' to have subtly different extensions in each of its two occurrences within the sentence. If this is so, their assent to these sentences can be explained without supposing that any participants agree to a contradiction. (For my purposes here, a "contextualist" is not someone who offers any particular theory of vagueness, but rather anyone who thinks that the hump responses in the present experiment are to explained by appealing to contextual shift in the extension of 'near the square'.)

Such a contextualist theory can come in one of two flavors: it might hold that 'near the square' has these different extensions because it has different contents in each of its uses, or it might hold that 'near the square' has the same content in each of its uses, but that nonetheless it has different extensions in different contexts. Following MacFarlane [14], I'll call the first flavor 'indexical contextualism' and the second 'nonindexical contextualism'. I discuss each in turn.[10]

Indexical Contextualism. Indexical contextualism about vague terms is defended in [24]. On this theory, different uses of vague terms can express different

[10] Besides the difference between indexical and nonindexical contextualism, there is another difference in the area: the difference between theories that posit sensitivity to context-of-use (sometimes called "contextualist") and theories that posit sensitivity to context-of-assessment (sometimes called "relativist"). I'll ignore that distinction; for my purposes here, I'm happy to lump the relativists in with the contextualists.

properties. This shiftiness is understood as the very same shiftiness exhibited by such indexical expressions as 'here', 'now', 'you', 'tomorrow', &c. For example, let's focus on 'you'. 'You', let's suppose, picks out a certain person: the person being addressed when it's uttered. Now, imagine someone uttering the following sentence: 'Mona sees you, and Louie sees you'. It should be clear that the two occurrences of 'you' in such an utterance might pick out different people; just imagine the context changing in the right way (that is, so that the first half of the sentence is addressed to someone different than the second half).

On an indexical contextualist theory, something just like this might be happening in the sentence 'The circle is near the square and not near the square'; the first occurrence of 'near the square' can pick out one property, and the second some other property. For this to be the case there would have to be some relevant shift in context between the two occurrences, and the indexical contextualist would have to provide some story about what the relevant context is and why it shifts.[11] Even with such a story in hand, though, the indexical contextualist runs into some difficulties with the experimental data.

The difficulty arises with the elided sentences: 'The circle both is and isn't near the square' and 'The circle neither is nor isn't near the square'. Each of these sentences contains only one occurrence of 'near the square'. It's clear, though, that indexicals, in these circumstances, can have only a single interpretation. Compare our earlier 'Mona sees you, and Louie sees you' to 'Mona sees you, and Louie does too'. Even with the same shift in context (that is, with the second half of the sentence addressed to someone different than the first half), the second sentence *must* report that Mona and Louie see the very same person. Since there's only one occurrence of 'you', it can only pick out one person.[12]

Thus, the indexical contextualist should predict that, although participants might agree to the non-elided sentences, they should not agree to the elided sentences, since the mechanism invoked to explain participants' agreement in the non-elided cases can't operate in the elided cases. Participants simply should not agree with elided sentences. At the very least, they should agree less than they do with the non-elided sentences. This prediction is not borne out. If we

[11] This requirement is not unique to the indexical contextualist; every contextualist needs such a story. I won't be concerned with the details of such stories here – see for example [17], [21], or [23]. (NB: Raffman and Shapiro are not indexical contextualists.)

[12] Similar phenomena arise around (at least) demonstratives, definite descriptions, and proper names. In each of the following pairs, the first member allows a shift where the second does not:

- – Mary's buying that, unless Murray buys that
  – Mary's buying that, unless Murray does
- – Put your bag on the table, and your books on the table
  – Put your bag on the table, and your books too
- – Esmerelda went to the store, and Esmerelda bought some fish
  – Esmerelda went to the store and bought some fish

consider each participants' maximum level of agreement, there is no significant difference between responses to elided and non-elided sentences.[13] Nor is there a difference in response types (flat, hump, &c.) between elided and non-elided cases.[14] If participants' agreement to these apparent contradictions, then, is to be explained by appealing to context, that context can't be operating in the way that context operates on indexicals.[15]

Nonindexical Contextualism. Is there another way, then, for context to come into play? The nonindexical contextualist thinks so. I think nonindexical contextualism, suitably filled in, provides one of the more plausible explanations for the results of the present study. The task of this section will be to present some constraints that the nonindexical contextualist must satisfy to explain the observed results.

To see how nonindexical contextualism works, let's consider an indexical case in more detail. Consider an utterance, by me, of the sentence 'I like to dance'. The occurrence of 'I' in that utterance refers to me, so the whole utterance has the content *Dave likes to dance*.[16] That content is (very) true, but it might have been false; it is true with regard to the world we find ourselves in, and false with regard to other possible worlds. So, in determining the extension (truth-value) of the utterance from its content, we need to take something more into account: we must consider at least which possible world we're in. The nonindexical contextualist finds a role for context in just this way – in the step from content to extension. They can offer various theories, still, about *which* contextual factors come into play; the key to nonindexical contextualism is *when* those factors do their work.[17] For the details of one particular nonindexical contextualist theory of vague predicates, see [6]; for general arguments that contextualists about vagueness should be nonindexical contextualists, see [1].

So what would a nonindexical contextualist offer as a take on the present study? Let's start with 'The circle is near the square and it isn't near the square'. The indexical contextualist held that this sentence ascribes one property ('near the square' in context 1) and the negation of *some other* property ('near the square' in context 2) to the circle; its content was thus baldly noncontradictory. But the nonindexical contextualist doesn't go this route; she'll say that the sentence ascribes one property (nearness-to-the-square) and the negation of *that very property* to the circle. In order to avoid contradiction, then, she must say that the one property has two different extensions with regard to two different contexts. Importantly, those contexts must both be at play in the interpretation

[13] As measured by a one-way ANOVA, $F(1, 148) = .24, p = .62$.
[14] $\chi^2(4, N = 149) = 1.98, p = .74$.
[15] This is similar to the argument in [26], except that Stanley fails to distinguish between indexical and nonindexical contextualism. See [1] for details.
[16] I ignore any possible context-sensitivity, of any sort, in 'likes to dance'.
[17] This way of framing the issue owes much to [10] and [14].

of the single sentence.[18] If context is ephemeral, dependent on, say, a transient mental state of the judger (as in [17]), then this should be possible. On the other hand, if context is coarser-grained, dependent on only things like world, approximate time, location, speaker, and the like, then we can see that context could not have changed mid-sentence, and so a contextualist explanation couldn't get off the ground.

By examining the elided conditions in the present study, we can see further constraints on a workable nonindexical contextualist theory. We've already seen that, for this explanation to work, the relevant features of the context in play must be relatively fine-grained. The elided conditions provide us evidence about *which* context it is that comes into play. Consider 'The circle both is and isn't near the square'. For a nonindexical contextualist explanation to work, the context relevant to determining the extension of 'near the square' cannot be the context in which 'near the square' is read by the participant. After all, there is only one such context, but the contextualist appeals crucially to a change in context between two extension-determinations.

I see two options for the nonindexical contextualist: 1) it may be that participants process this sentence into some form that contains two occurrences of 'near the square' or something (conceptual material, presumably) corresponding to 'near the square' – then each separate occurrence can be affected by the context in which it occurs – or 2) it may be that participants evaluate the conjuncts one at a time, retaining only the truth-value of each conjunct after its evaluation – then each evaluation can be affected by the context in which it occurs. Neither of these explanations is straightforwardly available to an indexical contextualist, lest she (falsely) predict that sentences like 'Mona sees you, and Louie does too' can exhibit the same kind of shift. The nonindexical contextualist, though, can avoid this prediction, by supposing that the duplication or repetition process operates on contents rather than characters or expressions.

Thus, nonindexical contextualism, suitably filled in as above, can offer an explanation of the present results. Below, I'll consider other possible explanations.

3.2 Noncompositional Theories

Another variety of not-really-a-contradiction explanation claims that the sentences in question are not compositionally interpreted; that 'The circle is near the square and it isn't near the square' directly expresses something like what's

[18] At least for indexical context-sensitivity (and why should nonindexical sensitivity differ?), it seems incontrovertible that multiple contexts can be involved in the interpretation of a single sentence. See note 12, or consider 'I am here now', which can be false if said very slowly while moving very quickly. Some authors, though, have missed this: for example Richard [18], who writes, 'Switching contexts in the middle of interpreting a sentence is clearly contrary to the spirit, not to speak of the letter, of Kaplan's approach to indexicals.' (I'm skeptical of his reading of Kaplan.) Other authors have played it down: see [10], which makes 'I am here now' come out as a logical truth in its logic of indexicals, or [14], which talks of context affecting whole propositions at once.

expressed by 'The circle is a borderline case of "near the square"'. Perhaps it's an idiom, or something like an idiom. Then participants' relatively high level of agreement could be explained without supposing that they agree to a contradiction.

The problem with such an account is that it's difficult to see why apparent contradictions would express borderline-case-ness. How would such an idiom get off the ground? Presumably because some other explanation canvassed here (in particular, one of the explanations in §§3.1, 3.4, or 3.5) was at one time correct; then language learners, for whatever reason, might have mistaken their elders' compositional utterances for direct claims of borderline-case-ness. This fills in the story, but it does so compositionally. Without some explanation very unlike this (lightning strike?), I don't see that a noncompositional theory can avoid essentially appealing to some compositional theory, and it seems that it will then take on the pros and cons of whatever compositional theory it chooses.

There will be a few extra cons, however. A non-compositional theory must explain why there is no significant difference in the frequency of observed hump responses between the four experimental conditions, and why there is no significant difference between the maximum responses given by participants in these conditions.[19] Do we have four closely-related idioms? If so, why? In addition, this strategy invokes an additional step: learners coming to acquire noncompositional uses of these once-compositionally-used expressions. Without further evidence, a noncompositional theory introduces needless complication; better to stick with a compositional story.

3.3 Error Theories

So much for explanations that work on the hypothesis that what participants are agreeing to isn't a contradiction. Among theories that concede that participants are agreeing to a contradiction, error theories of various sorts are available. An error theorist holds that, while participants are in fact agreeing to real contradictions, they are wrong to do so – these contradictions are simply false. Those who hold a supervaluationist or epistemicist theory of vague predicates might most naturally explain the present results via an error theory.

An error theory might work something like those presented in [5] and [25], according to which all competent speakers have dispositions to accept certain falsehoods involving vague predicates, or it might work in a more informal way, supposing participants to simply be mistaken, not in virtue of being competent speakers, but just in virtue of being confused, or not paying attention, or being misled by the experiment, or failing to report what they actually believe, or some such.

[19] See notes 9, 13, and 14; and note that there was also no significant difference between maximum responses to conjunctive and disjunctive sentences ($F(1, 148) = .53, p = .47$), nor any interaction effect on maximum responses between conjunction/disjunction and elided/non-elided (as measured by a two-way ANOVA, $F(1, 148) = 1.37, p = .25$).

Competence-based Error Theories. I'll turn now to the former sort of error theory. Eklund's view can directly explain why participants would make errors in these cases; it's part of his theory that competent speakers have a disposition to make errors in the use of vague predicates. But the errors he takes speakers to be disposed to make are not hump-style responses. Rather, he supposes that competent speakers are disposed to believe *tolerance principles* around their vague predicates. He takes his tolerance principle from Wright [29]; for a vague predicate F, the tolerance principle reads:

– Whereas large enough differences in F's parameter of application sometimes matter to the justice with which it is applied, some small enough difference never thus matters.[20]

But belief in a principle like this would not lead participants to give hump-style responses; rather, if it applied at all, it would lead participants to give flat responses, responses not affected by the small differences in the cases they were shown. So while Eklund predicts that participants will make a certain sort of error, he does not predict the hump-style responses given by many participants.

Sorensen [25] faces a similar problem: although he claims that competent speakers will believe contradictions involving vague predicates, he does not predict the present results. The "contradictions" Sorensen predicts speakers to believe are sentences of the form 'If a is F, then a's successor is F too', where a and its successor are consecutive members of a sorites sequence for F. Since Sorensen is an epistemicist, he thinks there is some sharp cutoff between the Fs and the non-Fs; when a and its successor straddle this sharp cutoff, he believes this conditional to be analytically false. Nonetheless, he thinks, we believe it.

This is essentially the same as Eklund's view, except for the decision to call these tolerance conditionals "contradictions". This sort of view, if it can be made to make any predictions at all about the present study, predicts *flat* responses, not *hump* responses. So again, this style of view cannot explain the present results.

I suppose someone might hold a view like this: being a competent speaker requires us to believe contradictions like 'a both is and isn't F' when a is a borderline case of F, but nevertheless such contradictions are always false. *That* view of course would predict the hump responses obtained in the present study. But why would competent speakers believe *those* falsehoods and not others? Any view of this sort would need to answer that question. Sorensen and Eklund go to great lengths to motivate their claims that speakers believe certain falsehoods; an error theorist of this type would need some story to fill a corresponding role. I know of no error theorist who holds this kind of theory, and so I know of no error theorist who's attempted to provide such a story.

[20] F's parameter of application is the dimension along which something can vary to make it more or less F; so 'tall"s parameter of application is height, 'bald"s is amount and arrangement of hair, &c.

Other Error Theories. I turn now to the other sort of error theory. This sort takes the participants who agreed with some contradiction to be mistaken for some reason other than their linguistic competence. So stated, there is a gap: on its own, this offers us no explanation for why participants would make *these* errors and not others. It could of course be supplemented with some theory about the conditions under which people are likely to make certain errors, and then that supplemental theory could be dealt with on its own merits.[21]

Two such supplemental theories are offered by an anonymous referee. First, it's possible that, although participants would naturally want to simply reject all the sentences, the mere fact of being asked about the same sentence again and again suggests that something else is wanted of them. This might lead participants to vary their responses. Indeed, it's likely that asking participants about the same sentence repeatedly leads at least some of them to vary their responses, to avoid being uncooperative. As the referee points out, though, this stops well short of explaining why participants would vary their responses in such a coordinated way; it would predict (correctly) few flat responses, but it would fail to predict the hump responses that predominated.[22]

Second, it's possible that being asked about their agreement or disagreement with the sentences, rather than the sentences' truth or falsity, suggested to the subjects that the issue at hand is a matter of personal opinion, causing them to respond to some proposition 'about which opinions could differ', rather than responding to the target sentence. I am skeptical about this hypothesis, for two reasons. The first reason is that it's not clear what this other proposition might be. In order to explain the present results, the proposition must meet two constraints: 1) it must be a plausible interpretation of the test sentences, and 2) it must be more likely to be agreed with in borderline cases. I don't know what might meet these constraints. The second reason is that agreement and disagreement are not restricted only or even primarily to matters of opinion. We quite often agree and disagree with statements of fact. As such, I doubt that asking about agreement and disagreement suggests to participants that the question is opinion-based, although there is certainly room to explore this issue further.

There may be other available hypotheses as to why participants would err in the task at hand in this experiment; each would have to be considered on its own merits.

[21] NB: It can't simply be that participants err randomly under certain conditions; there are very many possible response patterns that simply didn't occur, or that occurred very rarely, while the hump pattern occurred in the majority of responses.

[22] A partial explanation for the relatively large number of slope responses might be lurking around here; given that participants were presented with seven smoothly-shifting pairs and asked to judge each sentence from one to seven (a coincidental double use of seven), that may have suggested to some that a smooth shift in their responses from one to seven or from seven to one was called for. This, of course, is to gesture towards an error theory of the slope responses; but I don't see how the slope responses can be accounted for without an error theory of some sort. They are in that regard quite unlike hump (and for that matter flat) responses.

3.4 Fuzzy Theories

A fuzzy theory can both 1) allow that participants interpreted the sentences in question as contradictions, and 2) allow that participants might not be mistaken in partial assent to such sentences. This second feature is a virtue for a few reasons. First, as we've seen in §3.3, no existing error theory predicts speakers to be mistaken in this way; and second, it seems a bit odd to suppose that speakers are mistaken about what's near what, when they can see the relevant objects clearly, are deceived in no way about the distance between them, and are not under any time pressure to come to a judgment. A fuzzy theory can allow for non-mistaken (partial) assent to contradictions because on a fuzzy theory contradictions can be partially true, as we saw in §1.1.

At first blush, then, it appear that the fuzzy theorist has the resources to account for the responses observed. This appearance is strengthened if we look at the mean responses for each question (see Figure 2 on page 175): the clear cases on each end result in mean responses just above 2 – very low in agreement – and the mean responses rise gradually as one approaches pair C, where the mean response is just barely above 4, the midpoint in the agreement scale. These data are very much in line with what a fuzzy theorist would most likely predict.

Appearances, though, can be deceiving. Although the mean responses to each question create a pattern congenial to the fuzzy theorist, they do so for a strikingly non-fuzzy reason. This can be brought out by considering the difference between the *maximum* of the *mean* responses (4.1) and the *mean* of the *maximum* responses (5.3). The majority of responses were hump responses, but not all humps reach their peak in response to pair C, presumably due to slight disagreements between participants on which pairs were the clearest borderline cases. Recall Figure 3 on page 175.

If the fuzzy theorist's formalism maps directly on to participants' responses, we would expect participants' responses to these contradictions to peak somewhere around 4, the midpoint. After all, none of these sentences can ever be more than .5 true, on a fuzzy theory. But this is not what happens. As reported above, more participants peak at 7 – full agreement – than at any other response.[23] The mean of the maximum responses is 5.3 – significantly above 4.[24]

The fuzzy theorist, faced with these data, should conclude that the fuzzy formalism does not map directly onto participants' responses, then. Here's a hypothesis she might offer: perhaps responses as high as 7 – full agreement – can still indicate the speech act of .5-assertion. If this is so, then the fuzzy theorist can simply claim that participants who gave very high responses to these sentences were still only .5-asserting them.

[23] Actually, more than *twice* as many peak at 7 than at any other response, and over half of participants peak at either 6 or 7.

[24] In fact, this is so for each of the four conditions: for conjunction, non-elided, mean 5.2, $t(43) = 3.96, p < .001$; for conjunction, elided, mean 5.3, $t(39) = 4.719, p < .001$; for disjunction, non-elided, mean 5.7, $t(28) = 5.67, p < .001$; for disjunction, elided, mean 5.1, $t(35) = 2.81, p < .01$.

I don't see that this hypothesis is untenable, but it would take some filling in. Presumably a response of 7 can also indicate 1-assertion (full assertion), so this hypothesis leads the fuzzy theorist to suppose that a 7-point scale from 'Disagree' to 'Agree' is not sensitive to the different degrees of assertion participants might wish to make. But if not this sort of scale, then what *would* be sensitive to those degrees? It seems that the fuzzy theorist appealing to this hypothesis would need to address that question. With an answer to that question in hand, a study like the present one could be conducted, to see whether participants really do indicate .5-assertion to these sentences.

Alternatively, the fuzzy theorist could offer an error theory of some sort. She might allow that although the highest level of assertion *appropriate* to these sentences is .5, most participants in fact evinced a higher level of assertion, and simply claim that these participants are mistaken. As we've seen, such responses are unilluminating unless conjoined with some explanation of why participants would make *these* mistakes in *these* circumstances; but there is no reason why a fuzzy theorist couldn't propose such an explanation.

3.5 Dialetheisms

A dialetheic theory like that presented in [19] shares some of the features of a fuzzy explanation for the present data: it can allow that, in line with appearances, participants are responding to genuine contradictions; and it can allow that these participants are not mistaken. What's more, since a dialetheic theory predicts that the contradictions that occurred in this study are (fully) true, it naturally predicts levels of assent higher than the midpoint values most naturally predicted by fuzzy theorists.

This is because, according to this variety of dialetheic theory, the borderline contradictions in the present study are true.[25] The circle really is both near the square and not near the square, when it's a borderline case of 'near the square'. And similarly, it's neither near the square nor not near the square, in the same circumstances. Since participants in the present study were well-positioned to see this, and since they are competent with 'near the square', conjunction, disjunction, and negation, they agreed with the borderline contradictions because they recognized them as true.

A dialetheic explanation, then, faces a quite different puzzle from the other theories we've seen. The question a dialetheist must answer is not 'Why so much assent?' but 'Why so little?'. As we've seen, the mean of the maximum responses was 5.3. Even allowing for ceiling effects, this is unlikely to represent full agreement. But if participants were well-situated to recognize the truth, and the truth is contradictory, why would they not simply fully agree to borderline contradictions? A dialetheist owes some answer here.

Since I defend a dialetheic theory of vagueness elsewhere, I'll offer a sketch of one possible answer. It's been alleged among cross-cultural psychologists that

[25] It thus differs from the dialetheic theory proposed in [9], which holds borderline contradictions to always be false. A Hyde-style dialetheist would presumably resort to an error theory of some variety to explain the present results.

people from East Asian cultures are more open to contradictions than are people from Western cultures. These allegations, though, have often used a very wide sense of 'contradiction', much wider than that used here. For example, Peng and Nisbett [15] count all of the following as "tolerating contradictions":

- Preferring compromise to adversarial dispute resolution
- Preferring proverbs like 'too humble is half proud' to proverbs like 'one against all is certain to fall'
- Reconciling 'most long-lived people eat fish or chicken' with 'it's more healthy to be a strict vegetarian'

Clearly, their sense of 'contradiction' is not the sense in play here; so while they may have found a very real cultural difference, their data do not show anything about cultural acceptance of contradictions, in our sense.

In an attempt to connect this cross-cultural research more directly to the philosopher's idea of contradiction, Huss and Yingli [8] ran a cross-cultural study that asked participants in Canada and China for their responses to more paradigm contradictions: the liar paradox, a reductio argument, and most importantly for my purposes here, a borderline contradiction. In particular, they presented their participants with a vignette describing a borderline case of 'raining', and asked about the sentence 'It's raining and it's not raining'.

Despite the narrower focus, the results they found were broadly in line with Peng and Nisbett's research: Huss and Yingli's Chinese participants were much more willing to agree with the contradictions they saw than were their Canadian counterparts. This suggests that cultural differences matter for agreement with contradictions, in particular borderline contradictions. One possibility is that Westerners hold a cultural norm against agreeing with contradictions.[26]

Suppose this to be true. Then, despite their linguistic competence pushing them to accept the borderline contradictions, subjects in the present experiment (as well as Canadian subjects in Huss and Yingli's study) may well have had their assent reduced by cultural norms. The effect would be much the same if we were to ask participants for their grammatical (rather than semantic) intuitions about sentences like 'Which table did you leave the book on?'; although ending a sentence with a preposition is perfectly grammatical in English, the cultural norm against it may well drive participants to reduce their judgments of grammaticality.[27]

If it is true that Westerners have a cultural aversion to contradictions in general, we should expect the levels of assent given by university students in North Carolina to be somewhat lower than what would be generated purely by their linguistic competence; once we take this into account, the dialetheist has a straightforward explanation for the middling levels of assent. So it seems that the dialetheist has a plausible explanation for the observed results as well.

[26] Note that if contextualism of the sort described in §3.1 is right, Huss and Yingli's sentence was presumably not really interpreted as a contradiction either, at least by those who agreed with it. A contextualist should then probably say that Canadians are more likely to give such a sentence a contradictory reading than Chinese.

[27] See [12] for examples of this sort of response.

As an anonymous referee points out, one could also suppose that East Asians hold a cultural norm pushing in favor of contradictions; or that both Westerners and East Asians hold cultural norms pushing in favor of contradictions, but that the East Asian norm is stronger. Either of these hypotheses gibes with the cross-cultural results, but would not support the dialetheist interpretation of the present study. They might support a fuzzy interpretation or even a purely classical interpretation. Unfortunately, despite the wealth of data on cross-cultural psychology, not much is yet known about how cultural norms relate to contradictions in the sense that's relevant here. More research is called for, to get clearer on what cultural differences exist and how they are arrived at.

4 Conclusions

When it comes to (apparent) borderline contradictions, then, it seems that the nonindexical contextualist and the dialetheist offer the two most plausible explanations of the observed results. Before I close, I want to draw some attention to the similarities between these views that allow them to succeed where other views do not. I also want to draw attention to just how hard it will be to design an experiment that could distinguish between these theories.

Note that the nonindexical contextualist, to plausibly explain the results of this study, needed to invoke a relatively fine-grained notion of context. In particular, it seems that context must be able to change for a participant who sees nothing different and doesn't move. Context must thus be at least difficult to observe. Now, the nonindexical contextualist I've envisioned sticks to classical logic *at the level of extensions*. But since it's very difficult to tell when we've changed context, this means that the logic of properties we'll use to generate experimental predictions will blur across contexts. And when you blur classical logic in this way, the result is the paraconsistent logic LP. (See [13] and [4] for details and discussion.)

On the other hand, the dialetheist view I defend in [19] holds LP to be the correct logic of vagueness even in a single context.[28] Thus, it could be quite tricky to find an experimental wedge between the two views. The key to such a wedge would come from some operationalization of the notoriously slippery term 'context'. The contextualist and the dialetheist make different predictions about what will happen in a single context. I leave this issue for future work.

References

1. Åkerman, J., Greenough, P.: Vagueness and non-indexical contextualism. In: New Waves in the Philosophy of Language. Palgrave Macmillan, Oxford (2010)
2. Alxatib, S., Pelletier, J.: The psychology of vagueness: Borderline cases and contradictions. Mind and Language (201x) (forthcoming)

[28] NB: the dialetheist is under no obligation to use a fine-grained context, although she might find reason to.

3. Bonini, N., Osherson, D., Viale, R., Williamson, T.: On the psychology of vague predicates. Mind and Language 14, 377–393 (1999)
4. Brown, B.: Yes, Virginia, there really are paraconsistent logics. Journal of Philosophical Logic 28, 489–500 (1999)
5. Eklund, M.: What vagueness consists in. Philosophical Studies 125(1), 27–60 (2005)
6. Fara, D.G.: Shifting sands: An interest-relative theory of vagueness. Philosophical Topics 28(1), 45–81 (2000) (Originally published under the name "Delia Graff")
7. Fine, K.: Vagueness, truth, and logic. Synthèse 30, 265–300 (1975)
8. Huss, B., Yingli, Y.: Ethnologic: Empirical results (2007); Presented at the 2007 meeting of the Canadian Philosophical Association (Saskatoon, SK)
9. Hyde, D.: From heaps and gaps to heaps of gluts. Mind 106, 641–660 (1997)
10. Kaplan, D.: Demonstratives. In: Almog, J., Perry, J., Wettstein, H.K. (eds.) Themes From Kaplan, pp. 481–564. Oxford University Press, Oxford (1989)
11. Keefe, R.: Theories of Vagueness. Cambridge University Press, Cambridge (2000)
12. Labov, W.: When intuitions fail. In: McNair, L., Singer, K., Aucon, M. (eds.) Papers from the Parasession on Theory and Data in Linguistics, Chicago Linguistic Society, vol. 32, pp. 77–106 (1996)
13. Lewis, D.: Logic for equivocators. Noûs 16(3), 431–441 (1982)
14. MacFarlane, J.: Nonindexical contextualism. Synthèse 166, 231–250 (2009)
15. Peng, K., Nisbett, R.: Culture, dialectics, and reasoning about contradiction. American Psychologist 54, 741–754 (1999)
16. Priest, G.: Logic of paradox. Journal of Philosophical Logic 8, 219–241 (1979)
17. Raffman, D.: Vagueness and context-relativity. Philosophical Studies 81, 175–192 (1996)
18. Richard, M.: Attitudes in context. Linguistics and Philosophy 16, 123–148 (1993)
19. Ripley, D.: Sorting out the sorites. In: Berto, F., Mares, E., Tanaka, K. (eds.): Paraconsistent Logic (tentative title) (201x) (forthcoming)
20. Serchuk, P., Hargreaves, I., Zach, R.: Vagueness and use: Four experimental studies on vagueness. Mind and Language (201x) (forthcoming)
21. Shapiro, S.: Vagueness in Context. Oxford University Press, Oxford (2006)
22. Smith, N.J.J.: Vagueness and Degrees of Truth. Oxford University Press, Oxford (2008)
23. Soames, S.: Understanding Truth. Oxford University Press, Oxford (1998)
24. Soames, S.: Replies. Philosophy and Phenomenological Research 65(2), 429–452 (2002)
25. Sorensen, R.: Vagueness and Contradiction. Oxford University Press, Oxford (2001)
26. Stanley, J.: Context, interest-relativity, and the sorites. Analysis 63(4), 269–280 (2003)
27. Tye, M.: Sorites paradoxes and the semantics of vagueness. Philosophical Perspectives 8, 189–206 (1994)
28. Williamson, T.: Vagueness. Routledge, London (1994)
29. Wright, C.: On the coherence of vague predicates. Synthèse 30, 325–365 (1975)

Notes on the Comparison Class

Stephanie Solt*

Zentrum für Allgemeine Sprachwissenschaft, Berlin
solt@zas.gwz-berlin.de

Abstract. This paper investigates the role of comparison classes in the semantics of gradable adjectives in the positive form, focusing on the case where the comparison class is expressed overtly via a *for*-phrase (e.g. *John is tall for a jockey*). Two central questions are addressed: what information does the comparison class provide, and how is this information integrated compositionally? It is shown that the standard of comparison invoked by the positive form can be analyzed as a range of values whose width is based on the degree of dispersion in the comparison class. Compositionally, the comparison class can be analyzed as an argument of a null positive morpheme (contra Kennedy [13]), in parallel to recent proposals for the superlative (e.g. Heim [9]). The implications of the analysis for the choice between degree- and delineation-based analyses of gradable adjectives are discussed.

1 Introduction

A long tradition (Bartsch & Vennemann [3], Cresswell [5], Klein [14], von Stechow [17], Fults [8], van Rooij [16]) holds that sentences involving vague predicates, such as those in (1), should be analyzed with reference to a **comparison class** that in some way serves to provide a frame of reference or standard of comparison. For example, (1-a) might be interpreted as saying that Fred's height exceeds the standard for some set of individuals of which Fred is a member (adult American men, 8-year-old boys, basketball players, etc).

(1) a. Fred is tall
b. Sue's apartment is expensive
c. George doesn't have many friends

This view is made more plausible by the fact that the comparison class may apparently be made overt via a *for*-phrase, as in (2):

* My thanks to the reviewers for the workshop Vagueness in Communication, and for this volume, for their extensive and helpful comments. Thanks also to the audience of Vagueness in Communication, whose questions and observations have helped me clarify my thinking on this topic. All errors and remaining weaknesses are of course my own. Support for this research was provided by the European Science Foundation (ESF) and the Deutsche Forschungsgemeinschaft (DFG) under the auspices of the EUROCORES programme LogICCC.

R. Nouwen et al. (Eds.): ViC 2009, LNAI 6517, pp. 189–206, 2011.

(2) a. Fred is tall for an eight-year-old
b. Sue's apartment is expensive for an apartment on this street
c. For a politician, George doesn't have many friends

Results from psycholinguistic experimentation further support the reality of the comparison class in the interpretation of vague adjectives - even among children. Barner & Snedeker [2] presented 4-year-old children with a collection of doll-like objects of varying heights, which were given the novel name 'pimwits'. When asked to identify which were the tall and short pimwits, children classified roughly the tallest third of the array as 'tall' and the shortest third as 'short'. But when the distribution of heights of the objects was changed (i.e. by adding more tall or short pimwits), children's standards for *tall* and *short* changed correspondingly, indicating that the statistical properties of the comparison class provided were used in determining the extensions of these words.

The present paper takes the notion of a comparison class as a starting point, and addresses the question of how the standard of comparison is set relative to the comparison class. That is, what information does the comparison class provide, and how does this enter into the semantic representation? Here, I will focus in particular on examples featuring overt *for*-phrases, such as in (2), and argue that the same treatment can be extended to the corresponding bare cases, as in (1).

I approach these questions within a degree-based framework, according to which the truth conditions of sentences involving gradable adjectives are expressed in terms of relationships between degrees on a scale associated with some dimension of measurement (see especially Cresswell [5], as well as later work in this tradition such as Heim [10] and Kennedy [12,13]). This approach could perhaps be considered the current standard in the analysis of gradability and vagueness, and with some good justification. It is first of all widely accepted that semantics must at least sometimes make reference to the notion of degrees, the classic case of this being examples where degrees are explicitly mentioned, such as *Fred is 1,8 meters tall* or *John is 5 cm taller than Fred*. By adopting a degree-based framework more generally, it is possible to give a unified analysis to cases such as these as well as those where degrees are not mentioned (e.g. *Fred is taller than John*). Beyond this, degree-based approaches have been shown to allow the compositional analysis of a wide range of degree modifiers, including *very*, *too*, the comparative morpheme *-er*, the equative *as*, measure phrases, and others.

But the unmodified or positive form of gradable adjectives, where there is no overt degree morphology, poses a bit of a challenge to degree theories. Developing an adequate treatment of the positive form is necessary to establish the general applicability of what has proved to be an otherwise very fruitful approach to the analysis of gradablility, and a number of authors have tackled this problem (including Cresswell [5], von Stechow [18], Fults [8], Kennedy [13], Rett [15]). The present work is intended to contribute to this line of research.

Degree-based frameworks are not, however, the only option for the semantic treatment of gradable adjectives. A leading alternative is the delineation-based approach of Klein [14], in which the semantics of gradable expressions are stated in terms of relationships between individuals, not degrees. Though it is not my primary goal here, at the end of the paper I will briefly contrast how degree- and delineation-based theories fare with respect to the data discussed here, and consider the implications of the present analysis for the choice between these two approaches.

2 Comparison Classes and Standards

2.1 Standard as Range

Following Cresswell [5] and others, I take gradable adjectives to express relationships between individuals and degrees (3). As a first attempt, let us then imagine that in the case of the positive (unmodified) form of the adjective, a comparison class provides a standard of comparison in the form of a standard degree $d_{Std:C}$ that saturates the first (degree) argument of the gradable adjective (4)[1]:

(3) $[\![\text{tall}]\!] = \lambda d \lambda x.HEIGHT(x) \geq d$

(4) $[\![\text{Fred is tall for an 8-year-old}]\!] = 1$ iff $HEIGHT(fred) \geq d_{Std:8.yr.olds}$

How might $d_{Std:C}$ be determined?

A straightforward possibility suggested in the early literature on the topic (e.g. Cresswell [5]) is that the standard is an average over the comparison class. But Kennedy [13] points out that matters cannot be as simple as this, in light of the felicity of examples such as (5).

(5) Nadia's height is greater than the average height of gymnasts, but she still isn't tall for a gymnast

Based on (5) it seems that $d_{Std:C}$ would need to be a degree greater than the average; but it is not at all clear what this degree should be.

Taking the standard of comparison to be an average (or any other single point) provided by the comparison class also raises questions as to the proper treatment of positive/negative antonym pairs such at *tall* and *short*. It seems that pairs such as (6-a), (6-b) are interpreted with reference to the same comparison class (either overt or covert).

(6) a. Fred is tall (for an 8-year-old)
b. Fred is short (for an 8-year-old)

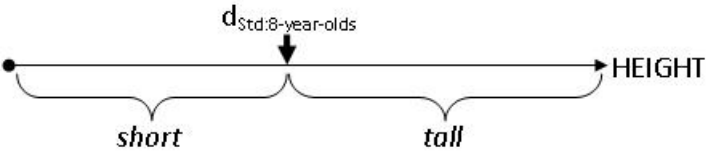
[1] Here I am in particular considering what Kennedy [13] calls relative gradable adjectives. Absolute gradable adjectives such as *straight* and *dry*, whose standards appear to reference endpoints on a scale, and which typically do not occur with *for*-phrases, exhibit different properties. I do not attempt to treat this class here.

Suppose, as is commonly done, that the entry for the negative antonym is identical to that for the positive antonym, with the exception that the 'greater than or equal to' operator is replaced by 'less than or equal to':

(7) $[\![\text{short}]\!] = \lambda d \lambda x.HEIGHT(x) \leq d$

If we take the standard of comparison in both cases to be a single point $d_{Std:C}$ provided by the comparison class, as in (8), the positive and negative antonyms are then defined essentially as complementaries, dividing the semantic space completely between them.

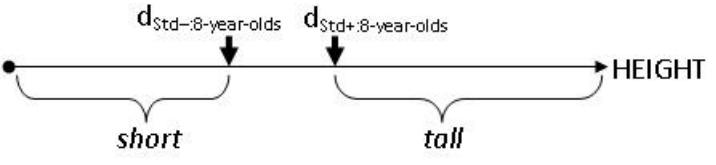
(8)



Intuitively, pairs such as *tall* and *short* are instead contraries, in that there is a range of heights for which both (6-a) and (6-b) would be judged false (Cruse [6]). This is of course supported by the felicity of conjunctions such as *Fred isn't tall, but he's not short either.*

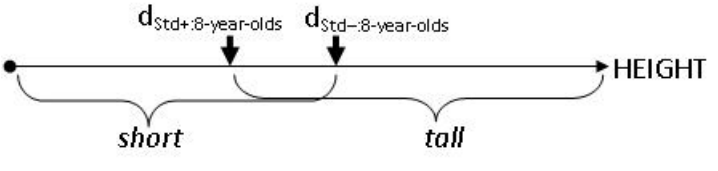
On the other hand, if *tall* and *short* are taken to invoke different standards, as in (9), these can be set in such a way as to establish a 'gap' between the positive and negative antonym:

(9)



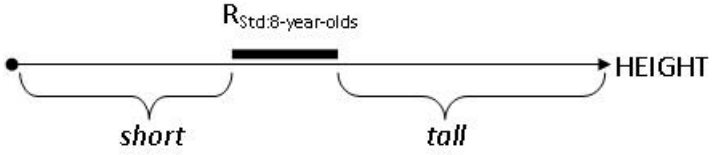
But now we face the more serious question of why these two standards always seem to stand in the same relationship to one another, namely $d_{Std+:C}$ always being higher than $d_{Std-:C}$. Short of stipulation, there is nothing obvious that rules out the possibility that, for some adjectives or in some contexts, the position of the two standards might be reversed, which would allow the truth of a sentence such as *Fred is both tall and short for an 8-year-old.*

(10)



These objections are overcome if the standard of comparison is taken to be a range $R_{Std:C}$ rather than a point, as proposed by von Stechow [18]:

(11)



The range specified in (11) encodes the intuitive gap between the positive and negative antonyms. Furthermore, if the range is taken to contain the average (mean or median) over the comparison class, we have an explanation for examples such as (5) above, in that the lower bound for the positive member of the pair will be higher than the average.

2.2 Standards and Distributions

Support for a range-based standard, and a clue to its relationship to the comparison class, is provided by a brief example. Consider the sentences in (12), based on examples from Kennedy [13].

(12) a. Sue's apartment is expensive for an apartment on this street
b. Paul's apartment is inexpensive for an apartment on this street

Suppose that it is the case that Sue's rent is 800 €, Paul's rent is 600 €, and the median rent on this street is 700 €. Are the sentences in (12) true? The answer, I believe, is that it depends. Specifically, it depends on the amount of variation in the rents of apartments on the street in question. If the vast majority of apartments on this street rent for between 650 € and 750 €, we are likely judge both (12-a) and (12-b) to be true, given that both Sue's and Paul's rents fall outside of this typical range (Sue's on the high side, Paul's on the low side). But now suppose that there is greater variation in the rents on the street (say, rents anywhere between 500 € and 1000 € are common). Then it seems that (12-a), (12-b) would no longer be judged true, despite the fact that neither the average rent nor the values corresponding to Sue and Paul have changed.

This example demonstrates that the comparison class provides statistical information that serves to determine the thresholds for adjectives such as *expensive* and *inexpensive*. Specifically, what is relevant is not only a central value, but also some measure of the extent of dispersion of values corresponding to members of the comparison class.

Returning to the previously introduced idea of a standard as a range of degrees, we can now be more explicit. The standard range R_{Std} can be defined as a central range whose width is dependent on the degree of dispersion in the comparison class. This may be formalized by borrowing two statistical concepts, the median and the median absolute deviation (MAD); the latter is a measure of dispersion around a median, parallel to the standard deviation as a measure of deviation around a mean.

For the examples in (12), we then have the following:

(13) $R_{Std:apt.on.this.street} = median_{x:apt.on.this.street(x)}COST(x)$
$\pm n \bullet MAD_{x:apt.on.this.street(x)}COST(x)$

Here R_{Std} is defined as a range around the median value over the comparison class, in this case apartments on this street. If we imagine the measures of members of the comparison class to be normally distributed, then R_{Std} corresponds to the central peak of the bell curve, and will be narrower or wider depending on how peaked or flat that curve is.

We could perhaps have used the mean rather than the median, but means are more sensitive to extreme high or low values. My own intuition is that it is the distribution of apartments, and not the distribution of prices, that determines the truth or falsity of examples such as (12); this makes a median more appropriate than a mean.

In (13), the parameter n reflects indeterminacy in the number of cases whose values are within R_{Std}. In a symmetric distribution, the central fifty percent of cases fall within one MAD of the median. If we are satisfied in letting *expensive (for an apartment on this street)* pick out the highest-priced quartile of apartments, and *inexpensive (for an apartment on this street)* pick out cheapest quartile, then n may be set at 1. If we wish to allow a greater number of cases to count as *expensive/inexpensive*, then n must be set at some value less than one. The latter seems to me intuitively correct. Recall also that the children in Barner & Snedeker's [2] study typically labeled the tallest third of objects as *tall*, which similarly would imply a value less than one for n.

The more general case is the following, where MEAS stands for a measure function (a function that relates individuals to degrees on the scale of some relevant dimension).

(14) $R_{Std:C} = median_{x \in C}MEAS(x) \pm n \bullet MAD_{x \in C}MEAS(x)$

The formula in (14) is admittedly complex, and it might be tempting to simplify it to something along the lines of (15), where $R_{Std:C}$ is defined as the median plus or minus some value calculated as a proportion of the median:

(15) $R_{Std:C} = median_{x \in C}MEAS(x) \pm n \bullet median_{x \in C}MEAS(x)$

But this is not adequate. The width of the range in (15) fails to factor in the degree of dispersion in the comparison class. As demonstrated in the discussion of the examples in (12), dispersion in the comparison class is relevant to judgments of truth and falsity. Thus if we want to capture speakers' intuitions on the interpretation of the positive form via the definition of a standard degree or set of degrees, we need to posit an entry along the lines of (14), and not the simpler (15). I will return to this point below.

Finally, it is worth pointing out that the formulation of the standard of comparison developed here is consistent with recent theories of vagueness which define the semantics of vague predicates relative to a 'significant' degree of the

property in question. Fara [7] in particular makes the notion of significance central to her theory of vagueness, proposing for example that *a lot* is interpreted as 'significantly more than some norm', where the norm depending on the situation might be what is expected, what is typical, what is needed or wanted, etc. Kennedy [13] builds on this view by proposing that the positive form of a gradable adjective is true of an individual if it has a sufficient degree of the given property to stand out in the context. Importantly, from a statistical perspective, what qualifies as significant is defined in terms of deviation around a central value, just as has been done in (14) above.

3 The Integration of the *for*-Phrase

3.1 Kennedy (2007) and the Presuppositionality of the *for*-Phrase

In the preceding section I argued that the comparison class introduced by the *for*-phrase provides statistical information on the basis of which a standard of comparison can be calculated. In the present section, I consider the question of how this information is integrated compositionally.

Kennedy [13] makes the important observation that *for*-phrases are presuppositional in nature. As examples, (16) presupposes that Fred is 8 years old (it would be infelicitous if he were older or younger); (17) presupposes that Kyle's car is a Honda; the infelicity of (18) can be attributed to presupposition failure.

(16) Fred is tall for an 8-year-old

(17) Kyle's car is expensive for a Honda

(18) ?? Kyle's BMW is expensive for a Honda

Thus the inclusion of a *for*-phrase actually has two effects semantically: it introduces a comparison class (and the statistical information it provides) while at the same time contributing a presupposition to the resulting sentence.

Kennedy captures this combined role with an analysis that takes the *for*-phrase to introduce a domain restriction on the gradable expression. On this account, gradable adjectives denote measure functions (functions from individuals to degrees); *tall*, for example, denotes a function from individuals to their heights (19). The *for*-phrase composes directly with the gradable adjective by contributing a domain restriction; *tall for an 8-year-old* thus comes to denote a function from 8-year-olds to their heights (20). Finally, a null degree morpheme POS (for 'positive') takes the adjective as argument, and returns a predicate over individuals, which includes a standard of comparison calculated as a function of the gradable adjective. Crucially, in the case involving a *for*-phrase, the standard-calculating function $\boldsymbol{s}$ operates on the domain-restricted expression, thereby incorporating the domain into the standard-setting procedure. Thus *POS tall for an 8-year-old* is a predicate true of an 8-year-old if his or her height exceeds the value that would be considered significant for an 8-year old (21).

(19) $[\![tall_{\langle ed \rangle}]\!] = \lambda x.tall(x)$

(20) $[\![tall\ for\ an\ 8\text{–}year\text{–}old_{\langle ed \rangle}]\!] = \lambda x : 8\ years\ old(x).tall(x)$

(21) $[\![POS\ tall\ for\ an\ 8\text{–}year\text{–}old_{\langle et \rangle}]\!] =$
$= \lambda y.(\lambda x : 8\text{–}years\text{–}old(x).tall(x))(y) > \mathbf{s}(\lambda x : 8\text{–}years\text{–}old(x).tall(x))$

Yet appealing though this approach is, it is less clear how it would deal with examples such as the following:

(22) a. For a lawyer, Bill is poor
b. For a lawyer, Bill has a small salary
c. For a lawyer, Bill is poorly paid
d. For a lawyer, Bill doesn't earn much money

Example (22-a) predictably has the presupposition that Bill is a lawyer, which under Kennedy's analysis could be captured as a domain restriction on the gradable adjective *poor*. But (22-b)–(22-d) share the same presupposition. In these cases this presupposition cannot easily be analyzed as a domain restriction, in that the subject of whom the presupposition holds (Bill) is not an argument of the gradable expression (*small*, *poorly* and *much*, respectively).

One might attempt to get around this problem by proposing (as suggested by Kennedy himself) that the domain restriction need not be identical with the denotation of the nominal in the *for*-phrase, but rather a function of it. In (22-b), for instance, suppose that the domain of *small* is restricted to the salaries of lawyers. Then after further semantic composition, *has a small salary for a lawyer* comes to denote a predicate that is true of an individual if he has a salary which falls within the extension of the domain-restricted predicate *small*; since only lawyers have salaries of lawyers, the result is that the subject (here, Bill) is by presupposition a lawyer.

But this would not suffice to rescue an example such as the following:

(23) Sara reads difficult books for an 8-year old

Here we have the characteristic presupposition that Sara is eight years old. Suppose that we again attempt to capture this with a restriction on the domain of *difficult* to books read by eight-year-olds. The sentential predicate is thus true of an individual if she reads books which are in the extension of the domain-restricted *difficult* (which by presupposition are books that are read by eight-year-olds). But here, there is no resulting presupposition that the subject herself is eight years old, the key fact being that one does not need to be eight years old to read books read by eight-year-olds.

A different sort of problem is posed by examples such as these:

(24) a. The store is crowded for a Tuesday
b. For a Sunday, there aren't many cars in the parking lot

Here we seem to have comparison classes of times, and correspondingly presuppositions on the time of utterance. For example, (24-a) presupposes that

the time of utterance is a Tuesday (as evidence, it would be infelicitous if uttered on a Friday). We might seek to capture this with a domain restriction over times:

(25) $[\![crowded\ for\ a\ tuesday_{\langle i,ed\rangle}]\!] = \lambda t : tuesday(t)\lambda x.crowded(x)(t)$

But it is not clear how the standard-setting function $\boldsymbol{s}$ could pick out a significant degree $\mathbf{s}(\lambda t : tuesday(t)\lambda x.crowded(x)(t))$ independently of the type e entity of which *crowded* is predicated.

Note also that the *for*-phrase can - and in some cases, must - be separated from the gradable expression, unexpected if they form a constituent:

(26) a. (For an 8-year-old,) Fred is tall (for an 8-year-old)
b. (For an amateur,) Martha is a good (*for an amateur) golfer (for an amateur)
c. (For a politician,) George doesn't have many (*for a politician) friends (for a politician)

Finally, on Kennedy's account, where the *for*-phrase composes with the gradable adjective before the latter combines with degree morphology, it is to be expected that *for*-phrases could occur with any degree modifier, and not only with the null positive morpheme POS. But examples such as the following are ungrammatical:

(27) a. *Fred is taller than Sam for an 8-year-old
b. *Fred is as tall as Sam for an 8-year-old
c. *Fred is that tall for an 8-year-old
d. *Fred is 1,2 m tall for an 8-year-old

To be certain, *for*-phrases are not exclusively limited to occurring with the adjective in its positive form. Bale [1] discusses examples of *for*-phrases in comparatives such as (28) where the adjective or the comparison class differ between main clause and *than*-phrase:

(28) a. ?Fred is taller for a boy than he is wide for a boy
b. John is taller for a man than Mary is for a woman

For-phrases are also at least marginally acceptable with *too* (e.g. *?Fred is too tall for a jockey*). But these further examples reinforce that the felicitous occurrence of a *for*-phrase is dependent on the degree morpheme and the rest of the degree construction (e.g., the nature the of the *than*-clause).

In short, once we consider a broader range of examples, it becomes clear that the *for*-phrase cannot be analyzed as introducing a domain restriction on the gradable adjective. But then what alternative will capture its dual role as standard-setter and presupposition trigger?

3.2 The *for*-Phrase and POS

In developing an alternate compositional analysis of the *for*-phrase, it is helpful to consider some parallel cases of degree constructions where a phrasal constituent serves to specify a threshold degree. First, in their tendency to be

extraposed (cf. (26)), *for*-phrases behave a lot like *than*-phrases in comparatives and *as*-phrases in equatives:

(29) a. Martha is a better (*than George) golfer (than George)
b. Martha is as good (*as George) a golfer (as George)

It is common to analyze *than*-phrases as arguments of the comparative morpheme *-er*. Bhatt & Pancheva [4] argue that their seemingly extraposed position marks the scope of *-er*, in that the *than*-phrase is merged counter-cyclically after *-er* has raised from its base-generated position. In the case of the positive (unmodified) form of the adjective, there is of course no overt degree morphology. But a tradition going back to Cresswell [5] holds that the semantics of the positive form involves a phonologically null degree morpheme POS (cf. the discussion of Kennedy's [13] analysis in Section 3.2; see also von Stechow [18], Heim [11], Fults [8]). The parallel between *for*-phrases on the one hand and *than*- and *as*-phrases on the other thus suggests that the *for*-phrase might similarly mark the scope of, and be interpreted in relation to, null POS.

Kennedy [13] argues against analyzing the *for*-phrase as an argument of POS, citing among other reasons that this does not explain its presuppositional behavior. But in this respect there is a relevant parallel in the superlative, which exhibits presuppositions very similar to those discussed here. On one reading (the so-called relative reading; see especially Szabolcsi [19], Heim [9]), superlatives such as those in (30) are interpreted as conveying that the subject has a higher degree of the property in question than any other member of some contextually relevant comparison class. The comparison class may optionally be made explicit with an *of*-phrase, as in (31):

(30) a. Fred is the tallest student
b. John read the longest book
c. Sue's apartment is the most expensive
d. George has the fewest friends

(31) a. Fred is the tallest of the students in the second grade class
b. John read the longest book of anyone in the class
c. Of the apartments on this street, Sue's is the most expensive
d. George has the fewest friends of any politician I know

Importantly, just as in the case of the *for*-phrase, the subject in the superlative examples is presupposed to be a member of the comparison class (regardless of whether or not this comparison class is made overt). For example, Fred in (31-a) is by presupposition a student in the second grade class; George in (31-d) is a politician I know, and so forth.

Heim [9] proposes that the superlative morpheme *-est* takes a covert comparison class argument, as reflected in the the following entry, where C is a variable over comparison classes, P denotes a relationship between degrees and individuals (type $\langle d, et \rangle$), and x is by presupposition an element of C:

(32) $[\![\text{-}est]\!] = \lambda C_{\langle et\rangle}\lambda P_{\langle d,et\rangle}\lambda x : x \in C.\exists d[P(x,d)\wedge$
$\forall y[y \neq x \wedge y \in C \rightarrow \neg P(y,d)]]$

Building on Heim's approach, a parallel entry can be proposed for the null morpheme POS, in which it takes a comparison class as argument, and introduces the standard R_{Std} whose definition was developed in the previous section:[2]

(33) $[\![POS]\!] = \lambda C_{\langle et\rangle}\lambda P_{\langle d,et\rangle}\lambda x : x \in C.\forall d \in R_{Std:C}\,[P(x,d)]\,,$

where $R_{Std:C} = median_{y\in C}(max(d)\,[P(y,d)])$
$\pm n \bullet MAD_{y\in C}(max(d)\,[P(y,d)])$

In cases with an overt *for*-phrase, I take this to provide the comparison class argument; this implies that *for* itself is semantically inert (though see below for an alternate possibility). As for how the *for*-phrase, and thus the comparison class, is compositionally integrated, I follow Bhatt & Pancheva's [4] analysis of the comparative morpheme *-er* in proposing that POS originates in the specifier position of the gradable adjective, but raises to a position right-adjoined to VP, at which point the *for*-phrase is merged in its specifier position. (On this analysis, the base position of the *for*-phrase is at the right edge of the VP; when it occurs sentence-initially, this is the result of further movement.)

For a simple example such as (34-a), we then have the LF in (34-b); the semantic derivation proceeds as in (35):

(34) a. Fred is tall for an 8-year-old
b. Fred $[_{VP}[_{VP}$ is t_i tall] $[_{DegP}$ POS_i [for an 8-year-old]]]

(35) $[\![is\ t_i\ tall]\!] = \lambda d\lambda x.HEIGHT(x) \geq d$

$[\![POS_i\ for\ an\ 8\ year\ old]\!] =$
$= \lambda P_{\langle d,et\rangle}\lambda x : 8.year.old(x).\forall d \in R_{Std:8.year.olds}\,[P(x,d)]$

$[\![is\ t_i\ tall\ POS_i\ for\ an\ 8\ year\ old]\!] =$
$= [\![POS_i\ for\ an\ 8\ year\ old]\!]\,([\![is\ t_i\ tall]\!])$
$= \lambda x : 8.year.old(x).\forall d \in R_{Std:8.year.olds}\,[HEIGHT(x) \geq d]\,,$

where $R_{Std:8.year.olds} =$
$= median_{y:8.year.old(y)}(max(d)\,[HEIGHT(y) \geq d])$
$\pm n \bullet MAD_{y:8.year.old(y)}(max(d)\,[HEIGHT(y) \geq d])$

On this account, *tall for an 8-year-old* is a predicate true of an 8-year-old if his height exceeds the median plus n•MAD in height over all 8-year-olds. While this is little different from what would obtain if the *for*-phrase were analyzed as a domain restriction on the adjective *tall*, a difference emerges when we consider cases in which the subject of the presupposition is not an argument of

[2] Fults [8] similarly concludes on the basis of syntactic and semantic tests that the *for*-phrase is an argument of POS.

the gradable adjective; recall that these cases were problematic for Kennedy's analysis. The crucial aspect of the present approach is that the presupposition is defined on the type e argument of POS, and not on the argument of the gradable adjective. To return to an earlier example, we have the following:

(36) a. Sara reads difficult books for an 8-year-old
b. Sara $[_{VP}[_{VP}$ reads t_i difficult books$]$ $[_{DegP}$ POS_i [for an 8-year-old]]]

(37) $[\![reads\ t_i\ difficult\ books]\!] = \lambda d \lambda x.DIF(books\ read\ by\ x) \geq d$

$[\![reads\ t_i\ difficult\ books\ POS_i\ for\ an\ 8\ year\ old]\!] =$
$= \lambda x : 8.year.old(x).\forall d \in R_{Std:8.year.olds}\,[DIF(books\ read\ by\ x) \geq d]\,,$

where $R_{Std:8.year.olds} =$
$= median_{y:8.year.old(y)}(max(d)\,[DIF(books\ read\ by\ y) \geq d])$
$\pm n \bullet MAD_{y:8.year.old(y)}(max(d)\,[DIF(books\ read\ by\ y) \geq d])$

Here we derive a predicate that is true of an 8-year-old if the difficulty of books he or she reads exceeds the median plus n•MAD over 8-year-olds in difficulty of books read. Thus just as in the case above, the presupposition that Sara is 8 years old is captured.

Finally, if the last argument of POS is allowed to range over times as well as over individuals, we can also accommodate examples such as (24), where we have a comparison class over times, and a corresponding presupposition regarding time of utterance. Here, it is assumed that POS raises to a higher position, immediately before the integration of the time argument:

(38) a. The store is crowded for a Tuesday
b. $[_{XP}\,[_{XP}$ The store is t_i crowded$]$ $[_{DegP}POS_i$ [for a tuesday]]]

(39) $[\![the\ store\ is\ t_i\ crowded\ POS_i\ for\ a\ tuesday]\!] =$
$= \lambda t : tuesday(t).\forall d \in R_{Std:tuesdays}[CROWDED(store)\ at\ t \geq d]$

where $R_{Std:tuesdays} =$
$= median_{t':tuesday(t')} max(d)[CROWDED(store)\ at\ t' \geq d]$
$\pm n \bullet MAD_{t':tuesday(t')} max(d)[CROWDED(store)\ at\ t' \geq d]$

Here, as in the previous cases, R_{Std} is defined in terms of median and MAD over the comparison class; the only difference is that the comparison class in this case is a set of times, such that R_{Std} represents a central range of degrees of crowdedness of the store on Tuesdays.

To summarize this section, modeling the analysis of the positive form on Heim's [9] analysis of the superlative allows a compositional analysis of *for*-phrases that captures the dual role of the comparison class they introduce: determining a standard of comparison, and introducing a presupposition. Furthermore, this analysis is able to handle cases not accounted for by Kennedy [13].

Note in conclusion that I have been maintaining the now-standard view of POS as phonologically null. An alternate possibility is that in cases with an overt

for-phrase, *for* is actually the spell-out of POS, with the noun phrase following *for* introducing the comparison class argument. Such an analysis would simplify the constituency (in that the first argument of *for*/POS would occur in a linearly adjacent position), and would eliminate the need to posit counter-cyclic merger along the lines of Bhatt & Pancheva [4]. I leave it as an open question as to whether this is the correct analysis; the fundamentals of the account developed above remain unchanged either way.

3.3 No *for*-Phrase

Up to this point I have been considering examples with overt *for*-phrases. Let us consider briefly the (more common) situation where a gradable adjective occurs in the positive form without a *for*-phrase to introduce a comparison class. In these cases it is reasonable to assume that there is an implicit comparison class that saturates the C argument of POS (cf. Heim [9] for a similar analysis in the case of the superlative).

Often the context is sufficient to specify the appropriate comparison class. For example, if an elementary school teacher remarks, upon meeting a new pupil, *Fred is tall*, the natural interpretation is that he is tall relative to boys of his age. But in other cases the context leaves multiple possibilities open, and in this case disagreement between speakers can arise. We might, for example, disagree as to whether a particular apartment could be called expensive, the source of the disagreement being that we have different frames of reference or comparison classes in mind (for example, apartments on this street vs. apartments rented by students). Presumably this is not all that common in practice, in that we seem to be able to understand each other without too much trouble when we use gradable adjectives.

In the literature on comparison classes (e.g. Klein [14]), they have typically been conceptualized as sets of individuals. But consideration of examples such as *the store is crowded for a Tuesday* has led us to broaden the view of comparison classes to include also sets of times. With this expanded view, sentences that at first do not seem to lend themselves to a comparison class analysis in fact are amenable to this approach. For example, *the store is crowded today* has a reading (probably the most natural one) that does not involve the comparison of 'the store' to other locations, but rather a comparison of 'today' to other days. This reading cannot be captured via a traditional view of a comparison class over individuals, but as shown above in (38) can be handled with a comparison class over times.

There may be other possibilities as well. Fara [7] cites the following example: I am throwing a huge party, and my refrigerator is full of beer I have bought for the guests. My friend looks in the refrigerator and exclaims "Wow, that's a lot of beer." Fara proposes that *a lot of beer* can be interpreted as 'significantly more beer than one typically finds in a refrigerator'. But this intuition could be restated using the language of comparison classes: considering situations of refrigerators stocked with beer, the present case is at the high end in terms of

amount of beer. In other words, we also seem to have comparison classes over something like situations.

Finally, there is a technical point that requires discussion. In the examples discussed in (34)–(38), POS is interpreted with wider scope than its base-generated position, the result of raising at LF. But with the semantics given in (33), once the comparison class argument of POS is saturated, it could combine *in situ* with a gradable adjective:

(40) $[\![\text{C-POS tall}]\!] = [\![\text{C-POS}]\!]([\![\text{tall}]\!])$
$= \lambda P \lambda x : x \in C . \forall d \in R_{Std:C} [P(x, d)](\lambda d \lambda x . HEIGHT(x) \geq d)$
$= \lambda x : x \in C . \forall d \in R_{Std:C} [HEIGHT(x) \geq d]$

The existence of this possibility gives rise to two questions. First, why does POS raise at all, given that it may be interpreted *in situ*? And secondly, recall from (26) that a *for*-phrase often cannot occur directly adjacent to the gradable expression; this is the case in particular with modified nominals, where the *for*-phrase appears either to the right of the noun or sentence initially:

(41) a. (For an 8-year old,) Fred is a tall (*for an 8-year old) boy
b. (For an amateur,) Martha is a good (*for an amateur) golfer (for an amateur)

If the *for*-phrase marks the semantic scope of POS, and POS can be interpreted locally to the gradable adjective, why do we not find a *for*-phrase in this position?

There is an obvious possibility here: when POS is interpreted *in situ*, the modified nominal provides the comparison class. This is assumed by Cresswell [5]. And by way of parallel, Heim [9] makes a similar claim about the superlative. Recall that the relative reading of the superlative discussed above involves the superlative morpheme *-est* taking scope outside of the DP; in this case, the comparison class is provided by the context, or by an *of*-phrase. For example, on the relative reading *John climbed the highest mountain* means that he climbed a higher mountain than did any other member of some contextually salient group. But *-est* can also remain within the DP, and in this case the comparison class is equated to the denotation of the modified nominal. The result is what Heim terms the absolute reading. For example, on the absolute reading of *John climbed the highest mountain*, the comparison class is mountains, and the resulting meaning is that John climbed the highest mountain of all, i.e. Mt. Everest.

We might propose a similar story in the case of POS: when it is interpreted *in situ*, the comparison class is set equal to the modified nominal, making a *for*-phrase superfluous. The issue with this, as pointed out by Kennedy [13], is that modified nominals do not exhibit the same presuppositional behavior as *for*-phrases. For example, (42-a) is a presupposition failure; but (42-b) is not:

(42) a. ?? That's not large for a mouse. It's a rat.
b. That's not a large mouse. It's a rat.

While I do not have a full explanation for this difference, there are a couple of relevant observations that can be made.

First, in the case of a modified nominal, the superlative also allows a non-presuppositional reading; for example, (43) does not seem to involve presupposition failure:

(43) Pluto isn't the smallest planet. It's not a planet at all.

And secondly, while (42-b) demonstrates that something like *is a large mouse* can have a non-presuppositional reading, it does not rule out the possibility of a presuppositional reading as well, the one predicted if the nominal saturates the comparison class argument of POS. For example, on the readings brought out by the continuations given, all of the examples in (44) can be analyzed as presupposing that 'that' is a mouse.

(44) a. That's a large mouse.
b. That's not a large mouse. It's fairly average in size as mice go.
c. Is that a large mouse? I don't know anything about how big mice get.

The source of the non-presuppositional readings in examples such as (42-b) and (43) is not clear. One possibility is that these examples should be analyzed as something other than predicational in nature. I will not attempt to pursue this here. But to address the issue raised above, I conclude tentatively that in the case of modified nominals, POS can in fact occur *in situ* with the nominal saturating the comparison class argument slot; this restricts the occurrence of a *for*-phrase to cases where POS has higher scope, and thus the comparison class must be specified in some other way.

4 Degrees and Individuals

The primary goal of this paper has been to explore the role of comparison classes in the semantics of gradable adjectives in their positive form, focusing on what information the comparison class contributes to the truth conditions, and how this may be given a formal, compositional implementation within a degree-based semantics. I have shown that the positive form can receive an analysis modeled on recent treatments of the superlative, in which the comparison class (either provided by a *for*-phrase or contextually supplied) is an argument of a null degree morpheme POS, and provides as a standard of comparison a range of degrees around a central value. The parallel between the positive and the superlative is a meaningful one, in that it suggests that there is not such a fundamental difference between the case where there is overt degree morphology (the superlative) and the case where there is not (the positive). That is, within a degree-based framework, the positive form does not require any exceptional treatment beyond the sort required for other types of degree constructions.

But there are a couple aspects of the present analysis that might be criticized as not entirely satisfying. First, as discussed Section 2, the definition of the

standard range R_{Std} must be stated in fairly complex terms, incorporating two statistical measures (the median and the median average deviation) as well as the parameter n. And secondly, in the formalization developed in Section 3, it must be stipulated in the semantics for the positive morpheme POS that the subject of the resulting predicate is a member of the comparison class. To be certain, this stipulation has a parallel in a similar restriction on the superlative, and it is reasonable to think the two patterns are related. But why things should be like this is less clear.

I would like to suggest that both issues stem from the same source. The analysis of the positive form developed above states its truth conditions in terms of relationships between degrees. But to get the facts right, we actually need to keep track of individuals. That is, to appropriately set the thresholds for the application of a gradable adjective such as *expensive* or *tall*, it is necessary to factor in how the individuals in the comparison class are distributed with respect to the dimension in question (e.g. cost or height): are they clustered closely together, or more dispersed? And we further need to establish that a particular individual (in most of our examples, the sentential subject) is a member of the comparison class. The complexities and stipulations discussed above are simply what is needed within a degree-based framework to establish these relationships between individuals. But perhaps these complications could be avoided if the truth conditions of sentences involving gradable adjectives in their positive form were stated in terms of individuals, and not degrees.

As discussed briefly in the introductory section of this paper, this is precisely at the core of the delineation-based approach to vagueness and gradability, a leading alternative to the degree-based framework. In the seminal work in this tradition, Klein [14] proposes that gradable adjectives denote simple one-place predicates (i.e., without any sort of degree argument), which differ from ordinary one-place predicates in being partial functions which are dependent on a contextually determined comparison class. Thus a gradable adjective such as *tall* partitions the comparison class into three disjoint sets: a positive extension, a negative extension and an extension gap (individuals of whom the predicate is neither true false).

It is beyond the scope of this paper to develop an alternative delineation-based analysis of the facts discussed here. But to sketch out in rough form what such an analysis might look like, we might follow Klein's general approach and take a gradable adjective and antonym pair to be interpreted relative to a comparison class, and to partition that set into three subsets.[3] For example:

(45) a. $[\![tall(C)]\!] = \{x : x \textit{ is one of the tallest } C's\}$
b. $[\![short(C)]\!] = \{x : x \textit{ is one of the least tall } C's\}$
c. $GAP = C - ([\![tall(C)]\!] \cup [\![short(C)]\!])$

[3] Klein states his analysis in terms of adjectives and their negations. He does not explicitly discuss the relationship between the negation of an adjective (e.g. *not tall*) and its antonym (e.g. *short*). This in itself is an interesting topic for further study. But as it is the latter that is relevant to the facts discussed here, I modify Klein's treatment somewhat.

If we establish some rough standard for what proportion of the comparison class must fall within each of these three sets (say, one third in each), we capture the intuition that the degree of dispersion in the comparison class is relevant to what counts as tall or short (cf. the discussion of example (12)).

An analysis along the lines of (45) would eliminate the need to build complex statistical measures into the semantics of the positive form. Beyond this, it would have the benefit of capturing with one mechanism the dual semantic role of the comparison class. First, the comparison class in essence sets the standard of comparison, because it is the members of this set that are being sorted or grouped. And secondly, the subject must be a member of that comparison class (e.g. only an 8-year-old can appear within a ranking of 8-year-olds).

The possibility of giving a simpler and more intuitive analysis of the positive form would seem to represent an advantage for the delineation-based framework over the degree semantics I have assumed in this paper. But the greater simplicity in this one area comes with a cost elsewhere. Specifically, the analysis represented in (45) removes reference to degrees from the lexical entry of the gradable adjective. But there is little doubt that the semantics of gradable adjectives must sometimes involve reference to degrees. The most obvious examples are constructions with measure phrases (*Fred is 1,80 meters tall*; *Sue's car cost 1000 euros more than Paul's*), but also perhaps what Kennedy [13] terms absolute gradable adjectives, i.e. adjectives such as *straight* or *full* whose standards seem to be defined in terms of endpoints on a scale, rather than orderings of individuals. Degrees must therefore enter the semantic representation somehow, perhaps via the semantics of measure phrases themselves (as proposed by Klein). The result will be a system in which truth conditions for gradable adjectives are sometimes stated in terms of orderings of individuals and sometimes in terms of degrees. While this is certainly not unthinkable, it contrasts with the more unified account possible within a degree-based theory. The challenge for the delineation-based approach would be to show that these two types of representations can be connected in a satisfying and compositional way.

5 Conclusion

Comparison classes play an important role in the interpretation of gradable adjectives such as *tall*, *expensive* and *crowded*. The choice of the comparison class - be it a set of individuals, of times or perhaps of situations - and the distribution of individuals within that class have a truth conditional effect. Whether one approaches the data from the perspective of a semantics based on degrees or from some alternative perspective, this effect must be represented.

Focusing on cases involving an overt *for*-phrase, I have in this paper developed a novel analysis of the comparison class which is able to overcome several issues with Kennedy's [13] recent analysis, while also extending to cases that have not previously been considered as falling within this type of approach. While these are undoubtedly not the last words that will be written on this subject, the present work has perhaps helped to move the discussion further.

References

1. Bale, A.C.: A universal scale of comparison. Linguistics and Philosophy 31(1), 1–55 (2008)
2. Barner, D., Snedeker, J.: Compositionality and statistics in adjective acquisition: 4-year-olds interpret *tall* and *short* based on the size distributions of novel noun referents. Child Development 79, 594–608 (2008)
3. Bartsch, R., Vennemann, T.: Semantic structures: A study in the relation between syntax and semantics. Athaenum Verlag, Frankfurt (1973)
4. Bhatt, R., Pancheva, R.: Late merger of degree clauses. Linguistic Inquiry 35, 1–45 (2004)
5. Cresswell, M.J.: The semantics of degree. In: Partee, B. (ed.) Montague Grammar, pp. 261–292. Academic Press, New York (1977)
6. Cruse, D.: Lexical Semantics. Cambridge University Press, Cambridge (1986)
7. Fara, D.: Shifting sands: an interest-relative theory of vagueness. Philosophical Topics 28, 45–81 (2000)
8. Fults, S.: The structure of comparsion: an investigation of gradable adjectives Phd dissertation, University of Maryland (2006)
9. Heim, I.: Notes on superlatives (1999) (unpublished manuscript)
10. Heim, I.: Degree operators and scope. In: Semantics and Linguistic Theory (SALT) X. CLC Publications (2000)
11. Heim, I.: Little. In: Semantics and Linguistic Theory (SALT) XVI. CLC Publications (2006)
12. Kennedy, C.: Projecting the adjective: the syntax and semantics of gradability and comparison. Outstanding dissertations in linguistics. Garland, New York (1999)
13. Kennedy, C.: Vagueness and grammar: The semantics of relative and absolute gradable predicates. Linguistics and Philosophy 30, 1–45 (2007)
14. Klein, E.: A semantics for positive and comparative adjectives. Linguistics and Philosophy 4(1-45) (1980)
15. Rett, J.: Degree modification in natural language. Phd dissertation, Rutgers University (2008)
16. van Rooij, R.: Implicit versus explicit comparatives. In: Egre, P., Klinedienst, N. (eds.) Vagueness and Language Use. Palgrave Macmillan, New York (2010)
17. von Stechow, A.: Comparing semantic theories of comparison. Journal of Semantics 3, 1–77 (1984)
18. von Stechow, A.: Times as degrees: früh(er) 'early(er)', spät(er) 'late(r)', and phrase adverbs (2006) (unpublished manuscript)
19. Szabolcsi, A.: Comparative superlatives. Papers in Theoretical Linguistics, MITWPL 8, 245–266 (1986)

Author Index