Language, Cognition, and Mind

Elena Castroviejo
Louise McNally
Galit Weidman Sassoon *Editors*

The Semantics of Gradability, Vagueness, and Scale Structure

Experimental Perspectives

EXTRAS ONLINE

Springer

Language, Cognition, and Mind

Volume 4

This series takes the current thinking on topics in linguistics from the theoretical level to validation through empirical and experimental research. The volumes published offer insights on research that combines linguistic perspectives from recently emerging experimental semantics and pragmatics as well as experimental syntax, phonology, and cross-linguistic psycholinguistics with cognitive science perspectives on linguistics, psychology, philosophy, artificial intelligence and neuroscience, and research into the mind, using all the various technical and critical methods available. The series also publishes cross-linguistic, cross-cultural studies that focus on finding variations and universals with cognitive validity. The peer reviewed edited volumes and monographs in this series inform the reader of the advances made through empirical and experimental research in the language-related cognitive science disciplines.

More information about this series at http://www.springer.com/series/13376

Elena Castroviejo · Louise McNally
Galit Weidman Sassoon
Editors

The Semantics of Gradability, Vagueness, and Scale Structure

Experimental Perspectives

Springer

Editors
Elena Castroviejo
Department of Linguistics
and Basque Studies
University of the Basque Country
(UPV/EHU)
Vitoria-Gasteiz
Spain

and

Ikerbasque, Basque Foundation
for Science
Bilbao
Spain

Louise McNally
Department of Translation
and Language Sciences
Universitat Pompeu Fabra
Barcelona
Spain

Galit Weidman Sassoon
Department of English Literature
and Linguistics
Bar Ilan University
Ramat Gan
Israel

ISSN 2364-4109 ISSN 2364-4117 (electronic)
Language, Cognition, and Mind
ISBN 978-3-319-77790-0 ISBN 978-3-319-77791-7 (eBook)
https://doi.org/10.1007/978-3-319-77791-7

Library of Congress Control Number: 2018936184

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG
part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Gradability, Vagueness, and Scale Structure: From the Armchair to the Lab

Elena Castroviejo, Louise McNally and Galit W. Sassoon

Abstract In this chapter we present an overview of three main issues that have surrounded the study of gradable properties—vagueness, measurement, and dimensionality—and how they have been pursued from the perspectives of philosophy, linguistics, and psychology. We then provide a brief summary of each chapter in the volume, together with a guide to how the chapters relate to each other thematically.

Keywords Semantics · Properties · Adjectives · Vagueness · Measurement
Dimensionality

1 Introduction

Many properties that we use to describe individuals or categories of individuals—dimension, texture, emotions, worth, are just a few examples—manifest themselves to a greater or lesser degree, and it is often relevant to group, order, or compare individuals according to the degrees they possess of the properties in question. This

E. Castroviejo (✉)
Department of Linguistics and Basque Studies,
University of the Basque Country (UPV/EHU),
Vitoria-Gasteiz, Spain
e-mail: elena.castroviejo@ehu.eus

E. Castroviejo
Ikerbasque, Basque Foundation for Science, Bilbao, Spain

L. McNally
Department of Translation and Language Sciences, Universitat Pompeu Fabra,
Barcelona, Spain
e-mail: louise.mcnally@upf.edu

G. W. Sassoon
Department of English Literature and Linguistics, Bar Ilan University,
Ramat Gan, Israel
e-mail: galit.weidman-sassoon@biu.ac.il

E. Castroviejo et al. (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure*,
Language, Cognition, and Mind 4, https://doi.org/10.1007/978-3-319-77791-7_1

fact has raised important questions in the fields of philosophy, linguistics, and the study of cognition.

First, there is the vagueness question. If a property (take size as an example) can be held to different degrees, on what basis do we partition a set of individuals based on their size—how do we draw the line between the large and the small? What is different about vague predicates like *long* versus those that are arguably not vague (or needn't be vague), such as *spotted*, for which drawing the line is much easier?

Second, there is the measurement question. By what means do we order or compare properties, particularly when these are not reducible to a single measurable dimension, as is the case with e.g. beauty? How can we best model the semantics of degree, measure, and comparison constructions in language?

Third, logically prior to the measurement question is the dimension selection question. Many properties can be compared on more than one dimension. For example, we can compare the intensity of a color, or its extension on a surface. This and other sorts of variability have led to the study of what has come to be known as *scale structure*, which can be understood in the broadest sense as the study of how lexical semantics—including not only the multidimensionality of properties, but also other factors as well—is connected to more formal properties of gradability, such as whether a property can be held to a minimal or maximal degree (i.e., is associated with a *bounded* or *unbounded* scale), and what the nature is of the *standard* or threshold for truthful ascription of the property.

These problems have been amply explored over the years from a highly theoretical/conceptual perspective ("the armchair"). However, with the increasing presence of experimental methods both in philosophy and linguistics, theoretical analyses are now being tested experimentally. The chapters in this volume, contributed by philosophers, linguists and psychologists, all reflect this move from the armchair to the lab. In the remainder of this introduction, we first present a very brief review of some of the most important perspectives on the semantics of gradability, vagueness, and scale structure found in the philosophical and linguistics literature, as well as in the study of cognition and categorization. We then offer an overview of each chapter.

2 Perspectives

2.1 The Perspective from Philosophy and Logic

The philosophical literature has focused in large part on the challenge of understanding how reasoning with vague concepts is possible at all. Classical logic embraces the law of the excluded middle, making it impossible for an entity to be neither P nor not P, or to be both P and not P simultaneously. But precisely this seems possible and sometimes even rather natural for vague predicates like *tall* or *long*. Given our world knowledge concerning pencils and their varying sizes, an 18-cm long pencil is obviously long, and a 5-cm long pencil is obviously not long. But it is not clear whether a 10- or 12-cm long pencil is long or not. When asked about it, speakers may well accept that such a pencil is “neither long nor not long,” or rather that such a pencil is “long and not long” (cf. Alxatib and Pelletier 2011; Égré et al. 2013).

Like contradictions, paradoxes also put the foundations of logic at risk, and therefore philosophers have long been engaged in a search for a solution to famous examples such as the Sorites paradox, illustrated in (1), where starting with natural premises such as **P1** and **P2**, the absurd conclusion **C** is reached through iteration. The question is whether a logic that derives this paradox is sound.

(1) **P1**: A 20-cm long pencil is long.
P2: Any pencil that is 1 mm shorter than a long pencil is long.
C: A 1-cm long pencil is long.

These manifestations of vagueness have triggered the development of rich and diverse theories of the logic underlying the use of vague predicates. These theories offer alternative theoretical mechanisms for the representation of borderline cases, imprecise boundaries, the Sorites paradox, and the context dependency of vague predicates. Epistemicists (Williamson 1994; Fara 2000) urge us to conserve classical logic, but it is hard to resist accepting that vague predicates have no true cut-off points for applicability, as supervaluationists maintain (Fine 1975; Kamp 2013; Keefe 2000). It is also hard to disallow contradictions, as subvaluationists do (Hyde 1997; Cobreros 2011a) or to resist accepting the tolerance of vague predicates to small differences, a property that gives rise to the intuition that premises like **P2** in (1) are true. Moreover, the acceptance of tolerance takes us further away from classical logic (Kamp 2013; Shapiro 2006; Soames 1999; Stanley 2003; van Rooij 2011; Cobreros 2011b; Cobreros et al. 2012; Ripley 2011).

In an effort to determine which of the competing theoretical accounts of vagueness is more successful at explaining natural language data, experimentation has become not only welcome but also necessary. Philosophical experiments on vague adjectives most commonly investigate classifications on Sorites-like distributions, where the set of entities covers a full range of values for a property in a certain range with no gaps in between any two values (cf. Alxatib and Pelletier 2011; Égré et al. 2013; Raffman 1994, 2005; for a more recent study, see Verheyen et al. 2016). These experiments usually assess judgments of the truth of vague sentences (e.g. *X is tall*) or of the acceptability of contradictory or tautological predicates (e.g. *P and/or not P*; see Égré and Zehr, this volume). The related problem of faultless disagreement has also been explored recently (see the study by Solt in this volume and references cited there). Contradictory sentences and cases of faultless disagreement share the fact that they put the consistency of natural language at risk; they differ in that in the former case, one and the same speaker asserts or accepts the truth of a proposition and its negation (e.g. *John is tall and not tall*), while in the latter, the inconsistent propositions are ascribed to two different speakers, neither of which can be considered wrong in their belief.

Philosophers have typically been less concerned with other manifestations of vagueness in language and thought, such as the connection between vagueness and gradability, or the manifestation of vagueness in degree modification and comparison, issues to which we now turn.

2.2 The Perspective from Linguistics

We have seen that vagueness poses a problem for a theory of truth and for the foundations of logic: Is a binary system yielding truth and falsity enough to cover borderline cases, and can such a system be consistent? The linguist's take on vagueness is quite different. The starting point is rather the question of how vagueness is linguistically realized. One straightforward answer is that it is encoded as gradability, and yet this is just the beginning of the story.

Let us start by noting that gradability is naturally realized on adjectives and adverbs. The grammatical diagnostic to determine whether an adjective is gradable consists in placing it in a comparative construction, as in (2).

(2) a. Maria is taller than John.
b. Your shirt is wetter than mine.

Gradable adjectives like *tall* and *wet* pose an interesting problem for linguists because they seem to have many possible interpretations depending on the phrases in which they occur (for example, which degree modifiers they combine with) and the situations in which these phrases are used. For instance, adjectives like *tall* cannot easily be modified by *slightly*, while adjectives like *wet* can (Rotstein and Winter 2004; Kennedy and McNally 2005; Kennedy 2007b; Sassoon 2012b). This and other grammatical distinctions reveal a particular way of integrating contextual information in the lexical meaning of such adjectives. While some adjectives—those known as *relative* or *open scale*—resort to a contextual standard (to assert that Maria is tall, we need to consider similar individuals and average over their heights), for others—the *absolute* or *closed scale* adjectives—it is enough to attend to their conventional meaning to be certain that they can be truthfully predicated of an individual (but see Sassoon and Toledo 2011; McNally 2011for qualifications).

Linguists are interested in determining what the core meaning components of gradable adjectives are that enable speakers to use them in different ways, and what information from the context of use speakers find relevant for interpretation (e.g. for standard selection), and what they ignore. For example, Aparicio et al. (2015) and Aparicio et al., this volume, report on a series of eye-tracking experiments which allow us to better understand the correlation between integration of contextual information and the scale structure of adjectives. A different line of research concerns how degree modifiers can only be used not only as indicators of lexical semantics, but also as sociolinguistic markers. Beltrama, this volume, presents a study that aims to characterize the social meanings associated with putatively illicit cases of modification such as *totally tall* (on this, see also Beltrama 2016, 2018a; Beltrama and Staum Casasanto 2017).

Linguists are also interested in formally representing the lexical meaning of vague predicates and the way they combine with their arguments and modifiers. On one view, let's call it the *vagueness approach*, gradable predicates such as *tall* should have the same semantic type as non-vague predicates such as *four-legged*; the difference is simply that the former have a gap in their extension, corresponding to the set of individuals that cannot be said to be in either the positive or the negative extension of the predicate. There are different versions of the vagueness approach, most of which rely on the notion of an ordering to derive the degree-like properties of such predicates (Cresswell 1976; Klein 1980; van Rooij 2011). This can be illustrated as follows.

Assume that an adjective like *tall* denotes a partial function with a positive extension, a negative extension, and an extension gap. To judge whether a sentence including a vague adjective is true, context provides a comparison class (a collection of relevant similar objects); the choice of comparison class influences which entities fall into the positive and negative extensions, and which fall into the gap. For example, in (3), our judgment will partly depend on the physical properties of Peter and on whether the comparison class is the set of boys, male adults, jockeys or basketball players.

(3) Peter is tall.

For some comparison classes, Peter will fall into the positive extension, while in others he will fall into the negative extension, and in still others he might fall into the gap. The variation in the positive versus negative extensions across different comparison classes determines the ordering of entities relative to the given predicate: An entity x will be ordered higher than another entity y iff we can find a comparison class (most relevantly, the one consisting just of x and y) for which x falls into the positive extension and y falls into the negative extension—the assumption being that for no choice of comparison class can either of the two be empty. Thus, entity orderings (or even degrees, where those are needed) are derived from the entity sets forming the basic denotations of predicates (see especially van Benthem 1982; Bale 2008; Burnett 2016).

A more popular view in recent years is one on which entity sets are derived, and degrees are the primitives from which these sets are derived. Gradable predicates include a degree argument which is bound after modification by certain degree expressions such as *very, completely*, etc. This *degree approach* has at least two implementations. Gradable adjectives are either analyzed as relations between degrees and individuals (type $\langle d, et\rangle$), as proposed by e.g. Seuren (1973), von Stechow (1984), Heim (1985), or Bierwisch (1989), or as measure functions of type $\langle e, d\rangle$, as proposed by e.g. Bartsch and Vennemann (1972), Kennedy (1999, 2007b). On this view, to avoid a type mismatch in the composition of (3), and to introduce a standard for the truthful application of the predicate, it is common practice to assume the presence of a null positive degree morpheme—*pos*—that combines with the gradable adjective and yields a predicate of individuals (von Stechow 1984; Kennedy 2007b, a.m.o.). See Solt and Gotzner (2012) for an experimental study that bears on the question of whether degrees are among the primitives of natural language ontology or whether it is sufficient to posit ordering relations.

Up to this point in the discussion, we have made the maximally simple assumption that gradable predicates are, crucially, unidimensional. An adjective like *tall* lexicalizes a scale related to height, which can be associated with a set of degrees, a dimension and an ordering. However, the empirical domain is much broader and more complex. For example, beyond dimensional adjectives, there are evaluative adjectives in the sense of Bierwisch (1989), such as *industrious*, for which there is no obvious measuring system. Should such adjectives include a degree argument? Is there a simple scale of industriousness, or rather are there various dimensions that can be evaluated when we assert that someone is or is not industrious?

Multidimensionality is precisely the focus of Sassoon (2012a, 2013a); Sassoon and Fadlon (2017), who build on Bartsch and Vennemann (1972) and Bartsch (1984, 1986) to develop and empirically test a quantificational analysis for such adjectives, examples of which also include *healthy* and *intelligent*. Quantification plays a role in the sense that, in some cases, in order for an entity to count as bearing a multidimensional property, *all* of its dimensions must be satisfied, while in other cases, it is argued, only *some* dimensions must be satisfied. Thus, for example, someone is healthy if she is healthy in every way, and sick if she is sick in some way.

The contribution by Solt to this volume looks at a different aspect of multidimensionality, namely its role in disagreements about orderings (Kennedy 2013; Bylinina 2014; Umbach 2016; McNally and Stojanovic 2017). Solt notes that even some physical properties are arguably multidimensional. For example, two individuals looking at photographs of roads might disagree about whether one road is bumpier than another because one places more weight on the number of bumps, while the other focuses on their size or severity. Solt presents evidence that the disagreement that can arise in such cases is distinct from the disagreement found with adjectives like *fun*, which arises due to variation in our subjective experience.

Another focus of study in the linguistic research on vagueness regards the categories in which gradability is realized. Gradability has been observed not only in adjectives and adverbs, but in verbs and nouns, as well (see Bolinger 1972; Hay et al. 1999; Vanden Wyngaerd 2001; Kennedy and Levin 2008; Piñón 2008; Kennedy 2012; Bochnak 2013b; Rappaport-Hovav 2014; Fleischhauer 2016, a.m.o.; see Wellwood 2014 for a recent attempt to unify gradability across categories under a single analysis). Consider, for example, the following data from Morzycki (2009) and Constantinescu (2011), respectively.

(4) a. a big idiot/stamp collector
    b. that idiot/*stamp collector of a doctor.

As in the case of gradable adjectives, to determine whether a noun is gradable, linguists pay attention to the characteristics of the syntactic constructions in which they appear (and the lexical semantics of the categories with which they are composed). In (4a), we are inclined to consider *idiot* and *stamp collector* as gradable because they are modified by *big*, which does not restrict the size of the referent of the noun, but rather the degree of idiocy or of stamp collectorhood. In (4b), on the other hand, the *N-of-an-N* construction, analyzed by Bolinger (1972) and Matushansky (2002), among others, as a diagnostic for gradability, gives different results for *idiot* and *stamp collector*, whereby only the former but not the latter would be analyzed as such. These and other tests have been used by Constantinescu (2011) to conclude that gradability in the nominal domain is not a uniform phenomenon.

Data such as (4a) have led Morzycki (2009) and others to assign nouns a degree argument as part of their lexical entry. For a set of different constructions involving natural kind nouns such as *duck* and social nouns like *philosopher* (as in "more a duck than a goose" and "more a linguist than a philosopher"), Sassoon (2017) also advocates a degree analysis whereby the conceptual structure of nouns and closeness to prototypes or stereotypes can be modeled in the same way as for gradable predicates. The contribution to this volume by de Vries tests the viability of this sort of analysis. Others, such as Constantinescu (2011) and Beltrama and Bochnak (2015), defend a degreeless analysis for intensifying constructions that builds on quantification over contexts or worlds in an epistemic modal base. In a similar vein are those analyses that completely rely on degrees of precision, as in Lasersohn's (1999) use of slack regulation and Morzycki's (2011) more recent implementation.

The comparability of properties which otherwise do not appear to be gradable raises a final theoretical question that has concerned linguists interested in natural language metaphysics, namely: What is a degree? The advocates of the degree approach to gradability assume that degrees are primitive parts of the ontology of natural language. Nouwen and Dotlačil, this volume, go further and present experimental evidence that degrees are not only atomic entities, but also, just like regular individual objects, can cumulate as pluralities. Other options for the treatment of degrees include Grosu and Landman's (1998) account, on which degrees consist of a measure value, a measure domain and the object measured. In contrast, among those who do not treat degrees as primitives, we find degrees being equated to equivalence classes of individuals (Cresswell 1976; Klein 1980) and to state kinds (Anderson and Morzycki 2015). Finally, in various works, Moltmann (2004, 2007, 2009) claims that tropes, roughly realizations of properties, are enough to account for gradability phenomena. See also Scontras (2014) for a recent in-depth development of the semantics of amounts and degrees.

To understand the linguistic realization of gradability, analyzing crosslinguistic data is essential. Beck et al. (2004, 2009), Kennedy (2007a); Bobaljik (2012), and Bochnak (2013a) are examples of theoretical works that either compare the syntax-semantics mappings of linguistic constructions, such as the comparative, in parametrically distinct languages, or else provide a formal account of these constructions in under-represented and under-studied languages. Experimental crosslinguistic research on gradability phenomena is also ongoing: Examples include Pancheva and Tomaszewicz (2011), O'Connor et al. (2012) and Tucker et al., this volume.

Let us conclude this subsection by zooming out to the level of the utterance. As noted in e.g. Doran et al. (2009) and Beltrama and Xiang (2012), statements including gradable predicates can give rise to scalar implicatures. That is, gradable predicates are sometimes part of a scale of lexical items of varying strengths, and thus asserting that something is, for example, good may give rise to the (cancellable) implicature that it is not excellent. However, gradable adjectives do not behave exactly like other sorts of expressions that can be ordered on scales (such as numerals or quantifiers), in the consistency with which they yield scalar implicatures (van Tiel et al. 2016). The question is why. Building on Krifka's (2002) observation that scalar implicature with numerals is sensitive to contextually-relevant granularity (for example, whether exact numbers or e.g. multiples of 100 are under discussion), McNally (2017) suggests that adjectival scales might be much more sensitive to granularity than numerals. Cummins' contribution to this volume supports the view that, rather than granularity as such, what matters is the set of salient scalar alternatives in the context.

The linguistic and logical complexity posed by vague and gradable expressions raises questions concerning the mechanisms required for their processing and the manifestation of these in the online processing of language. Such questions, including those regarding the stages at which different types of information (e.g. grammatical vs. visual context) intervene and assist or affect processing, are typically addressed by psychologists and psycholinguists. Linguists and philosophers theorizing on language have historically drawn a static picture focusing on grammatical rules as opposed to online processes of structure building or parsing and the dynamics of semantic composition as a sentence unfolds. It is only recently that they have begun testing theories of gradability, vagueness, and scale structure using online processing measurements of reaction time (RT), eye tracking or eletrophysiological activity in the form of event-related potentials (ERP). For more on this, see the next subsection.

2.3 The Perspective from the Study of Cognition and Categorization

A rich psychological and psycholinguistic tradition has studied the structure of concepts and the processes of online categorization and their connections to the lexical semantics of nouns and verbs (see e.g. Rosch and Mervis 1975; McCloskey and Glucksberg 1978; Hampton 1979; Osherson and Smith 1981; Barsalou 1993; Hampton 1998, 2007; Verheyen et al. 2010). Two chapters in this volume fall into this tradition: The study by Verheyen and Storms, which focuses on intersubjective variation in classification under vague concepts; and that by de Vries, which investigates subtle distinctions in the way subjects categorize different types of nouns, inspired by influential accounts of vagueness and gradability in the nominal domain (Kamp and Partee 1995; Morzycki 2009). The contribution by Schumacher, et al., addresses the online processing of categories in the context of modifiers such as *real* and *fake*, using ERPs. A related line of psycholinguistic work on dimension indeterminacy and multidimensionality is also developing (see Sassoon and Fadlon 2017 for a judgment study and Sassoon et al. (t.a.) for an ongoing ERP project).

In contrast, the exploration of adjectives—the prime examples of vagueness and gradability—has generally lagged in this tradition behind that of nouns or verbs (Cappelletti et al. 2008, note 1). A notable exception is the early study by Rips and Turnbull (1980), who reported experimental evidence to the effect that relative standards for property ascription are not stored. Rather, in each context of use of an adjective like *big*, the characteristic size of one of the categories of the entity argument (e.g. child or female) determines the standard; when no such category is particularly salient, additional factors are likely to affect the standard choice, such as the size of the speaker. Interestingly, they also found that this dependency of the standard on a comparison class or speaker was not manifest in absolute adjectives. Further studies exploring adjectival property ascription (including in first language acquisition), especially what sorts of factors figured into judging objects as e.g. big versus little, were carried out later in the 1980s by, e.g. Smith and colleagues (Smith et al. 1986, 1988, and references therein).

The linguistic work on scale structure in the first half of the 2000s triggered a burst of experimental research into the cognitive basis for scale structure and its connections to vagueness. For example, Frazier et al. (2008) considered adjectives like *dirty*, whose standard is identified with the minimum on their scale (a minimal amount of dirt suffices for something to count as dirty) and adjectives like *clean*, whose standard is identified with the maximum on their scale (maximal cleanliness is required for something to count as clean). Frazier et al. (2008) substantiated reported intuitions that minimum standard adjectives like *dirty* are more acceptable with minimizers like *slightly* and *a little* than are maximum standard adjectives like *clean*. The former were regarded as acceptable in 85% of the cases, like the non-modifed forms, as opposed to 60% acceptance of modified upper-closed total adjectives (see also Bogal-Allbritten 2012). But more important, an eye-tracking study investigated the online processing of scale structure. This study further showed that the total first pass reading times of regions of sentences with maximum-standard adjectives modified by *slightly* were longer than those of similar regions with minimum-standard adjectives modified by *slightly*, suggesting that the processing of scales and endpoint standards is an obligatory part of semantic composition of phrases containing absolute adjectives, rather than a non-compulsory sort of late pragmatic processing.

Results reported in Syrett (2007) showed that adults treat absolute adjectives differently than relative adjectives. When presented with two rods of different lengths, but neither of which was obviously long, participants easily complied with a request to hand the experimenter “the long one,” suggesting that they can readily appeal to a contextually given comparison class to find a standard for membership that satisfies the existence presupposition associated with *the*. In contrast, when faced with two containers which were not full, participants tended to reject a request to hand the experimenter “the full one,” suggesting that they use a maximum standard of membership and are reluctant to shift this standard even in the presence of a presuppositional demand to this end. This and other work by Syrett and colleagues (e.g. Syrett et al. 2010; Syrett and Lidz 2010) further explored the differences between relative and absolute adjectives in the course of first language acquisition (for additional acquisition studies on related questions, see also e.g. Barner and Snedeker 2008; Tribushinina and Gillis 2012).

In the intervening years, additional experimental research—not only by psycholinguists but also by theoretical linguists and philosophers—has focused on the understanding of relative and absolute adjectives, including their interaction with degree modification (e.g. Schmidt et al. 2009; Solt 2011; McNabb 2012; Solt and Gotzner 2012; Liao and Meskin 2017; Liao et al. 2016; Solt 2016; Hansen and Chemla 2017), and their online processing (Aparicio et al. 2015). This research supports various theoretical distinctions between relative and absolute adjectives that also manifest themselves in their online processing. Aparicio et al., this volume, contribute to this ongoing endeavor via an eye tracking study of minimum-standard adjectives and an offline study of the pragmatics of their use.

Finally, beyond psycholinguistic analyses that involve the notion of comparison (e.g. Scontras et al. 2012), research on the interpretation and processing of comparative constructions is starting to develop, especially in the domain of so-called "comparative illusions" (Wellwood et al. 2009, 2017; O'Connor 2015), but also more generally, as in Grant (2013). Tucker et al., this volume, continues this line of research.

3 The Chapters in This Volume

With this brief introduction in hand we now turn to the contributions of this volume to the state of the art concerning the offline and online study of gradability, scale structure and vagueness. We have not divided the volume into different thematically unified sections because, as the following overview of the book chapters reveals, we do not think they lend themselves easily to such a classical division. Rather, the volume has a family resemblance structure where each paper shares a slightly different set of properties with each other paper. Nonetheless, for the reader's convenience we would like to indicate several different ways in which the papers may be grouped so as to help readers navigate in the volume according to the topics that most interest them.

First, the papers can be grouped according the lexical category of the items under investigation. Chapters 2, 5 and 6 are concerned with nouns, while Chaps. 3, 4, 7, 8, 10 and 11 investigate adjectives; Chap. 9 focuses on fractions which, while structurally nominal, functionally behave more like quantifiers.[1]

Second, the papers can also be divided up according to the linguistic constructions that they address. While Chaps. 2, 5, 6 and 8 specifically examine the positive forms of adjectives and nouns, Chaps. 3, 10 and 11 discuss comparative constructions, and Chap. 9 deals with modifiers of fractions that are formally close to comparatives (e.g. *more than half*). Chapter 7 is concerned with (non-comparative) degree-modification of adjectives. Finally, Chap. 4 examines the effect of certain sorts of (positive form) adjectival modification on noun-based categorization.

Third, various subgroupings are possible based on the specific aspects of interest in the given items or constructions. While Chaps. 3–5, 10 and 11 address formal semantic aspects of words and constructions, Chaps. 2, 8 and 9 are concerned with general pragmatic factors, and Chaps. 6 and 7 specifically examine the effects of speaker group on linguistic behavior. Chapters 2–6 constitute an additional subgroup, as all are centered on issues related to the multidimensionality of concepts.

[1] For research of the verbal domain, see, for example, Bochnak (2013b), Rappaport-Hovav (2014), and Fleischhauer (2016).

Yet another subgrouping is possible according to the language(s) studied: Most of the papers report studies of English, but Chap. 4 reports on German data, Chaps. 5, 6 and 11 present studies of Dutch (including Flemish), and Chap. 10 provides a contrastive study of Polish and English.

Finally, the chapters can be divided according to the methodologies used. Chapters 2, 3, 5–7, 9 and 11 use offline methods, while the remaining chapters use online methods, including measurement of response times (Chap. 10), eye movements (Chap. 8) and neural activity along the scalp (Chap. 4).

The volume begins with a study by **Paul Égré** and **Jérémy Zehr** ("Are gaps preferred to gluts? A closer look at borderline contradictions") of seemingly contradictory assertions involving vague predicates. Vague adjectives, as noted above, admit borderline cases. One manifestation of borderline cases is that individuals can be ascribed both an adjective and its negation at the same time, as in e.g. *X is tall and not tall*, as well as neither the adjective nor its negation, as in *X is neither tall nor not tall* (Ripley 2011; Alxatib and Pelletier 2011; Serchuk et al. 2011; Égré et al. 2013). Égré and Zehr hypothesize that there is a preference for "gappy" descriptions (*neither A nor not A*) over "glutty" descriptions of the form *A and not A*. Though this hypothesis is supported by the results, they show that both kinds of descriptions are acceptable.

The analysis they propose for the data adopts the distinction offered by Cobreros et al. (2012) between strict and tolerant meanings for vague adjectives, and a specific implementation of the Strongest Meaning Hypothesis, in line with Alxatib and Pelletier (2011). However, in contrast to previous literature, Égré and Zehr argue in favor of local, rather than global, pragmatic accommodation of strict and tolerant truth operators. Assuming the strongest meaning of, e.g., *tall* to convey "tall by every standard," *neither*-descriptions are consistent, while *and*-descriptions are not. A penalty is exerted on the acceptability of the latter due to the need to resort to a more tolerant interpretation such as "tall by some standard." Given that evaluative multidimensional adjectives are associated with context dependent sets of dimensions (e.g. the adjectives *conservative* and *liberal* can relate to politics, religion, sex, family structure, dress code, music, and/or theoretical views), Égré and Zehr point out, in agreement with earlier observations by Kamp and Partee, that such adjectives are, intuitively, more acceptable in forms like *X is A and not A* than dimensional adjectives are, because they can be interpreted as, e.g., "X is A in some respects and not A in other respects." More generally, they question whether the preference for *neither*-descriptions over *and*-descriptions is systematic, or whether it is likely to vary depending on the adjectival type (relative vs. absolute, or unidimensional vs. multidimensional).

The increased level of indeterminacy and contextual variance that multidimensional adjectives manifest is also exhibited in a higher acceptability of so-called faultless disagreements concerning their application. **Stephanie Solt**'s contribution ("Multidimensionality, subjectivity and scales: Experimental evidence") reports on a study of precisely this phenomenon. Solt asked participants to decide whether only one of two speakers in disagreements such as (5) can be right, or whether both can be right, that is, whether the disagreement was a matter of fact or opinion.

(5) A: Look Tommy's shirt is dirtier than the one his little brother Billy is wearing.
B: No, Billy's shirt is dirtier than Tommy's.

The results indicate that multidimensionality is indeed a source of subjectivity in comparative forms. However, Solt's study also reveals that what it means to be multidimensional and what sorts of factors underlie disagreement is more complex than suggested by earlier work such as e.g. Sassoon (2013b), Lasersohn (2005), or Bylinina (2014). Multidimensional adjectives such as *good*, *intelligent*, or *beautiful* yielded different results not only from unidimensional adjectives such as *tall*, *old*, *expensive*, or *empty*, but also from adjectives such as *dirty*, *smooth*, *light*, or *sharp*. The first group clearly permitted faultless disagreements; the second group tends not to permit faultless disagreement at all. However, the third group yielded clearly mixed judgments as to whether a disagreement involving them would be a matter of fact or opinion—for example, we might disagree as to whether Tommy's shirt is dirtier than Billy's because we choose different ways of measuring dirtiness (e.g. overall presence of dirt vs. presence of a small amount of highly noticeable dirt), and depending on the choice of measure, the ordering of the shirts might be different. However, in some cases it may happen that all of the different choices of measure result in the same ordering, leading to the intuition that the disagreement is a matter of fact. Solt concludes that judge dependence is crucial to the first group (indeed, she argues that while the properties in question are clearly conceptually multidimensional, they do not always behave grammatically as if they were multidimensional), while in the case of the third group different dimensions can be distinguished, selected, and integrated in a contextually determined manner for the purposes of comparison.

The processing of adjectivally modified nouns such as *real diamond* or *fake diamond* also seems to depend on the identification and highlighting of a set of dimensions. However, in this case the dimensions consist of prototypical or stereotypical features of the head noun. Crucially, with *fake*-type adjectives, in contrast to *real*-type adjectives, these features of the meaning of the head noun are negated. Thus, the meaning of a nominal containing *fake* can be understood as, e.g., "in some sense *x* is N and in some sense *x* is not N;" that is, it seems to involve quantification over dimensions. Following Peirce (1910), and very much in line with Solt's and Égré and Zehr's observations, **Petra Schumacher**, **Patrick Brandt** and **Hanna Weiland-Breckle** propose in their contribution ("Online processing of *real* and *fake*: The cost of being too strong") that this hidden meaning is the result of a repair that circumvents contradiction.

To test this hypothesis and its consequences for the neural signature of the processing of such modified nouns, Schumacher et al. measured ERPs of participants during the processing of *fake*-modified nouns, as compared to baselines formed by nouns modified by ordinary negative adjectives like *flawed*. They observed a Late Positivity, which is characteristic of referential shifts or reconceptualization—for example, it has been observed during processing of metonymic uses of noun phrases, as in *The ham sandwich paid*. Schumacher et al. argue that since *fake*-type modification involves an intermediate representation that is semantically contradictory, the Late

Positivity reflects an interface repair mechanism that deals with the contradiction. In contrast, processing of *real*-type adjectives, as compared with simpler baselines formed by ordinary positive adjectives like *white*, evoked no comparable processing costs. This finding aligns with the cost-free processing of e.g. *She read Dickens before she met him*, where different aspects of the meaning of *Dickens* are highlighted, but no reconceptualization occurs. Thus, the chapter locates the processing of *fake*- and *real*-type adjectives within a typology of the neural signatures of different types of semantic and pragmatic operations, explaining the processing differences as effects of recovery from inconsistent interpretations through dimension shifting.

Like Schumacher et al., **Hanna de Vries** ("Gradable nouns as concepts without prototypes") is concerned with concepts expressed by nouns. However, her paper addresses a different issue. In a foundational paper, Kamp and Partee (1995) argued against the association of certain predicates with a prototype. For example, adjectives like *tall* or *intelligent* are associated with upper-open scales, and their meanings do not seem to be represented correctly by means of any prototypical values on those scales: There is no upper bound such that higher degrees of height decrease an entity's tallness. Rather, the taller one is, the better. These adjectives are also vague and gradable. In contrast, nouns such as *bird* are associated with a prototype, but are not gradable in the same way that adjectives and adverbs are.

Considering this typology, de Vries argues that nouns like *genius*, which are also associated with upper-open scales (e.g. intelligence), do not have a prototype representation, either. She starts by characterizing prototypicality in terms of "maximal embodiment," that is, in terms of manifesting, in the case of gradable properties, ideal values for those properties. Maximal embodiment cannot be satisfied in cases where (1) having more of a property is considered better, and (2) there is no maximal value of the property in question. Thus, if subjects appear to have prototypicality judgments for concepts that do not satisfy maximal embodiment, like "genius," these judgments must reflect other factors.

De Vries tests the hypothesized difference between nouns like *genius*, which are associated with upper open scales, and nouns like *bird*, which are associated with upper closed prototypicality scales, using classical methodologies developed within the cognitive psychological research on conceptual structure. One experiment shows that factors like familiarity and, especially, attitude largely explain the prototypicality judgments in *genius*-type nouns, but are unrelated to prototypicality in *bird*-type nouns. The other experiments look at a related question: Do unbounded properties in fact play a greater role than bounded properties in subjects' decisions about categorization with nouns like *genius*, as opposed to nouns like *bird*? Subjects were asked to generate properties for a range of *genius*- and *bird*-type nouns. De Vries then measures to what extent membership in the class described by each type of noun is linked with unbounded property dimensions (for example, the more intelligent one is, the more likely they are to be classified as a genius). She finds a strong tendency towards the use of unbounded dimensions for nouns like *genius*, but not for nouns like *bird*.

The contribution by **Steven Verheyen** and **Gert Storms** ("Education as a source of vagueness in criteria and degree") examines yet another factor that plays a role in how we understand and navigate the dimensions and boundaries associated with nominal predicates. Building on previous studies that suggest that upbringing plays a role in categorization behavior, Verheyen and Storms look specifically at level of education—i.e. whether individuals have completed *only* compulsory education or *also* higher education—as a factor in the classification behavior of individuals with different levels of education for categories such as fruits, vegetables, fish, insects, sports, sciences, tools, and furniture.

Verheyen and Storms's study starts with a distinction proposed in Devos (1995, 2003) between vagueness in criteria and vagueness in degree. In the former, there is indeterminacy with respect to the conditions of application of the predicate to the noun. For instance, is chess a sport? There could be disagreement depending on whether the relevant criteria include physical activity or competition. In the latter, there is indeterminacy with respect to the extent of application given fixed conditions. Consider hiking, for example. While we can be certain that it meets the criterion of physical activity, we could argue about whether it meets this criterion sufficiently. Devos suggested that vagueness in criteria is primarily involved in categorization involving nouns, while vagueness in degree is mainly involved in categorization involving adjectives. Verheyen and Storms challenge this idea and implement a mathematical model capable of measuring both factors. The results of this study show that (1) both vagueness in criteria and vagueness in degree are found in nouns, and (2) criteria and degree differences are systematically related to subjects' properties, such as their level of education. Compared to subjects with only compulsory education, subjects with higher education endorse fewer items and use different conditions of application, especially in nouns they are more familiar with, such as those related to science categories.

Social criteria also play a role in the data explored by **Andrea Beltrama** ("Intensification, gradability and social perception: The case of *totally*"). Grammatical distinctions in gradable adjectives—i.e. whether the standard for ascribing the adjective is a minimum, a maximum, or is contextually determined—typically condition the acceptability of degree modifiers. For instance, whereas *very* can modify relative adjectives such as *tall*, *completely* only targets absolute adjectives such as *full* or *empty*. However, such restrictions cease to apply when such modifiers target scales that are grounded in the attitude of the speaker, rather than in the lexical meaning of the subsequent adjective. It is in these contexts that such intensifiers are perceived as having an especially salient social meaning. Beltrama's chapter explores precisely this phenomenon. It falls within the new domain of what we might call experimental socio-semantics, in which social meaning is assumed to be amenable to systematic and formal scrutiny, and compositional meaning is taken to affect an "expression's suitability to serve as a vehicle for social meaning."

More specifically, the chapter presents a study of the social meaning conveyed by the degree expression *totally*, whose interpretation varies depending on, roughly, whether the constituent it modifies is a closed-scale adjective or not. If the adjective is associated with a closed scale, *totally* is understood lexically, i.e. as entailing a maximal degree. If the adjective is not closed-scale, *totally* is understood as a "speaker-oriented" intensifier that does not entail maximal degree. Beltrama tested the reactions of participants (concerning factors such as solidarity and status) when presented with the two types of *totally* as modifiers of different constituents, and also in comparison to other intensifiers (*completely*, *really*) and unmodified adjectives. The results support the author's hypothesis that *totally* is more likely to be interpreted as a carrier of social meaning on its speaker-oriented interpretation than on its lexical interpretation. Beltrama further discusses the social meaning carried by *totally* when modifying extreme adjectives such as *awesome*. The inherent emotive meaning conveyed by such adjectives is considered as a potential factor in explaining their unpredicted behavior, as reflected in the experimental results.

Beltrama's study points to the salience of the distinction between relative and absolute adjectives for (socio) linguistic phenomena. **Helena Aparicio**, **Ming Xiang** and **Christopher Kennedy** ("Informativity and grammar in referential effects of contrast involving adjectivally modified NPs") consider this distinction in relation to language processing. Their work builds on seminal psycholinguistic research on the interpretation of gradable adjectives by Julie Sedivy and collaborators (Sedivy et al. 1999; Sedivy 2003, 2005) which, in a series of Visual World eye-tracking studies, investigated the effect that contextual information has on incremental semantic processing. Participants in their experiments received verbal instructions such us "pick up the tall glass" while looking at displays of four objects. The instructions contained a restrictive prenominal adjective, which can trigger quantity-based pragmatic reasoning about a set of referents that contrast along the adjectival dimension. Each instruction was tested against two types of displays that either supported a contrastive interpretation of the adjective by including a contrastive element in the display (i.e. an object that could be described by the head noun in the instruction but not the adjective, e.g., a short glass), or lacked such contrastive object, rendering the use of the adjective redundant. Their results showed that the presence of a contrasting object in the visual display facilitated the lexical processing of the adjective, as revealed by the fact that the target object was identified significantly faster (even before information about the head noun was available to the participants) in those displays that contained a contrastive set of objects compared to those that did not. These results suggest that semantic processing is incremental and that the processing of attributive relative adjectives like *tall* is facilitated when the visual context supports a restrictive interpretation of the predicate—what is often called the Referential Effect of Contrast (REC).

Aparicio et al. aim to test whether pragmatic reasoning alone is sufficient to explain RECs, or whether lexical properties of the different classes of adjectives (essentially, whether the threshold that establishes the positive form of the adjective is an endpoint of the scale vs. determined through an extensional comparison class of individuals) also contribute to these effects. With this goal in mind, they carried out two experiments. The first is an extension of the Visual World study reported in Aparicio et al. (2015), in which the authors tested color, relative and Maximum Standard Absolute adjectives, to Minimum Standard Absolute Adjectives (MinAAs, such as *spotted*, *bent*, *striped*). The results show that unlike relative, color or Maximum Standard Absolute Adjectives (see Aparicio et al. 2015), MinAAs do not exhibit RECs.

The second experiment addresses the question of how informative the different classes of adjectives tested by Aparicio et al. (2015) and the eye-tracking study reported in this volume are perceived to be when used restrictively versus redundantly. In this study, participants rated the informativity of the instructions used in the eyetracking studies given the two displays (i.e. contrastive and non-contrastive) tested. The results indicate that color adjectives, relative adjectives and Maximum Standard Absolute Adjectives show a difference in ratings such that redundant uses were perceived as too informative compared to restrictive uses, whereas ratings pertaining to MinAAs did not show any difference between conditions. Putting together these experimental results and the work in Aparicio et al. (2015), a correlation is found between giving rise to RECs and penalizing overspecification. However, it is argued that informativity alone cannot account for the different properties of all the RECs reported in Aparicio et al. (2015). The authors conclude that even though informativity is clearly an important driver of RECs, the lexical semantics of the adjective classes also contributes to further shape RECs, a result that reveals interesting connections between scale structure, contrast effects and informativity that are worthy of future investigation and theorizing.

The remaining three papers in the volume take us away from lexical semantics, each presenting studies that focus on a different formal semantic and pragmatic property that has been associated with gradability. In "Modified fractions, granularity and scale structure," **Chris Cummins** further explores Krifka's (2002) hypothesis that the granularity of the scalar alternatives associated with an expression influence the pragmatic inferences that hearers draw. Krifka observed that when one hears, for example, *There were 81 people at the meeting*, one is likely to infer that exactly 81 people were there (arguably via a standard scalar implicature), but if one hears that there were 80 people at the meeting, the coarser granularity of a scale in tens rather than in units facilitates more approximative interpretations (and, arguably, influences related scalar inferences). Cummins et al. (2012) tested this hypothesis on the domain of modified numerals, showing that a sentence like *There's room for more than 80 people* yields an upper bound inference of something like *not more than 100*, even though such upper bound inferences do not arise with modified numerals of finer granularity (for instance, *There's room for more than 81 people* does not yield the inference that there is not room for more than 82). Cummins' study in this volume extends the exploration of this phenomenon to modified fractions such as *more than one third* or *less than five sixths*.

The chapter reports on a series of judgment studies in which subjects were presented with sentences describing quantities in terms of modified fractions with different numerators and denominators (thirds, fourths, fifths, and so on). For each modified fraction, subjects had to freely supply the numerical percentage range they thought the fraction corresponded to (e.g. they might provide 35–60% for *more than one third*). Different experiments tested specific questions, such as whether making particularly salient a particular fractional scale (e.g. fifths), and thus a particular set of scalar alternatives, influenced the upper bound that subjects offered. Cummins' results suggest that, rather than granularity being an explanation for inference patterns in and of itself, it is the pragmatically salient alternatives—no matter what their granularity—that influence scalar inference. Though numerical expressions at a finer level of granularity generally make salient a different set of alternatives than do those at a coarser level of granularity, in the case of fractions, the situation proves to be more complex: some fractions (such as quarters and tenths) are highly salient even in cases that are unexpected based on a direct extension of the analysis of granularity effects for non-fractional expressions. Cummins' work underscores not only the key role of scalar alternatives in interpretation, but also the complexity involved in determining exactly what these alternatives are in any given case.

Another ongoing issue in the formal semantic representation of gradable predicates involves the relation between polar opposites, including comparative morphology itself (*more A/A-er...than...*, *less A...than...*). Büring (2007) argues that the semantics of adjectives like *short* should be characterized in terms of decomposition into the semantics for *tall* plus the semantics for an abstract morpheme paraphraseable as *little*, but e.g. Heim (2008) has pointed out problems for such an analysis. In "Decomposition and processing of negative adjectival comparatives," **Daniel Tucker**, **Barbara Tomaszewicz**, and **Alexis Wellwood** contribute to this debate by bringing processing data to bear on the question. Their experiments take as a premise the Interface Transparency Thesis (Lidz et al. 2011), according to which "the verification procedures employed in understanding a declarative sentence are biased towards algorithms that directly compute the relations and operations expressed by the semantic representation of that sentence." This thesis leads to the prediction that if the semantic representation of an adjective or a comparative involves decomposition (as would be the case of *short*(*er*) on Büring's analysis), it should take longer to process than the base (e.g. *tall*(*er*)). The authors were also interested in testing whether there were any differences in the processing of comparative statements if the statement was presented in mathematical notation (e.g. $A < B$), in which the comparative morpheme is arguably not decomposable, instead of natural language (e.g. *A is shorter than B*).

Tucker et al. designed a picture-matching task in which subjects were given a comparative statement in either natural language or mathematical notation and had to decide whether the statement truthfully described an image presenting the two compared objects, or not. Variants of the experiment were done in English and in Polish. The reaction times recorded on the task reveal that processing sentences with negative polar comparatives (e.g. *shorter than*) was systematically slower than the processing of positive comparatives, lending initial support to the decomposition hypothesis. However, the same effect was registered when the comparative was expressed in mathematical notation (i.e., $A < B$ took longer to process than $A > B$), in contrast to the results reported in Deschamps et al. (2015), where no such difference was found. Tucker et al.'s chapter thus leaves open the question as to whether decomposition is supported. They hypothesize that mathematical notation might show the same effect due to translation into natural language during processing; if that is not the case, then an alternative explanation for the slower processing of negative polar adjectives is probably called for.

Perhaps one of the biggest unresolved questions in the analysis of comparatives and, indeed, all gradable expressions, involves the status of degrees. If degrees are crucial to an analysis of gradability in language and thus have a place in natural language metaphysics, what are they like? **Rick Nouwen** and **Jakub Dotlačil** ("Plural comparison?") focus specifically on the possibility that not only do degrees constitute a subsort of entity, but moreover the domain of degrees has the same mereological structure as other (sub)domains of entities. On this view, thus, there are pluralities of degrees.

Nouwen and Dotlačil suggest that positing pluralities of degrees could help overcome problems faced by the analysis of sentences like *The participants typed faster than each of them wrote*, where a plural or quantified expression appears in the *than*-clause of a comparative. Specifically, in some cases such sentences appear to involve quantifier raising out of the *than*-clause (so that, for example, the just mentioned sentence could be understood as equivalent to "Each participant's writing speed is such that all of the participants typed faster than that speed"), despite the fact that the general theory of quantifier scope does not predict *than*-clause-internal quantifiers to be able to raise. As an alternative, they suggest that the reading can be explained if the *than*-clause introduces a plurality of degrees, and comparison is cumulative. To test this hypothesis, they carried out a judgment task in which subjects had to decide whether a given sentence truthfully described a particular situation of plural comparison, presented in the form of a graph. Their results are better explained by a semantics that includes the possibility of pluralities of degrees and cumulative comparison than one in which such an option is not available.

Acknowledgements Most of the chapters in this volume are based on work presented at the workshop "Gradability, Scale Structure, and Vagueness: Experimental Perspectives," held at CSIC's Centro de Ciencias Humanas y Sociales (CCHS) in Madrid, in May 2015; two additional contributions were invited. We thank the members of the Linguistics and Cognitive Science Group (LyCC), led by Violeta Demonte, for their support in the organization of the workshop. We are also grateful to our contributors and to the other participants in the workshop for the lively discussion we had there. We also acknowledge the various sources of funding that helped underwrite the workshop and the preparation of this volume. Castroviejo and Sassoon were partially supported by grant FFI2012-34170, from the Ministry of Economy and Competitiveness (MINECO). Castroviejo also acknowledges grant FFI2015-66732-P (Ministry of Economy, Industry and Competitiveness, and the European Regional Development Fund, FEDER, EU), the IT769-13 Research Group (Basque Government), and UFI11/14 (University of the Basque Country). McNally acknowledges Spanish grants FFI2013-41301-P (MINECO) and FFI2016-76045-P (AEI/MINEICO/FEDER, UE), AGAUR (Catalan government) grant 2014SGR698, an ICREA Foundation Acadèmia award, and the Department of Translation and Language Sciences at Universitat Pompeu Fabra. Sassoon acknowledges the support of the German-Israeli Foundation for Scientific Research and Development (GIF). Thanks also to our abstract and manuscript reviewers, to Helen van der Stelt and Jolanda Voogd at Springer, and especially to Chungmin Lee for his initial encouragement to produce this volume.

References

Alxatib, S., & Pelletier, F. J. (2011). The psychology of vagueness: Borderline cases and contradictions. *Mind and Language*, *26*(3), 287–326.

Anderson, C., & Morzycki, M. (2015). Degrees as kinds. *Natural Language and Linguistic Theory*, *33*(3), 791–828.

Aparicio, H., Xiang, M., & Kennedy, C. (2015). Processing gradable adjectives in context: A visual world study. In S. D'Antonio, M. Moroney, & C. Little (Eds.), *Proceedings of Semantics and Linguistic Theory* (Vol. 25, pp. 413–432).

Aparicio, H., Xiang, M., & Kennedy, C. (2018). Informativity and grammar in referencial effects of contrast involving adjectivally modified NPs. In E. Castroviejo, L. McNally, & G. W. Sassoon (Eds.), *Gradability, vagueness, and scale structure: Experimental perspectives*. Berlin: Springer.

Bale, A. (2008). A universal scale of comparison. *Linguistics and Philosophy*, *31*(1), 1–55.

Barner, D., & Snedeker, J. (2008). 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*(3), 594–608.

Barsalou, L. W. (1993). Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A. F. Collins, S. E. Gathercole, M. A. Conway, & P. E. Morris (Eds.), *Theories of memory* (pp. 29–101). East Sussex, UK: Lawrence Erlbaum Associates.

Bartsch, R. (1984). The structure of word meanings: Polysemy, metaphor, metonymy. In F. Landman & F. Veltman (Eds.), *Varieties of formal semantics* (pp. 25–54). Dordrecht: Foris.

Bartsch, R. (1986). Context-dependent interpretations of lexical items. In J. Groenendijk, D. de Jongh, & M. Stokhof (Eds.), *Foundations of pragmatics and lexical semantics, GRASS 7* (pp. 1–26). Dordrecht: Foris.

Bartsch, R., & Vennemann, T. (1972). The grammar of relative adjectives and comparison. *Linguistische Berichte*, *20*, 19–32.

Beck, S., Krasikova, S., Fleischer, D., Gergel, R., Hofstetter, S., Savelsberg, C., et al. (2009). Crosslinguistic variation in comparison constructions. *Linguistic Variation Yearbook*, *9*(1), 1–66.

Beck, S., Oda, T., & Sugisaki, K. (2004). Parametric variation in the semantics of comparison: Japanese versus English. *Journal of East Asian Linguistics*, *13*(4), 289–344.

Beltrama, A. (2016). *Bridging the gap: Intensifiers between semantic and social meaning*. Dissertation, University of Chicago.

Beltrama, A. (2018a). Totally between discourse and subjectivity: Exploring the pragmatic side of intensification. *Journal of Semantics 35*(2), 219–267.

Beltrama, A. (2018b). Intensification, gradability and social perception: The case of totally. In E. Castroviejo, L. McNally, & G. W. Sassoon (Eds.), *Gradability, vagueness, and scale structure: Experimental perspectives*. Berlin: Springer.

Beltrama, A., & Bochnak, M. (2015). Intensification without degrees cross-linguistically. *Natural Language and Linguistic Theory*, *33*(3), 843–879.

Beltrama, A., & Staum Casasanto, L. (2017). *Totally* tall sounds *totally* younger: Intensification at the socio-semantics interface. *Journal of Sociolinguistics, 21*(2), 154–182.

Beltrama, A., & Xiang, M. (2012). Is good better than excellent? An experimental investigation on scalar implicatures and gradable adjectives. In *Proceedings of Sinn und Bedeutung* (Vol. 17, pp. 81–98).

van Benthem, J. (1982). Later than late: On the logical origin of the temporal order. *Pacific Philosophical Quarterly*, *63*, 193–203.

Bierwisch, M. (1989). The semantics of gradation. In M. Bierwisch & E. Lang (Eds.), *Dimensional adjectives* (pp. 71–261). Berlin: Springer.

Bobaljik, J. D. (2012). *Universals in comparative morphology: Suppletion, superlatives, and the structure of words*. Cambridge, MA: MIT Press.

Bochnak, M. (2013a). *Cross-linguistic variation in the semantics of comparatives*. Dissertation, University of Chicago.

Bochnak, M. (2013b). Two sources of scalarity within the verb phrase. In B. Arsenijević, B. Gehrke, & R. Marín, (Eds.), *Studies in the composition and decomposition of event predicates* (pp. 99–123). Dordrecht: Springer.

Bogal-Allbritten, E. (2012). Slightly coerced: Processing evidence for adjectival coercion by minimizers. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society* (Vol. 48.1, pp. 77–92). Chicago: Chicago Linguistic Society.

Bolinger, D. (1972). *Degree words*. The Hague: Mouton.

Büring, D. (2007). Cross-polar nomalies. In T. Friedman & M. Gibson (Eds.), *Proceedings of Semantics and Linguistic Theory* (Vol. 17, pp. 37–52). Ithaca, NY: Cornell University.

Burnett, H. (2016). *Gradability in natural language: Logical and grammatical foundations*. Oxford: Oxford University Press.

Bylinina, L. (2014). *The grammar of standards: Judge-dependence, purpose-relativity, and comparison classes in degree constructions*. Dissertation, Utrecht University.

Cappelletti, M., Fregni, F., Shapiro, K., Pascual-Leone, A., & Caramazza, A. (2008). Processing nouns and verbs in the left frontal cortex: A transcranial magnetic stimulation study. *Journal of Cognitive Neuroscience*, *20*(4), 707–720.

Cobreros, P. (2011a). Paraconsistent vagueness: A positive argument. *Synthese*, *183*(2), 211–227.

Cobreros, P. (2011b). Supervaluationism and classical logic. In R. Nouwen, R. van Rooij, H.-C. Schmitz, & U. Sauerland (Eds.), *Vagueness in communication*. Lecture Notes in Computer Science (Vol. 6517, pp. 51–63). Berlin: Springer.

Cobreros, P., Égré, P., Ripley, D., & van Rooij, R. (2012). Tolerance and mixed consequence in the supervaluationist setting. *Studia Logica*, *100*(4), 855–877.

Constantinescu, C. (2011). *Gradability in the nominal domain*. Dissertation, Leiden University.

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

Cummins, C. (2018). Modified fractions, granularity and scale structure. In E. Castroviejo, L. McNally, & G. W. Sassoon (Eds.), *Gradability, vagueness, and scale structure: Experimental perspectives*. Berlin: Springer.

Cummins, C., Sauerland, U., & Solt, S. (2012). Granularity and scalar implicature in numerical expressions. *Linguistics and Philosophy*, *35*(2), 135–169.

Deschamps, I., Agmon, G., Lewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition*, *143*(2), 115–128.

Devos, F. (1995). Still fuzzy after all these years. A linguistic evaluation of the fuzzy set approach to semantic vagueness. *Quaderni di Semantica*, *16*(31–32), 47–82.

Devos, F. (2003). Semantic vagueness and lexical polyvalence. *Studia Linguistica*, *57*(3), 121–141.

Doran, R., Baker, R. E., McNabb, Y., Larson, M., & Ward, G. (2009). On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics*, *1*(2), 211–248.

Égré, P., & Zehr, J. (2018). Are gaps preferred to gluts? A closer look at borderline contradictions. In E. Castroviejo, L. McNally, & G. W. Sassoon (Eds.), *Gradability, vagueness, and scale structure: Experimental perspectives*. Berlin: Springer.

Égré, P., De Gardelle, V., & Ripley, D. (2013). Vagueness and order effects in color categorization. *Journal of Logic, Language and Information*, *22*(4), 391–420.

Fara, D. G. (2000). Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics*, *28*(1), 45–81.

Fine, K. (1975). Vagueness, truth and logic. *Synthese*, *30*(3), 265–300.

Fleischhauer, J. (2016). *Degree gradation of verbs*. Düsseldorf: Düsseldorf University Press.

Frazier, L., Clifton, C., & Stolterfoht, B. (2008). Scale structure: Processing minimum standard and maximum standard scalar adjectives. *Cognition*, *106*(1), 299–324.

Grant, M. A. (2013). *The parsing and interpretation of comparatives: More than meets the eye*. Dissertation, University of Massachusetts, Amherst.

Grosu, A., & Landman, F. (1998). Strange relatives of the third kind. *Natural Language Semantics*, *6*(2), 125–170.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *18*(4), 441–461.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition*, *65*(2), 137–165.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science*, *31*(3), 355–384.

Hansen, N., & Chemla, E. (2017). Color adjectives, standards, and thresholds: An experimental investigation. *Linguistics and Philosophy*, *40*(3), 239–278.

Hay, J., Kennedy, C., & Levin, B. (1999). Scalar structure underlies telicity in degree achievements. In T. Matthews & D. Strolovitch (Eds.), *Proceedings of Semantics and Linguistic Theory* (Vol. 9, pp. 127–144). Ithaca, NY: Cornell University.

Heim, I. (1985). *Notes on comparatives and related matters*. Unpublished manuscript, University of Texas.

Heim, I. (2008). Decomposing antonyms. In *Proceedings of Sinn und Bedeutung* (Vol. 12, pp. 212–225).

Hyde, D. (1997). From heaps and gaps to heaps of gluts. *Mind*, *106*(424), 641–660.

Kamp, H. (2013). The paradox of the heap. In K. von Heusinger & A. G. B. Ter Meulen (Eds.), *Meaning and the dynamics of interpretation. Selected writings of Hans Kamp* (pp. 263–319). Leiden: Brill.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*(2), 129–191.

Keefe, R. (2000). *Theories of vagueness*. Cambridge: Cambridge University Press.

Kennedy, C. (1999). *Projecting the adjective: The syntax and semantics of gradability and comparison*. New York: Garland Press.

Kennedy, C. (2007a). Modes of comparison. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society* (Vol. 43, pp. 141–165). Chicago: Chicago Linguistic Society.

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

Kennedy, C. (2012). The composition of incremental change. In V. Demonte & L. McNally (Eds.), *Telicity, change, state: A cross-categorial view of event structure* (pp. 103–121). Oxford: Oxford University Press.

Kennedy, C. (2013). Two sources of subjectivity: Qualitative assessment and dimensional uncertainty. *Inquiry*, *56*(2–3), 258–277.

Kennedy, C., & Levin, B. (2008). Measure of change: The adjectival core of degree achievements. In L. McNally & C. Kennedy (Eds.), *Adjectives and adverbs: Syntax, semantics and discourse* (pp. 156–182). Oxford: Oxford University Press.

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

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, *4*(1), 1–45.

Krifka, M. (2002). Be brief and vague! And how Bidirectional Optimality Theory allows for verbosity and precision. In D. Restle & D. Zaefferer (Eds.), *Sounds and systems: Studies in structure and change: A Festschrift for Theo Vennemann* (pp. 439–458). Berlin: Mouton de Gruyter.

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

Lasersohn, P. (2005). Context dependence, disagreement, and predicates of personal taste. *Linguistics and Philosophy*, *28*(6), 643–686.

Liao, S.-Y., & Meskin, A. (2017). Aesthetic adjectives: Experimental semantics and context-sensitivity. *Philosophy and Phenomenological Research*, *94*(2), 371–398.

Liao, S.-Y., McNally, L., & Meskin, A. (2016). Aesthetic adjectives lack uniform behavior. *Inquiry*, *59*(6), 618–631.

Lidz, J., Halberda, J., Pietroski, P., & Hunter, T. (2011). Interface transparency and the psychosemantics of *most*. *Natural Language Semantics*, *6*(3), 227–256.

Matushansky, O. (2002). A beauty of a construction. In L. Mikkelsen & C. Potts (Eds.), *Proceedings of the 21st West Coast Conference on Formal Linguistics* (pp. 264–277), Somerville, MA: Cascadilla Press.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory and Cognition*, *6*(4), 462–472.

McNabb, Y. (2012). *The syntax and semantics of degree modification*. Dissertation, University of Chicago.

McNally, L. (2011). The relative role of property type and scale structure in explaining the behavior of gradable adjectives. In R. Nouwen, R. van Rooij, U. Sauerland, & H.-C. Schmitz (Eds.), *Vagueness in communication*. Lecture Notes in Computer Science (Vol. 6517, pp. 151–168). Berlin, Heidelberg: Springer.

McNally, L. (2017). Scalar alternatives and scalar inference involving adjectives: A comment on van Tiel, et al. 2016. In J. Ostrove, R. Kramer, & J. Sabbagh (Eds.), *Asking the right questions: Essays in honor of Sandra Chung* (pp. 17–28). https://escholarship.org/uc/item/8255v8sc.

McNally, L., & Stojanovic, I. (2017). Aesthetic adjectives. In J. Young (Ed.), *The semantics of aesthetic judgments* (pp. 17–37). Oxford: Oxford University Press.

Moltmann, F. (2004). Properties and kinds of tropes: New linguistic facts and old philosophical insights. *Mind*, *113*(449), 1–41.

Moltmann, F. (2007). Events, tropes, and truthmaking. *Philosophical Studies*, *134*(3), 363–403.

Moltmann, F. (2009). Degree structure as trope structure: A trope-based analysis of positive and comparative adjectives. *Linguistics and Philosophy*, *32*(1), 51–94.

Morzycki, M. (2009). Degree modification of gradable nouns: Size adjectives and adnominal degree morphemes. *Natural Language Semantics*, *17*(2), 175–203.

Morzycki, M. (2011). Metalinguistic comparison in an alternative semantics for imprecision. *Natural Language Semantics*, *19*(1), 39–86.

Nouwen, R., & Dotlačil, J. (2018). Cumulative comparison: Experimental evidence for degree cumulation. In E. Castroviejo, L. McNally, & G. W. Sassoon (Eds.), *Gradability, vagueness, and scale structure: Experimental perspectives*. Berlin: Springer.

O'Connor, E. (2015). *Comparative illusions at the syntax-semantics interface*. Dissertation, University of Southern California.

O'Connor, E., Pancheva, R., & Kaiser, E. (2012). Evidence for online repair of Escher sentences. In *Proceedings of Sinn und Bedeutung* (Vol. 17, pp. 363–380).

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*(1), 35–58.

Pancheva, R., & Tomaszewicz, B. (2011). *Experimental evidence for the syntax of phrasal comparatives in Polish* (Vol. 17.1, pp. 185–194). University of Pennsylvania Working Papers in Linguistics.

Peirce, C. S. (1910). *The Charles S. Peirce papers*. Cambridge, MA: Harvard University Press.

Piñón, C. (2008). Aspectual composition with degrees. In L. McNally & C. Kennedy (Eds.), *Adjectives and adverbs: Syntax, semantics and discourse* (pp. 183–219). Oxford: Oxford University Press.

Raffman, D. (1994). Vagueness without paradox. *The Philosophical Review*, *103*(1), 41–74.

Raffman, D. (2005). How to understand contextualism about vagueness: Reply to Stanley. *Analysis*, *65*(3), 244–248.

Rappaport-Hovav, M. (2014). Building scalar changes. In A. Alexiadou, H. Borer, & F. Schäfer, (Eds.), *The syntax of roots and the roots of syntax* (pp. 259–281). Oxford: Oxford University Press.

Ripley, D. (2011). Contradictions at the borders. In R. Nouwen, R. van Rooij, U. Sauerland, & H.-C. Schmitz (Eds.), *Vagueness in communication*. Lecture Notes in Computer Science (Vol. 6517, pp. 169–188). Berlin: Springer.

Rips, L. J., & Turnbull, W. (1980). How big is big? Relative and absolute properties in memory. *Cognition*, *8*(2), 145–174.

van Rooij, R. (2011). Vagueness and linguistics. In G. Ronzitti (Ed.), *Vagueness: A guide* (pp. 123–170). Berlin: Springer.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

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

Sassoon, G. W. (2012a). The double nature of negative antonymy. In A. Aguilar-Guevara, A. Chernilovskaya, & R. Nouwen (Eds.), *Proceedings of Sinn und Bedeutung* (Vol. 16.2, MIT Working Papers in Linguistics, pp. 543–556). Cambridge, MA: MIT Press.

Sassoon, G. W. (2012b). A slightly modified economy principle: Stable properties have non-stable standards. In *Proceedings of the Israel Association of Theoretical Linguistics (IATL)* (Vol. 27, MIT Working Papers in Linguistics, pp. 163–181). Cambridge, MA: MIT Press.

Sassoon, G. W. (2013a). A typology of multidimensional adjectives. *Journal of Semantics*, *30*(3), 335–380.

Sassoon, G. W. (2013b). *Vagueness, gradability and typicality: The interpretation of adjectives and nouns*. Leiden: Brill.

Sassoon, G. W. (2017). Comparisons of nominal degrees. *Language*, *93*(1), 153–188.

Sassoon, G. W., & Fadlon, J. (2017). The role of dimensions in classification under predicates predicts their status in degree constructions. *Glossa: A Journal of General Linguistics, 2*(1), 42, 1–40.

Sassoon, G. W., & Toledo, A. (2011). *Absolute and relative adjectives and their comparison classes*. Unpublished manuscript, University of Amsterdam and Utrecht University.

Sassoon, G. W., Meir, N., Fadlon, J., Anaki, D., & Schumacher, P. (t.a.). *The acceptability, processing and neural signature of nominal gradability*. Unpublished manuscript, Bar Ilan University, The Hebrew University, Jerusalem, and University of Cologne.

Schmidt, L. A., Goodman, N. D., Barner, D., & Tenenbaum, J. B. (2009). How tall is *tall*? Compositionality, statistics, and gradable adjectives. In *Proceedings of the Annual Conference of the Cognitive Science Society* (Vol. 31, pp. 2759–2764).

Schumacher, P. B., Brandt, P., & Weiland-Breckle, H. (2018). Online processing of *real* and *fake*: The cost of being too strong. In E. Castroviejo, L. McNally, & G. W. Sassoon (Eds.), *Gradability, vagueness, and scale structure: Experimental perspectives*. Berlin: Springer.

Scontras, G. (2014). *The semantics of measurement*. Dissertation, Harvard University.

Scontras, G., Graff, P., & Goodman, N. D. (2012). Comparing pluralities. *Cognition*, *123*(1), 190–197.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23.

Sedivy, J. C. (2005). Evaluating explanations for referential context effects: Evidence for Gricean mechanisms in online language interpretation. In J. Trueswell & M. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 345–364). Cambridge, MA: MIT Press.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

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

Seuren, P. (1973). The comparative. In F. Kiefer & N. Ruwet (Eds.), *Generative Grammar in Europe* (pp. 528–564). Dordrecht: Reidel.

Shapiro, S. (2006). *Vagueness in context*. Oxford: Oxford University Press.

Smith, L., Cooney, N. J., & McCord, C. (1986). What is “high”? The development of reference points for “high” and “low”. *Child Development*, *57*, 583–602.

Smith, L., Rattermann, M. J., & Sera, M. (1988). “Higher” and “lower“; Comparative and categorical interpretations by children. *Cognitive Development*, *6*, 131–145.

Soames, S. (1999). *Understanding truth*. Oxford: Oxford University Press.

Solt, S. (2011). How many *Mosts*? In *Proceedings of Sinn und Bedeutung* (Vol. 15, pp. 565–579).

Solt, S. (2016). On measurement and quantification: The case of *most* and *more than half*. *Language*, *92*(1), 65–100.

Solt, S. (2018). Multidimensionality, subjectivity and scales: Experimental evidence. In E. Castroviejo, L. McNally, & G. W. Sassoon (Eds.), *Gradability, vagueness, and scale structure: Experimental perspectives*. Berlin: Springer.

Solt, S., & Gotzner, N. (2012). Who here is tall? Comparison classes, standards and scales. In *Pre-Proceedings of the International Conference Linguistic Evidence 2012* (pp. 79–83). Tübingen: Eberhard Karls Universität Tübingen.

Stanley, J. (2003). Context, interest relativity and the Sorites. *Analysis*, *63*(280), 269–281.

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

Syrett, K. (2007). *Learning about the structure of scales: Adverbial modification and the acquisition of the semantics of gradable adjectives*. Dissertation, Northwestern University.

Syrett, K., & Lidz, J. (2010). 30-month-olds use the distribution and meaning of adverbs to interpret novel adjectives. *Language Learning and Development*, *6*(4), 258–282.

Syrett, K., Kennedy, C., & Lidz, J. (2010). Meaning and context in children's understanding of gradable adjectives. *Journal of Semantics*, *27*(1), 1–35.

van Tiel, B., van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, *33*(1), 137–175.

Tribushinina, E., & Gillis, S. (2012). The acquisition of scalar structures: Production of adjectives and degree markers by Dutch-speaking children and their caregivers. *Linguistics*, *50*(2), 241–268.

Tucker, D., Tomaszewicz, B., & Wellwood, A. (2018). Decomposition and processing of negative adjectival comparatives. In E. Castroviejo, L. McNally, & G. W. Sassoon (Eds.), *Gradability, vagueness, and scale structure: Experimental perspectives*. Berlin: Springer.

Umbach, C. (2016). Evaluative propositions and subjective judgments. In C. Meier & J. van Wijnbergen-Huitink (Eds.), *Subjective meaning: Alternatives to relativism* (pp. 127–168). Berlin: De Gruyter Mouton.

Vanden Wyngaerd, G. (2001). Measuring events. *Language*, *77*(1), 61–90.

Verheyen, S., & Storms, G. (2018). Education as a source of vagueness in criteria and degree. In E. Castroviejo, L. McNally, & G. W. Sassoon (Eds.), *Gradability, vagueness, and scale structure: Experimental perspectives*. Berlin: Springer.

Verheyen, S., Hampton, J. A., & Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the Rasch model. *Acta Psychologica*, *135*(2), 216–225.

Verheyen, S., Dewil, S., & Égré, P. (2016). Subjective meaning in gradable adjectives: The case of 'tall' and 'heavy'. Submitted.

de Vries, H. (2018). Gradable nouns as concepts without prototypes. In E. Castroviejo, L. McNally, & G. W. Sassoon (Eds.), *Gradability, vagueness, and scale structure: Experimental perspectives*. Berlin: Springer.

Wellwood, A. (2014). *Measuring predicates*. Dissertation, University of Maryland.

Wellwood, A., Pancheva, R., Hacquard, V., Fults, S., & Phillips, C. (2009). The role of event comparison in comparative illusions. Poster Presented at the *22nd Annual CUNY Conference on Human Sentence Processing*, Davis, CA.

Wellwood, A., Pancheva, R., Hacquard, V., & Phillips, C. (2017). The anatomy of a comparative illusion. Submitted.

Williamson, T. (1994). *Vagueness*. London: Routledge.

Are Gaps Preferred to Gluts? A Closer Look at Borderline Contradictions

Paul Égré and Jérémy Zehr

Abstract This paper examines the acceptance of so-called borderline contradictions involving vague adjectives. A close look at the available data from previous studies points toward a preference for “gappy” descriptions of the form “x is neither P nor not P” over “glutty” descriptions of the form “x is P and not P”. We present the results of an experiment in which we tested for that difference systematically, using relative gradable adjectives. Our findings confirm that both kinds of descriptions are accepted, but indeed that “neither”-descriptions are to a large extent preferred to “and”-descriptions. We examine several possible explanations for that preference. Our account relies on the distinction proposed by Cobreros et al. (J Philos Logic, 1–39, 2012) between *strict* and *tolerant* meaning for vague adjectives, as well as on a specific implementation of the strongest meaning hypothesis endorsed by Cobreros et al. as well as Alxatib and Pelletier (Mind Lang 26(3): 287–326 2011a). Our approach, however, argues in favor of local pragmatic strengthening instead of global strengthening in order to derive that preference.

Keywords Borderline contradictions · Strict-tolerant logic · Gradable adjectives
Vagueness · Presupposition · Local accommodation · Pragmatic strengthening
Three-valued logic

P. Égré (✉)
Département d’Études Cognitives et Département de Philosophie de l’École
Normale Supérieure, Institut Jean-Nicod, ENS, CNRS, EHESS, PSL Research
University, Paris, France
e-mail: paul.egre@ens.fr

P. Égré
Swedish Collegium for Advanced Study, Uppsala, Sweden

J. Zehr
University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: jeremy.e.zehr@gmail.com

E. Castroviejo et al. (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure*,
Language, Cognition, and Mind 4, https://doi.org/10.1007/978-3-319-77791-7_2

1 Introduction

Several experiments in the last decade indicate that sentences that correspond to literal contradictions from the standpoint of classical logic are accepted to a significant extent by naive subjects to describe *borderline cases* of a vague predicate, in a way that they are not for so-called *clear cases* (Ripley 2011a; Alxatib and Pelletier 2011a; Serchuk et al. 2011; Égré et al. 2013). Thus, Ripley (2011a) found that sentences like (1-a) and (1-b) are accepted significantly more by subjects presented with a circle-square pair at middling distance from one another, compared to a circle and a square touching each other, or to a circle and a square appearing further away from each other. Similarly, Alxatib and Pelletier (2011a) found that sentences of the form (2-a) and (2-b) are checked "True" significantly more by participants when pertaining to a man appearing of height $5'11''$ than for a man appearing of height $6'6''$ or for a man of height $5'4''$.

(1) a. The circle is and isn't near the square.
b. The circle neither is nor isn't near the square. Ripley (2011a)

(2) a. Man x is tall and not tall.
b. Man x is neither tall nor not tall. Alxatib and Pelletier (2011a)

Similar sentence forms are intuitively unacceptable when involving *precise* predicates (see (3)), setting aside cases of presupposition failure as in (4). Possibly, (4-b) might be acceptable, but only to convey that $\sqrt{2}$ is outside the domain of application of the predicate "prime number", assuming the latter applies only to integers (see Zehr 2014 for more on the difference between vagueness and presupposition failure)[1]

(3) a. # 9 is and isn't a prime number.
b. # 9 neither is nor isn't a prime number.

(4) a. # $\sqrt{2}$ is and isn't a prime number.
b. (?) $\sqrt{2}$ neither is nor isn't a prime number.

Prima facie, the acceptability of contradictory sentences in borderline cases (so-called "borderline contradictions", following Ripley 2011a's terminology) may not appear so surprising, since a borderline case of a vague predicate is often characterized as a case for which one feels equally attracted toward applying and toward denying the predicate (e.g. Peirce 1902). From a behavioral point of view, however,

[1] Sentence (4)-b may be judged outright false of course, since $\sqrt{2}$ *is not* a prime number. It seems to us acceptable in a context in which a teacher, let us say, would want to cut short a dispute between two pupils, one of them arguing that $\sqrt{2}$ is a prime integer, and the other arguing that $\sqrt{2}$ is not a prime integer, both mistakenly thinking it is an integer. Both pupils would wrongly presuppose that $\sqrt{2}$ is an integer, and the point of (4)-b would be to reject that presupposition. See Sect. 4.4.2 below for more on the analysis of such presuppositional sentences.

that characterization is compatible with subjects systematically rejecting descriptions of the form “*x* is *P* and not *P*”, or “*x* is neither *P* nor not *P*”, while judging “*x* is *P*” half the time, and denying “*x* is *P*” the other half.

The results of Ripley (2011a); Alxatib and Pelletier (2011a) suggest that that picture is inadequate, however. From a semantic point of view, those findings are not easily accommodated by either supervaluationist or subvaluationist theories of vagueness, which predict sentences of the form “*x* is *P* and not *P*” and “*x* is neither *P* nor not *P*” to remain contradictions in borderline cases (see Ripley 2013; Alxatib and Pelletier 2011a; Égré et al. 2013). They are more easily accommodated in paraconsistent-friendly frameworks or in fuzzy logic, however, that is in theories in which classical contradictions do not automatically receive the value False. One framework of particular to us is the so-called strict-tolerant framework (Cobreros et al. 2012), in which both kinds of sentences can be true tolerantly in borderline cases.

In this paper, we propose to dig further into the explanation of borderline contradictions. In what follows we shall refer to sentences of the form “*x* is *P* and not *P*” as *conjunctive* descriptions of borderline cases, or *conjunctions* for short, and to sentences of the form “*x* is neither *P* nor not *P*” as *negative disjunctive* descriptions, or *disjunctions* for short. More often, we will refer to them simply as “and”-descriptions and “neither”-descriptions. Although we shall sometimes use the expressions “glutty” descriptions and “gappy” descriptions, we will favor the “and” versus “neither” denominations, which in a sense are the most theory-neutral. The phenomenon we are interested in is whether the two kinds of descriptions are equally accepted in borderline cases, or whether there is a preference for one kind of description over the other. Our intuition tells us that disjunctions might be preferred to conjunctions, that is, it might be easier to describe a borderline case as “neither tall nor not tall” than as “tall and not tall”.

A closer look at extant results suggests that this is likely to be the case (see the review in Sect. 2). The difference, however, has not been an object of attention in previous studies. We propose to test for that preference. One motivation to do so is that the difference can potentially cast further light on the selection between two kinds of meaning for vague predicates. Alxatib and Pelletier (2011a, b) and Cobreros et al. (2012, 2015b) both treat vague predicates as pragmatically ambiguous between a strong and a weak meaning (aka *strict* and *tolerant* meanings, see Cobreros et al. 2012). The strong or strict meaning of a predicate like “tall” is intuitively equivalent to “clearly tall” and the weak or tolerant meaning to “relatively tall” (“not clearly not tall”). Alxatib and Pelletier (2011a, b) and Cobreros et al. (2012, 2015b) both formulate the hypothesis that the strong meaning ought to be selected first, but do not look at whether the principle should entail a preference for negated disjunctions over conjunctions. In Sect. 3, we present the results of an experiment confirming our main intuition. In Sect. 4, we propose algorithm intended to account for that result. This algorithm too implements the idea that the strict meaning is selected before the tolerant meaning, but importantly it rests on the idea that pragmatic strengthening is done locally, rather than globally for whole sentences.

2 Gaps and Gluts: A Brief Review

Borderline cases of a vague predicate are commonly described either as cases leaving a gap between the positive extension of the predicate and its negative extension, or in a dual manner, as cases where the positive extension of the predicate and its negative extension overlap, thereby creating a glut (see Fine 1975; Égré et al. 2013). Intuitively, a negative disjunctive description of the form "*x* is neither *P* nor not *P*" matches the idea of a gap between the positive and the negative extension, and a conjunctive description of the form "*x* is both *P* and not *P*" the idea of a glut. The first question we seek to clarify is whether "glutty" and "gappy" descriptions of borderline cases are used to the same extent.

Ripley (2011a) presented participants with two kinds of conjunctive and disjunctive descriptions, which he calls elided versus non-elided. In an elided conjunction, the conjunction is internal to the VP ("The circle both is and isn't near the square"), whereas in a non-elided conjunction, it is sentential ("The circle is near the square and it isn't near the square"). In both what Ripley calls non-elided disjunctions ("The circle neither is near the square nor isn't near the square") and elided disjunctions ("The circle neither is nor isn't near the square"), the disjunction is VP-internal, but what varies is whether part of the VP is elided or not. Ripley presented participants with 7 pairs of a circle and square at varying distances from one another, and asked them to rate each description for each stimulus on a 1–7 scale, with 1 labeled 'Disagree' and 7 labeled 'Agree'. In Fig. 1, we give a summary of his data, where we aggregated scores for so-called elided versus non-elided description types. Prima facie, we see no preference for one description type over the other, neither globally, nor in the specific case of stimulus C which receives the highest assent to both description types.

On the other hand, we do observe an overall preference for disjunctions over conjunctions in the experiment run by Alxatib and Pelletier (2011a). Alxatib and Pelletier's methodology was different from Ripley's, since participants had to check True, False, or Can't Tell to four kinds of description including "Tall", "not Tall", "Tall

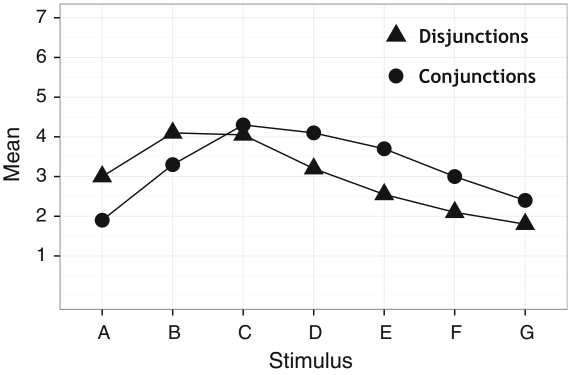


Fig. 1 Ripley 2011's data: triangles represent aggregate scores for disjunctions, and dots aggregate scores for conjunctions. Mean scores are highest for non-extreme distances (B and C)

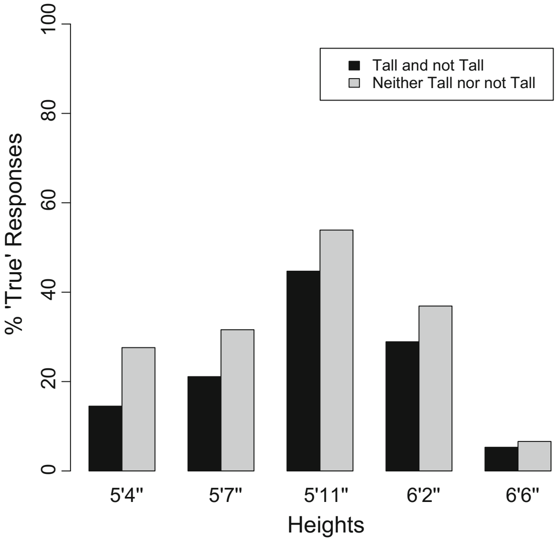


Fig. 2 Alxatib and Pelletier (2011a)'s data: proportion of 'True' checks in conjunctions and disjunctions

Table 1 Alxatib and Pelletier (2011a)'s data for the central stimulus of height 5′11″ (*n* and *b* refer to "neither"- vs. "both"-sentences, *T*, *F*, *C* to "True", "False" and "Can't tell" responses)

	Tn	Fn	Cn	Total
Tb	22	12	0	34
Fb	13	18	0	31
Cb	6	2	3	11
Total	41	32	3	76

and not Tall", and "Neither tall nor not tall". In Fig. 2, we reproduce the proportions of 'True' checks to the latter two kinds, namely conjunctions and disjunctions. What we can observe is a higher level of 'True' checks to disjunctions, in each stimulus. For the stimulus of middling height in their stimulus set, Alxatib and Pelletier (2011b) report the data reproduced in Table 1, where Tn, Fn, and Cn stand for the numbers of participants responding "True", "False" and "Can't tell" to "#2 is neither tall nor not tall", and where Tb, Fb, and Cb give the corresponding numbers for "#2 is both tall and not tall". As they point out, a McNemar-Bowker test for symmetry gives a value of 8.04, with $p < 0.05$, suggestive of a difference between the two descriptions.

Serchuk and colleagues also investigated borderline contradictions, this time comparing more explicitly than Ripley the effect of having two kinds of negation, VP-internal or sentential. Unlike Ripley or Alxatib and Pelletier, they did not use perceptual stimuli, but asked participants to imagine borderline cases, describing those in terms of the operator "clearly". For instance, in one of their scenarios, a woman named Susan is described as being "somewhere between women who are clearly rich and women who are clearly non-rich". For disjunctions, they used *unnegated*

Table 2 Serchuk et al. (2011)'s data on disjunctive and conjunctive descriptions of borderline cases

Description	'True'	'False'	Other
Either *x* is *P* or *x* is not *P*	113	137	100
x is *P* or it is not the case that *x* is *P*	141	88	121
x is *P* and *x* is not *P*	66	195	89
x is *P* and it is not the case that *x* is *P*	25	247	78

disjunctions of two kinds: "Either Susan is rich or Susan is not rich", and "Susan is rich or it is not the case that Susan is rich". Each time, however, the disjunction is sentential and not VP-internal: they did not present participants with "Susan is rich or not rich". They used a similar pair of probes for conjunctions, namely "Susan is rich and Susan is not rich", and "Susan is rich and it is not the case that Susan is rich" (here too, they did not use "Susan is rich and not rich"). Participants had to check exactly one answer among six possible answers for each sentence, within the set consisting of {"true", "not true but also not false", "partially true and partially false", "false", "both true and false", "true or false but I don't know which"}. In Table 2, we present Serchuk et al.'s data in a table in which we collapsed the responses other than True and False under a third category "Other". In this case, we cannot directly compare "True" responses to "neither" and "and" descriptions, because Serchuk et al. did not use any sentence of the form "neither... nor". However, we can get an indirect idea of the participants' judgments about "neither" sentences, granted that they would stand in an inverse relation to those for "either" sentences. We shall therefore compare the proportion of "True" responses for conjunctions to the proportion of "False" responses for disjunctions. There is a higher proportion of False answers to the disjunction "either x is P or x is not P" than of "True" answers to "x is P and not P", and similarly when we compare "False" answers to "x is P or it is not the case that x is P" with "True" answers to "x is P and it is not the case that x is P". A Fisher test based on that comparison yields a significant difference for the former two sentences ($p < 10^{-10}$) as well as for the latter two ($p < 10^{-16}$).

Finally, we mention the results of an unpublished study, by Solt and Gotzner (2010). They showed participants picture series depicting either suitcases of various sizes, bluejeans of varying price, or cities of different distances from Berlin. The goal of the experiment was to see which pictures in each series would be classified by participants as satisfying the respective adjectives "big" (*gross*), "expensive" (*teuer*), "far" (*weit*) in comparison to their negation or the polar antonym. In one condition participants had to decide which pictures satisfied the adjective, and which satisfied the syntactic negation. In another condition a different group had to decide which pictures satisfied the adjective, and which satisfied the polar antonym. Their results indicate that only a small minority of participants left no gap between the adjective and its opposite (whether polar or syntactic), or ascribed the same item to both descriptions. Although that study does not investigate the acceptance of complex sentences like the ones we are interested in, it supports the idea of a preference for gaps over gluts when people are asked to judge of a sentence and its negation separately.

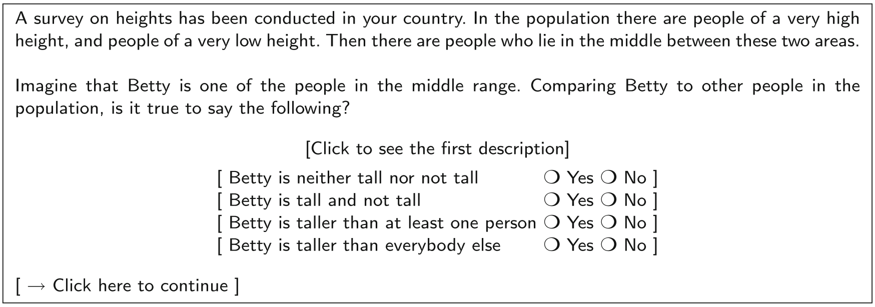
What the present review points toward, therefore, is a preference for the description of borderline cases in terms of negated disjunctions ('neither P nor not P') over conjunctions ('P and not P'). Because the data we reviewed are partial and unsystematic, we proceed to test that hypothesis more systematically in what follows. An important caveat is that we limited our study to sentences of the form "*x* is P and not P" and "*x* is neither P nor not P". We did not test sentences in which the conjunction or the disjunction are clearly sentential ("*x* is P and *x* is not P"/"neither is *x* P nor is *x* not P") nor sentences in which the negation is propositional ("it is not the case that"). From Serchuk et al.'s data, we see that there is an effect of those variations on acceptance, but what we were interested in is whether there is a contrast between conjunctive and disjunctive sentences for those cases in which the sentences seem best accepted in the first place.

3 Experiment

In order to test the preference for "gappy" descriptions over "glutty" descriptions of borderline cases, we designed an experiment intended to compare the acceptance of both kinds of descriptions. The experiment we report here is the third and main one in a series of three, and it replicates the results of the two previous, pilot versions. We presented participants with different scenarios involving different adjectives, each time involving the verbal description of a borderline case, drawing inspiration from Serchuk et al. (2011), who basically asked participants to imagine borderline cases based on similar verbal descriptions. We then asked participants to judge the adequacy of contradictory descriptions of the forms in (5), leaving them the possibility to either accept or reject the description.

(5) a. *Neither*-descriptions: "BORDERLINE is neither ADJECTIVE nor NOT ADJECTIVE"
b. *And*-descriptions: "BORDERLINE is ADJECTIVE and NOT ADJECTIVE"

Our prediction was that we should see a higher rate of acceptance of "neither"-descriptions over "and"-descriptions, but also that either should be significantly more accepted than outright false sentences. In order to get a homogenous set of results, however, our first step was to define general principles for the selection of adjectives and for the construction of our scenarios.



A survey on heights has been conducted in your country. In the population there are people of a very high height, and people of a very low height. Then there are people who lie in the middle between these two areas.

Imagine that Betty is one of the people in the middle range. Comparing Betty to other people in the population, is it true to say the following?

[Click to see the first description]

[Betty is neither tall nor not tall ○ Yes ○ No]
[Betty is tall and not tall ○ Yes ○ No]
[Betty is taller than at least one person ○ Yes ○ No]
[Betty is taller than everybody else ○ Yes ○ No]

[→ Click here to continue]

Fig. 3 Example of a trial. [Bracketed texts] appeared dynamically: each description appeared after answering the preceding one or clicking Click to see the first description. The set of answers was validated by clicking on Click here to continue (appearing after the last click)

3.1 Design

The experiment was a block design: each participant faced two series of descriptions. In one block, the descriptions were as exemplified in (5). In the other block, NOT ADJECTIVE was replaced with a corresponding antonym (e.g. *short* in place of *not tall*). In this paper, we do not report the results for these antonymic descriptions: we treat them as fillers, even though they served as critical conditions for the purposes of another study investigating the relation between the two types of negations (syntactic vs. lexical).

The design took the form of an acceptability task: participants were presented with a fictive scenario verbally depicting a borderline case on a given gradable dimension and had to tell whether they would accept four descriptions of this borderline case, as exemplified in Fig. 3. On each trial, they saw the scenario first, followed by the instruction *Click to see the first description*. The instruction would disappear after clicking and simultaneously reveal the first of the four descriptions to be assessed. Each time, clicking on one of *Yes* and *No* would dynamically let the next description appear on screen. For the last description, checking *Yes* or *No* would reveal a link needed to validate all four answers of the current trial, reading *Click here to continue*. As long as the descriptions were visible on the screen, the participants were able to modify their judgments. Once they clicked on *Click here to continue*, they would see the next trial, with a new scenario targeting a distinct adjective.

3.2 Materials

3.2.1 Scenarios

Serchuk et al. (2011) used verbal descriptions of borderline cases in their experiments, as exemplified in (6), and then asked their participants to judge a list of sentences about these borderline cases. What follows is an example of the sentences they used:

(6) "Imagine that on the spectrum of rich women, Susan is somewhere between women who are clearly rich and women who are clearly non-rich".

To the extent that their results revealed acceptance of contradictory descriptions, they show that participants readily represent borderline cases when explicitly asked to do so. This prompted us to describe borderline cases in similar verbal ways. However, in contrast with Serchuk et al.'s descriptions, we did not let the target adjectives (e.g. *rich* in (6)) appear in the scenarios, to avoid any priming effect. Indeed, the description in (6) might prime a *clearly* reading of *rich* that would lead to systematically exclude the borderline cases from the positive extension of *rich*. Alternatively, it might prime a contrastive, looser use of *rich* leading one to categorize the borderline cases in the positive extension. If any of these effects were real, they could bias our results in a way that does not appear to be the case in Alxatib and Pelletier (2011a)'s pictorial context. Because of that, we formulated our scenarios by referring to the *scale* associated with the adjectives, using morphologically unrelated nouns whenever possible (cf. Fig. 3 where we used the word *heights* to refer to the scale associated with *tall*). We hypothesized that participants would naturally represent the individuals in the middle of these scales as borderline cases for the target adjectives. Note that contrary to pictorial representations, verbal descriptions have the advantage of letting subjects build their own, ideal representations of what a borderline case might be on the given scale. The particular scenarios we used are reported in Appendix 1.

3.2.2 Selection of Adjectives: Four Principles

Not all adjectives come with borderline cases, and not all adjectives seem to define their borderline area in the same way. In order to have a set of adjectives as homogeneous as possible, we selected them along four criteria that we present here. Our selection was partly guided by principles investigated in the works of Roche (2012) and Ruytenbeek (2013), both of them conducted under the supervision of Benjamin Spector, concerning the negation of adjectives.[2]

(i) *Gradable*

As exemplified in the scenario in Fig. 3 all the borderline cases we described were presented as lying in the central region of *a given scalar dimension*.[3] All the adjectives

[2] See Ruytenbeek et al. (2017) for a published follow-up to that work.

[3] In this paper, we take this *non-extreme* property to be definitory of borderline cases. By contrast, one might consider that an *extremely* insane person in a *perfect* physical shape could be a borderline case

we used are gradable and therefore define a scale along which their arguments are non-trivially ordered. We say that an adjective is gradable if and only if at least one of the following constructions is perceived as natural.

(7) a. X is ADJECTIVE-er than Y.
b. X is more ADJECTIVE than Y.

For example *tall* is gradable (8-a), whereas *underage* is not (8-b).

(8) a. Bill is taller than Sue.
b. #Bill is more underage than Sue.

Criterion 1: We excluded any non-gradable adjective from our list.

(ii) *Relative Versus Absolute*

Two kinds of gradable adjectives have been distinguished in the literature: relative (gradable) adjectives like *tall* and absolute (gradable) adjectives like *full* (Unger 1975; Kennedy 2007). Following one of Kennedy's tests, we say that a gradable adjective is absolute if the following entailment holds, and that it is relative otherwise:

(9) X is the ADJECTIVE one. $\rightsquigarrow$ X is ADJECTIVE (generally speaking)
a. Relative: My glass is the tall one. $\not\rightsquigarrow$ my glass is tall (generally speaking)
b. Absolute: My glass is the full one. $\rightsquigarrow$ my glass is full (generally speaking)

For example *tall* is relative because the inference is *not* systematic (9-a), but *full* is absolute because the inference *is* systematic (9-b). For the purpose of our study we chose to include only relative gradable adjectives. Our reason to do so was the following: intuitively, absolute adjectives denote an endpoint on a scale. For example, *full* denotes the maximum extent to which a recipient can be filled with substance.[4] A relative adjective like *tall*, on the other hand, does not select any context-invariant standard on a scale. Instead, it refers to a context-sensitive standard. In the way we constructed our stimuli, we always asked participants to imagine people or objects standing in the "middle range" of two extreme regions denoted by "very high" and "very low" along the corresponding dimension. For absolute gradable

for *healthy* if judgments for the sentence "this person is healthy" can be unclear and ambivalent. However, we excluded multi-dimensional adjectives from our sample as far as possible (see the discussion of evaluativity below), in order to rule out borderline cases arising from the relative weight of competing dimensions.

[4] The maximum standard to consider a glass "full" is still context-dependent: for instance, McNally (2011) discusses how wine need not reach the top of the glass for it to be considered full. A glass half-filled with wine can be called "full of wine" in some contexts, thereby threatening the generalization in (9). We don't think that undermines the point we are making here, however, as further tests can be used to corroborate the classification of "full" as absolute. However, the reader is invited to replace *full* by *empty* in our examples, as "empty" appears to show less context-sensitivity (in relation to glasses at least).

adjectives, we would have to adapt the descriptions. For example: if we were to ask participants to consider how they would describe glasses in the middle range between those that are "very filled" and those that are "filled very little", we would very likely fail to target the borderline region for what counts as "full". Intuitively, the borderline region for "full" is a region that is close enough to the maximum degree to which a recipient can be filled. Another reason we had for not including absolute adjectives in this experiment was that we also included antonyms, and here again, we can expect antonyms of absolute adjectives to not behave exactly like antonyms of relative adjectives (see Burnett (2016) for discussion). In summary, we did not include absolute adjectives in our experiment mostly to have a homogenous set of descriptions and adjectives to test.[5]

Criterion 2: We excluded absolute gradable adjectives from our list.

(iii) *Non-Evaluative, One-dimensional*

As Kennedy (2013) notes, all relative adjectives appear to be subjective. That is to say, they seem to systematically allow for faultless disagreement: substituting *tall* for ADJECTIVE in (10), as in (10-a), does not necessarily imply that either Mary or Sue is wrong, but doing the same with *prime* as in (10-b) *does* imply that either Mary or Sue is wrong. While Mary and Sue can truly diverge on what heights they consider to be tall, there is an objective standard for primeness and therefore one of them has to be wrong.

(10) *X* thinks *Z* is ADJECTIVE but *Y* does not.

a. Subjective: Mary thinks Paul is tall but Sue thinks he is not $\not\rightsquigarrow$ either Mary or Sue is wrong.

b. Not subjective: Mary thinks this number is prime but Sues thinks it is not $\rightsquigarrow$ either Mary or Sue is wrong.

In agreement with Kennedy's generalization, all the adjectives in our list satisfy the test of subjectivity.

However, some subjective adjectives are also *evaluative*, while others are not. We use the category *evaluative* in the sense of Kennedy (2013). That is, we call an adjective *evaluative* if substituting it for ADJECTIVE in (11) results in a non-deviant sentence.[6]

(11) *Z* finds that *X* is ADJECTIVE-er than *Y*.

[5] Some authors consider that the status of what we call borderline cases for absolute adjectives is due to a different phenomenon from the one at play with relative adjectives. For instance Kennedy (2007) claims that calling an almost-full glass "full" is a manifestation of *imprecision*, but that there is in fact a sense in which such a glass is uncontroversially *not* full. By contrast, for relative adjectives, there would be no non-arbitrary way to settle the question for borderline cases for they are a manifestation of *vagueness* proper.

[6] See Sæbø (2009) who first proposed tests of this sort. This sense of "evaluative", although related, is more specific than Rett (2007)'s sense, who calls an expression "evaluative if it makes reference to a degree that exceeds a contextually specified standard".

a. Evaluative: Mary finds that Paul is smarter than Joe.

b. Not evaluative: # Mary finds that Paul is taller than Joe.

As pointed by Kennedy (2013), evaluativity is a kind of subjectivity, as it correlates with faultless disagreement. However, evaluativity as diagnosed in (11) generally implies multi-dimensionality. In the case of "smart", evaluativity appears to correlate with the availability of different possible *respects* or *criteria* for building a scale on which one later establishes a threshold. Thus, one could entertain a relativistic approach where *X is* ADJECTIVE *and* NOT ADJECTIVE is interpreted as *X is* ADJECTIVE *according to some criterion and* NOT ADJECTIVE *according to some other criterion* (see Kamp and Partee (1995) for this observation, and Solt, this volume, for further remarks on that). To forestall any such explanation of our results,

Criterion 3: We excluded any evaluative adjective from our list.

In agreement with this criterion, the majority of the adjectives in our sample are associated with a unique salient dimension of comparison (such as "age" for "old", "height" for "tall", etc). One exception in our list is "rich", for which we mention "wealth", but for which we hint at a one-dimensional representation by talking of "degree of wealth".

(iv) *Individual-versus Stage-level*

The last criterion we used is based on the distinction between *individual*-level predicates and *stage*-level predicates. After Carlson (1977), we call an adjective *individual-level* if substituting it for ADJECTIVE in (12) results in a deviant sentence, and *stage-level* otherwise.

(12) There are two NOUN ADJECTIVE.

a. Individual-level: ? There are two men tall.

b. Stage-level: There are two men happy.

The properties attributed by individual-level predicates seem to be more inherent to their argument than those attributed by stage-level predicates. Once again, a relativistic approach could try to explain the acceptance of contradictory descriptions under a reading like *X is* ADJECTIVE *on some occasions and* NOT ADJECTIVE *on some other occasions*. To avoid that, we also relied on the following criterion:

Criterion 4: We excluded any stage-level adjective from our list.

Sentence types: Objet-oriented and human-oriented

It is important to stress that we used the four mentioned criteria fundamentally as a heuristics to build our stimuli. Our intention was not to test for those properties, nor to rigorously control for them, but to avoid confounds such as the ones we just discussed. Furthermore, the 8 adjectives that we used were distributed into two types of sentences: 4 adjectives were predicated of a subject denoting a human being (*tall*, *rich*, *heavy*, *old*) and 4 adjectives were predicated of a subject denoting an object or a property related to a human being (*loud*,*fast*, *large*, *wide*). Note that this division is not inherent in the meaning of the adjectives ("heavy" for instance is applicable to both persons and objects), but we imposed it arbitrarily to create variety in the comparison classes. We therefore did not treat the object-oriented/human-oriented distinction as a controlled, independent variable, for we did not expect that it should impact our main prediction. We only found it useful to have a balanced set of examples. We presented half the participants with the four adjectives of the former type *first* and the four adjectives of the latter type *second*; the other half of the participants saw the reverse order. This parameter was crossed with the type of negation (syntactic vs. polar) and each participant responded to *syntactic* descriptions built with either human-oriented or object-oriented adjectives exclusively.

3.3 Participants

We recruited 148 participants online and anonymously via the Amazon Mechanical Turk platform. There a link would redirect them to the Ibex servers, on which the experiment was developed and hosted. Before going through the actual trials, participants first had to complete a pre-questionnaire consisting of seven simple questions. They were also presented with a post-questionnaire that was used for the study on the syntactic vs antonymic negation. Those forms are reported in Appendix 2. Accuracy on the pre-questionnaire and on controls was very good and no participant was excluded.[7] Though we didn't explicitly require our participants to be speakers of English, near-ceiling accuracy shows that all our participants had a good understanding of the English language, which was a precondition to evaluate the sentences that we used as our stimuli. Each participant was assigned one of four groups: two groups of participants judged either 4 human-oriented or 4 non-human-oriented syntactic descriptions *before* judging polar descriptions (again, treated as fillers in this paper), and the two other groups judged them *after* the series of polar descriptions.

3.4 Results

The barplot in Fig. 4 reports the average acceptance of each of the four types of descriptions exemplified in Fig. 3.[8] The plot in Fig. 5 shows how the participants behaved regarding the acceptance of "and"- and "neither"-descriptions based on the number of trials in which they gave similar answers: few participants ever rejected both descriptions (left-most bar) or ever accepted the "and"-description while rejecting the "neither"-description (second bar from the left). The very few participants

[7] In this case, exclusion of participants would lead the regression models that we ran to not converge, because of an insufficient variability on the controls.

[8] The results were descriptively similar in all conditions. A graph presenting the results for each condition is included in Appendix 3.

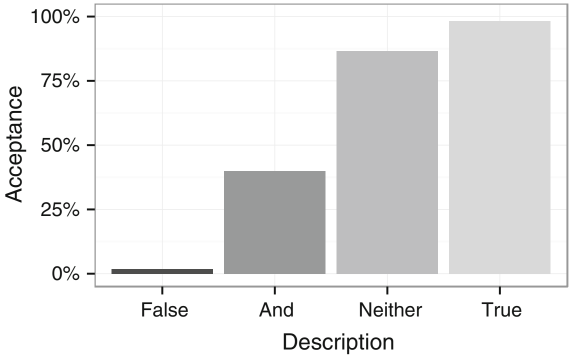


Fig. 4 Mean acceptance by description type

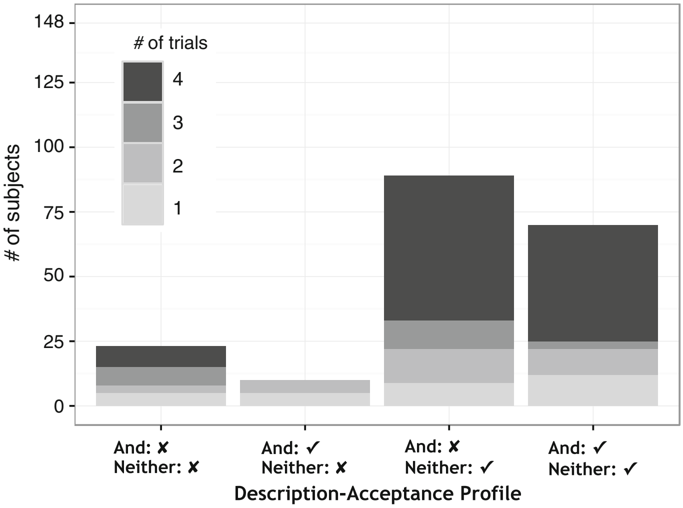


Fig. 5 Distribution of participants regarding the acceptance of "and"- and "neither"-descriptions based on the number of trials showing similar answers

who ever showed the latter behavior did it on only one or two trials, out of four. For most of the participants and most of the trials, the participants either accepted the "neither"-description while rejecting the "and"-description (third bar from the left), or they accepted both descriptions (right-most bar).

We ran mixed effect logistic regression models to analyze the data. We used the $glmer$ function from the $lme4$ package (version 1.1–11) for R (version 3.1.2) with the optimizer "bobyqa" to compute the most complex models that would converge, following Barr et al. (2013). Several models with the same level of complexity would converge, but they all predicted the observation of a "Yes" response depending on Description (*Neither* vs. *And* vs. *Control True* vs. *Control False*) with random intercepts for Participant and Adjective. They differed in whether they also included random slopes for Adjective, and whether they included random intercepts and random slopes for Block (*Before* vs. *After* judging descriptions with antonyms) or for Adjective Type (*Human* vs. *Non-Human* oriented).[9]

Throughout these models, we consistently found that the "and"-descriptions were accepted significantly more often than the control false descriptions, and significantly less often than the "neither"-descriptions. The models indicate that the "neither"-descriptions were accepted less often than the control true descriptions. We report the outputs of each model in Appendix 3.

3.5 Summary

The results confirm the previous observations from the experimental literature that we reported, according to which speakers do accept contradictory "neither"- and "and"-descriptions to describe borderline cases, given that those were accepted significantly more often than the control false descriptions. However, they provide us with a more nuanced picture of these judgments: first of all, it seems that neither of the two types of contradictory descriptions is generally as acceptable as plainly true (control) descriptions (a point that was left open in the existing experiments); and secondly, they show that "neither"-descriptions are significantly preferred to "and"-descriptions, eliciting a difference that was hinted at in the results of Alxatib and Pelletier (2011a) but not evidenced in those of Ripley (2011a). Note that the general preference for "neither"-descriptions over "and"-descriptions cannot be due to two different types of population, where one type of population (a majority) would accept only "neither"-descriptions and the other one (a minority) would accept only "and"-descriptions. Comparing the two bars corresponding to acceptance of "and"-descriptions in Fig. 5 (the second bar from the left and the right-most bar), it appears that for every participant, the acceptance of the "and"-description almost always co-occurred with the acceptance of the "neither"-description. By contrast, the "neither"-descriptions were often accepted alone (third bar from the left). As a consequence, on the one hand, these observations support the project of developing a system accounting for the acceptance of both kinds of descriptions but on the other hand, such a system should also account for the preference of "neither"-descriptions over "and"-descriptions.

4 Explaining the Asymmetry

Our two main predictions in undertaking this experiment were confirmed: first, we see that both "and"-descriptions and "neither"-descriptions are accepted significantly more than false control sentences using the same predicates, and secondly we see a

[9] See Appendix 3 for a list of the models.

marked preference for "neither"-descriptions over "and"-descriptions. This finding raises two main questions: the first is how we can explain the preference for "neither"-descriptions over "and"-descriptions, even as restricted to the class we considered in our study. The second is whether we can expect it to be robust across other kinds of adjectives than the ones we considered. In this section we focus on the first issue. We first discuss whether an explanation in terms of some domain-general bias is plausible. We propose instead a specific implementation of the strongest meaning hypothesis, based on the notions of strict versus tolerant meanings and the notion of local strengthening.

4.1 Omission Bias and Consistency Bias

One way in which one could be tempted to explain the preference for "neither"-sentences over "and"-sentences to describe a borderline case is as an instance of a domain-general bias toward omission rather than commission (see Spranca et al. 1991; Bonini et al. 1999). The explanation might go like this: a borderline case of a tall person is one for which participants feel uncertain whether to apply the predicate "tall" as opposed to "not tall". Participants may feel uncertain because, as postulated by Bonini et al. (1999), they may think that as a matter of fact only one of those descriptions is correct, but both descriptions compete on their mind. Intuitively, "neither"-descriptions can be taken to adequately express the participants' reluctance to ascribe either "tall" or "not tall". Instead of committing themselves to either the predicate or its negation, the participants express a preference for omitting both.

That explanation has a ring of truth, but it can't be quite adequate. One assumption that appears inadequate is that if participants felt that only one of the predicates "tall" and "not tall" ought to apply, then they should massively check "No" for descriptions like "x is tall and not tall". Unlike Alxatib and Pelletier (2011a), we did not leave participants with the option to say "Can't tell" in response to the sentences. But as our data indicate, however, about half the participants checked "Yes" to "and"-descriptions at least once (based on the "And: ✓" bars in Fig. 5). This confirms that an explanation in the style of Bonini et al. cannot be right: if we follow that explanation we can no longer explain why participants accept glutty descriptions to the extent that they do.

In light of our data, a different way of articulating the omission bias hypothesis may be as follows: participants do not feel that there is a fact of the matter dictating that only one description in terms of "tall" or "not tall" should be the correct one, but they may feel that once you commit yourself to one description, you should avoid using the other. In other words, participants may simply have a bias toward consistency. Although they feel that "tall" and "not tall" are equally applicable of a borderline case, they find more adequate to say that neither description applies, to avoid an inconsistency, rather than to say that both descriptions apply. In other words, the participants' behavior would simply reflect a preference for incompleteness over inconsistency.

That explanation sounds more convincing, but it still strikes us as ad hoc. Firstly, many participants accepted both kinds of contradictory descriptions on the same trial: in these cases, their bias toward consistency, which under this view is the reason why they accept the "neither"-description, would have to immediately fade in front of the "and"-description in order to allow for the acceptance of the latter. Secondly, it assumes that participants would interpret "tall" in a fixed way across both occurrences of "tall" and "not tall". What if participants had different interpretations in mind depending on the occurrence of "tall"? They may very well understand a sentence like "Betty is tall and not tall" to mean: "Betty is tall_1 and Betty is not tall_2", where "tall_1" picks a lower threshold for "tall" than "tall_2". If really participants think that no single way of drawing a line between "tall" and "not tall" people is correct, then this would be a very rational interpretation for them to use. Note that if this were the case, then "Betty is neither tall nor not tall" ought to mean "Betty is neither tall_2 nor not tall_1", which is logically equivalent to "Betty is tall_1 and Betty is not tall_2". Contextual variability easily avoids the inconsistency attached to either kind of description, but the preference for one kind of description over the other remains to be explained. If anything, following Grice's Maxim of Manner, one would expect to see a preference for the conjunctive description, since it is briefer and it appears morphologically simpler. In any event, appeal to a bias toward incompleteness over inconsistency loses its ground here.

In summary, we think that an explanation in terms of an omission bias can't be adequate, because it should predict that "and"-descriptions are not accepted at all. And an explanation in terms of a preference for incompleteness over inconsistency seems to us ad hoc and limited by assuming that participants would interpret "tall" rigidly, in a way that does not seem supported by the relative acceptance of sentences of the form "Betty is tall and not tall". Finally, a contextualist treatment can explain the consistency of borderline contradictions, but it does not straightforwardly explain the preference for "neither"- over "and"-descriptions.

4.2 *Strict Meaning Versus Tolerant Meaning*

In order to explain the findings of our study, we therefore turn to a distinct set of assumptions, and basically adopt the working assumption of Alxatib and Pelletier (2011a) and Cobreros et al. (2012) according to which vague adjectives are pragmatically ambiguous between two interpretations, a *tolerant* and a *strict* interpretation. This is similar to the idea of contextual variability, but putting more systematic constraints on the relation between two kinds of meaning.[10]

[10] Alxatib and Pelletier talk of sub- and super-interpretation. As pointed out by Cobreros et al. (2012), we can talk of sub- and super-interpretations, but provided we do not mistake the resulting logic of vague predicates for the subvaluationist and supervaluationist logics respectively, which are not truth-functional, unlike the strict-tolerant logic used in Cobreros et al. (2012). See Ripley (2013) and Alxatib et al. (2013) for discussion and comparison. The approach, while akin to the

We shall not review all the extant evidence for the distinction between strict and tolerant interpretations, but we only point out one of the earlier findings by Alxatib and Pelletier (2011a), which is that a significant proportion of their participants who checked True to the description of a borderline tall man as "tall and not tall" also checked False to the separate descriptions "tall" and "not tall".[11] The way that particular finding is explained by Alxatib and Pelletier (2011a) as well as Cobreros et al. (2012) is by appeal to the Strongest Meaning Hypothesis (SMH), namely the hypothesis that among two ambiguous meanings, the default should be to choose the strongest non-trivial meaning. In the case in question, the strongest non-trivial meaning that can be given to a sentence like "Betty is tall and not tall" is the conjunction of the tolerant meanings of "Betty is tall" and "Betty is not tall" respectively, whereas the strongest non-trivial meaning that can be given to the conjuncts separately is the strict meaning of "Betty is tall" and "Betty is not tall". For a speaker obeying the SMH, it is therefore consistent to accept "Betty is tall and not tall", while separately rejecting "Betty is tall" and "Betty is not tall", because the interpretation of "tall" and "not tall" switches from tolerant to strict from the conjunction to the conjuncts.

What the SMH encapsulates is a general bias toward the most informative meaning. We think the SMH can also be used to account for the preference of "neither"-descriptions over "and"-descriptions in our experiment. Our basic idea is simple: the idea is that when asked to decide whether a sentence containing a positive adjective or its negation is true or false, participants should have a bias toward selecting the strict meaning of the adjective and of its negation first. Only secondarily will they consider the tolerant meaning. To get the details right, however, we need to spell out some assumptions.

4.3 Local Strengthening

We adopt the three-valued presentation of strict and tolerant meanings given in Cobreros et al. (2015a). Given a propositional or first-order language, and a three-valued model, we call a sentence strictly true if it takes the value 1, and tolerantly true if it takes a value at least $\frac{1}{2}$ in that model, and we adopt the strong Kleene rules for the connectives. Given a vague predicate P, we represent the fact that a is a borderline case of P by assigning the value $\frac{1}{2}$ to Pa in that model. We now state our specific assumptions in order to explain the preference of "neither"-descriptions over "and"-descriptions.

(i) *Local Operators*

Our first assumption is that predicates are *locally* interpreted as tolerant or as strict at the subsentential rather than the sentential level. We basically posit strict and tolerant

contextualist strategy outlined at the end of the previous section, involves no mechanism of indexing. See Ripley (2011b) for more on this and the varieties of contextualism.

[11] See also Égré et al. (2013) for discussion, where a related phenomenon is discussed under the name "Hump Effect". The term "conjunction effect" now strikes us as more general and adequate.

Table 3 Truth-tables for the S and T operators

ϕ	$T\phi$	$S\phi$
0	0	0
$\frac{1}{2}$	1	0
1	1	1

operators S and T whose semantics is defined as in Table 3. The S and T operators correspond to Łukasiewicz's necessity and possibility operators in three-valued logic (see Malinowski 2007 for an overview). The S operator also corresponds to Bochvar's meta-assertion operator, sometimes written A, t or B (Bochvar 1937; Horn 1985; Beaver 2001; Spector 2012), which plays a role in the theory of presupposition projection (more on this in Sect. 4.4).

Local modulation of strength appears to us as a natural idea, given that strict and tolerant meanings seem to have adverbial reflects.[12]

(13) Betty is (somehow) tall and (somehow) not tall

(14) Betty is neither (clearly) tall nor (clearly) not tall

We do not say that *somehow* and *clearly* literally correspond to the strict and tolerant operators that we just introduced. Rather, what (13) and (14) show is that local modifications of the adjectives are a productive linguistic operation. In the rest of the discussion, we will assume the existence of two covert linguistic operators *strictly* and *tolerantly* that can appear in place of *somehow* and *clearly* and that respectively have the effects of T and S.

The algorithm must be such as to output the following interpretations for the above sentences[13]:

(15) $T(tall(a)) \wedge T(\neg tall(a))$

(16) $\neg(S(tall(a)) \vee S(\neg tall(a)) \equiv \neg S(tall(a)) \wedge \neg S(\neg tall(a))$

The assumption of local strengthening is a significant departure from most recent accounts of borderline contradictions, where the pragmatic meaning of a sentence is computed globally for the whole sentence (see Cobreros et al. 2012; Alxatib et al. 2013; Cobreros et al. 2015b). It is however suggested by (Kamp and Partee 1995, 156) in their discussion of the assertibility of literal contradictions and literal tautologies involving multi-dimensional predicates. They consider that in a sentence

[12] See for example Serchuk et al. (2011), who call "confusion hypothesis" the hypothesis of a systematic strengthening of vague adjectives by a covert "definitely" operator, following the terminology of Williams (2006) (based on a expression first due to P. Greenough).

[13] Since the algorithm we propose operates on high-level linguistic representations, it equally derives the expected interpretation regardless of the form of the logical translation of the *neither* descriptions.

such as "Bob is a man and not a man", each occurrence is interpreted differently, "as if it were modified by something like "in some respects"". We discuss the locality assumption as well as the quantification over respects in greater detail below.

(ii) *Predicate Negation*

To flesh out assumption (i), we supplement it with another one, which concerns the behavior of negation. We take it that the operators do not get embedded inside a negated predicate (as in *not tall*) except if specifically marked (as in *not* $\underline{tall}_{Focus}$ where the short break between *not* and *tall* might signal the presence of the *strictly* operator).

(iii) *Bottom-up Strengthening and Backtracking*

Our third assumption is that strongest meanings are computed incrementally in the course of building the syntactic representation of a sentence. The idea is that first, the leaves of a syntactic tree are given the strongest meanings. Then, given two subconstituents of a larger constituent, their meaning composition gets a check for nontriviality. We call a meaning trivial if it is necessarily empty or necessarily tautological. If it is nontrivial, the algorithm proceeds according to classical rules in order to deliver a semantic verdict relative to the model at hand. If a triviality is reached, then one needs to backtrack and reassign the leaves of the tree the next strongest meaning available in order to reiterate the algorithm, until the algorithm ends and gives a semantic verdict.

(iv) *Least Effort*

We note that Kamp and Partee envisage the acceptability of a sentence like "Bob is a man and not a man" as involving backtracking in a similar way: for them, the access to different respects is based on perceiving a contradiction from a uniform meaning, and on the need to abide by Gricean maxims. In order to explain the preference for "neither"-descriptions of borderline cases over "and"-descriptions, however, we need an additional assumption, which is that the simpler of two computational procedures should be generally preferred to the more complex. Or to put it differently, if a run of the algorithm involves backtracking, it will involve a more costly representation of meaning, and participants will be less likely to compute it. That is, if a triviality is reached in the course of building the meaning of a sentence, a participant can always be lazy and deliver a verdict according to the interpretation reached, instead of repairing to get a nontrivial meaning. Fundamentally, this means that backtracking is optional, an assumption which remains compatible with Gricean principles.

Illustration

Let us see how our algorithm works on the non-adverbial versions of sentences (13) and (14). The sentence (13) should first be interpreted as *Betty is strictly tall and strictly not tall* ($S(tall(a)) \wedge S(\neg tall(a))$) in virtue of (i), (ii) and (iii). That sentence is a contradiction, hence a trivial sentence. The next nontrivial interpretation we can obtain after backtracking is *Betty is tolerantly tall and tolerantly not tall* ($T(tall(a)) \wedge T(\neg tall(a))$) in virtue of (iii). Note that according to (iv), given the model we assumed, participants will either be lazy and judge $S(tall(a)) \wedge S(\neg tall(a))$ to have value 0 relative to their model of the situation, or they will have worked out the meaning of the sentence to be $T(tall(a)) \wedge T(\neg tall(a))$, and they will give it the value 1. The case of sentence (14) on the other hand involves no backtracking at all. The meaning we get for the sentence is *Betty is neither strictly tall nor strictly not tall* ($\neg S(tall(a)) \wedge \neg S(\neg tall(a))$), which is nontrivial and gets the value 1 in the model. Importantly, this implementation naturally accounts for the quasi-absence of acceptance of the "and"-descriptions to the exclusion of the "neither"-descriptions (second bar from the left in Fig. 5). Participants who are willing to rescue an "and"-description from contradiction must do more effort than they have to do for a "neither"-description, while representing the strict meaning in both cases. We predict as unlikely, therefore, for participants to accept the former kind of description while rejecting the latter.[14]

4.4 Comparisons

4.4.1 Global Strengthening

In order to establish whether our algorithm is plausible, we first need to compare it with extant algorithms, and then to see whether it makes further adequate predictions. To the best of our knowledge the closest kin to our algorithm is sketched in remarks made by Alxatib and Pelletier (2011a) about the computation of sub- and super-interpretation. Although Alxatib and Pelletier do not outline a general algorithm, they make some suggestive remarks, for instance concerning the meaning of double negations of gradable adjectives. For example, a sentence like:

(17) Betty is not not tall.

Can be used to convey that Betty is not short, but also that she is not "tall *tall*", in other words that she is borderline tall. Our algorithm can derive that meaning, since we basically get *Betty is not strictly not tall* ($\neg S(\neg tall(a))$) as the strengthened meaning.[15] This means exactly that Betty is tolerantly tall. As already mentioned, a globalist algorithm like that in Cobreros et al. (2012) cannot derive that interpretation, since $\neg\neg tall(a)$ is necessarily equivalent to $tall(a)$ both under its strict and under

[14] 10 participants did accept the "and"-description while rejecting the "neither"-description at least once (5 participants did so on one trial, 5 participants did so on two trials). This could be noise produced by inattention on these trials, even though our participants were highly accurate. It is also conceivable, in principle, that on one or two trials those 10 participants exceptionally went through the whole procedure described above for the "and"-description but stuck with a classical, logically contradictory interpretation for the "neither"-description.

[15] Note that in (17), *strictly* would again appear under the first *not* and typically trigger a short break in the prosodic contour. Relatedly, when we paraphrased the meaning of (17), we used "not tall *tall*" to mean "not strictly tall". The repetition of the adjective seems to be another focus-related strategy to embed the *strictly* operator under negation (see (ii) above).

its tolerant interpretation. The same holds of Cobreros et al. (2015b)'s modified algorithm, which introduces no pragmatic difference between a sentence and its double negation. Our assumption that predicate negation and sentential negation should be treated differently plays a crucial role here.

Another problem that has been raised by Alxatib et al. (2013) for the original globalist account in Cobreros et al. (2012) concerns sentences such as:

(18) Betty is tall and Betty is not tall, or Betty is rich.

When interpreted strictly, the sentence means merely that Betty is strictly rich. But intuitively, it in fact says that Betty is borderline tall or clearly rich, which is stronger. Cobreros et al. (2015b) propose a more elaborate algorithm capable of deriving that meaning. Our algorithm also derives it straightforwardly: the incremental processing of (18) will treat the constituent subsentence "Betty is tall and Betty is not tall", but as illustrated earlier embedding the strict operator would yield a trivial constituent, so backtracking reinterprets to mean that Betty is tolerantly tall and tolerantly not tall, and "Betty is rich" will be interpreted strictly, and their union will be nontrivial.

We note finally that none of the other algorithms of pragmatic meaning we cited would predict a preference for "x is neither P nor not P" over "x is P and not P". Alxatib et al. (2013)'s algorithm based on fuzzy logic, for example, would predict (13) and (14) to be equally acceptable, like Cobreros et al. (2012) or Cobreros et al. (2015b). The difference lies mostly in the fact that pragmatic strengthening in those cases is done globally, and not by the insertion of local operators.

4.4.2 Local Accommodation for Presupposition

As pointed out to us by Benjamin Spector, the use of the local operators S and T bears a strong analogy with the use of accommodation operators in the theory of presupposition projection. Thus, Spector (2012) shows how local accommodation in a three-valued treatment of presupposition can be handled by means of the S operator. The operator returns the value 1 when a sentence is true (gets the value 1) and the value 0 when the sentence is false or undefined (gets the value 0 or #). Spector's observation is that one can find pairs of sentences that have exactly the structure of conjunctive versus disjunctive borderline contradictions, except that they involve presuppositional instead of (merely) vague expressions. His example is the following (Spector 2012, 2016):

(19) a. John stopped smoking and John did not stop smoking.
b. Neither John stopped smoking, nor did he not stop smoking.

Spector notes that (19-b) is acceptable in a situation in which John never smoked, in a way that (19-a) isn't.[16] Indeed, this appears to be what (19-b) means, namely

[16] (19-a) may be acceptable if John is a borderline case of someone who stopped smoking. But set aside the vagueness of "stop" and "smoke" (assume they are fully crisp predicates).

that neither description is applicable to John, because he never smoked. The proper way to account for the acceptability of (19-b) is in terms of local accommodation, for which Spector's analysis is as follows:

(20) not (S(John stopped smoking) or S (John didn't stop smoking)).

When John never smoked, "John stopped smoking" is undefined, and by the semantics of S the whole sentence gets the value 1. Contrariwise, (19-a) stays false or undefined, however we apply the accommodation operator to its second conjunct, and irrespective of the value of the first conjunct. In other words, the sentence is subject to no pragmatic repair, unlike (19-b).

The analogy is indeed striking, but there remains an important difference between vagueness and presupposition, which is that conjunctive descriptions are acceptable with vague predicates in borderline cases to at least some extent, whereas they seem always unacceptable in cases of presupposition failure (see Spector 2012; Zehr 2014 for more on the comparison). In Spector's account of the presuppositional case, what matters is that (19-a) can never get the value 1, whereas (19-b) can. Interestingly, Spector (2012) mentions the potential usefulness of a dual operator such as T to handle vague sentences.[17] This is exactly what happens under our assumptions, since "Betty is tall and not tall" can get the value 1 as per the use of T. Regarding our main finding, however, the key part in our explanation lies in the supposition that the pragmatic process needed to give the sentence that value is more costly than it is for "Betty is neither tall nor not tall". So we acknowledge the analogy between Spector's pair and pairs consisting of conjunctive versus disjunctive borderline contradictions, but we think more is needed to account for the relative rather than absolute preference of one kind of sentence over the other in the vagueness case.

4.4.3 Quantification Over Standards

We deliberately set aside multidimensional adjectives from our sample of adjectives since "Betty is healthy and not healthy" is easily interpreted to mean that Betty is healthy in some respect (for instance blood pressure), and not healthy in some others (for instance cholesterol) (see Kamp and Partee 1995). Prima facie, a shift in respects of comparison is not what is operative with one-dimensional adjectives like "tall". However, the tolerant interpretation of a sentence like "Betty is tall and not tall" does convey that Betty is tall by some (acceptable) standard, and not tall by some (distinct, but equally acceptable) standard, relative to the same respect of comparison.

[17] Spector writes the operator in question W, for "weak truth", and writes the dual B. Note that we developed our account independently of Spector's, that is from the vantage of the strict-tolerant account of vagueness, without thinking about presupposition accommodation. Spector was motivated primarily by the phenomenon of local accommodation, and with no heed to the specific asymmetry between "neither" and "and"-sentences we discuss in the vagueness case.

And conversely, “Betty is neither tall nor not tall” appears to convey that not every standard makes Betty tall, and that not every standard makes Betty not tall.[18]

Galit Sassoon points out to us that Shamir (2013) found that one-dimensional adjectives “are not as good as multidimensional adjectives with modifiers over respects (such as e.g., “in some/most/every respect” or “except in some respects”), but they are not crushingly bad either, and their interpretation seems to build on quantification over standards of membership” (p.c.). That is, “tall in some respect” can be used to mean “tall by some standard”, and “tall in every respect” “tall by every standard”. Moreover, Sassoon (2013) found that in the case of multidimensional adjectives, “positive” adjectives (like “safe”) are predominantly universally modified, but “negative” adjectives (like “dangerous”) are predominantly existentially modified, even though either type admits both interpretations. Recently, Sassoon and Fadlon (2016) noticed that the “positive” one-dimensional adjectives that they used as fillers in their experiment show a similar preference for strong modifiers like “all” over weak modifiers like “some”. For Sassoon, this fact coheres with our finding and account of the data.

We note that if “tall” predominantly means “by every standard, tall”, then “not tall” could end up meaning “by every standard, not tall” or “tall, but not by every standard”, depending on whether negation should take narrow scope or wide scope over “every”. Either way, “tall and not tall” would be predicted to be anomalous under the universal interpretation of its first occurrence. On the other hand if “tall” can be existentially modified to mean “by some standard, tall”, as we get by the insertion of the T operator, then “tall and not tall” becomes admissible again (provided the existential modifier takes wide scope over negation, as we postulate for T). Sassoon’s observation raises a further question, however, which is whether multidimensional negative adjectives like “dangerous”, whose predominant reading is existential over respects, might invert the pattern we observed in one-dimensional adjectives, that is, whether we might see an interaction between the predominant reading of the adjective (existential vs. universal over respects), and the relative acceptability of “and”-contradictions over “neither”-contradictions. For instance, would “this is dangerous and not dangerous” be relatively more accepted than “this is neither dangerous nor not dangerous” compared to “this is safe and not safe” relative to “this is neither safe nor not safe”? This is an interesting question to ask, which we leave for further investigation.

[18] And indeed, while the account of strict and tolerant we adopted here is cashed out in trivalent terms, the original account of Cobreros et al. (2012) defines the tolerant meaning of an adjective in terms of an existential quantification and the strict meaning in terms of a universal quantification, not directly over acceptable standards, but in ways that are intertranslatable with that approach.

5 Conclusions and Perspectives

We reported on two main findings in this paper. The first is a new confirmation of the fact that classical contradictions both of the "neither" type and of the "and" type are accepted and used by naive speakers to describe borderline cases. This confirms that both gaps and gluts are operational in the representation of borderline cases (see Égré et al. 2013 for a related point). The second and more interesting finding is that for a representative class of relative gradable adjectives, "gappy" descriptions of the form "neither *P* nor not *P*" are preferred to "glutty" descriptions of the form "*P* and not *P*". The sense in which "neither"-descriptions are "gappy" is that they express that a particular case falls in the underlap between two strict extensions, and the sense in which "and"-descriptions are "glutty" is that they express that the same case falls in the overlap between two tolerant extensions. In agreement with the earlier accounts of the pragmatic meaning of vague predicates presented in Alxatib and Pelletier (2011a), Cobreros et al. (2012), Égré et al. (2013), Cobreros et al. (2015b), we have argued that we can account for that preference if indeed there is a bias toward selecting the strict meaning first, but moreover, and more centrally regarding the implications of our the theory, if pragmatic strengthening is done locally rather than globally.

Several questions remain. One is whether we can obtain independent confirmation of our explanation of the preference for "neither"-descriptions over "and"-descriptions in terms of the latter involving more steps of computation. Our account has implications regarding the online processing of these sentences. An interesting avenue for future research would be to collect response times and eye-tracking data: we expect slower response times, and possibly longer fixation and more regression eye-movements when accepting the "and"-descriptions than when accepting the "neither"-descriptions, under the assumption that the backtracking steps that we posit to access a tolerant reading have the same effects as found in the interpretation of garden-path sentences (Frazier and Rayner 1982).[19]

Another issue concerns the generality of our finding for other adjectival types. Does our algorithm predict the same preference for "neither"-descriptions over "and"-descriptions for absolute adjectives like "empty"? Relatedly, what are the *facts* for absolute adjectives? To answer the first question, we need a representation of absolute adjectives in our model. Burnett (2014, 2016) argues that the strict denotation of "empty" coincides with the classical denotation of "empty" (that is, it should denote the zero degree of being filled on the relevant scale), though the tolerant denotation of "empty" can include recipients with a tiny bit of stuff in them. Symmetrically, Burnett takes the tolerant extension of "not empty" to be identical to the classical extension, but she assumes the strict extension to be a proper subset of the classical extension (hence "not empty", read strictly, does not mean "strictly speaking, not empty", but something like: "clearly not empty", as Burnett stresses). The strict/tolerant duality that we observe for relative adjectives is therefore preserved for absolute adjectives: the strict extension of "not empty" is the complement

[19] Thanks to G. Sassoon for making that suggestions.

of the tolerant extension of "empty" and conversely. As a result, and as for relative adjectives, the gap defined by the strict extensions of "empty" and "not empty" corresponds to the glut defined by the tolerant extensions of "empty" and "not empty". Using our algorithm, Burnett's account would therefore predict the same preference for a description of the form "neither empty nor not empty" over a description of the form "empty and not empty" when describing a glass with a tiny bit of liquid in it. Although both are pragmatically non trivial and true, the former should be preferred, by the principle of using strict meaning first.

Regarding the second question, presently we cannot rule out the possibility for "and"-descriptions to be preferred to "neither"-descriptions for absolute adjectives, contrary to the predictions just discussed. If that were the case, one ought to question the interpretation of the second member of the "neither"-descriptions. It seems to us that a description of the form "neither empty nor not empty" has a natural interpretation paraphrasable as "neither empty nor, *strictly speaking*, not empty". Using our *strictly* operator, this reading amounts to "neither *strictly* empty nor not *strictly* empty", which is a manifest contradiction. However, the present remarks are based entirely on conjectures, and we need to confront them to actual data in order to make progress. We are in the process of running a separate study on absolute adjectives in order to gain further insights about such examples. Meanwhile, we think our results on relative adjectives already give us a compelling argument in favor of some form of local pragmatic strengthening in the computation of the meaning of sentences involving vague vocabulary.

Acknowledgements We thank the Agence Nationale de la Recherche under program TrilLogMean ANR-14-CE30-0010-01 for support, as well as grants ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL* for research carried out at the Departement of Cognitive Studies of ENS. PE thanks the Swedish Collegium for Advanced Study who hosted him as a visiting fellow during the time this paper was written (Feb-June 2016). JZ thanks the National Science Foundation grant BCS-1349009 attributed to Florian Schwarz, and Florian Schwarz himself, for their support. We are particularly grateful to Galit Sassoon, to Benjamin Spector, and to three anonymous referees for detailed comments and suggestions. We also thank Sam Alxatib, Elitzur Bar-Asher Siegal, Pablo Cobreros, Chris Cummins, Matti Eklund, Vladimir Goranko, Yael Greenberg, Yossi Grodzinsky, Nat Hansen, Hans Kamp, Louise McNally, Marie-Christine Meyer, Peter Pagin, Roumy Pancheva, David Ripley, Robert van Rooij, Hans Rott, Uli Sauerland, Stephanie Solt, Steven Verheyen, Åsa Wikforss for various discussions. We also thank the participants of the "Vagueness and Three-valued Logics" course given by PE at the LLCC at the University of Jerusalem in 2015, as well as audiences in Madrid, Stockholm, Uppsala, and at the ESSLLI 2016 summer school in Bolzano (Course on "Trivalent Logics and Natural Language Meaning"). We warmly thank the editors, Elena Castroviejo Miró, Louise McNally, and Galit Sassoon for their editorial work and feedback.

Appendix 1: Scenarios

Scenarios for Human-Oriented Adjectives

Rich

A survey on wealth has been conducted in your country. In the population there are people with a very high degree of wealth, and people with a very low degree of wealth. Then there are people who lie in the middle between these two areas.

Imagine that Sam is one of the people in the middle range. Comparing Sam to other people in the population, is it true to say the following?

Tall

A survey on heights has been conducted in your country. In the population there are people of a very high height, and people of a very low height. Then there are people who lie in the middle between these two areas.

Imagine that Sam is one of the people in the middle range. Comparing Sam to other people in the population, is it true to say the following?

Old

A survey on age has been conducted in your country. In the population there are people whose age is very high, and people whose age is very low. Then there are people who lie in the middle between these two areas.

Imagine that Sam is one of the people in the middle range. Comparing Sam to other people in the population, is it true to say the following?

Heavy

A survey on weight has been conducted in your country. In the population, there are people of a very high weight, and people of a very low weight. Then there are people who lie in the middle between these two areas.

Imagine that Sam is one of the people in the middle range. Comparing Sam to other people in the population, is it true to say the following?

Scenarios for Object-Oriented Adjectives

Fast

A survey on people's cars has been conducted in your country. In the population, there are people who own very high speed cars, and people who own very low speed cars. Then there are people who own cars that lie in the middle between these two areas.

Imagine that Sam is one of the people owning a car in the middle range. Comparing Sam's car to the cars of other people in the population, is it true to say the following?

Large

A survey on people's houses has been conducted in your country. In the population, there are people who own houses with a lot of space, and people who own houses with very little space. Then there are people who own houses that lie in the middle between these two areas.

Imagine that Sam is one of the people owning a house in the middle range. Comparing Sam's house to the houses of other people in the population, is it true to say the following?

Loud

A survey on people's voice has been conducted in your country. In the population, there are people whose voice has a very high intensity, and people whose voice has a very low intensity. Then there are people whose voice lie in the middle between these two areas.

Imagine that Sam is one of the people whose voice lie in the middle range. Comparing Sam's voice to the voices of other people in the population, is it true to say the following?

Wide

A survey on people's feet has been conducted in your country. In the population, there are people with a very high foot breadth, and people with a very low foot breadth. Then there are people whose foot breadth lie in the middle between these two areas.

Imagine that Sam is one of the people with a foot breadth in the middle range. Comparing Sam's feet to the feet of other people in the population, is it true to say the following?

Appendix 2: Questionnaires

Pre-questionnaire

Before proceeding to the actual experiment, please answer these simple questions.

There are 7 days in a week	❍ Yes ❍ No
Barack Obama is the current President of the USA	❍ Yes ❍ No
Abraham Lincoln was born in 2003	❍ Yes ❍ No
Marilyn Monroe died in 1780	❍ Yes ❍ No
Nicolas Sarkozy was one of the Presidents of the United States	❍ Yes ❍ No
California is part of the USA	❍ Yes ❍ No

Post-questionnaire

Please answer these few questions about the expriment. We are interested in what you actually remember at this point, so please do not reload the previous pages. Thank you.

Was there a scenario describing a population of pregnant women? ❍ Yes ❍ No

Were the descriptions that you saw in the first half of the experiment of a different form from those in the second half? ❍ Yes ❍ No

[Can you give an example of the type of descriptions in the first half?

I

Can you give an example of the type of descriptions in the second half?

I]

The last two questions and their input fields would appear only if the participants reported a difference between the descriptions in the first and in the second halves of the experiment.

Appendix 3: Results Per Group and Regression Models

We ran several models, differing with respect to the complexity of their random effect structures. They all included a random intercept, but no random slope for Participant (N = 148). As explained in the Design subsection of Sect. 3, each participant was assigned to one of four groups, determined by two factors. Block Order indicates whether the participant responded *before* or *after* judging another set of descriptions with antonyms, and Adjective Type indicates whether the participant responded to descriptions directly referring to a human being as their subject or to an object or a property of a human being.

Table 4 Outputs for models with the formula $Response = yes \sim Description + (1|Participant) + (1 + Description|Adjective) + (1 + Description|Block)$

Neither	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.5971	0.3419	10.521	<2e-16 ***
True	4.2426	1.6634	2.551	0.0108 *
And	−4.3489	0.4941	−8.801	<2e-16 ***
False	−10.5246	0.7412	−14.200	<2e-16 ***
True				
(Intercept)	7.839	1.334	5.875	4.22e-09 ***
Neither	−4.242	1.377	−3.082	0.00206 **
And	−8.591	1.688	−5.090	3.58e-07 ***
False	−14.767	1.429	−10.330	< 2e-16 ***
False				
(Intercept)	−6.9275	0.6254	−11.077	< 2e-16 ***
And	6.1759	0.8515	7.253	4.08e-13 ***
Neither	10.5247	0.7349	14.321	< 2e-16 ***
True	14.7673	1.6375	9.018	< 2e-16 ***

We included a random intercept for Adjective (N = 8) in all our models. We were also able to fit models including a random slope for Adjective. One set of such models added a random intercept plus a random slope for Block Order, and another set of such models added a random intercept plus a random slope for Adjective Type. For all these models, we ran three versions: one with *Neither* as baseline for Description, one with *True* as a baseline and one with *False* as a baseline. The outputs of the models for the former structure (with random intercepts and slopes for Block Order) are presented in Table 4, the outputs of the models for the latter structure (with random intercepts and slopes for Adjective Type) are presented in Table 5. By removing the random slope for Adjective, we were also able to fit models including a random intercept and slope both for Block Order and Adjective Type with *Neither* as a baseline. The contrast with *Control True* remains significant (Table 6).

To further investigate the source of the significant contrasts between *Neither* and *Control True*, and with regard to the apparently mixed descriptive results in Fig. 6 above, we ran additional models on subsetted data. We first considered four subsets: the responses of all the participants in the human-oriented adjective groups, those of all the participants in the non-human-oriented adjective groups, those of all the participants who responded in the first block and those of all the participants who responded in the second block. All these models had a random intercept for Participant, and a random intercept plus slope for Adjective, Block Order and Adjective Type. None of them yielded a significant difference between *Neither* and *True* ($0.2 < p < 0.25$). We then subsetted the data to each minimal group of participants.

Table 5 Outputs for models with the formula $Response = yes \sim Description + (1|Participant) + (1 + Description|Adjective) + (1 + Description|AdjectiveType)$

Neither	Estimate	Std. error	z value	Pr(>|z|)
(Intercept)	3.4863	0.3649	9.556	< 2e-16***
True	3.5304	1.1144	3.168	0.00154 **
And	−4.1911	0.4507	−9.298	< 2e-16 ***
False	−10.1704	0.6343	−16.035	< 2e-16 ***
True				
(Intercept)	7.0164	0.9039	7.763	8.33e-15 ***
Neither	−3.5301	1.0223	−3.453	0.000554 ***
And	−7.7211	1.3308	−5.802	6.57e-09 ***
False	−13.7003	1.2777	−10.723	< 2e-16 ***
False				
(Intercept)	−6.6840	0.5847	−11.43	<2e-16 ***
And	5.9791	0.5519	10.83	<2e-16 ***
Neither	10.1703	0.5871	17.32	<2e-16 ***
True	13.7010	1.3338	10.27	<2e-16 ***

Table 6 Output for one model with the formula $Response = yes \sim Description + (1|Participant) + (1 + |Adjective) + (1 + Description|Block) + (1 + Description|AdjectiveType)$

Neither	Estimate	Std. error	z value	Pr(>|z|)
(Intercept)	3.6215	0.4070	8.898	<2e-16 ***
True	4.2418	1.8012	2.355	0.0185 *
And	−4.3742	0.5169	−8.463	<2e-16 ***
False	−10.5672	0.7154	−14.771	<2e-16 ***

The models failed to converge for the responses from the first block for the non-human oriented descriptions, and for the responses from the second block for the human-oriented descriptions. The other two models indicated a significant contrast between *Neither* and *Control True* ($p < 0.01$).

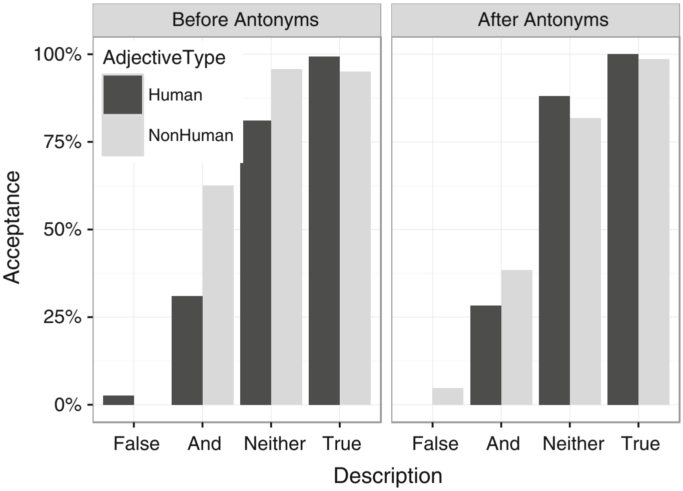


Fig. 6 Mean acceptance by description type and by group

References

Alxatib, S., & Pelletier, F. (2011a). The psychology of vagueness: Borderline cases and contradictions. *Mind & Language*, *26*(3), 287–326.

Alxatib, S., & Pelletier, J. (2011b). On the psychology of truth-gaps. In R. Nouwen, R. van Rooij, U. Sauerland, & H.-C. Schmitz (Eds.), *Vagueness in communication*. Lecture Notes in Computer Science (Vol. 6517, pp. 151–168). Berlin: Springer.

Alxatib, S., Pagin, P., & Sauerland, U. (2013). Acceptable contradictions: Pragmatics or semantics? A reply to Cobreros, et al. *Journal of Philosophical Logic*, *42*(4), 619–634.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Beaver, D. I. (2001). *Presupposition and assertion in dynamic semantics*. Stanford, CA: CSLI publications.

Bochvar, D. A. (1937). On a three-valued calculus and its applications to the paradoxes of the classical extended functional calculus. *Mathematicheskii sbornik*, *4*(46), 287–308. English translation published in 1981 by M. Bergmann, *History and Philosophy of Logic*, *2*(1981), 87–112.

Bonini, N., Osherson, D., Viale, R., & Williamson, T. (1999). On the psychology of vague predicates. *Mind & Language*, *14*(4), 377–393.

Burnett, H. (2014). Penumbral connections in comparative constructions. *Journal of Applied Non-Classical Logics*, *24*(1–2), 35–60.

Burnett, H. (2016). *Gradability in natural language: Logical and grammatical foundations*. Oxford: Oxford University Press.

Carlson, G. N. (1977). *Reference to kinds in English*. Dissertation, UMass/Amherst.

Cobreros, P., Égré, P., Ripley, D., & van Rooij, R. (2012). Tolerant, classical, strict. *The Journal of Philosophical Logic, 41*(2), 347–385.

Cobreros, P., Égré, P., Ripley, D., & van Rooij, R. (2015a). Vagueness, truth and permissive consequence. In D. Achouriotti, H. Galinon, & J. Martinez (Eds.), *Unifying the philosophy of truth* (pp. 409–430). Springer.

Cobreros, P., Égré, P., Ripley, D., & van Rooij, R. (2015b). Pragmatic interpretations of vague expressions: Strongest meaning and nonmonotonic consequence. *Journal of Philosophical Logic*, *44*(4), 375–393.

Égré, P., De Gardelle, V., & Ripley, D. (2013). Vagueness and order effects in color categorization. *Journal of Logic, Language and Information*, *22*(4), 391–420.

Fine, K. (1975). Vagueness, truth, and logic. *Synthese*, *30*(3–4), 265–300.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210.

Horn, L. R. (1985). Metalinguistic negation and pragmatic ambiguity. *Language*, *61*(1), 121–174.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*(2), 129–191.

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

Kennedy, C. (2013). Two sources of subjectivity: Qualitative assessment and dimensional uncertainty. *Inquiry*, *56*(2–3), 258–277.

Malinowski, G. (2007). Many-valued logic and its philosophy. In *Handbook of the history of logic* (Vol. 8, pp. 13–94).

McNally, L. (2011). The relative role of property type and scale structure in explaining the behavior of gradable adjectives. In R. Nouwen, R. van Rooij, U. Sauerland, & H.-C. Schmitz (Eds.), *Vagueness in communication*. Lecture Notes in Computer Science (Vol. 6517, pp. 151–168). Berlin, Heidelberg: Springer.

Peirce, C. S. (1902). Vague. In J. M. Baldwin (Ed.), *Dictionary of philosophy and psychology* (p. 748). New York: Macmillan.

Rett, J. (2007). Antonymy and evaluativity. In *Semantics and Linguistic Theory* (Vol. 17, pp. 210–227).

Ripley, D. (2011a). Contradictions at the borders. In R. Nouwen, R. van Rooij, U. Sauerland, & H.-C. Schmitz (Eds.), *Vagueness in communication*. Lecture Notes in Computer Science (Vol. 6517, pp. 169–188). Berlin: Springer.

Ripley, D. (2011b). Inconstancy and inconsistency. In P. Cintula, C. G. Fermüller, L. Godo, & P. Hájek (Eds.), *Reasoning under vagueness: Logical, philosophical, and linguistic perspectives* (pp. 41–58). College Publications.

Ripley, D. (2013). Sorting out the Sorites. In K. Tanaka, F. Berto, & E. Mares (Eds.), *Paraconsistency: Logic and applications*, (pp. 329–348). Dordrecht: Springer.

Roche, L. (2012). *La négation des adjectifs*. Master's thesis, ENS/EHESS/Paris V.

Ruytenbeek, N. (2013). *An experimental approach of negated gradable adjectives*. Master's thesis, ENS/EHESS/Paris V.

Ruytenbeek, N., Verheyen, S., & Spector, B. (2017). Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives. *Glossa: A Journal of General Linguistics*, *2*(1), 92, 1–27.

Sæbø, K. J. (2009). Judgment ascriptions. *Linguistics and Philosophy*, *32*(4), 327–352.

Sassoon, G. W. (2013). A typology of multidimensional adjectives. *Journal of Semantics*, *30*(3), 335–380.

Sassoon, G. W., & Fadlon, J. (2016). *The role of dimensions in classification under predicates predicts their status in degree constructions*. Unpublished manuscript, Bar Ilan University.

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

Shamir, A. (2013). *Disjunctivity*. Master's thesis, The Hebrew University of Jerusalem.

Solt, S., & Gotzner, N. (2010). *Expensive, not expensive, or cheap?* Paper presented at the 11th Szklarska Poreba Workshop.

Spector, B. (2012). Vagueness, (local) accommodation, presupposition and restrictors. Course Notes on Trivalent Semantics for Vagueness and Presupposition, Vienna.

Spector, B. (2016). Presupposition (and vagueness) projection at the propositional level. ESSLLI 2016 Course notes on Trivalents Logics and Natural Language Meaning, Day 2.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, *27*(1), 76–105.

Unger, P. (1975). *Ignorance: A case for scepticism*. Oxford: Oxford University Press.

Williams, J. R. G. (2006). An argument for the many. *Proceedings of the Aristotelian Society*, *106*(1), 411–419.

Zehr, J. (2014). *Vagueness, presupposition and truth-value judgments*. Dissertation, ENS, Paris.

Multidimensionality, Subjectivity and Scales: Experimental Evidence

Stephanie Solt

Abstract This paper investigates the subjective interpretation of the comparative forms of certain gradable adjectives, exploring in particular the hypothesis put forward in several recent works that such 'ordering subjectivity' derives from the multidimensional nature of the adjectives in question. Results of an experimental study are presented which demonstrate that ordering subjectivity is more widespread than previously recognized, and that in this respect, gradable adjectives divide into not two but three groups: objective, subjective and mixed. Evidence is also offered that adjectival multidimensionality itself is a heterogenous phenomenon. On the basis of these observations as well as the experimental findings, it is argued that there are two separate sources of ordering subjectivity: multidimensionality and judge dependence. This proposal is formalized within a semantic framework in which gradable adjectives lexicalize families of measure functions indexed to contexts and in some cases judges.

Keywords Adjective · Comparative · Gradability · Measurement Multidimensionality · Subjectivity · Faultless disagreement · Predicate of personal taste · Context dependence

1 Introduction

It is well known that certain adjectival predicates are subjective or judge-dependent, in that two competent speakers can disagree as to whether the predicate applies, without either appearing to have said something incorrect or false (see Kölbel 2004; Lasersohn 2005, 2009; Stephenson 2007; Sæbø 2009; Moltmann 2010; and other work cited below). Such 'faultless disagreement' is observed most classically with so-called predicates of personal taste such as *tasty* and *fun*, but also with evaluative adjectives more generally (e.g. *beautiful*) and with the unmodified positive forms of vague gradable adjectives (e.g. *tall*):

S. Solt (✉)
Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS), Berlin, Germany
e-mail: solt@leibniz-zas.de

E. Castroviejo et al. (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure*,
Language, Cognition, and Mind 4, https://doi.org/10.1007/978-3-319-77791-7_3

(1) a. Speaker A: The chili is tasty! **faultless**
Speaker B: No, it's not tasty at all!

b. Speaker A: The Picasso is beautiful! **faultless**
Speaker B: No, it's ugly!

c. Speaker A: Anna is tall! **(potentially) faultless**
Speaker B: No, she's not!

Recently, attention has turned to a second sort of subjectivity, which characterizes the comparative forms of some but not all gradable adjectives (Kennedy 2013; Bylinina 2014, 2017; Umbach 2016; McNally and Stojanovic 2017). By way of example, two competent speakers might faultlessly disagree as to which of two dishes is tastier (2a), or which of two paintings is more beautiful (2b), but not about which of two individuals is taller (2c). In what follows, I will refer to the phenomenon exemplified in (2a-b) as **ordering subjectivity**.

(2) a. Speaker A: The chili is tastier than the soup! **faultless**
Speaker B: No, the soup is tastier!

b. Speaker A: The Picasso is more beautiful than the Miró. **faultless**
Speaker B: No, the Miró is more beautiful.

c. Speaker A: Anna is taller than Zoe. **factual only**
Speaker B: No, Zoe is the taller of the two!

For the leading semantic approach to gradability, namely the degree-based analysis of Cresswell (1977), Kennedy (1997), Heim (2000) and others, ordering subjectivity is problematic. In such a framework, gradable adjectives lexicalize measure functions that map individuals to degrees on scales: *tall* is based on a HEIGHT measure function, *beautiful* on a BEAUTY function, and so forth (3). Comparative constructions are then analyzed as expressing relations between the degrees assigned to two individuals (4).

(3) a. $[\![\text{tall}]\!] = \lambda d \lambda x . \mu_{HEIGHT}(x) \succeq d$

b. $[\![\text{beautiful}]\!] = \lambda d \lambda x . \mu_{BEAUTY}(x) \succeq d$

(4) The Picasso is more beautiful than the Miró.
$\mu_{BEAUTY}(Picasso) \succ \mu_{BEAUTY}(Miro)$

The mostly unspoken assumption underlying lexical entries of this form is that each dimension of measurement DIM is uniquely associated with a measure function μ_{DIM} whose output encodes the ordering of individuals relative to DIM. But examples such as (2a-b) suggest that this can't be right. Rather, it seems that measure functions must in some way be relativized to speakers, thereby allowing disagreement as to orderings.

The objective of this paper is to work towards an account of ordering subjectivity within a degree-based semantic framework. In particular, I will investigate a proposal put forth in several recent works that a—or the—source of ordering subjectivity is the **multidimensionality** of the predicates in question (Kennedy 2013; Bylinina 2014, 2017; Umbach 2016; McNally and Stojanovic 2017). Whereas the attribution of a predicate such as *tall* is based on a single underlying dimension, namely height, that of a predicate such as *beautiful* is based on multiple underlying component dimensions; for (1b) and (2b), for example, the dimensions of beauty might involve line, color, balance, and so forth. Subjectivity is proposed to arise because different individuals may weight these component dimensions differently, potentially resulting in a reversal of the relative ordering of two individuals. Exploring this line of explanation will prompt us to take a closer look at what it means for an adjective to be characterized as multidimensional.

Whichever approach one chooses to pursue, a crucial step in developing an adequate formal theory of ordering subjectivity (or subjectivity more generally) is to clarify which gradable adjectives are interpreted subjectively in their comparative forms. For dimensional adjectives such as *tall* and evaluative adjectives such as *beautiful* and *tasty*, the picture seems clear: in the former case, statements about orderings are objective, while in the latter, they are necessarily subjective. But this is far from exhausting the broad and varied spectrum of gradable adjectives. Of particular interest are adjectives such as *clean/dirty*, *smooth/rough* and *sharp/dull*. These differ from adjectives such as *tall* in that they lack commonly used measurement units. But they also different from those such as *beautiful* and *tasty* in that they appear to describe physical properties of objects in the world, rather than judgments based on internalized experiences. Can two individuals disagree faultlessly about which of two shirts is dirtier? which of two surfaces is rougher? which of two knives is sharper? As intuitions here are shaky, these questions were pursued experimentally, with the finding that ordering subjectivity is more widespread than has been previously recognized, and furthermore that in this respect, gradable adjectives pattern into not two but three subgroups: objective, subjective and mixed.

The primary proposal that is developed in this paper, which is based on the above two lines of investigation, is that there are two distinct sources of ordering subjectivity, namely multidimensionality and judge dependence. This proposal is formalized within a semantic framework in which gradable adjectives lexicalize not a single measure function but rather a set of such functions indexed to contexts and in some cases judges. Constraints on this set determine whether their comparative forms can be interpreted objectively, subjectively or in both ways. An ancillary conclusion that emerges is that adjectival multidimensionality is not a homogeneous phenomenon but rather has several distinct subtypes.

The structure of the paper is as follows: Sect. 2 presents the experiment and discusses some related phenomena. Section 3 briefly reviews existing semantic theories of subjectivity, with a view to assessing how well they are able to account for the experimental findings. Section 4 delves into the phenomenon of multidimensionality, offering evidence for its heterogenous nature. Section 5 presents the formal proposal, and Sect. 6 concludes.

2 Experiment: Faultless Disagreement Paradigm

The present study employs a novel faultless disagreement paradigm to diagnose the presence of ordering subjectivity among a wide range of gradable adjectives, with the goal of establishing a firmer empirical basis for formal semantic theories of the phenomenon.

2.1 Participants

Participants were 91 native speakers of English, recruited via the online participant marketplace Amazon Mechanical Turk (MTurk). Recruiting was limited to MTurk workers with U.S. IP addresses. Native language was confirmed via a question at the end of the survey; no participants were excluded on the basis of this question.

2.2 Materials

Stimuli were based on 35 gradable adjectives, which were divided into the following categories according to their status as dimensional versus evaluative, as well as the type of interpretation of the adjective in its positive form and the corresponding structure of the scale it lexicalizes[1]:

- Dimensional gradable adjectives, more specifically relative gradable adjectives with numerical measures (**RelNum**): *tall, short, old, new, expensive*
- Relative gradable adjectives without numerical measures (**RelNo**): *sharp, dull, dark, light, hard, soft*
- Absolute gradable adjectives with scales closed on both ends (**Abs2**): *full, empty*
- Absolute gradable adjectives with scales closed on one end (**Abs1**): *wet, dry, straight, curved, rough, smooth, clean, dirty, salty*
- 'Evaluative' adjectives (**Eval**): *good, bad, beautiful, pretty, ugly, easy, interesting, boring, tasty, fun, intelligent, happy, sad*

Adjectives were assigned to these categories based on tests described in the literature, as follows. Relative gradable adjectives were identified as those for which both the

[1] In work on the semantics of gradable adjectives, it is now common to distinguish between context-dependent **relative** gradable adjectives and (more) context-independent **absolute** gradable adjectives (Kennedy and McNally 2005; Kennedy 2007). This distinction is proposed to derive from the structure of the scale lexicalized by the adjective: members of the absolute class have scales with maximum and/or minimum points, with these providing the standard for the adjective in its positive form, while members of the relative class have scales that are open on both ends, necessitating a contextual standard. A secondary objective of the present experiment was to explore the correlation between subjectivity and the relative/absolute distinction. Findings in this area are reported in Solt (2016), and due to space considerations will not be discussed here.

adjective and its antonym are acceptable in the frame *x is Adj but y is Adj-er*, and for which neither adjective nor antonym allows modification by *slightly*. Absolute gradable adjectives were identified as those for which either adjective or antonym is infelicitous in the above frame and/or can co-occur with *slightly*. Within the latter class, the division into doubly versus singly closed scales (Abs2 vs. Abs1) was based on judgments reported in the literature. An adjective was considered to have a numerical measure if its comparative form can be modified by a measure phrase. The evaluative category was selected to include adjectives of the sort discussed in the literature under the terms 'evaluative' (see especially Bierwisch 1989) or 'predicate of personal taste' (Lasersohn 2005 and many others). This is a mixed class, encompassing value, taste and aesthetic judgments, emotion words, and psychological predicates; its members are united, and distinguished from those of the other four categories, in that they do not denote external physical properties.

For each adjective, one or more dialogues were created, each featuring a disagreement between two speakers. For example:

(5) A: John and Fred look similar but John is taller than Fred.
B: No, Fred is the taller one of the two.

(6) A: Tommy's shirt is dirtier than the one his little brother Billy is wearing.
B: No, Billy's shirt is dirtier than Tommy's.

(7) A: The vase on the table is more beautiful than the one on the bookshelf.
B: No, the vase on the bookshelf is more beautiful.

Adjectives were split across 4 lists, which were tested sequentially. Some adjectives occurred on more than one list, in different dialogue contexts. Each list contained 8-12 test items and 12 fillers. Fillers were split equally between two types: (i) those expressing factual disagreements (example: A: The judge found Frank guilty. B: No, the judge found Frank innocent.); (ii) those expressing differences of opinion, including statements based on vague nominal predicates (e.g. *jerk*), deontic and epistemic modals, statements of likelihood, and moral judgments. Sample size was 20–25 per list. See the Appendix for the full list of critical items.

2.3 Procedure

The study was executed online via Amazon MTurk, and employed a forced choice task in which participants saw brief dialogues of the form in (5)–(7), and were asked to classify the nature of the disagreement between the two speakers. The task was introduced as follows:

(8) This study is about disagreements between people. Sometimes when two people disagree, only one of them can be right, and the other must be wrong. For example, in this short dialogue, Speaker A and Speaker B can't both be right, because Rosa can't have been born in both July and April.

> Speaker A: Rosa was born in July.
>
> Speaker B: No, Rosa was born in April.

> But sometimes when people disagree, there is no right or wrong answer - it's just a matter of opinion. Here's an example:

> Speaker A: Susan looks a lot like her sister.
>
> Speaker B: No—they don't look alike at all!

> In this HIT, you will see a series of short dialogues between two speakers A and B. Your task is to say whether there is a right or wrong answer, or whether it's a matter of opinion. Please answer based on your intuitions; do not think too long about each question.

Participants were then presented with a list of test and filler dialogues in pseudo-random order; their task was to classify each using one of two response options: "only one can be right; the other one must be wrong" and "it's a matter of opinion". The first of these was coded as a judgment of FACT; the second as a judgment of OPINION.

At the end of the questionnaire, participants were asked age and native language(s), and were given an opportunity to comment on the task. Participants were paid $0.75 for participation.

2.4 Results

The proportion of FACT judgments for each individual adjective and for the five subclasses in aggregate are displayed in Fig. 1. A mixed effect logistic regression model was fitted to the results using the *lme4* package (Bates et al. 2014) in *R* (R Core Team 2015), with response (FACT vs. OPINION) as dependent variable, adjective type as fixed effect, and random intercept for subject. The reference level was RELNUM.

Significant differences were found between RELNUM and ABS1 ($z = -7.016$, $p < 0.001$), RELNO ($z = -8.208$, $p < 0.001$) and EVAL ($z = -12.127$, $p < 0.001$). The difference between RELNUM and ABS2 was not significant ($z = -1.242$, $p = 0.214$). Among the classes found to differ significantly from RELNUM, subsequent post hoc testing via the *multcomp* package (Hothorn et al. 2008) using Tukey correction for multiple comparison found the following significant differences: ABS1 versus EVAL (z-ratio = 11.049, $p < 0.001$), RELNO versus EVAL (z-ratio = 9.054, $p < 0.001$) and ABS1 versus RELNO (z-ratio = 3.803, $p < 0.01$). Regarding the last contrast, however, an examination of the results for individual adjectives shows no clear separation between the two classes (see Fig. 1), suggesting that the overall difference found might be an artifact of the specific adjectives tested.

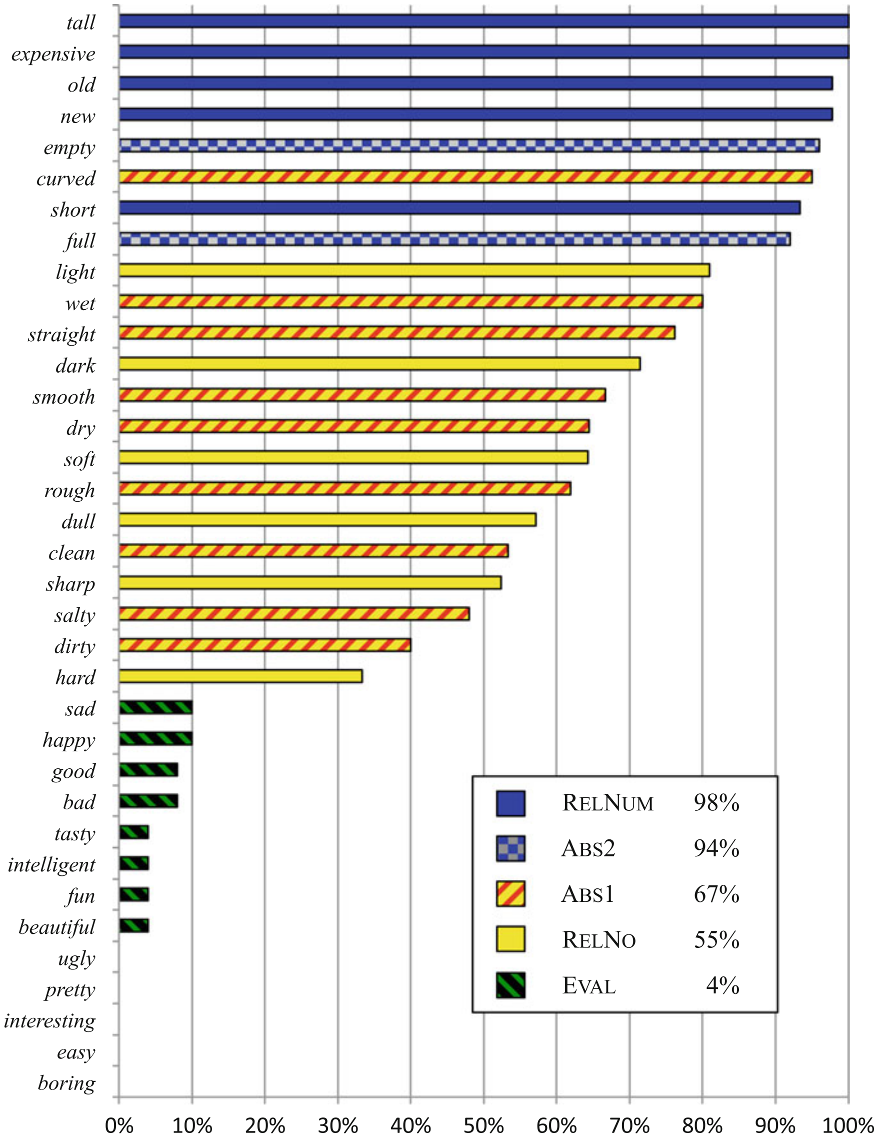


Fig. 1 Results of experiment—percent 'FACT' judgments

2.5 Discussion and Further Observations

With regards to adjectives of the *tall* and *beautiful* classes, our findings are as predicted. For *tall* and the other adjectives tested that have corresponding numerical measurement systems, subjects almost universally judged disagreements about comparative statements to be factual in nature. Note that the absolute double-closed scale pair *full/empty* might be assimilated to this group, in that degrees of fullness (or emptiness) can be quantified in percentages (e.g. *90% full*, *three quarters empty*). Conversely, for *beautiful*, *tasty*, and other adjectives that were classified as evaluative, disagreements about orderings are almost universally judged to be matters of opinion.

The more interesting finding is the existence of a large group of adjectives with mixed behavior, eliciting both FACT and OPINION judgments. This group includes in particular relative gradable adjectives without corresponding measurement systems, as well as absolute gradable adjectives with singly closed scales. Among this group, we observe a range from those adjectives that skew more towards factual readings (e.g. *straight/curved*) to those that skew towards faultless readings (e.g. *clean/dirty*, *salty*).

With respect to ordering subjectivity, we thus find that gradable adjectives divide into not two but rather three groups: objective, subjective and mixed. As a caveat, it is possible that further research might determine that these groups are not as distinct as they appear to be here, or that the dividing lines between them are not precisely where the present experiment shows them to be. That is, we cannot at this point rule out the possibility that adjectives in the objective group might in certain contexts allow subjective interpretations of their comparative forms, or conversely that members of the subjective class might in the right sort of contexts allow objective readings. However, one previously unrecognized finding appears quite clear: there is a large group of adjectives for which the interpretation of the comparative form is neither purely objective nor purely subjective.

Interestingly, the three-way division that emerges on the basis of the present faultless disagreement test is echoed in other phenomena. The most obvious of these involves measurability. Adjectives in the objective group have corresponding measurement units (in fact, the RELNUM group was defined as such). Those in the subjective group almost universally lack such units, and furthermore, for adjectives such as *fun*, *tasty*, *interesting/boring* and *beautiful/ugly*, it is hard to imagine how such units could be created (an exception in this group perhaps being *intelligent*, depending on whether one is willing to accept IQ points as a true measure of intelligence). Finally, adjectives in the mixed group fall somewhere in between. They too largely lack measurement units, but for adjectives such as *hard/soft*, *dark/light* and *clean/dry*, I think one has the intuition that it might be possible (say, in a laboratory setting) to establish such units.

A related phenomenon involves proportional comparisons. As discussed by Sassoon (2010), both dimensional and evaluative adjectives allow modification by proportional expressions such as *twice as*, and this extends to members of the intermediate group as well (see (9)–(11)). But when we turn to precise expressions of proportion such as *2.3 times as*, the picture changes (see (12)–(14)): these are possible for dimensional adjectives, and quite comically infelicitous for members of the evaluative class; for the mixed group they seem marginally possible, when we imagine we are in a situation (again, say, a lab) where the dimension in question is precisely measured:

(9) a. The Eiffel Tower is twice as tall as the Great Pyramid.
b. The laptop is five times as expensive as the tablet.

(10) a. The Serta mattress is twice as hard as the Sealy mattress.
b. The blue shirt is five times as dirty as the green one.

(11) a. Anna is twice as beautiful as Zoe.
b. The roller coaster was ten times as fun as the ferris wheel.

(12) a. The Eiffel Tower is 2.05 times as tall as the Great Pyramid.
b. The laptop is 4.9 times as expensive as the tablet.

(13) a. ? The Serta mattress is 1.9 times as hard as the Sealy mattress.
b. ? The blue shirt is 5.1 times as dirty as the green one.

(14) a. # Anna is 2.3 times as beautiful as Zoe.
b. # The roller coaster was 9.8 times as fun as the ferris wheel.

Thus the pattern observed with respect to interpretation of the comparative form appears to be part of a broader set of facts that relates to the possibility of precise, quantitative measurement.

The remainder of this paper is devoted to developing an account of these patterns. The next section briefly reviews existing semantic theories of subjectivity, focusing on their ability to explain the experimental results. One important proposal to come out of this work is that of multidimensionality as a source of subjectivity, particularly ordering subjectivity; this topic is explored in the section that follows.

3 Theories of Subjectivity

Adjectival subjectivity is the topic of a large body of research in formal semantics. The earliest of this work focused on predicates of personal taste such as *tasty* and *fun*, and pursued the general approach of accounting for their subjectivity by relativizing the interpretation of the adjective to a judge whose opinion or perspective is expressed. Theories in this area can be divided into two broad classes, which differ in how dependence on a judge is linguistically encoded. The relativist analysis (Lasersohn 2005) includes a judge parameter to the index of interpretation, along with the usual time and world parameters (15a). The contextualist approach (Stojanovic 2007; Sæbø 2009), by contrast, assumes that predicates of this sort feature an additional judge or experiencer argument (15b).

(15) a. $[\![\text{tasty}]\!]^{w,t,j} = \lambda x.x$ tastes good to j in w at t

b. $[\![\text{tasty}]\!]^{w,t} = \lambda y \lambda x.x$ tastes good to y in w at t

Elaborations on and combinations of these two approaches are found in Stephenson (2007) and Bylinina (2014, 2017), among others, while authors including Moltmann (2010) have proposed analyses that do not rely on the notion of a judge.

In the form presented, neither of the formulas in (15) accounts for ordering subjectivity. *Tasty* is a gradable adjective, having comparative and superlative forms (*tastier*, *tastiest*) and allowing composition with degree modifiers (*rather/very/extremely tasty*). But the above analyses localize subjectivity at the level of the unmodified positive form, thus providing no explanation for the possibility of subjective judgments regarding the ordering of two entities along a dimension such as tastiness. This might however be remedied fairly simply, by starting with a gradable entry of the form in (3) and relativizing the measure function to a judge.

A more fundamental issue is that the above analyses do not provide an explanation for the finding that adjectives exhibiting ordering subjectivity divide into two groups, depending on whether or not they also allow factual readings for the comparative. If subjective adjectives are those whose interpretation is dependent on a judge index or argument, we are faced with the question of why some of them—but not others—can also be interpreted as making factual statements, i.e. statements that can be evaluated as objectively true or false. In fact, it is not clear how they can acquire factual interpretations at all.

From a different perspective, earlier authors including Kamp (1975) and Klein (1980) observed that certain gradable adjectives (e.g. *clever*) are dependent on multiple underlying dimensions for their ascription, one consequence of which is variability in judgments about the relative ordering of two individuals. More recent work (see especially Sæbø 2009; Kennedy 2013; Bylinina 2014; McNally and Stojanovic 2017; Umbach 2016) has connected this insight to the topic of subjectivity.

A central observation that has come out of this later work is that a wide range of gradable adjectives are subjective in their positive forms, including not only classical personal taste predicates but also other evaluative adjectives as well as vague gradable adjectives more generally; but only the first two of these are also subjective in their comparative forms (see again (1) vs. (2)). The conclusion is that there are two distinct loci of subjectivity. For vague gradable adjectives such as *tall*, subjectivity is localized not in the lexical meaning of the adjective itself but rather in the semantics of the positive morpheme *pos* that provides the threshold of applicability for the adjective in its unmodified form. For adjectives such as *tasty*, *fun* and *beautiful*, it derives from the lexical semantics of the adjective.

Kennedy (2013) proposes that this difference in which adjectival forms can be interpreted subjectively corresponds more fundamentally to two distinct types of subjectivity, the first deriving from uncertainty in the determination of the contextual standard for the application of a vague adjective, the second deriving from what he terms the "shared semantics of qualitative assessment." He notes however that the two sorts of subjectivity might nonetheless be unified as deriving from a more basic property of "dimensional uncertainty." For adjectives of the *tall* class, it is uncertainty as to the dimensions involved in standard calculation, while for those of the *tasty* sort, it is uncertainty as to how the dimensions of qualitative assessment are integrated by different judges.

Kennedy makes the further important observation that many gradable adjectives are ambiguous between an objective/dimensional reading and a subjective/qualitative reading. For example, to say that the cake is heavy might be to say something about its objectively measurable weight, or alternately about the subjective experience of tasting it. This suggests an account of the mixed group found in the present experiment in terms of ambiguity (though we will see below that there are also other possibilities).

The notion of multidimensionality as a source of subjectivity is taken up further by McNally and Stojanovic (2017) in the context of an investigation of aesthetic adjectives such as *beautiful*. They observe that "[d]eciding whether an adjective describing a multidimensional property holds of some individual involves not only determining a threshold of applicability but also determining the relative weight of each of the dimensions that contribute to the property in question. Here, again, there will be room for disagreement between speakers" (2017, p. 21). And further: "Two speakers may disagree about whether Ayumi is healthier than Mihajlo because they may disagree about whether one component of health or another (e.g. the state of the cardiovascular system vs. the immune system) should carry more weight" (2017, pp. 21–22). Multidimensionality is however only one source of subjectivity, others being experiential semantics (characterizing adjectives such as *tasty* and *interesting*) as well as evaluativity in the sense of expressing an attitude of positive or negative evaluation on the part of the speaker (e.g. *good*, *bad*, *beautiful*).

Bylinina (2014) proposes a formal analysis of adjectival subjectivity that explicitly incorporates multidimensionality. Her account is based in part on the observation that the class of adjectives exhibiting ordering subjectivity can itself be further subdivided: subjective readings for the comparative are possible for both adjectives such as *fun*, *tasty* and *interesting* that refer to internalized experiences as well those such as *intelligent* that do not; but only the former allow a judge or experiencer PP:

(16) a. The chili was tasty to me.
b. The book was interesting to/for me.
c. ?? Anna is intelligent to/for me.

Bylinina proposes that the interpretation of both sorts of adjectives is dependent on a judge index, but that the judge plays a different role in the two cases. Members of the *tasty* class have an experiencer argument that is equated to the judge. In the case of adjectives such as *intelligent*, she draws on work by Sassoon (to be discussed further below) in proposing that their subjectivity derives from multidimensionality: degrees of intelligence, for example, can be conceptualized as the lengths of vectors in a multidimensional space, with the weights assigned to component dimensions being relativized to judges. Her formalization is the following (where Q is a dimension contributing to intelligence, w_Q^j is the weight assigned by j to Q, $m_{x,Q}$ is the measure of an individual x with respect to Q and s_Q is the standard of applicability for Q).

(17) $[\![m_{x,\textbf{intelligent}}]\!]^{c;w,t,j} = \lambda x. \sqrt{\sum_{Q} [w^{j}_{Q}(m_{x,Q} \succ s_{Q})]^{2})}$

Umbach (2016) takes a somewhat similar approach, analyzing the evaluative adjective *beautiful* in terms of a generalized measure function that maps entities to points in a multidimensional attribute space.

In summary, several authors have argued convincingly that a source of adjectival subjectivity, and specifically ordering subjectivity, is the multidimensional nature of the properties in question. But note that each of these accounts has treated multidimensionality-based subjectivity as a variety of judge dependence: two judges may weight an adjective's dimensions differently, potentially giving rise to disagreements about orderings. This brings up a more general point. In all of the works discussed in this section, the focus has been on 'subjectivity' in the sense of the diverging perspectives of distinct speakers. This perhaps stems from the initial focus on personal taste predicates such as *tasty* and *fun*, which so clearly express individuals' judgments or tastes. When we expand our focus to the full range of adjectives considered in the present work, it becomes clear that differences between judges are not the only source of variable judgments regarding orderings; rather, it seems that a single speaker's judgments may also be potentially uncertain or changeable. Consider for example two shirts, one which is clean except for a grass stain on the sleeve, the other slightly dingy overall. Which one should I consider dirtier, and which cleaner? I think my answer has to be 'it depends'—on what type of shirt and how it will be used, on what sort of dirt we are most concerned about, and so forth. The same might be said, for example, regarding which of two surfaces is rougher, or which of two fences is straighter. Variability of this sort cannot be accounted for by relativization to a judge, but rather seems to reflect a more general sort of context dependence.

In the next section, I take a more in-depth look at the nature of adjectival multidimensionality. This will form the basis for the formal account in Sect. 5, which also seeks to clarify the relationship between multidimensionality and judge dependence.

4 Identifying Multidimensionality

If we are to investigate the hypothesis that a source of subjectivity (including ordering subjectivity) is the multidimensional nature of the predicates in question, then we must have a way of identifying which adjectives are multidimensional. This turns out to be less straightforward than it might initially seem.

4.1 Sassoon's Theory of Multidimensionality

As noted above, it has long been recognized that some gradable adjectives are multidimensional (see especially Kamp 1975 and Klein 1980; for discussion of multidimensionality more broadly, see also Bartsch and Vennemann 1972; Bartsch 1984, 1986; Landman 1989). But the most in-depth investigation of multidimensionality is found in the work of Sassoon (2007, 2011, 2012, 2013, 2015), who develops a comprehensive semantic theory that encompasses both multidimensional adjectives and nouns, and that extends to topics including the nature of the adjectival antonymy relationship and the semantics of comparison and degree modification. In Sassoon's theory, multidimensional adjectives such as *healthy*, *sick*, *identical*, and *intelligent* are associated with dimensions that can be specified overtly or bound by explicit or implicit logical binding operators. For conjunctive adjectives such as as *healthy*, the default binding operator is universal quantification: to be healthy is to be healthy in all contextually relevant respects (18a). For disjunctive adjectives such as *sick*, the default is existential quantification: to be sick is to be sick in some relevant respect(s) (18b). Adjectives such as *intelligent* are mixed, with pragmatics determining the binding operation.

(18) a. *healthy*: $\lambda x.\forall Q \in DIM(healthy) : Q(x)$

b. *sick*: $\lambda x.\exists Q \in DIM(sick) : Q(x)$

Comparatives might then be analyzed as involving the counting of or quantification over dimensions: one individual might be evaluated as healthier than another if she is healthy in a larger number of relevant respects, if for relevant respects generally she is healthier, or if she is healthier in some particular contextually salient respect (Sassoon 2015).

Multidimensionality manifests itself grammatically in a number of ways: individual dimensions may be specified via prepositional phrases headed by *with respect to* or *in* (19) or inquired about via a *wh-phrase* (20); dimensions may be quantified over (21); and quantificational force may be restricted by exception phrases (22).[2] None of these are possible with (uni-)dimensional adjectives such as *tall*.

(19) a. The patient is healthy with respect to blood pressure.

b. The boxes are identical in size and weight.

c. # Zoe is tall with respect to height.

(20) a. In what respects is the patient healthy/sick?

b. In what respects are the boxes identical?

c. #? In what respect is Zoe tall?

(21) a. The patient is healthy in every/most/three/some (important) respect(s).

b. The boxes are identical in every/most/three/some respect(s).

c. # Zoe is tall in every/most/three/some respect(s).

(22) a. The patient is healthy/not sick except for high blood pressure/asthma/a slight cold.

[2] Which quantifiers are felicitous, and whether an exception phrase is possible with an adjective in its positive or negated form, depend to some extent on whether the adjective is conjunctive or disjunctive. I will attempt as much as possible to abstract away from these details here.

b. The boxes are identical except for size/color.

c. # Zoe is tall except for …

Sassoon backs up these judgments with extensive corpus and experimental data, particularly relating to the pattern in (22).

Multidimensionality of the sort described here has also been proposed to play a role in other linguistic patterns, such as the acceptability of so-called borderline contradictions (see Égré and Zehr, this volume).

4.2 Varieties of Multidimensionality

Among the multidimensional adjectives that Sassoon investigates are a number that were found in the present research to exhibit ordering subjectivity: *good, bad, beautiful, ugly, happy, intelligent, tasty, clean* and *dirty*. More generally, when we look at the mixed and purely subjective groups that emerged from the experiment, we see that many are multidimensional at least in a conceptual sense. Whether an individual or experience might be characterized as fun, interesting, boring, or easy—or more fun/interesting/boring/easy than another—is clearly dependent on multiple aspects or properties of the entities under consideration. Even the adjective *salty* can be put in this class: while one might think that degree of saltiness is dependent on a single dimension, namely salt content, research in psychophysics has in fact found that perceptions of saltiness are impacted by a variety of other factors, including consistency, texture and fat content (see e.g. Christensen 1980; Pflaum et al. 2013; Suzuki et al. 2014).

However, when we attempt to confirm the multidimensional status of such adjectives via tests based on the constructions in (19)–(22), and thereby clarify which of the adjectives exhibiting ordering subjectivity are multidimensional, the results are quite mixed. Consider to start the personal taste predicates *tasty* and *fun*, both of which patterned as purely subjective in our experiment:

(23) a. The chili was tasty with respect to …

b. In what respect/way was the chili tasty?

c. The chili was tasty in every/?most/??three/some respect(s).

d. The chili was tasty except for the consistency/being too salty/??

(24) a. The roller coaster was fun with respect to …

b. In what respect was the roller coaster fun?

c. The roller coaster was fun in ?every/?most/??three/some respect(s).

d. The roller coaster was fun except for the wind/the rattling/??

Compared to the corresponding examples with *healthy*, *sick* and *identical*, it seems more difficult to continue the sentences in (23a), (24a), or to answer the questions in (23b), (24b).[3] What are the respects of tastiness and fun that contribute to the attribution of these predicates? If anything, the questions seem to favor a rhetorical interpretation, challenging the interlocutor to name even one ground for calling the chili tasty or the roller coaster fun. Similarly, universal and existential quantification over dimensions is moderately acceptable ((23c), (24c)), producing emphatic and hedging effects, respectively, but precise counting of dimensions (??*fun/tasty in three respects*) is rather odd. Finally, it is certainly possible to distinguish a few particular aspects of the properties in questions to form the basis of exception phrases (e.g. saltiness and consistency in the case of *tasty*); but after these the task becomes more difficult (see (23d), (24d)), suggesting that there is a considerable residual meaning that cannot be easily separated into discrete dimensions.

A similar issue emerges with other evaluative predicates, where we see that even when examples parallel to (19)–(22) sound felicitous, they do not necessarily involve specification of or quantification over dimensions. Take for example *beautiful*, another of the adjectives that fell in the purely subjective group in our experiment. A Google search yields thousands of examples of the phrases *beautiful in every respect* and *beautiful in every way*. But many of these have the character of those in (25), where the listed aspects seem to be not component dimensions of the predicate *beautiful* but rather component parts of a complex entity or event that is the subject of predication.

(25) a. The wedding was beautiful in every respect … the weather, the venue, the bride's dress, and most of all, the people!
b. This newly built home is beautiful in every way, featuring a welcoming great room with stone fireplace, a light-filled open-plan kitchen, and a spacious master bedroom suite.

Something similar is seen with exception phrases: *Zoe is beautiful except for …* is most naturally continued with something like *her crooked nose/her small eyes/her hair*/etc.; but nose, eyes, hair and the like are not dimensions of beauty but rather parts of the individual described. To be sure, dimensional uses can be found, as when we characterize a painting as *beautiful except for the color* (McNally and Stojanovic 2017). But the simpler the object of predication, the more difficult it is to construct such examples. As an extreme case, imagine a paint chip in a particular shade of blue. I might characterize the color as beautiful, but it is hard to imagine specifying the dimensions that make it so (?*this color is beautiful with respect to …*) or less so (?*this color is beautiful except for …*). Replacing *beautiful* with *ugly* makes these judgments in my opinion even sharper. Sassoon (2013) acknowledges and discusses non-dimensional uses of exception phrases with multidimensional adjectives, but without really exploring the difficulty of creating true dimensional examples for those such as *beautiful*.

[3] For myself, examples of this sort are quite bad; a reviewer, however, found them more acceptable. Such between-speaker variation is itself indicative of the difficulty in classifying an adjective as multidimensional versus unidimensional.

Here I do not mean to claim that adjectives such as *tasty*, *fun* and *beautiful* can never have a multidimensional interpretation (in Sassoon's sense); the possibility of dimensional exception phrases and the like is enough to show this cannot be right. The multidimensional interpretation might in particular be more available to experts in the relevant domains (think for example of a food writer or art critic), who have a trained ability to introspect into the factors underlying their judgments. The point is rather that such adjectives, while without doubt multidimensional at the conceptual level, also have an interpretation—perhaps the most salient one—on which they behave grammatically as if they were unidimensional.[4]

Consider now the adjectives in our mixed group. Of these, *clean* and *dirty* are discussed as multidimensional by Sassoon, and this is supported by the above-described tests:

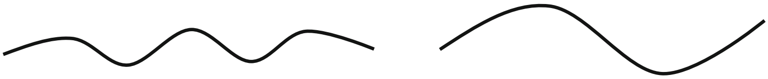
(26) a. In what respect(s) was the shirt clean/dirty?
b. The shirt was clean/dirty in every/most/?three/some respect(s).
c. The shirt was clean/wasn't dirty except for the musty smell/a few grass stains/being slightly dingy.

But when we look at other members of this group, the results are quite different. Taking *except* phrases as an example, it is difficult to construct true dimensional completions of examples such as the following:

(27) a. The line was(n't) straight/curved except for …
b. The leather was(n't) smooth/rough except for …
c. The knife was(n't) sharp/dull except for …
d. The soup was(n't) salty except for …

Yet there is nonetheless a sense in which adjectives such as these are multidimensional. This is most clearly brought out by considering cases of potential disagreement. For example, we might disagree—or simply find it difficult to decide—which of the two lines below is straighter or more curved, the issue being how exactly we should measure degree of straightness or curvature: is it a matter of the number of curves? the sharpness of each? the total area of deviation from perfect straightness? There seems to be no principled correct answer.

(28)



To take a more concrete example, imagine two city streets, one paved and completely smooth except for a few largish speed bumps and potholes, the second with an all-over cobblestone surface. Which is bumpier? Again the answer seems to be 'it depends', the issue once more being how different sorts of bumps, dips and other deviations from complete flatness should be integrated to derive an overall degree of

[4] I thank the reviewers for pointing out the need to clarify this point.

bumpiness.[5] I believe similar examples might be constructed for other members of the mixed class, including *rough/smooth*, *sharp/dull* and perhaps even *wet/dry*. This is not multidimensionality in quite the same sense as that characterizing adjectives such as *healthy*, whose meanings can readily be broken down into discrete independent dimensions (e.g. blood pressure, cholesterol, etc.) that we can name, count and quantify over. But adjectives of the *curved* and *bumpy* type share with those of the *healthy* type the property that their attribution depends on multiple aspects of the physical characteristics of entities, which must be integrated in some way to produce the overall meaning of the adjective.

We have seen that there are adjectives that are in some sense multidimensional but that are not entirely felicitous in the constructions in (19)–(22). The reverse is also true: certain adjectives that are generally considered to be dimensionally ambiguous rather than multidimensional are relatively acceptable with *respect*. Examples are *large* and *long*:

(29) In which respect is London larger than New York?
Land area ✔ Population size ✘

(30) The sofa is larger than the bench in every respect.

(31) a. The trip to Tübingen is longer than the trip to Konstanz.
b. In which respect—travel time or distance in kilometers?

This suggests that *which respect* questions at least might in fact offer a test for the contextual dependence of the communicated dimension, rather than for multidimensionality.

In summary, the preceding discussion suggests that adjectival multidimensionality is not a homogenous phenomenon. There are gradable adjectives such as *healthy* and *identical* that are multidimensional in what might be called a quantificational sense: their component dimensions are readily named, easily separated, and grammatically active, and for the positive form of the adjective at least, a variety of tests suggest that they are integrated by means of quantificational operators. But there are other sorts of intuitively multidimensional adjectives—examples being *bumpy*, *curved*, *salty* and (in my judgments) *fun* and *tasty*—for which the individual component dimensions are much less grammatically, or even conceptually, accessible. The attribution of such predicates certainly depends on multiple aspects or properties of the object of predication; but (ordinary) speakers are quite likely not aware of or able to name these aspects and properties. Furthermore, that such adjectives tend to pattern as unidimensional rather than multidimensional on the above-described tests suggests that their dimensions do not compose via universal or existential quantification but rather are integrated in some other manner to create a single, complex dimension. The dividing line between these two variants of multidimensionality is not entirely sharp; quite plausibly, some adjectives (e.g. perhaps *beautiful*) allow both sorts of

[5] The pair *flat/bumpy* was not included in the present experiment, but I hypothesize that they would behave similarly to pairs such as *smooth/rough*; as *bumpy* provides a particularly nice example, I allow myself the liberty of using it here.

interpretations, or combine the two on a single usage. Given this, I will continue to use the term 'multidimensional' to describe both sorts of adjectives.

For the purposes of the present paper, the crucial observation is that both varieties of multidimensionality—the quantificational variety and the complex dimension variety—appear to give rise to the possibility of subjective judgments regarding orderings. Capturing this observation is a central goal of the formal analysis proposed below.

4.3 Multidimensionality and Evaluation

There is a further distinction among the class of adjectives that are multidimensional in the broad sense, which is subtle but I believe nonetheless real, and which is relevant to the adequate formal analysis of such adjectives.

For classic examples of multidimensional adjectives such as *healthy/sick* and *identical* as well as those such as *clean/dirty*, *straight/curved* and *flat/bumpy*, the overall meaning of the adjective is in a sense built up directly from its component dimensions, integrated in some contextually determined way. The degree of sickness of an individual is determined by the nature and perhaps severity of his relevant illnesses; the bumpiness of a road by the size/shape/etc. of the bumps and dips on it; the straightness or curvedness of a line, by the number or shape or other mathematical properties of the curves on it.

For so-called evaluative adjectives, namely those of the sort that made up the EVAL group in the present experiment, there is something more that this. Specifically, while the adjective's meaning is based in some way on multiple underlying properties of the object of predication, there is also an inherent human element. Some are experiential in nature, as diagnosed by the possibility of modification by an experiencer PP (e.g. *tasty to me*; *fun for me*; see Sect. 3); experiential meaning requires an experiencer. Others express an aesthetic or taste judgment. Yet others convey an emotion, and are thus necessarily rooted in the perceptions or feelings of an individual. And while it is arguably not an inherent aspect of their meaning, on their typical uses most are evaluative in the sense of expressing a positive or negative value judgment; value judgments (like taste and aesthetic judgments) require an individual who judges. To borrow a term used by McNally and Stojanovic (2017), all of these sorts of adjectives require the "intermediation of a sentient individual" in their attribution.

The claim that I would thus like to make is that multidimensional adjectives can stand in two distinct types of relations to their component dimensions. For those such as *healthy*, *clean/dirty* and *flat/bumpy*, the adjective's overall meaning can be expressed directly as a function of its dimensions (though the function is context dependent, and might not be fully transparent to the ordinary speaker). But for adjectives such as *fun*, *tasty* and *beautiful*, what we have called dimensions are more properly factors that contribute to an agent's subjective experience with or evaluation of an entity or event. That is, the adjective's meaning is not a direct function of its dimensions; rather, 'dimensions' serve as the basis for a taste, value or aesthetic judgment, and it is this that might more properly be considered the meaning of the adjective.

This above claim is similar to one made by the moral philosopher Hare (1952), who argues that evaluative terms such as *good* have the special function in language of commending, and cannot be defined in terms of other words which themselves do not have this function without losing the means of performing the commending function. A good strawberry, for example, may be one that is large, red and juicy; but *good* as applied to strawberries cannot be defined as meaning ‘large, red and juicy’. Hare further argues for the need to distinguish the meaning of evaluative words from the criteria for their application; the latter vary with the class of items to which the word is applied (i.e. what makes a good car is different from what makes a good strawberry), while the meaning, whose core is the commending function, remains constant. Criteria as discussed by Hare are close in spirit to what we have called the dimensions of evaluative adjectives (see also Umbach 2016 for related discussion).

It is rather difficult to design diagnostics for the distinction suggested above, but a possible one is based on follow-up questions. For at least some adjectives of the *healthy/clean/bumpy* sort, a speaker can be asked to clarify her assertion by means of a *what respect/way* question.

(32) a. Fred is healthier/sicker than Tom.
b. The blue shirt is cleaner/dirtier than the green one.
c. Weserstrasse is bumpier that Friedelstrasse.
i. In what respect / way?

But for assertions based on the comparative forms of evaluative adjectives and personal taste predicates, such a question about respects is, as I have suggested above, slightly infelicitous. Instead, a more natural way to question the speaker’s assertion is to ask for her reasons for it, for example with *What makes you say that?*

(33) a. The chili is tastier than the soup.
b. The roller coaster was more fun than the ferris wheel.
c. The Picasso is more beautiful than the Miró.
i. #In what respect / way?
ii. Why do you say so / what makes you say that?

This suggests a recognition that for adjectives of the latter sort, the objective properties of the subject(s) of predication contribute to the attribution of the adjective only indirectly, through their effect on the perceptions or judgments of the speaker.

4.4 Summary

We have seen here that a wide variety of gradable adjectives are multidimensional in a conceptual sense, being dependent on multiple properties of an object for their attribution, and thereby distinguishable from straightforward (uni-)dimensional adjectives, which lexicalize a single, typically measurable dimension. But the multidimensional class can itself be further subdivided. In some such adjectives (or perhaps more accurately, uses of such adjectives), the component dimensions are readily accessible and grammatically active, while in others they are integrated in a way that is not transparent to the average speaker. And I have argued that the meaning of some conceptually multidimensional adjectives can be expressed as a direct function of their dimensions, while for others, their dimensions play a more indirect role in their meaning, as factors contributing to some sort of judgment by a sentient individual. Importantly, all of these varieties of multidimensionality result in ordering subjectivity, though I will propose that they do so in different ways.

5 Proposal

In this section, I outline a theory of gradable adjective meaning that formalizes the observations from the prior two sections, and that provides the basis for explaining the availability of objective and subjective readings of the comparative forms of different sorts of adjectives.

5.1 Scalar Semantics

I begin with the definition of a scale S as triple of the following form:

(34) $S = \langle D, \succ, DIM \rangle$, where

- DIM is a dimension of measurement
- D is a set of degrees
- $\succ$ is an ordering relation on D

Differing from some other authors, I assume here that D can but need not be the real numbers, and that the ordering relation $\succ$ can but need not be a total order on D.

A measure function μ_{DIM} can then be defined as a function from a domain of measurement Dom (e.g. the domain of individuals or of events, or a subset thereof) to some scale S tracking dimension DIM.

Building on proposals by Sassoon (2010) and Kennedy (2013), I further propose that gradable adjectives have underspecified semantics, lexicalizing not a single measure function but a family of functions indexed to contexts. Each context c in the set of contexts C specifies a world, time and judge as well as other aspects of the situation of utterance; here I explicitly assume that two contexts $c, c' \in C$ may differ in the measures assigned to individuals, even if the physical properties of objects in the world remain the same. The general template for gradable adjective meaning is thus the following:

(35) $[\![\text{Adj}]\!]^c = \lambda d \lambda x . \mu^c_{DIM}(x) \succeq d$

To put this differently, gradable adjectives on this view lexicalize dimensions rather than particular scales or measure functions. A dimension is a property that an entity can have more or less of. A measure function corresponding to that dimension is a mapping from individuals to degrees that represent the extent that each individual has the property in question. As a very simple example to demonstrate that these two things are not equivalent, the single dimension *HEIGHT* may be tracked by a function that maps individuals to their height in inches, or alternately by a function that maps individuals to their height in centimeters. For a simple unidimensional adjective such as *tall*, this might be the only sort of variation that is possible; but for other classes, there are further possibilities. Below I will argue that the availability of objective versus subjective readings for the comparative derives from constraints on the possible variation in the family of functions $\{\mu^c_{DIM} : c \in C\}$ that is the semantic content of the adjective.

5.2 Sources of Objectivity

Above we noted the link between measurability—i.e. the possibility of associating entities with numerical measures—and objective rather than subjective interpretations for the comparative. Building on this insight, I propose that **objective readings are possible in those cases where the set of measure functions lexicalized by the adjective is such that it allows a principled, order-preserving mapping to the real numbers**. This has the effect of externalizing orderings of individuals, aligning them across speakers to the fixed order of the number line.

There are several routes to such a mapping. The most straightforward of these arises when the adjective lexicalizes measure functions that are **additive with respect to concatenation**, meaning that the measure assigned to two individuals concatenated in the relevant way is the sum of their two individual measures (see Krifka 1989; Sassoon 2010; Lassiter 2011 and references therein). The dimension of height is a classic example: the height of two individuals stacked one on top of the other is the sum of their individual heights. Other dimensions that satisfy additivity include weight, depth, width, length, volume and duration. Even cost arguably falls in this class: while items are often cheaper if purchased in quantity, the fact that we recognize this as a discount is an indication that we perceive cost as inherently additive. Additivity provides the possibility of numerical measurement: some standard element is selected as the basis of a unit of measurement, and the measure of any individual can then stated in terms of multiples of this standard. A 6-meter-tall tree, for example, is one whose height is equivalent to the concatenation of six copies of a 1-meter standard element.

Formally, additivity may be encoded via a constraint on the set of measure functions $\{\mu^c_{DIM} : c \in C\}$ that is lexicalized by the adjective. For readers interested in the technical details, the constraint is that in (36) (where $\oplus$ is the relevant concatenation operation). A sample denotation for an adjective satisfying this constraint is (37).

(36) **Additive measure functions:**
$\forall c \in C$ and $\forall a, b \in Dom,\ \mu^c_{DIM}(a \oplus b) = \mu^c_{DIM}(a) + \mu^c_{DIM}(b)$

(37) $[\![\text{tall}]\!]^c$=$\lambda d \lambda x.\mu^c_{HEIGHT}(x) \succeq d$,
where $\forall c \in C$ and $\forall a, b \in Dom$,
$\mu^c_{HEIGHT}(a \oplus b) = \mu^c_{HEIGHT}(a) + \mu^c_{HEIGHT}(b)$

Beyond additivity, there are other possible routes to numerical measurement. First, there are dimensions for which **natural, speaker-external phenomena** serve as the basis for measurement units. Examples of this include temperature as well as temporal dimensions. In the case of time, the rotation of the earth and its orbit around the sun provide the basis for the units ‘day’ and ‘year’; subdivision and concatenation of these units yield further units such as ‘hour’, ‘minute’, and ‘week’. For temperature, the freezing and boiling points of water provide two anchor points on the scale, which can then be divided into equal increments, for instance by equal increases in the level of mercury in a thermometer. Units derived in this way provide another sort of principled mapping from entities to numbers.

A further class of dimensions that support numerical measurement consists of those that are **derivable from measurable dimensions in a context-independent way**. The dimension of fullness provides a good example: the degree of fullness of a container (say, a bottle or gas tank) can be expressed as the volume of its contents divided by its capacity, i.e. the volume it is able to hold. A half full tank, for example, is one whose contents have half the volume of its capacity. Other dimensions in this class might be purity (defined as volume of impurities relative to total volume) and speed (distance traveled divided by duration). In each of these cases, numerical measures can be derived on the basis of the component measure functions, which enables proportional or ratio measure expressions, as in *20% full*, *90% pure*, and *5 kilometers per hour faster/slower*.

Formally, adjectives falling in this class are those that satisfy the constraint in (38). As an example, the corresponding lexical entry for the adjective *full* is given in (39):

(38) **Context independent derived measure functions**:
$\forall c \in C$ and $\forall x \in Dom$,
$\mu^c_{DIM}(x) = f(\mu^c_{DIM_1}(x), \mu^c_{DIM_2}(x), \ldots, \mu^c_{DIM_n}(x))$,
where $\mu^c_{DIM_1},\ \mu^c_{DIM_2}, \ldots \mu^c_{DIM_n}$ are objective measure functions

(39) $[\![\text{full}]\!]^c$=$\lambda d \lambda x.\mu^c_{FULLNESS}(x) \succeq d$,
where $\forall c \in C$ and $\forall x \in Dom$,
$$\mu^c_{FULLNESS}(x) = \frac{\mu^c_{VOLUME}(content(x))}{\mu^c_{VOLUME}(capacity(x))}$$

In all of these cases, entities can be associated in a principled way with numerical values that reflect their position with respect to the relevant dimension DIM. The prediction is that the comparative form of the corresponding adjectives will be interpreted objectively, and this is consistent with our experimental findings for *tall/short* and *expensive* (additive dimensions), *old/new* (time expressions) and *full/empty* (function of additive measure functions). We would predict similar results for other adjectives in these classes.

5.3 Sources of Subjectivity

Let us turn now to adjectives whose comparative forms can be interpreted subjectively, as diagnosed by the possibility of faultless disagreement. The overall approach that I pursue is that **ordering subjectivity arises when the set of measure functions lexicalized by the adjective is such that a difference in context can result in a difference in the relative ordering of two individuals**. Building on the previously discussed observations by Bylinina (2014), as well as the discussions in Sects. 3 and 4, I propose that this can come about in two ways, namely through multidimensionality and dependence on a judge.

Multidimensionality. Above we discussed the insight that certain adjectives exhibiting ordering subjectivity are multidimensional. Underspecification in or uncertainty about the component dimensions and how they should be integrated results in the potential for disagreement as to orderings. Take for example the pair *clean/dirty*. Intuitively, the degree of cleanness or dirtiness of an object is a function of the amount and type of dirt on it, perhaps in proportion to its size. But which sorts of dirt (broadly construed) we are concerned with, and how different sorts should be weighted relative to one another, are matters of potential disagreement, and there does not seem to be a principled correct choice. On one way of making this more specific, shirt a might work out to be dirtier than shirt b, while on another equally valid choice, the reverse relation might obtain.

To formalize this, I follow Sassoon (2013) and Bylinina (2014) in proposing that adjectives of this sort are associated in each context c with a set of component dimensions $DIM_1^c, DIM_2^c, \ldots, DIM_n^c$. Departing somewhat from these authors, I further assume that to each dimension DIM_i^c there corresponds a measure function $\mu^c_{DIM_i^c}$, the outputs of which are integrated by some function f^c. We have already seen something similar in the form of the lexical entry for *full*. But in that case, subjectivity did not arise, because both the component dimensions and the manner of their combination were fully specified. Ordering subjectivity arises when this requirement is relaxed, such that one or both of these factors becomes context dependent. (40) specifies the form of such functions, and (41) gives a plausible if undoubtedly overly simplistic entry for *dirty* in this form.

(40) **Context-dependent derived measure functions**:
$\forall c \in C$ and $\forall x \in Dom$,
$\mu^c_{DIM}(x) = f^c(\mu^c_{DIM_1^c}(x), \mu^c_{DIM_2^c}(x), \ldots, \mu^c_{DIM_n^c}(x))$

(41) $[\![\text{dirty}]\!]^c = \lambda d \lambda x . \mu^c_{DIRTINESS}(x) \succeq d$,
where $\forall c \in C$ and $\forall x \in Dom$,
$$\mu^c_{DIRTINESS}(x) = \frac{\sum_{i=1}^{n} k_i^c \cdot \mu^c_{AMOUNT}(dirt_i^c(x))}{\mu^c_{SIZE}(x)}$$

Here the individual dimensions that underlie the adjective's meaning may themselves be objectively measurable. Subjectivity derives from the potential for variation in the choice of these dimensions and how they are combined.

Note that in the above formulation I have not made a distinction between the quantificational and complex dimension varieties of multidimensionality discussed in the previous section, though I leave open the possibility that this may ultimately prove necessary.

Judge dependence. The entries in (40) and (41) do not explicitly reference a judge. Rather, measure functions are indexed to contexts; distinct orderings in two contexts c and c' may derive from a difference between judges (the judge being part of the context), but also from other contextual factors. This is as it should be, given the earlier observation that uncertainty or variability regarding the ordering of individuals relative to a multidimensional property such as dirtiness can persist in the judgments of a single speaker. However, we have also seen that many gradable adjectives denote properties whose ascription necessarily involves a human element, or what was earlier called the mediation of a sentient individual. These include value judgments (*good/bad*), aesthetic judgments (*beautiful/ugly*), taste ascriptions (*tasty*), experiential properties (*interesting/boring*) and internal states (*happy/sad*). Such adjectives do not directly describe properties of objects and events in the world, but rather our perceptions of, judgments about and experience with the objective world. For this class, I propose that their dependence on sentient mediation be represented in their semantics. I thus build on the existing tradition of work on subjectivity in taking these to involve measure functions parameterized to a judge.

Adapting for concreteness the relativist approach, we may represent this formally as follows:

(42) **Judge dependent measure functions**:
$[\![\text{Adj}]\!]^{c;j} = \lambda d \lambda x . \mu^{c;j}_{DIM}(x) \succeq d$
where $\mu^{c;j}_{DIM}(x)$ should be interpreted as
'the degree to which j judges x in context c to have property DIM'

(43) $[\![\text{beautiful}]\!]^{c;j} = \lambda d \lambda x . \mu^{c;j}_{BEAUTY}(x) \succeq d$

Again it is possible that this class must be further subdivided, for example to distinguish between adjectives with experiential semantics such as *tasty* and *interesting* and evaluative predicates such as *beautiful* (per McNally and Stojanovic 2017, cf. Sect. 3). I do not attempt to address this here.

The formulations in (42) and (43) do not represent adjectives such as *beautiful* as explicitly multidimensional. The rationale for this derives from the observations in Sect. 4. As was noted there, the dimensions underlying adjectives such as *beautiful* and *tasty* are not as accessible grammatically or even conceptually as for paradigm cases such as *healthy*. More fundamentally, I argued in that section that dimensions play a different role for such adjectives than for those such as *healthy* and *clean/dirty*, being not direct components of the adjective's meaning but rather grounds for an agent's taste, value or aesthetic judgment. This suggests that in these cases the meaning of the adjective should not be represented as a function of its dimensions. On the basis of these observations, as well as general considerations of parsimony, I thus tentatively conclude that subjective adjectives of the judge-dependent type are only multidimensional at the conceptual level, but should be represented as unidimensional in their semantics.

As further evidence for the need to distinguish between underspecification of meaning due to multidimensionality (as in (40)) and judge dependence (as in (42)), there are adjectives that appear to be ambiguous between the two types of subjective interpretations. Consider again the adjective *bumpy*, and the two city streets from our earlier example, one smooth except for isolated potholes and speed bumps, the other with a cobblestone surface. On the basis of this description alone (or perhaps pictures of the two streets), there is room for uncertainty or between-speaker disagreement as to which of the sentences in (44) is true, the issue being how to weight the different sorts of bumps and dips to arrive at an overall measure of bumpiness. But the disagreement in (45) implies something further, namely that the speakers have experienced the two streets in question (e.g. by riding a bike over them):

(44) a. Weserstrasse is bumpier than Friedelstrasse.
b. Friedelstrasse is bumpier than Weserstrasse.

(45) a. I find Weserstrasse bumpier than Friedelstrasse.
b. I find Friedelstrasse the bumpier of the two.

Thus *bumpy* appears to allow both a simple multidimensional interpretation and an experiential (and thus judge dependent) interpretation. The adjective *sharp* provides a similar example: on examining two knives under a microscope, one might be uncertain as to which is sharper, the issue again being how precisely degrees of sharpness should be determined. But to assert that *I find the first knife sharper than the second one* requires that I have used both of them, and says something about my own subjective and experience-based perception of how the two should be ordered. The difference between the examples with and without *find* suggests that an adjective can be multidimensional—and thus exhibit ordering subjectivity—without explicitly having a judge or experiencer as part of its semantics.

5.4 Mixed Predicates

Having discussed the characteristics of the measure functions that support objective and subjective judgments about orderings, let us turn now to one of the central findings from the present experimental research, namely that many of the adjectives tested—among them *hard/soft*, *sharp/dull*, *clean/dirty*, *rough/smooth* and *salty*—allowed both types of interpretations. The framework for adjectival semantics proposed in this section is able to account for the existence of this mixed group, and also for the difference between these adjectives and members of the purely subjective group.

Mixed behavior may first of all arise as the result of an ambiguity between measurable/objective and subjective/experiential interpretations, a possibility suggested by Kennedy (2013). This does not require us to posit a lexical ambiguity for the adjectives in question; rather, it is already allowed for by the underspecified, context-dependent template for gradable adjective meaning in (35). An explanation in terms of ambiguity is plausible in particular for adjectives whose meaning relates to perceptual dimensions such as sight, hearing and taste. A good example is *salty*, which might be interpreted objectively in terms of salt content or subjectively in terms of an experiencer's perception of a substance's taste properties. On the former interpretation *salty* aligns to adjectives of the *full* class (and thus could potentially be associated with a numerical measure), while on the latter it aligns to the judge or experiencer-dependent *tasty/beautiful* class. Correspondingly we found mixed judgments in the experiment. Other adjectives tested that might fall in this group include *light/dark* and *hard/soft*.

This is however not the only possible source of mixed behavior. Rather, the logical forms of adjectives of the multidimensional sort themselves allow objective as well as subjective interpretations. The intuition behind the formalisms in (40) and (41) is that the lexical entries of adjectives such as *clean* and *dirty* underspecify the component dimensions and their manner of combination that go into the assessment of an entity's degree of (say) cleanness or dirtiness. Ordering subjectivity arises when two speakers disagree about how these should be specified, or when a single speaker is uncertain as to how to specify them. But contexts can only vary so much: while there may be room for disagreement as to how different sorts of dirt and so forth should be weighted to arrive at an overall degree of dirtiness for an entity, a shirt that is covered with oil stains must be evaluated as dirtier than one that is clean except for a few smudges of dirt near the hem. In formal terms, for all contexts $c \in C$, the order of the degrees assigned to these two shirts relative to the dimension dirtiness remains the same. As a special case of objective judgments with this class, the present account correctly predicts that a shirt that is completely free of dirt will necessarily be assessed as cleaner/less dirty than one with some amount of contextually relevant dirt: regardless of how dirt types are weighted in a particular context, the mathematical form of the lexical entry in (41) has the consequence that the former shirt will be mapped to the 0 point on the dirtiness scale, while the latter will be assigned some positive value.[6]

Returning to the experimental results, we might hypothesize that in the case of multidimensional adjectives such as *clean/dirty*, *smooth/rough*, *dull/sharp* and perhaps others, subjects who gave FACT and OPINION judgments made different assumptions about the two entities under consideration, or about the relevant context. The first group may potentially have assumed that the entities were different to

[6] I thank an anonymous reviewer for pointing out this example.

such a degree, or the context specified to such a degree, that all available measure functions would yield the same ordering. The second group may conversely have assumed that the individuals were sufficiently close in their properties, or that the context was sufficiently underspecified, that that there were available interpretations (i.e. measure functions) that would yield different orderings. I believe this is a plausible explanation for the experimental findings, though it would benefit from further experimental investigation.

Crucially, though, for judge-dependent adjectives of the *beautiful* and *tasty* sort, objective interpretations for the comparative cannot be derived in the same way. For members of the multidimensional class, the range of possible variation in degrees assigned to entities is constrained by the possible choices for the component dimensions $DIM_1^c, DIM_2^c, \ldots, DIM_n^c$ and the function f^c; for certain pairs of entities or situations of utterance, these constraints have the effect of eliminating the possibility of variation in orderings. But for *beautiful* and the like, varying judgments about orderings derive directly from the varying perceptions and tastes of distinct agents or experiencers, which are not constrained in any formal way. Correspondingly, we predict members of this class to be interpreted purely subjectively in the comparative, and this is exactly what we found. Thus drawing a distinction between multidimensionality and judge dependence as sources of ordering subjectivity further helps to explain why some adjectives that are subjective in this sense also allow objective interpretations, while others do not.

Table 1 Classes of gradable adjectives

		Interpretation of comparative	
		Objective	Subjective
(a) Measurable		✔	
• Additive	*tall/short, expensive*		
• Externally anchored	*new/old, hot/cold*		
• Context-independent derived	*full/empty, pure/impure, fast/slow*		
(b) Multidimensional (context-dependent derived)	*clean/dirty, straight/curved*	✔	✔
(c) Judge-dependent	*tasty, fun, beautiful/ugly interesting/boring, happy/sad*		✔
Ambiguous between (b) & (c)	*bumpy/flat, sharp/dull*	✔	✔
Ambiguous between (a) & (c)	*salty, hard/soft, dark/light*	✔	✔

5.5 *Summary*

The observations from this section are summarized in Table 1, which presents a classification of gradable adjectives by the formal properties of their (families of) measure functions, and the corresponding availability of objective versus subjective readings for the comparative form. The table is populated with examples taken from the above discussion. I have not, however, attempted a full classification of all the adjectives experimentally tested, and here there are questions that could be raised; as an example, on the criteria discussed above the adjective *intelligent* would seem to be multidimensional rather than judge-dependent, but unlike others in this class it elicited purely subjective readings for its comparative form. Additional research would be beneficial to understanding if the categorization proposed here must be refined, and in further developing diagnostics to assign adjectives to the appropriate category or categories.

6 Conclusions

The starting point for this paper was the observation that the comparative forms of certain gradable adjectives are interpreted subjectively, a pattern that is problematic for standard theories of gradability. The hypothesis was explored that subjectivity of this sort derives from the multidimensional nature of the properties in question. I have attempted to make two empirical contributions in this work. The first is to demonstrate experimentally that ordering subjectivity is more widespread than previously recognized, and furthermore that adjectives with this property pattern into two groups, depending on whether or not they also allow objective readings for their comparative forms. The second is to show that multidimensionality is a complex and multifaceted phenomenon, and that not all gradable adjectives that are conceptually multidimensional should be represented as explicitly multidimensional in the semantics. From a theoretical perspective, I have argued that the facts are best captured by positing two distinct sources of ordering subjectivity, multidimensionality being one, the second being parameterization to a judge, i.e. a sentient individual whose judgments, tastes or emotions are expressed. Formally, I have developed this insight in a theory of gradability on which the availability of objective versus subjective readings of the comparatives derives from the formal properties of the measure functions lexicalized by gradable adjectives.

There are a number of important issues that I have not been able to address completely in the context of the present work. In particular, while the experiment reported here demonstrated the existence of two distinct classes of subjective adjectives, it does not provide direct evidence for the underlying distinction I have proposed, namely multidimensionality versus judge dependence. It is to be hoped that the predictions of this proposal can be tested more directly in future experimental research. Here, an issue is that it is challenging to find adequate diagnostics for multidimensionality and especially judge dependence that might serve as the basis for an experimental paradigm (cf. the discussion in Sect. 4). With regards to diagnosing the presence of an explicitly represented sentient judge, the varying acceptability of different types of follow-up questions discussed in Sect. 4.3 might provide a starting point; another potential direction involves the possibility of 'coordination by stipulation' discussed by Kennedy and Willer (2016), which plausibly is sensitive to the distinction proposed in the present paper. Alternately, a promising more indirect approach is found in the work of Kaiser and Herron Lee (2017a, b), who show that predicates of personal taste and multidimensional adjectives pattern differently in the complement position of Experiencer-Theme verbs such as *hear*, which make salient an experiencer who may be associated with the corresponding role introduced by the adjective. This area is ripe for further research.

Looking more broadly, in focusing on the comparative I have made no attempt to address the interpretation of gradable adjectives of various sorts in their positive forms. One question that merits further investigation relates to the semantically multidimensional class. As was discussed above, there is considerable evidence that the positive forms of a subclass of such adjectives involve quantification over dimensions, something that is not easily expressed in the present formalization, in which explicitly multidimensional adjectives are analyzed in terms of derived measure functions. This suggests there remains work to be done in integrating the present findings with those from previous work on multidimensionality. I have furthermore not attempted to apply the present analysis to the subjectivity of the positive form of gradable adjectives, or to relate the discussion of ordering subjectivity to the large body of insights on subjectivity more generally. Clearly there are connections here, but attempting to explore them would take us too far from the central topic of the paper. I therefore leave the broader implications of these findings for the study of subjectivity as a topic for the future.

Acknowledgements I would like to express my gratitude to the editors of this volume Elena Castroviejo-Miró, Louise McNally and Galit Sassoon, as well as three anonymous reviewers, whose insightful comments and suggestions on earlier versions have led to a much improved paper. Thanks also to Carla Umbach, Nat Hansen, Kjell Johan Sæbø and the audiences at ExpGrad, SuB20 and the ZAS for very helpful discussion, and to Nadja Reinhold and Lisa Reimann for assistance with the experimental research. All errors are of course my own. This work was supported by the German Science Foundation (DFG) under grant SO1157/1-1.

Appendix

This appendix provides the full stimuli (critical items) used in the experiment.

List 1 A: Anna’s apartment is dirtier than Paul’s.
B: No, Paul’s place is dirtier.

A: Frank is shorter than Jimmy.
B: No, Jimmy is the shorter one.

A: John and Fred look similar but John is a little taller than Fred.
B: No! Fred is the taller one of the two.

A: Lea and Marie are both sad but Marie is sadder.
B: No, Lea is sadder.

A: Lilian’s car is newer than Noemi’s car.
B: No, Noemi’s is definitely the more recent one.

A: My painting is prettier than yours.
B: No! My painting is definitely prettier than yours.

A: Susan just got out of the water, but her hair is already drier than mine.
B: No, it’s not - your hair is definitely drier than Susan’s.

A: The green towel is wetter than the red one.
B: No, the red towel is wetter.

A: This building is older than the building Julia lives in.
B: No, Julia’s building is older.

A: The mug is cleaner than the spoon.
B: The spoon is cleaner than the mug.

A: This cat is happier than that dog.
B: No, the dog is the happier one of them.

A: Those sneakers are uglier than the Converse sneakers you tried on earlier.
B: No, the Converse sneakers were uglier.

List 2 A: Can I borrow your pencil? Mine is duller than the one you have.
B: No, my pencil is even duller than yours.

A: Caryl and Tina both have blond hair, but Caryl’s is lighter than Tina’s.
B: No, Tina’s hair is definitely lighter than Caryl’s.

A: Give those kids the green ball to play with, it’s softer than the red one.
B: No, the red ball is softer.

A: I would rather use the yellow pillow—it’s harder than the white one.
B: No, the white pillow is the harder one of the two.

A: Math is easier than Geography.
B: Geography is a lot easier than Math!

A: Take the red knife, it’s sharper than the one you’re using.
B: No, the knife I have now is sharper than the red one.

A: The fence in front of Mr. Harington’s house is straighter than the one in front of Mr. Rave’s house.
B: No, Mr. Rave’s fence is straighter.

A: The old Ipod Touch 4G is more expensive than the new Ipod Touch 5G.
B: No, the new one is the more expensive device.

A: The second line on that graph is more curved than the first one.
B: No, the first line is more curved than the second.

A: The walls in the dining room are darker than the walls in the living room.
B: No, the walls in the living room are darker.

A: This small piece of paper is smoother than that big piece of paper.
B: No, the big piece is smoother.

A: This stone right in front of us is rougher than that one in the back.
B: No, the stone in the back is rougher.

List 3 A: Frank is shorter than his friend Jimmy.
B: No, Jimmy is the shorter one.

A: John and Fred look similar but John is taller than Fred.
B: No, Fred is the taller one of the two.

A: Lillian and Nicole both have the same kind of cellphone, but Lillian's is newer than Nicole's.
B: No, Nicole's phone is newer than Lillian's.

A: Look—Tommy's shirt is dirtier than the one his little brother Billy is wearing.
B: No, Blly's shirt is dirtier than Tommy's.

A: My apartment building is older than the building Julia lives in.
B: No, Julia's building is older.

A: Susan just got out of the water, but her bathing suit is already drier than mine.
B: No, it's not—your bathing suit is drier than Susan's.

A: The green towel on the hook is wetter than the blue one.
B: No, the blue towel is wetter.

A: The lecture we heard last week was more boring than today's lecture.
B: No, today's lecture was more boring.

A: The mug you just handed me is cleaner than the one on the counter.
B: No, the one on the counter is cleaner.

A: The necklace Susan is wearing today is uglier than the one she had on yesterday.
B: No, the one she was wearing yesterday was uglier.

A: The program we watched about India was more interesting than the one about Japan.
B: No, the program about Japan was the more interesting of the two.

A: The vase on the table is more beautiful than the one on the bookshelf.
B: No, the vase on the bookshelf is more beautiful.

List 4 A: I just read the essay John wrote and it is worse than Bill's.
B: No it isn't. The one Bill wrote is worse.

A: Look at Sue's new bike—it's better than Anne's.
B: No, Anne's bike is better.

A: The cream cake is tastier than the chocolate cookies.
B: No, the chocolate cookies are tastier.

A: The math professor is more intelligent than the physics professor.
B: No, I disagree. The physics professor is the more intelligent one.

A: The movie theater is emptier today than it was yesterday.
B: No it wasn't. It was emptier yesterday.

A: The roller coaster was more fun than the ferris wheel.
B: No it wasn't! The ferris wheel was more fun.

A: The vegetable soup is saltier than the chicken soup.
B: No, the chicken soup is saltier.

A: The wine bottle is fuller than the champagne bottle.
B: No, the champagne bottle is fuller.

References

Bartsch, R. (1984). The structure of word meanings: Polysemy, metaphor, metonymy. In F. Landman & F. Veltman (Eds.), *Varieties of formal semantics* (pp. 25–54). Dordrecht: Foris.

Bartsch, R. (1986). Context-dependent interpretations of lexical items. In J. Groenendijk, D. de Jongh, & M. Stokhof (Eds.), *Foundations of pragmatics and lexical semantics*, GRASS 7 (pp. 1–26). Dordrecht: Foris.

Bartsch, R., & Vennemann, T. (1972). The grammar of relative adjectives and of comparison. *Linguistische Berichte*, *20*, 19–32.

Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using 'Eigen' and S4. R package version 1.1-7. http://CRAN.R-project.org/package=lme4.

Bierwisch, M. (1989). The semantics of gradation. In M. Bierwisch & E. Lang (Eds.), *Dimensional adjectives: Grammatical structure and conceptual interpretation* (pp. 71–261). Berlin: Springer.

Bylinina, L. (2014). *The grammar of standards: Judge-dependence, purpose-relativity, and comparison classes in degree constructions*. Dissertation, Utrecht University.

Bylinina, L. (2017). Judge-dependence in degree constructions. *Journal of Semantics*, *34*(2), 291–331.

Christensen, C. M. (1980). Effects of solution viscosity on perceived saltiness and sweetness. *Perception and Psychophysics*, *28*(4), 347–353.

Cresswell, M. J. (1977). The semantics of degree. In B. H. Partee (Ed.), *Montague Grammar* (pp. 261–292). New York: Academic Press.

Hare, R. M. (1952). *The language of morals*. Oxford: Clarendon Press.

Heim, I. (2000). Degree operators and scope. In J. Brendan & T. Matthews (Eds.), *Proceedings of the 10th Semantics and Linguistic Theory Conference (SALT 10)* (pp. 40–64). Ithaca, NY: CLC Publications.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363.

Kaiser, E., & Lee, J. H. (2017a). Predicates of personal taste and multidimensional adjectives: An experimental investigation. In *Proceedings of the 35th West Coast Conference on Formal Linguistics (WCCFL35)*. Somerville, MA: Cascadilla Proceedings Project.

Kaiser, E., & Lee, J. H. (2017b). Experience matters: A psycholinguistic investigation of predicates of personal taste. In D. Burgdorf, J. Collard, S. Maspong, & B. Stefánsdóttir (Eds.), *Proceedings of Semantics and Linguistic Theory 27 (SALT 27)*. Washington, DC: Linguistic Society of America.

Kamp, H. (1975). Two theories about adjectives. In E. L. Keenan (Ed.), *Formal semantics of natural language* (pp. 121–155). Cambridge: Cambridge University Press.

Kennedy, C. (1997). *Projecting the adjective: The syntax and semantics of gradability and comparison*. Dissertation, University of California at Santa Cruz.

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

Kennedy, C. (2013). Two sources of subjectivity: Qualitative assessment and dimensional uncertainty. *Inquiry*, *56*(2–3), 258–277.

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

Kennedy, C., & Willer, M. (2016). Subjective attitudes and counterstance contingency. In M. Moroney, C.-R. Little, J. Collard, & D. Burgdorf (Eds.), *Proceedings of Semantics and Linguistic Theory 26 (SALT26)* (pp. 913–933). Washington, DC: Linguistic Society of America.

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, *4*(1), 1–45.

Kölbel, M. (2004). Faultless disagreement. In *Proceedings of the Aristotelian Society* (Vol. 104, pp. 53–73). New Series.

Krifka, M. (1989). Nominal reference, temporal constitution and quantification in event semantics. In R. Bartsch, J. van Benthem, & P. van Emde Boas (Eds.), *Semantics and contextual expression* (pp. 75–115). Dordrecht: Foris.

Landman, F. (1989). Groups II. *Linguistics and Philosophy*, *12*(6), 723–744.

Lasersohn, P. (2005). Context dependence, disagreement, and predicates of personal taste. *Linguistics and Philosophy*, *28*(6), 643–686.

Lasersohn, P. (2009). Relative truth, speaker commitment, and control of implicit arguments. *Synthese*, *166*(2), 359–374.

Lassiter, D. (2011). *Measurement and modality: The scalar basis of modal semantics*. Dissertation, New York University.

McNally, L., & Stojanovic, I. (2017). Aesthetic adjectives. In J. O. Young (Ed.), *Semantics of aesthetic judgment* (pp. 17–37). Oxford: Oxford University Press.

Moltmann, F. (2010). Relative truth and the first person. *Philosophical Studies*, *150*(2), 187–220.

Pflaum, T., Konitzer, K., Hofmann, T., & Koehler, P. (2013). Influence of texture on the perception of saltiness in wheat bread. *Journal of Agricultural and Food Chemistry*, *61*(45), 10649–10658.

R Core Team. (2015). R: A language and environment for statistical computing. *R foundation for statistical computing*, Vienna, Austria. http://www.R-project.org/.

Sæbø, K. J. (2009). Judgment ascriptions. *Linguistics and Philosophy*, *32*(4), 327–352.

Sassoon, G. W. (2007). *Vagueness, gradability and typicality, a comprehensive semantic analysis*. Dissertation, Tel Aviv University.

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

Sassoon, G. W. (2011). Adjectival vs. nominal categorization processes: The rule vs. similarity hypothesis. *Belgium Journal of Linguistics*, *25*, 104–147.

Sassoon, G. W. (2012). The double nature of negative antonymy. In A. Aguilar Guevara, A. Chernilovskaya, & R. Nouwen (Eds.), *Proceedings of Sinn und Bedeutung 16. MIT Working Papers in Linguistics* (pp. 543–556).

Sassoon, G. W. (2013). A typology of multidimensional adjectives. *Journal of Semantics*, *30*(3), 335–380.

Sassoon, G. W. (2015). *A degree-approach account of multidimensional gradability*. Unpublished manuscript, Bar-Ilan University.

Solt, S. (2016). Ordering subjectivity and the absolute/relative distinction. In N. Bade, P. Berezovskaya, & A. Schöller (Eds.), *Proceedings of Sinn und Bedeutung* (Vol. 20, pp. 676–693).

Stephenson, T. (2007). Judge dependence, epistemic modals, and predicates of personal taste. *Linguistics and Philosophy*, *30*(4), 487–525.

Stojanovic, I. (2007). Talking about taste: Disagreement, implicit arguments, and relative truth. *Linguistics and Philosophy*, *30*(6), 691–706.

Suzuki, A. H., Zhong, H., Lee, J., & Martini, S. (2014). Effect of lipid content on saltiness perception: A psychophysical study. *Journal of Sensory Studies*, *29*(6), 404–412.

Umbach, C. (2016). Evaluative propositions and subjective judgments (pp. 127–168). In C. Meier & J. van Wijnbergen-Huitink (Eds.), *Subjective meaning: Alternatives to relativism*. Berlin: de Gruyter.

Online Processing of “Real” and “Fake”: The Cost of Being Too Strong

Petra B. Schumacher, Patrick Brandt and Hanna Weiland-Breckle

Abstract Strengthening literal meanings of linguistic expressions appears central to communicative success. Weakening on the other hand would appear not to be viable given that literal meaning already grossly underdetermines reality, let alone possibility. We discuss productive weakening in *fake*-type adjectival modification and present evidence from event-related brain potentials that such weakening has neurophysiological consequences and is qualitatively different from other mechanisms of modification. Specifically, the processing of *fake*-type constructions (e.g., “a fake diamond”) evokes a Late Positivity as characteristic of certain types of referential shift or reconceptualization. We argue that *fake*-type composition involves an intermediate representation that is semantically contradictory and that the Late Positivity reflects an interface repair mechanism that redresses the contradiction. In contrast, composition involving reputedly over-informative *real*-type adjectives evokes no comparable processing costs.

Keywords Contradiction · Repair · Weakening · Inference · Privative adjectives Comprehension · ERP · Late positivity

P. B. Schumacher (✉) · H. Weiland-Breckle
Department of German Language and Literature I, University of Cologne, Cologne, Germany
e-mail: petra.schumacher@uni-koeln.de

H. Weiland-Breckle
e-mail: h.weiland-breckle@uni-koeln.de

P. Brandt
Institut für Deutsche Sprache, Mannheim, Germany
e-mail: brandt@ids-mannheim.de

E. Castroviejo et al. (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure*,
Language, Cognition, and Mind 4, https://doi.org/10.1007/978-3-319-77791-7_4

1 Introduction

It appears to be a basic trait of the relation between language on the one hand and reality and possibility on the other hand that the former immensely underdetermines the latter. As Russell (1940:87) puts it:

> Owing to the fact that words are general, the correspondence of fact and sentence which constitutes truth is many-one, i.e., the truth of the sentence leaves the fact more or less indeterminate.

Grice's (1975, 1989) theory of implicature explains how strengthening the literal meanings of linguistic expressions is decisive for actual communicative success. The opposite of strengthening, i.e., interpreting expressions in terms that are weaker or less specific than their conventional meanings, is not generally taken to be a crucial part of the machine that allows us to effectively associate forms and meanings. However, Carston (1997:106) gives rather straightforward examples where weakening ("loosening" in her terms) applies so as to arrive at a meaningful interpretation, cf. (1) and (2).

(1) a. France is hexagonal.
b. I love bald men.
c. This steak is raw.
(2) a. Have you eaten my chocolate heart?
b. Here is my new flatmate. [referring to a newly acquired cat]

The cases in (1) can be described as idealization and exaggeration respectively. Particularly pertinent to the discussion here are the cases in (2) that appear to necessitate the negation of certain properties associated with the head noun (e.g., regarding (2a), being organic and alive) or, alternatively, regarding as relevant only certain entailed properties (e.g., being shaped like a heart). In other words, what happens here is that the coded meaning is weakened in such a way that it makes sense in combination with the modifying adjective. Beyond such putatively metaphoric examples, which could possibly be relegated to rhetoric, certain superficially plain cases similarly do appear to enforce weakening, as illustrated by the example of nominal modification in (3).

(3) Fred bought a fake diamond.

Adjectives like *fake, former* or *alleged* are dimensional adjectives whose interpretation depends on context like that of many other adjectives. However, while the overwhelming majority of adjectives serves to narrow down the denotation of the noun they modify (cf. for instance the intersective adjective in (4)), *fake*-type adjectives enforce moving beyond the literal denotation of the noun they combine with and thus violate the principle that it is the head of a phrase that determines its overall meaning (cf. Kamp and Partee's (1995) pertaining discussion of their "Head Primacy Principle"). More specifically, interpreting a noun phrase involving a *fake*-type adjective involves weakening its literal meaning by way of negating (or maybe ignoring) certain properties customarily entailed by the meaning of the head noun.

For instance, a “fake diamond” will lack certain properties of diamonds while it will instantiate others. Eventually and crucially, a fake diamond is not a diamond, which is however what the basic head principle of compositional interpretation would dictate.

(4) illustrates the basic case of adjectival modification by means of intersective or subsective adjectives which stands in contrast to the case of modification by *fake*-type adjectives.[1] The combination in (4) represents the intersection of two sets where it also forms a subset with the set denoted by the head noun (cf. (5)). This contrasts with *fake*-type modifications that yield a situation where *fake diamonds* are not a subset of *diamonds* (cf. (6)).

(4) Lily sold a green dress.
(5) $\|$ green dress $\| = \|$ green $\| \cap \|$ dress $\|$ & $\|$green dress$\| \subset \|$dress$\|$.
(6) $\|$ fake diamond $\| \not\subset \|$ diamond $\|$.

The meaning of a *fake*-type construction seems to involve an extra quantification (“in some sense x is a P and in some sense x is not a P”). Building on Peirce (1910), we would like to propose that this hidden meaning is really the result of a repair that circumvents a violation of the semantically basic law of contradiction.[2] Concretely, this indicates that the derivation of the eventual meaning of the *fake*-type construction proceeds roughly as in (7), involving an intermediate representation that violates the law of contradiction stating that it is impossible that both p and not p.

(7) (a) This is a fake diamond.
(b) This is a diamond and this is not a diamond.
(c) In some sense this is a diamond and in some sense this is not a diamond.

Which particular aspects are negated depends on the context of utterance, as well as which other aspects are possibly highlighted. Moreover, certain adjectives like “alleged” or “potential” merely challenge the presence of particular dimensions (cf. Franks 1995). Note that in terms of strength, we arrive at an overall weaker meaning as a result of this repair.[3]

[1] In the literature, the term *privative* is often used for what we call *fake*-type adjectives here. As we believe the case to be more general (viz., “constructional” in the case of e.g. animal-for-statue alternations, cf. below) and as we do not intend to engage in the discussion of adjective classification (cf. Kamp and Partee 1995; Morzycki 2015), we stick to the neutral term in what follows. Lakoff and Johnson (1980:120f) propose that privative adjectives like *fake* do negate purposive and functional properties (as opposed to perceptual and motor activity properties) as making up the multidimensional Gestalt associated with *fake*-modified nominals. Partee (2010) on the other hand gives an argument based on Polish NP split phenomena that *fake*-type adjectives that can occur in predicative position should be regarded as subsective adjectives additionally entailing coercion to the effect of loosening the overall interpretation.

[2] Peirce (MS 678:34, 1910) states: “…that which characterizes and defines an assertion of possibility is its emancipation from the Principle of Contradiction”.

[3] Contradictions are the strongest possible meanings as they exclude everything. Brandt (2016) argues that the hidden aspectual-temporal, modal or comparative meaning of inchoative, middle, excessive and directional complement constructions should be regarded as the result of the same interface repair mechanism that circumvents a violation of the law of contradiction by introducing an extra quantification (or rather by semantically dislocating a problematic meaning component).

Let's consider another type of adjectives. *Real* and similar adjectives are also context-dependent. However, combinations with *real*-type adjectives result in the highlighting of particular aspects of the denotation of the head noun:

(8) Rachel bought a real diamond.

To say that something is *a real diamond* (8) we first have to know what a diamond is in order to determine what it means to carry some prototypical features of diamonds or not—in contrast to for instance saying that something is a green dress where interpretation requires the intersection of green things and things that are dresses. As Austin (1962: 70–71) put it

> the function of 'real' is not to contribute positively to the characterization of anything, but to exclude possible ways of being *not* real—and these ways are both numerous for particular kinds of things, and liable to be quite different for things of different kinds. It is this identity of general function combined with immense diversity in specific applications which gives to the word 'real' the, at first sight, baffling feature of having neither one single 'meaning', nor yet ambiguity, a number of different meanings.

To a certain extent, compositions involving *real* are then also about negating or ignoring certain dimensions but crucially they do not cause a contradiction as is the case of *fake*-type adjectives. *Real* needs a basis for comparison to single out a particular dimension which can come in different flavors—either in the form of the intrinsic properties of the head noun (8) or in the form of context-specific knowledge (9).

(9) Suzy likes big dogs with a thick fur. For her, only the chow chow is a real dog.

A *real*-type construction highlights standard or prototypical properties of the head noun in the particular context of use and thus intimately depends on the head noun for interpretation.

Note also that we do not think that the use of *real* and similar adjectives reflects mere redundancy since these adjectives are used for reasons of highlighting, and thus strengthen prototypical aspects of the concept. There is thus added value of uttering *a real diamond* versus *a diamond* (cf. Horn 1984; Grice 1989).

What *real* and *fake* have in common is that compositional processing of the constructions involving them relies on inferential reasoning which at the same time updates the context (i.e. the head noun) in fundamental ways. Processing of an NP like *a real X* depends on the identification of a prototype certain aspects of which are highlighted, but crucially not at the expense of negating certain aspects of the head noun. *A fake X* calls for the negation of the head noun's properties, which results in a momentary contradiction that needs to be repaired. In the following experiment, we investigate whether these cases of weakening and highlighting rely on the same inferential process or whether they recruit distinct mechanisms. Given the characterization of the two types of adjectives above, there might well be qualitative differences in the underlying processes. While a salient property is highlighted in the composition with *real*-type adjectives, the meaning adjustment triggered by the combination with *fake*-type adjectives gives rise to a reconceptualization of the original denotation.

The negation of certain properties of the head noun (and the consequent reconceptualization) should be more computationally demanding than highlighting a particular property.

2 Processing Considerations

To our knowledge the processing of *fake* and *real*-type adjectives has not yet been investigated. However, other combinatorial operations and processes of meaning adjustment which have been examined in the past will be summarized in this section.

Context-dependent reconceptualization has been investigated in situations similar to *the chocolate heart* from (2a) above. Take for instance the productive animal-for-statue alternation such as the use of *lion* in *the stone lion* or *dove* in *the wooden dove* (10a). The interpretation of these animal-for-statue uses - similar to the case of *fake*-type constructions - involves negating or ignoring certain properties of the head noun and thus weakening its meaning (see also Franks 1995; Kamp and Partee 1995; Morzycki 2015). At the same time, certain meaning aspects appear to be promoted, resulting in this case in a statue or artifact reading.[4] This particular composition thus shows clear commonalities with *fake*-type combinations, with the important exception that the adjectives involved do not necessarily trigger privation on the basis of some intrinsic property but can be used in a feature-negating sense in certain combinations (cf. also Franks 1995 on proper versus functional privatives).

(10) a. The wooden **dove** was on the table.
     b. The wooden **trunk** was next to the bed.

Such manipulation of meaning—involving weakening in particular—has been shown to be computationally demanding relative to a non-extending baseline where a material adjective is properly attached to a noun denoting an object likely to be made up of the particular material (10b). In particular, processing differences between (10a) and (10b) have been observed using the methodology of event-related brain potentials (ERPs). ERPs reflect the neuronal activity triggered by sensory, cognitive or motor events. They provide a multi-dimensional signal including information about the polarity and magnitude of the signal’s amplitude, temporal aspects relative to the onset of a critical stimulus and the topographical distribution recorded at the surface of the scalp (because activity is nowadays typically recorded from at least 24 electrodes placed across the participants scalp). Crucially, ERPs are relative measures between a critical condition and a minimally differing control condition. Silent reading of constructions like (10a) in a word by word manner evoked a Late Positivity time-locked to the onset of the head noun compared to (10b). This Late Positivity stands for a positive-going deflection for the animal-for-statue reading

[4] Note that nothing excludes that weakening and strengthening apply with respect to one and the same construction. For the cases at hand at least, weakening should apply first (as it makes no sense to strengthen contradictory statements).

(10a) between roughly 500–800 ms after noun onset and a maximum distribution over left posterior electrode sites and is considered to reflect costs from narrowing the animal to the statue reading (cf. Schumacher 2013).

A similar Late Positivity, albeit with a broader topographical distribution, was also registered for container-for-content alternations as exemplified by (11a), where the predicate *spill* requires a complement representing a substance and the intended meaning of the expression *bucket* is the content of the bucket. In this case, the interpretation can be strengthened in favor of the telic function of the noun, i.e. holding some kind of content z, because the semantic type of the y-argument of *spill* cannot be a physical object but must be a substance (12). This type mismatch requires a repair operation, which compared to (11b) is reflected by a Late Positive deflection in the ERP recording relative to the noun phrase (cf. Schumacher 2013).[5]

(11) a. What did Eric spill on the stairs? He spilled **the bucket** on the stairs.
b. What did Eric buy at the flee market? He bought **the bucket** at the flea market.

(12) λy λx λe. spill(e, x, y) & Eric(x) & bucket(y) & telic = λz λe’. contain(e’, y, z)

A final piece of evidence supporting the idea of compositional repair operations comes from the property-for-person alternation in (13a) best known from Nunberg’s (1979) discussion of *the ham sandwich* case. Here, a salient property is used to refer to a person. When uttered in the context of a doctor’s office or hospital, *the hepatitis* can be understood as an individual situationally associated with (the symptoms of) hepatitis. This meaning transfer to the person reading results in processing costs observable in a Late Positivity for the noun phrase in (13a) contrasted with a control like (13b) (Schumacher 2014).

(13) a. The doctor asks his assistant again who had called that early. The assistant responds that **the hepatitis** had called that early.
b. The doctor asks his assistant again who had called that early. The assistant responds that **the therapist** had called that early.

Interestingly, not all meaning alternations exert costs during comprehension. For example, content-container (14a vs. 14b) or producer-product (15a vs. 15b) alternations show no processing differences between the two intended readings (cf. Schumacher 2013; Weiland-Breckle and Schumacher 2017). This has been reported on the basis of ERP data, self-paced reading and eye-tracking during reading (e.g., Frisson 2009 for an overview, Schumacher 2013). The absence of additional processing demands indicates that certain meaning adjustments do not—or no longer—require repair. One explanation that has been advocated in Sauerland and

[5] Question-answer sequences were adopted in this study which was conducted in German verb-final constructions to assure that the mismatch between the predicate and the noun was measurable at the noun. Accordingly, the question introduced the predicate and generated a high expectation for a content-denoting expression, thereby creating a semantic type conflict as early as at the point of encountering the noun in the answer.

Schumacher (2016) is that the latter alternations represent more systematic, conventionalized alternations that have come to be more conceptually integrated and are underspecified for their potential readings, while the examples from (10)—(13) have not reached this representational level yet but rather come with a core meaning and the extended meaning must be constructed on the basis of contextual information, i.e. it must be repaired.

(14) a. She tipped over **the milk**.
b. She drank **the milk**.

(15) a. He read **Dickens**.
b. He met **Dickens**.

The literature further suggests that there are different types of repair. When the interpretation of an entity must be extended to an event, as is the case for complement coercion, processing costs are observable in reading time, eye tracking and ERP measures. The predicate *begin* in (16a) subcategorizes for a complement of the type event which requires the reader to construct an activity that can be performed with the memo, like reading or typing the memo.

(16) a. The secretary began **the memo** about the new office policy.
b. The secretary typed **the memo** about the new office policy.

Eye tracking makes it possible to record eye gaze and eye fixations during reading, which manifests discrete processing profiles for minimally differing sentences. Eye tracking measures reveal more regressions back into and longer total reading times of the critical noun region for (16a) compared to (16b), which are associated with constructing the additional event structure (e.g., Traxler et al. 2002; McElree et al. 2006). ERP recordings of complement coercion register a different signature than the repair cases discussed above: an N400, this is a negative-going waveform with a peak latency around 400 ms after the onset of the noun phrase in (16a) relative to (16b) (Baggio et al. 2010). This substantiates the claim that the underlying reanalyses are qualitatively different. Corroborating support comes from coercion at the phrasal level between an adjective and a noun since (17a) requires a type shift from mountain to e.g. climbing of the mountain and registers more enhanced regressions and re-reading times in eye tracking than the control condition in (17b) (Frisson et al. 2011).

(17) a. The athlete is convinced that the difficult **mountain** will require all his strengths.
b. The athlete is convinced that the difficult **exercise** will require all his strengths.

Overall, the processing evidence indicates that different operations that manipulate meanings are at work during composition. In certain cases, underspecified lexical representations facilitate compositional processes in such a way that no extra processing demands are observable. In cases of a type shift from entity to event, building of additional structure exerts processing costs, reflected in late eye tracking measures and an N400 effect. Pragmatic strengthening evokes a qualitatively different ERP profile with a Late Positivity. We consider this Late Positivity to reflect repair processes that lead to updating of representational structure and reconceptualization; for example, the fact that *the wooden dove* is not a dove with respect to some dimensions of a dove (but quite certainly with respect to the dimension of shape) creates a new representation void of the contextually illicit dimensions.

3 Experimental Investigation

In the following ERP study, we examine the time course and underlying processes of compositional adjective-noun processing involving *fake*- and *real*-type adjectives in comparison to combinations of intersective and subsective adjectives with nouns that do not rely on inferential processing. We predict that composition involving *fake*-type adjectives is computationally demanding because it requires negating or ignoring certain aspects of the head noun's denotation. As mentioned above, this operation that is inherent to the adjective has consequences for interpretation, yielding an output like (18b) that appears contradictory at first sight and must be repaired to one of the more narrow meanings illustrated in (18').

(18) a. This is a fake diamond.
b. This is a diamond and this is not a diamond.

(18') a. This is a diamond (in certain respects) and this is not a diamond (in other respects).
b. This sparkles like a diamond but does not cost nearly as much as a diamond.
c. This cuts like a diamond but does not look like one.
d. etc.

This repair should be reflected in a Late Positive deflection at the point when adjective and noun are combined with each other and reconceptualization takes place. In addition, since privation is an inherent property of the adjective, we might observe processing demands relative to the adjective as well in anticipation of upcoming repair processes. Note that such costs are not predicted for combinatorial processing with *real*-type adjectives, since its function is to highlight certain properties but not negate them. Furthermore, early repair processes (i.e. time-locked to the adjective) have not been predicted for the animal-for-statue ("wooden dove") alternation as there is nothing semantically problematic about the adjective as such in these constructions.

Since *fake* and *real* also differ in terms of their polarity, they will be matched to different control adjectives. *Fake* will be contrasted with adjectives that have a negative denotation and *real* with positive adjectives. In this way we want to assure that any effect observable for *fake* is not caused by negative denotation per se but rather a consequence of repair and updating. Some theories claim that negative terms are composed as negations of their positive counterparts which has syntactic and semantic ramifications (Kennedy 2001; Heim 2008; Sassoon 2013). To this end we selected positive and negative control adjectives on the basis of polarity judgments by an independent group of participants (see 3.1.2). ERP research has only looked at polarity in combination with affect and mood; Herbert et al. (2008) found for instance a more pronounced N400 for unpleasant versus pleasant adjectives. This adds to the general view that the processing of positive information is easier than that of negative information.

We had one more control condition that consisted of an anomalous adjective-noun combination. It is well-known from the ERP literature that semantic violations engender an N400 effect (for an overview see Kutas and Federmeier 2011) and the comparison of the anomalous combination with the control conditions thus served as a basic contrast to assure that participants read the stimuli for comprehension and showed the typical effect for semantic anomalies.

3.1 Methods

3.1.1 Participants

Twenty-eight right-handed, monolingual speakers of German (5 men and 23 women; mean age: 23.6 years; range: 19–30 years of age) from the University of Cologne participated in this study after giving written informed consent. All participants had normal or corrected-to-normal vision and reported no history of neurological or psychiatric disorders.

3.1.2 Materials

This study was conducted in German. As described above we used a 2 × 2 experimental design with the factors composition (critical vs. control adjectives) and polarity (positive vs. negative valency) which yielded the following four conditions illustrated in Table 1:

(19) a. Susi legt einen falschen Diamanten auf den Tisch.
Susi puts a fake diamond on the table
“Susi puts a fake diamond on the table.”

b. Susi legt einen unreinen Diamanten auf den Tisch.

Table 1 2 × 2 Experimental design

	Critical condition	Control condition
Negative polarity	*Fake*-type adjectives (19a)	Negative adjectives (19b)
Positive polarity	*Real*-type adjectives (20a)	Positive adjectives (20b)

Susi puts a flawed diamond on the table
"Susi puts a flawed diamond on the table."

(20) a. Susi legt einen echten Diamanten auf den Tisch.
Susi puts a real diamond on the table
"Susi puts a real diamond on the table."

b. Susi legt einen weißen Diamanten auf den Tisch.
Susi puts a white diamond on the table
"Susi puts a white diamond on the table."

We first carried out a polarity rating task on the adjectives used in the four conditions in which 79 participants who were not included in the ERP study (24 men, mean age: 24.8 years, range: 20–34 years) rated 150 adjectives distributed across three lists for their polarity on a 7-point scale with 1 marking negative and 7 positive polarity. This included a total of 115 subsective and intersective adjectives as well as 23 *fake*-type adjectives and 12 *real*-type adjectives that were all used in the ERP study. The mean polarity ratings and standard deviations are reported in Table 2, illustrating a clear polarity effect whereby negative and *fake*-type adjectives were evaluated as negative adjectives and positive and *real*-type adjectives as more positive. The intuition that *fake* and *real* differ along the dimension of polarity was thus born out.

Based on these ratings, we constructed 64 items per condition. Since the classes of *fake*- and *real*-type adjectives are relatively small, we had to repeat those adjectives but always changed the rest of the proposition. We used 23 *fake*-type adjectives (angeblich/mutmaßlich—"alleged", ehemalig/einstig/früher—"former", erdichtet/erfunden/fiktiv/fingiert—"ficitious", nachgeahmt/falsch—"fake", gespielt/gestellt/scheinbar/vermeintlich/vorgetäuscht—"pretend", imitiert/nachgebildet/nachgemacht—"imitated", künstlich—"artificial", plagiiert—"plagiarized", voraussichtlich/zukünftig—"prospective") and 12 *real*-type adjectives (echt/richtig/wirklich—"real", amtlich/offiziell—"official", tatsächlich/aktuell—"actual", original/ursprünglich—"original", wahr/wahrhaftig—"true", zweifelsfrei—"doubtless") Table 3 exemplifies the matching across conditions.

An additional 64 anomalous stimuli were constructed to create a semantic mismatch at the noun (21).

Table 2 Mean polarity ratings from the pretest and standard deviation in parentheses (1 is negative, 7 positive)

Adjective	Polarity rating
Fake–type adjective	3,22 (0.91)
Negative adjective	2.11 (0.97)
Real-type adjective	5.24 (0.99)
Positive adjective	5.55 (0.92)

Table 3 Sample adjectives and their polarity matched controls

Fake–type adjectives	Negative adjectives	*Real*-type adjectives	Positive adjectives	Noun
1. falsch ("fake")	unrein ("flawed")	echt ("real")	weiß ("white")	Diamant ("diamond")
2. nachgemacht ("imitated")	kaputt ("damaged")	wahr ("true")	vergoldet ("gold-plated")	Rolex ("Rolex")
3. gespielt ("pretend")	lähmend ("paralyzing")	wahrhaftig ("true")	angenehm ("pleasant")	Angst ("anxiety")
4. künstlich ("artificial")	stinkend ("evil-smelling")	echt ("real")	flauschig ("fluffy")	Pelz ("fur")
5. mutmaßlich ("alleged")	brutal ("brutal")	wirklich ("real")	jugendlicher ("youthful")	Mörder ("murderer")
6. fingiert ("fictious")	verschwunden ("lost")	amtlich ("official")	beglaubigt ("certified")	Testament ("will")

(21) Susi legt einen flüssigen Diamanten auf den Tisch.
Susi puts a liquid diamond on the table
"Susi puts a liquid diamond on the table."

The total set of these 320 items was interspersed with an additional 160 filler sentences that did not contain adjective—noun combinations. The items were pseudo-randomized and presented in six different orders across all participants. To make sure that participants attended to the stimuli, each sentence was followed by a word recognition task for which each sentence was assigned a probe word. Half of the items from every condition were meant to evoke a true response to a probe word from the sentence stimulus and the other half a false response to a probe word that had not been mentioned in the sentence.

3.1.3 Procedure

Each participant was seated in a dimly lit, sound-proof booth for the recording of the electroencephalogram (EEG). Stimuli were presented visually in the center of a computer screen with off-white letters against a black background. Each trial began with a fixation star that was displayed for 500 ms in the middle of the monitor and followed by a blank screen for 150 ms. Stimuli were presented word by word, with each word being displayed for 400 ms with an interword interval of 150 ms. After a blank screen of 500 ms, three question marks occurred for 500 ms, followed by the presentation of a probe word for the word recognition task. Participants were required to respond as quickly and accurately as possible by pressing a "yes" or "no" button on a gamepad. The assignment of the left and right response buttons was counterbalanced across participants. After the probe task, a blank screen was presented for 1000 ms before the beginning of the next trial. After giving written informed consent, participants were prepared for the ERP recording, completed a brief practice session to get acquainted with the experimental protocol and were then presented with the experimental session.

3.1.4 EEG Recording and Preprocessing

The EEG was recorded from 24 Ag/AgCl scalp electrodes placed according to the international 10–20 system and mounted in an elastic cap (*Easycap*, Munich, Germany). Electrodes were referenced to the left mastoid and re-referenced offline to linked mastoids. AFz served as ground electrode. To monitor for artifacts resulting from eye movements and blinks, two sets of electrode pairs were placed at the outer side of each eye for the horizontal electrooculogram (EOG) and above and below the participant's right eye for the vertical EOG. Electrode impedances were kept below 5 kΩ. All EEG and EOG channels were amplified with a *BrainAmp DC amplifier* (Munich, Germany) and digitized with a rate of 500 Hz.

Before averaging, the EEG data were band pass filtered offline (0.3–20 Hz) to remove unsystematic pre-stimulus differences triggered by slow signal drifts. Then automatic (±40 μV for the EOG electrodes) and manual rejections were performed to exclude trials containing ocular and amplifier saturation artifacts. Trials with incorrect answers or time-outs to the word recognition task were also excluded from the analysis. The application of all of these rejection criteria amounted to the exclusion of 6.8% of the data points. Average ERPs were investigated from the onset of the adjective until 1000 ms after the noun.

3.1.5 Data Analysis

The statistical analysis is based on the mean amplitude value per condition in predetermined time windows. A repeated measures analysis of variance (ANOVA) was performed with the factors COMPOSITION (critical vs. control adjective), POLARITY (positive vs. negative) and the topographical factor REGION OF INTEREST (ROI). The electrodes were grouped by topographical ROIs which entered the analysis with four levels: left anterior (F3, F7, FC1, FC5, T7), left posterior (C3, CP1, CP5, P3, P7), right anterior (F4, F8, FC2, FC6, T8), right posterior (C4, CP2, CP6, P4, P8). Analyses were carried out in a hierarchical order. Whenever there was an interaction with the factor COMPOSITION, pairwise comparisons were conducted for *fake*-type versus negative adjectives on the one hand and *real*-type versus positive adjectives on the other. The anomaly contrast was computed separately with the factors CONDITION (anomalous vs. negative) and ROI. Huynh–Feldt adjustment was applied when the analysis involved factors with more than one degree of freedom in the numerator. The analyses were performed using the ez-package (Lawrence 2013) in R (R Core Team 2015).

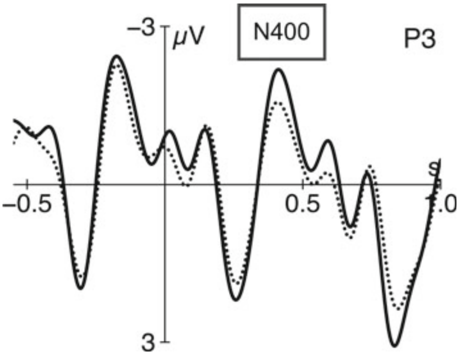


Fig. 1 Grand average ERPs at a selected electrode time-locked to the onset of the noun for the anomalous adjective-noun combinations (solid line) and the negative adjective condition (dotted line). Noun onset is at vertical bar. Negativity is plotted upwards. All figures are filtered with a 0.8 low pass filter for presentational purposes only

3.2 Results

The control contrast between anomalous adjective-noun combinations and the valency-controlling adjective-noun combinations elicited a negativity with a maximum peak around 400 ms after noun onset in the anomaly condition relative to the negative adjective condition. This effect is illustrated in Fig. 1. It represents an N400 typically observed for prediction errors when two entities like “liquid” and “diamond” cannot be combined with each other in a meaningful way. This contrast thus demonstrates that our setup evoked standard ERP signatures during compositional processing. The statistical analysis in the time window from 350–450 ms after head noun onset registered a condition x ROI interaction ($F(3, 81) = 3.01, p < 0.05$), which was reflected by significant effects of anomaly in the posterior ROIs (left posterior: $F(1, 27) = 5.01, p < 0.05$), right posterior: $F(1, 27) = 4.29, p < .05$).

Figure 2 shows the contrast between positive and negative adjectives and reflects a more pronounced negative amplitude between 300–400 ms after adjective onset (onset is at −550 ms in the figure) for the *negative versus positive adjectives.*

Figure 3 presents the comparison of *fake*-type adjectives with their negative controls. The graph begins at the onset of the adjective (−550 ms on temporal axis) and the vertical axis marks the onset of the noun (0 ms). The figure spans till 1000 ms after noun onset. Figure 3 reveals first that the negative adjectives elicited a more pronounced negativity between −250 and −150 ms (i.e. 300–400 ms after adjective onset) compared to *fake* adjectives. Subsequently, the processing of *fake* adjectives shows two positive-going deflections contrasted with the negative adjectives, the first one peaking between 150–300 ms and the second one between 600–700 ms. We propose below that the first effect is a Late Positivity triggered by the inherent instruction of the adjective (that surfaces 700–850 ms after the onset of the adjective),

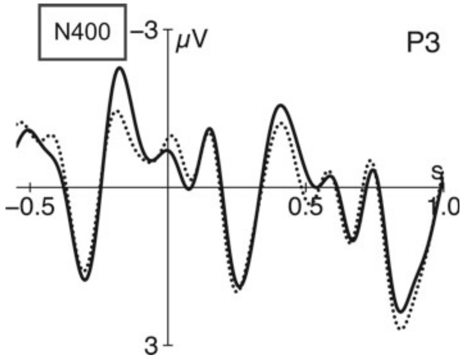


Fig. 2 Grand average ERPs at a selected electrode for the negative (solid) and positive adjective condition (dotted). Figure spans from adjective onset (−550 ms) till 1000 ms after noun onset. The onset of the noun is at the vertical bar. Negativity is plotted upwards

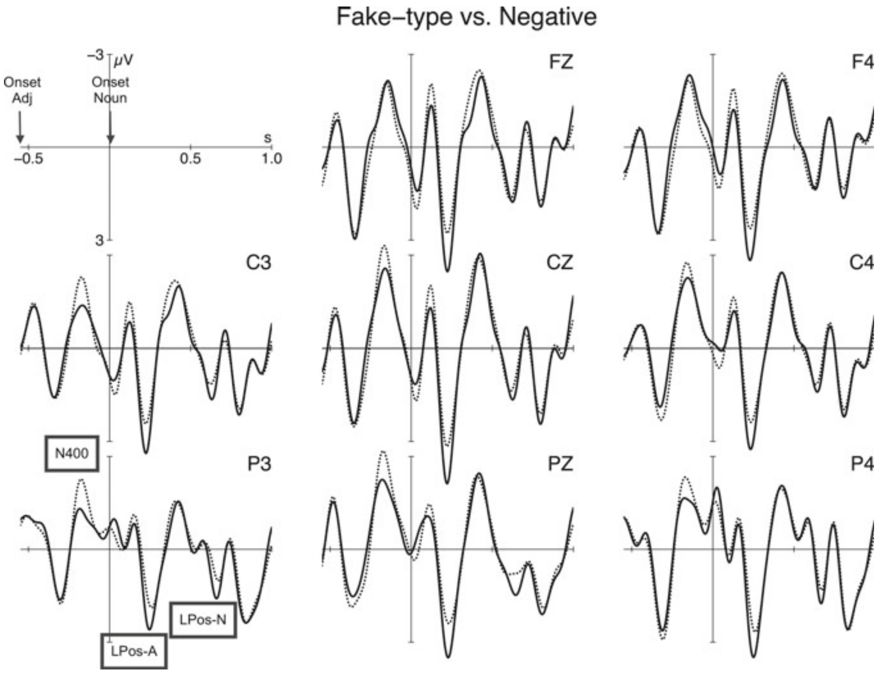


Fig. 3 Grand average ERPs at selected electrodes for the comparison of *fake*-type adjectives (solid line) with their negative controls (dotted line). Adjective onset is at −550 ms and noun onset at 0 ms (the vertical axis). Negativity is plotted up. "LPos-A" marks the positivity relative to adjective onset and "LPos-N" to head noun onset

while the second effect is a Late Positivity attributed to adjective-noun combinatorics time-locked to the onset of the head noun.

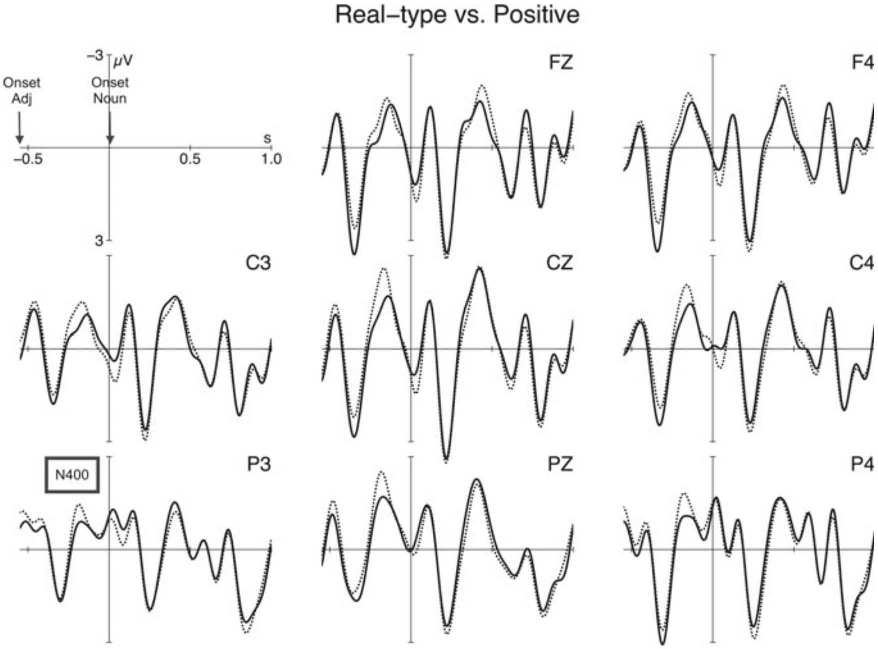


Fig. 4 Grand average ERPs at selected electrodes for the comparison of real-type adjectives (solid line) versus positive adjectives (dotted line). Adjective onset is at −550 ms and noun onset at 0 ms (the vertical axis). Negativity is plotted upwards

The comparison of *real*-type adjectives and positive adjectives is displayed in Fig. 4. It shows that these two conditions only differ in a more enhanced negativity between −250 and −150 ms (i.e. 300–400 ms after adjective onset) for the positive control compared to *real* adjectives. No later effects are observable that could be attributed to specific processing costs associated with the highlighting function of *real*-type adjectives.

These patterns were confirmed by statistical analyses. In the time window 300–400 ms after the onset of the adjective, statistical analyses by means of ANOVA revealed main effects of composition ($F(1, 27) = 8.37, p<0.008$) and polarity ($F(1, 27) = 9.74, p<0.005$) as well as interactions of composition x ROI ($F(3, 81) = 5.57$, $p<0.003$), polarity x ROI ($F(3, 81) = 3.05$, $p<0.05$) and composition x polarity x ROI ($F(3,81) = 3.64, p<0.03$). Following up on the highest interaction, the pairwise comparison between *fake versus negative adjectives* showed reliable effects in the left posterior ROI ($F(1, 27) = 7.64, p<0.02$) and the comparison *real versus positive adjectives* registered a significant difference over all ROIs (left anterior (marginal): $F(1, 27) = 4.07, p<0.06$, right anterior: $F(1, 27) = 9.15, p<0.006$, left posterior: $F(1, 27) = 7.67, p<0.02$, right posterior: $F(1, 27) = 10.22, p<0.004$).

In the temporal window between 150–300 ms after the onset of the noun (i.e. 700–850 ms after adjective onset), the analysis registered a composition x polarity interaction ($F(1, 27) = 33.82$, $p < 0.001$). Resolution of this interaction displayed a reliable effect of composition for *fake versus negative adjectives* ($F(1, 27) = 14.49$, $p < 0.001$) and none for the comparison of *real versus positive adjectives* ($F < 3.1$) in this time window.

Finally, 600–700 ms after noun onset, there was a main effect of polarity ($F(1, 27) = 6.49$, $p < 0.02$) and a composition x polarity x ROI interaction ($F(3, 81) = 3.13$, $p < 0.05$). Resolution of this interaction showed an effect of composition for *fake versus negative adjectives* over the left posterior ROI ($F(1, 27) = 4.41$, $p < 0.05$) and no reliable effects for *real versus positive adjectives* ($F < 0.4$).[6]

4 Discussion

We pursued the following research questions: What are the cognitive consequences for inferential processing during adjective-noun combinatorics? And are there differences between pragmatically highlighting certain prototypical properties of entities falling under nominals modified by *real*-type adjectives on the one hand and redressing an inherent contradiction (p and ¬ p) in the case of modification by *fake*-type adjectives on the other hand? The data showed that these two types of adjectives differ in that *fake*-type adjectives are more computational demanding than *real*-type adjectives.

While there seems to be good reason to consider strengthening a part of the compositional process, weakening is not generally taken to be operative in meaning composition, although certain classes of cases including animal-for-statue alternations and in particular *fake*-type adjectives clearly suggest the opposite. This study provided evidence that weakening in composition is neurophysiologically real as well as qualitatively different from strengthening. In particular, *fake*- as well as animal-for-statue type constructions exert more processing costs than intersective or subsective adjectival modifications, including *real*-type adjectives that might be expected to be pragmatically problematic due to being over-informative. We proposed that the computational cost involved in *fake*-type constructions reflects a repair mechanism that renders interpretable structures violating the semantically basic law of contradiction, i.e., the requirement that “not both p and not p”, cf. Brandt (2016).

In the following we first chart the time-course of adjective-noun comprehension before we take a closer look at the processing of the two critical adjective types.

In the current investigation, ERP effects were observed in three discrete time windows. The first window spans from 300–400 ms after the onset of the adjective. At this point, lexical processing of the adjectives takes place which is affected by both polarity and the type of composition. Negative adjectives show a more pronounced N400 than positive adjectives (cf. Fig. 2), which is in line with the findings of Herbert et al. (2008). Furthermore, the intersective and subsective control adjectives show a

[6] Note that for reasons of space, we did not include the separate analysis for the midline electrodes, which registered comparable results.

more enhanced N400 relative to the inference inducing adjectives (Figs. 3 and 4). This may be a result of the fact that *fake*- and *real*-type adjectives are more vague and depend on the head noun for interpretation to a much larger extent while the inter- and subsective adjectives can be interpreted in and of themselves and are thus lexically more determinate.

The second window of interest spans from 150–300 ms after the onset of the noun, which corresponds to 700–850 ms after the onset of the adjective. This positivity emerges for *fake*-type adjectives only, indicating that it is triggered by specific requirements of this class of adjectives—such as the requirement to negate certain aspects of the head noun. Since this operation is already available at the adjective, the repair may either be anticipated (thus reflecting a Late Positivity to the onset of the adjective) or be a very early reflex of combinatorial processing of the adjective-noun combination. In support of the former view, a third effect emerges between 600 and 700 ms after noun onset, which reflects another positivity. We consider this to be the point at which updating is fully completed.

Importantly, the positivities surface only in the *fake* conditions and not in the *real* conditions. *Real*-type adjectives highlight prototypical properties of the entity denoted by the head noun. They require the instantiation of a prototype or contrast set (cf. Austin 1962). Pragmatic theory predicts inferential processing to set up a comparison class and arrive at the added value interpretation. The ERP data suggest that no processing effort accrues for the respective kind of inferencing. Selecting a dimension from the set of dimensions of the head noun for the purpose of singling it out comes for free. In contrast, explicitly negating certain dimensions during composition with *fake*-type adjectives is costly. We propose that extra processing effort is needed due to the contradiction inherent to this particular composition.

The data thus indicate that repair processes triggered by violations of the law of contradiction evoke a Late Positive potential and not an N400. The N400 has generally been associated with prediction-based processing showing a more pronounced amplitude when a prediction error occurs (cf. Kutas and Federmeier 2011). This may well happen during adjective-noun combination, as indicated by our control contrast of the anomalous adjective-noun combination (“liquid diamond”) with an acceptable combination (“flawed diamond”). The semantically illicit combination evoked an N400 after the onset of the head noun. Crucially, the repair induced by the inherent negation in *fake*-noun combinations does not result in such a prediction error (it is the adjective’s inherent instruction to the processor that a contradiction of the sort “p and not p” is to be expected). Rather, the combination evokes a different ERP effect, namely a Late Positivity.

Such a Late Positivity has also been observed during the processing of other types of meaning adjustment that require negation of certain dimensions of the respective noun, including animal-for-statue (*wooden dove*), container-for-content (*spilling the bucket*) and property-for-person (*the hepatitis called*) readings that involve referential updating or reconceptualization. In these cases, the repair is induced by a feature mismatch between the predicate and the noun, i.e. the conflict arises compositionally. In contrast, *fake*-type adjectives inherently carry the conflicting potential. This may explain why an earlier Positivity is observed that is time-locked to the *fake* adjective.

Such early repair processes are not predicted for the animal-for-statue alternation as there is nothing semantically problematic about the adjective as such in these combinations. The presence of the early Positivity can thus be accounted for on the basis of the contrast between inherently private adjectives like *fake* on the one hand and combinatorially emerging privation (*wooden turtle*) on the other hand (cf. Franks 1995).

We consider the Late Positivity an instance of referential updating or reconceptualization by which the new concept (*fake diamond*) qua combination requires modification of the representation of the head noun. This definitely involves rebutting certain dimensions associated with the combined constituents. Additionally, it may involve the emergence of novel dimensions that are specific to the newly constructed concept (cf. Hampton 1987). On the basis of the current experimental design, we cannot distinguish between these two accounts and future research will have to investigate this issue further.

With regard to *real*-type adjectives, the data also suggest that the combination of this type of adjective with a noun should not be analyzed as an instance of redundancy or over-informativeness. Previous work by Engelhardt et al. (2011) registered N400 effects for over-informative prenominal modifiers. Such an effect was absent in our study supporting the view that added value is ascribed to the modification with *real*-type adjectives.

In sum, we have provided evidence that some inferences during adjective-noun combinations are computationally costly. The highlighting of a particular dimension of the head noun is effortless compared to the modification with inter- or subsective adjectives, while negating certain aspects of the head noun in order to resolve an apparent contradiction engenders a Late Positivity. The underlying repair results in an updated discourse representation of the denotation of the head noun, which mirrors processes observed in other cases of context-dependent meaning adjustment.

Acknowledgements We would like to thank Claudia Kilter, Fabienne Bartsch, Filiz Özden and Johanna Kasper for their assistance during data collection. This research was supported by the DFG-funded XPrag.de project of the first author (SCHU 2517/6-1).

References

Austin, J. L. (1962). *Sense and sensibilia*. Oxford: Oxford University Press.

Baggio, G., Choma, T., van Lambalgen, M., & Hagoort, P. (2010). Coercion and compositionality. *Journal of Cognitive Neuroscience, 22*(9), 2131–2140.

Brandt, P. (2016). Fehlkonstruktion und Reparatur in der Bedeutungskomposition. *Linguistische Berichte, 248,* 395–433.

Carston, R. (1997). Enrichment and loosening: Complementary processes in deriving the proposition expressed. *Linguistische Berichte, 8,* 103–127.

Engelhardt, P. E., Demiral, B., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition, 77*(2), 304–314.

Franks, B. (1995). Sense generation: A "quasi-classical" approach to concepts and concept combination. *Cognitive Science, 19,* 441–505.

Frisson, S. (2009). Semantic underspecification in language processing. *Language and Linguistics Compass, 3*(1), 111–127.

Frisson, S., Pickering, M. J., & McElree, B. (2011). The difficult mountain: Enriched composition in adjective-noun phrases. *Psychonomic Bulletin & Review, 18*(6), 1172–1179.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge: Harvard University Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts* (pp. 41–58). New York: Academic Press.

Hampton, J. A. (1987). Inheritance of attributes in natural concept conjunctions. *Memory and Cognition, 15,* 55–71.

Heim, I. (2008). Decomposing antonyms. *Proceedings of Sinn und Bedeutung, 12,* 212–225.

Herbert, C., Junghofer, M., & Kissler, J. (2008). Event related potentials to emotional adjectives during reading. *Psychophysiology, 45*(3), 487–498.

Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature in context. In D. Schiffrin (Ed.), *Meaning, form and use in context: Linguistic applications* (pp. 11–42). Washington: Georgetown University Press.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition, 57*(2), 129–191.

Kennedy, C. (2001). Polar opposition and the ontology of ‘degrees’. *Linguistics and Philosophy, 24*(1), 33–70.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62,* 621–647.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Lawrence, M. A. (2013). ez: Easy Analysis and Visualization of Factorial Experiments. R Package Version 4.2-2. http://CRAN.R-project.org/package=ez.

McElree, B., Frisson, S., & Pickering, M. J. (2006). Deferred interpretations: Why starting Dickens is taxing but reading Dickens isn’t. *Cognitive Science, 30*(1), 181–192.

Morzycki, M. (2015). *Modification*. Cambridge: Cambridge University Press.

Nunberg, G. (1979). The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy*, *3*(2), 143–184.

Partee, B. (2010). Privative adjectives: Subsective plus coercion. In R. Bäuerle, U. Reyle, & T. E. Zimmermann (Eds.), *Presuppositions and discourse* (pp. 273–285). Amsterdam: Elsevier.

Peirce, C. S. (1910). *The Charles S. Peirce papers* (Microfilm ed.). Cambridge: Harvard University.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Russell, B. (1940). *An inquiry into meaning and truth*. London: Allen & Unwin.

Sassoon, G. W. (2013). A typology of multidimensional adjectives. *Journal of Semantics, 30*(3), 335–380.

Sauerland, U., & Schumacher, P. B. (2016). Pragmatics: Theory and experiment growing together. *Linguistische Berichte, 245,* 3–24.

Schumacher, P. B. (2014). Content and context in incremental processing: “The ham sandwich” revisited. *Philosophical Studies, 168*(1), 151–165.

Schumacher, P. B. (2013). When combinatorial processing results in reconceptualization: Towards a new approach of compositionality. *Frontiers in Psychology*, *4*(677).

Traxler, M. J., Pickering, M. J., & McElree, B. (2002). Coercion in sentence processing: Evidence from eye-movements and self-paced reading. *Journal of Memory and Language, 47*(4), 530–547.

Weiland-Breckle, H., & Schumacher, P. B. (2017). Artist-for-work metonymy: Type clash or underspecification? *Mental Lexicon, 12*(2), 219–233.

Gradable Nouns as Concepts Without Prototypes

Hanna de Vries

Abstract Kamp and Partee's (Cognition, 57:129–191, 1995) typology of concepts combines features from a classical theory of concepts with features from Prototype Theory. They argue for the existence of a class of concepts that have graded membership but lack prototypes; a crucial characteristic of such concepts (like TALL) is that they involve properties without an upper bound (there is no limit to how tall something can be, for example). In this paper, I explore the links between Kamp and Partee's typology and the linguistic domain of nominal gradability. I claim that the class of nouns that are linguistically gradable (i.e., denote a predicate with a degree argument, as diagnosed by various monotonicity-based tests) corresponds precisely to those concepts that have graded membership but lack a prototype. Based on several experiments, I show that the concepts expressed by such nouns are primarily associated with unbounded properties, while the concepts expressed by non-gradable nouns are associated with bounded or all-or-nothing properties. For example, while participants strongly associate both *student* and *nerd* with intelligence, they judge that nerdiness increases with intelligence (with no upper limit) while qualifying as a typical student simply requires *some* above-standard degree of intelligence.

Keywords Gradable nouns · Prototype theory · Typicality · Familiarity
Multidimensionality · Scale structure

1 Introduction

For anyone who has ever taken an undergraduate course on conceptual semantics, the following story should sound familiar. Once upon a time, people held to a Lockean, 'classical' view of concepts, according to which the meaning of words like *bachelor* or *bird* can be broken down into a list of necessary and sufficient criteria ('an unmarried man', or 'a feathered, winged creature that lays eggs') and concept membership was a simple, black-and-white matter of meeting all these criteria. Then, in

H. de Vries (✉)
Department of Language and Linguistic Science, University of York, York, UK
e-mail: devr.hanna@gmail.com

E. Castroviejo et al. (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure*,
Language, Cognition, and Mind 4, https://doi.org/10.1007/978-3-319-77791-7_5

the 1970s, Eleanor Rosch and her colleagues proved this view wrong with a series of psychological experiments, so that nowadays we all believe that concepts are *prototypes*—abstract embodiments of the quintessential bachelor or the most 'bird-like' bird—and concept membership is graded, determined by the degree of resemblance to such prototypes.

But of course, things quickly turn out to be more complicated than this. Membership of concepts like ODD NUMBER is still decided on the basis of definitional criteria—the number 12 is clearly non-odd, despite its high similarity to odd numbers like 11 and 21. Neither is possible to maintain a clear boundary between concepts with prototype structures and concepts with definitional structures: membership of a concept like GRANDMOTHER, for example, is clearly a black-and-white matter, but at the same time we do possess a notion of the 'quintessential grandmother'—say, a grey-haired, twinkly-eyed lady who spends her days knitting, petting cats and handing out homemade biscuits. In addition, people like Armstrong et al. (1983) have argued that graded typicality judgements do not necessarily reflect the underlying conceptual structure at all: in a series of studies, they show that even concepts like EVEN NUMBER yield graded typicality judgements, despite the fact that all their participants agreed that no even number could be 'more' or 'less' even than any other even number. Finally, Osherson and Smith (1981) note that reducing concepts to typicality structures predicts the wrong meanings for complex concepts: assuming that the meaning of PET FISH has to be compositionally derived from the meanings of PET and FISH, the classical approach simply defines a *pet fish* as an entity that meets both the criteria for PET and the criteria for FISH, while the prototype approach wrongly predicts that membership of the concept PET FISH should be determined by the degree of resemblance to something like a furry, cuddly herring (see also Kamp and Partee 1995; Fodor and Lepore 1996; Storms et al. 1999; Hampton and Jonsson 2012, and many others). In short, it seems a full theory of conceptual cognition needs elements of both Prototype Theory and the classical theory, as well as an account of the way membership judgements are influenced by factors unrelated to prototypicality (e.g. Malt and Smith 1982; Armstrong et al. 1983; Barsalou 1987; Kamp and Partee 1995; Prinz 2012).

One of the most influential papers offering a concrete hybrid model of the classical and prototype theories of conceptual cognition is Kamp and Partee (1995). Kamp and Partee assume that not all concepts are associated with prototypes; also, those that do are not reduced to their prototypes, but may combine a typicality structure with 'classical', definitional membership criteria.

In this paper, I explore the Kamp and Partee typology of concepts from a linguistic perspective. By empirically investigating the conceptual semantics of 'gradable nouns' like *idiot, nerd, genius* or *sports fan*, I will provide support for several of Kamp and Partee's claims: (1) not all concepts that have graded membership have a prototype, and vice versa; (2) prototypicality is defined in terms of what I will call 'maximal embodiment' of a concept, and if such maximal embodiment is not possible (because some of the properties that are central to the concept can have potentially infinite values) the concept will lack a prototype. Furthermore, I will claim that the class of nouns that are linguistically gradable (which I will define as 'denoting a predicate with a degree argument', as evidenced by their ability to appear in constructions like (1)) correspond precisely to those concepts that have graded membership but lack a prototype.

(1) a. Sally is a huge sports fan.
b. James is an incredible nerd.

The paper is structured as follows. In Sect. 2, I will introduce a (slightly modified version of) Kamp and Partee's typology of concepts. I will present some linguistic evidence that this typology has a direct correspondent in the semantics of the corresponding natural language expressions, in particular the well-known distinction between *absolute* and *relative* gradable adjectives. I hypothesise that relative adjectives like *tall* and *old* correspond to gradable concepts without prototypes, while absolute adjectives like *open* and *clean* correspond to gradable concepts with prototypes. Furthermore, I claim that nouns like those in (1) fall into the former conceptual class.

In Sect. 3 I move on to the meat of this paper: an experimental investigation of the conceptual semantics of gradable nouns, intended to test the hypothesis that their corresponding concepts lack prototypes. I present evidence that what appear to be prototypicality effects with these nouns (such as Albert Einstein representing the quintessential GENIUS) are correlated with both familiarity and emotional attitude, and propose that they are better analysed in terms of 'accessibility', where certain instances are seen as representative of a concept not because they embody the associated properties to a higher degree, but because they are more salient for some reason. (In other words, while we can straightforwardly predict membership ratings for instances of BIRD based on these instances' resemblance to e.g. a robin or blackbird, we cannot predict membership ratings for geniuses based on their degree of similarity to Einstein.) I also present evidence that the properties that are most strongly associated with concepts like NERD and GENIUS are inherently 'limitless' properties, in the sense that it is always possible to come up with an instance of the concept that embodies these properties to an even higher degree; following Kamp and Partee's reasoning, this points towards the absence of a prototype.

Finally, Sect. 4 will reflect on the implications of our experimental findings for the various questions I started out this paper with, and on the Kamp/Partee typology itself; I will also discuss some remaining issues and point out some directions for future research.

2 Kamp and Partee's Typology of Concepts

In a well-known paper, Kamp and Partee (1995) provide a tentative typology of concepts that is based on three semi-independent criteria: (i) whether a concept is vague or sharp ([±V]), (ii) whether it does or does not have a prototype ([±P]), and (iii) whether its extension is or is not determined by prototypicality (±PE]). The following table, adapted from Kamp and Partee, shows examples for each category. 'Sharp' defines properties, like *odd* or *even*, whose meaning can be precisely defined:

a number is either odd or even, and no number can be more odd or even than another one. In contrast, ‘vague’ properties are characterised by a ‘grey area’ of borderline cases; a property like *tall* is vague because it is impossible to define an exact boundary between things that are tall and things that are not.

	−P	+P	
		−PE	**+PE**
+V	tall, not red	adolescent	red, chair
−V	inanimate, odd number	grandmother, bird	n/a

The distinction between [±P] and [±PE] concepts is Kamp and Partee’s answer to some of the questions raised in the introduction to this paper. They combine aspects of both the classical theory of concepts and prototype theory by assuming that concepts *have* prototypes but are not reduced to them: so, concepts can have both a prototype structure and ‘classical’, definitional membership criteria. [−P] concepts do not have prototypes at all (more on the reasoning behind this later). For those concepts that have a prototype ([+P]), resemblance to this prototype usually determines concept membership, but not necessarily so. For natural kinds like BIRD and other concepts like ADOLESCENT, concept membership is essentially definitional: one is an adolescent if and only if one is neither young enough to be a child not old enough to be an adult, and one is a bird if and only if one has bird DNA. Nevertheless, instances of ADOLESCENT and BIRD can easily be rated according to their resemblance to an abstract prototype, whose properties might have little in common with the definitional criteria that define membership (for example, self-centeredness and an overpowering smell of Axe body spray as typical properties for ADOLESCENT).

Kamp and Partee also assume that the prototype of a concept C is defined as the abstract entity which is most C-like, i.e. most embodies the dimensions of C; I will call this the *maximal embodiment* interpretation of prototypicality, and examine it in a bit more detail in Sect. 3.2. Kamp and Partee argue that it follows from this that TALL cannot have a prototype: tallness has no upper bound, so there is no entity (not even an abstract one) which is *most tall*.[1] The other reason that concepts like TALL do not have a prototype in Kamp and Partee’s typology is that “it can be applied to an indefinite variety of things”—what counts as tall in a human is not the same as what counts as tall in a skyscraper.

[1] Peter Gärdenfors (e.g. Gärdenfors 2004) argues that prototypicality is represented as centrality in a conceptual space. Given that a conceptual space represents similarity, the center of such a space has the property of bearing simultaneously the most similarity to all other points in the space, and the least similarity to all points outside it (cf. Rosch and Mervis 1975 for experimental results that suggest the same). But in the case of a limitless property (like tallness), there is no such point—for any given instantiation of TALL, for example, there is always another one that bears even less resemblance to the contrasting concept of SHORT (put plainly: something or someone may always be taller). So the conclusion that TALL cannot have a prototype follows also in Gärdenfors’s approach.

2.1 Vagueness and Gradability

In the following sections, I will examine the relation between the different concept classes proposed by Kamp and Partee on the one hand, and several linguistic properties of the corresponding natural language expressions on the other. In the present section, I will look at linguistic gradability; in the next section, at the difference between nominal and adjectival concepts.

An (adjectival) predicate like *tall* is gradable if it can hold of an object to a greater or lesser degree; linguistically, this is reflected by their ability to appear in degree constructions like the following:

$$\text{(2) John is} \left\{\begin{array}{l}\text{6 feet tall.}\\ \text{very tall.}\\ \text{taller than Bill.}\end{array}\right\}$$

In order to account for the semantics of this predicate, an influential approach (e.g. Cresswell 1976; von Stechow 1984; Kennedy 1997; Heim 2000, and many others) assumes that they denote relations between entities and degrees (‘degree predicates’), as follows:

(3) $\mathbf{tall}_{d,et} = \lambda d \lambda x[\mathbf{tall}(d)(x)]$

Degree constructions like the one in (2) involve a comparison between the degree to which a given entity possesses a certain property and some other degree d'; the value of d' is provided by expressions like *six feet* or *than Bill*, or by a so-called *standard degree*. For example, the sentence *John is tall* is usually analysed in terms of a covert comparison between John’s height and a standard degree s_{tall}, which is contextually determined and roughly represents the average height of John’s peer group.

(4) $[\![\textit{John is tall}]\!] = \exists d[\mathbf{tall}(d)(j) \wedge d > s_{tall}]$

Another class of gradable predicates involves standard degrees that are fixed rather than contextually determined (Cruse 1980; Kennedy and McNally 2005). This is the class of *absolute* predicates (in contrast to predicates like *tall*, which are called *relative*): predicates like *open*, *empty*, *safe* and *transparent*. Relative and absolute adjectives show distinct linguistic behaviour, for example in the type of modification they are compatible with: relative adjectives do not allow endpoint modifiers like *completely*, while intensifiers like *very* or *incredibly* are marginal with many absolute adjectives (Bolinger 1972):

(5) a. Sally is incredibly/*barely/*completely/*maximally tall.
b. The window is *incredibly/barely/completely/maximally open.

A common approach to the relative/absolute distinction assumes that the former lack both a minimal and a maximal value,[2] while the latter have either or both. If an

[2] The idea that relative adjectives have no minimal value may seem strange if we consider adjectives like *tall*, *old*, *cheap* or *fast*; after all, we have clear mathematical notions of zero height, zero age or zero speed. However, zero does not appear to be part of the *linguistic* set of degrees corresponding

adjective has minimal or maximal values, its associates standard degree coincides with one or the other (Kennedy and McNally 2005; see also Cruse 1980; Rotstein and Winter 2004; Kennedy 2007). For example, the statement *the glass is empty* is true if and only if the glass is maximally empty; in contrast, the statement *the door is open* is true if and only if the door has a non-zero degree of openness (Yoon 1996). If an adjective has neither a maximal nor a minimal value, its standard degree has to be determined by context.

All in all, we can distinguish four classes of adjectives in this way:

(6) a. Relative adjectives: neither a minimal nor a maximal value (*tall*, *intelligent*, *old*)
b. Absolute adjectives with both a minimal and a maximal value (*open*, *closed*, *empty*, *transparent*)
c. Absolute adjectives with a minimal but no maximal value (*dirty*, *dangerous*, *sick*)
d. Absolute adjectives with a maximal but no minimal value (*clean*, *safe*, *healthy*)

(Note that, for each member of the classes (a) and (b), its antonym is a member of the same class, whereas the antonyms of the adjectives in (c) are members of class (d) and vice versa.)

The Kennedy/McNally approach to the relative/absolute distinction allows us to established our first link between the semantics of gradable expressions and the Kamp/Partee typology. Recall the two reasons for Kamp and Partee to propose that concepts like TALL do not have a prototype: there is no limit to tallness, and its interpretation is context-dependent. In the above approach, these two properties are tied together: all and only all completely unbounded properties have a standard that is determined by context.

The properties of relative adjectives like *tall*, *wide* and *intelligent* clearly place them in the [+V−P] class. How do the concepts denoted by absolute adjectives like *open* and *empty* fit into Kamp and Partee's typology?

First, for those adjectives (like *open*, *empty* and *clean*) that have maximal values, it makes sense to assume that their corresponding concepts have prototypes. Neither of the reasons Kamp and Partee give for TALL's lack of a prototype apply to a concept like OPEN: it can be maximally embodied (once an aperture of some sort reaches a certain level of openness, it simply cannot get any more open), and its meaning does not vary with context the way *tall* does: regardless whether we are talking of doors, mouths, or boxes, *open* retains the same core meaning. A seven-foot human counts

to these adjectives: for example, I find statements like *# My house is extremely slow* or *# I am going faster than that house* to be quite anomalous. Put differently, having a speed, height, or age of absolutely zero is the same as having no speed, height or age at all (Lehrer 1985). But for any speed, height or age greater than zero—even for a tiny value—we will always be able to imagine someone or something who is, for example, only half that age or height or moves even slower. In other words, the linguistic scale of 'tallness' asymptotically approaches zero (so to speak) and, hence, is usually considered unbounded on both sides.

as tall, a seven-foot giraffe does not; but doors, mouths and boxes that feature an inch-wide aperture all count as open.

On the other hand, at least part of Kamp and Partee's reasoning about TALL does extend to the concepts (like DIRTY) denoted by adjectives that have minimal but no maximal value: like height, dirtiness is an unbounded property, and cannot be maximally embodied. I will therefore assume that, like relative adjectives, absolute adjectives that lack a maximal degree correspond to concept without prototypes.

Whether absolute predicates are sharp or vague is controversial (e.g. Lewis 1979; Kennedy 2007; van Rooij 2011; Sassoon 2012; Lassiter and Goodman 2013; Burnett 2014). On the one hand, as suggested above, the boundary between things that are *P* and things that are *not P* seems quite clearly defined with absolute adjectives: even a slightly open door counts as *open*, not as *closed*. On the other hand, there are many examples in the literature of presumably vague uses of absolute adjectives: for example, while the sentence *The restaurant is full tonight* could mean that the restaurant is serving the maximal number of guests, it also has a salient interpretation that can be paraphrased as something like "The restaurant is fuller than expected", where the standard is supplied by context (e.g. the usual number of guests for this restaurants on this particular night of the week) (Rotstein and Winter 2004). So, depending on which side of the debate we pick, we may classify absolute adjectives with maximal values as either [−V+P−PE], on a par with concepts like GRANDMOTHER and BIRD, or as [+V+P+PE], on a par with concepts like RED and FURNITURE.

Is OPEN more like GRANDMOTHER or more like RED? An important parallel between OPEN and RED is that there is a close link between their prototypes and their membership criteria: the prototype of RED captures what it means to be red, and the prototype of OPEN captures what it means to be open. In contrast, as we have seen, the prototypes of GRANDMOTHER and BIRD are independent of what being a grandmother or a bird means: being a grandmother just means being a woman who has grandchildren, not having grey hair and twinkly eyes and being fond of knitting.

I propose the following way out in order to avoid making a commitment on the vagueness of absolute adjectives, which is not really my concern here. Instead of distinguishing concepts on the basis of vagueness ([±V]), I will distinguish them on the basis of (conceptual) gradability: a concept is gradable ([+G]) if it expresses a property that can hold of an object to a greater or lesser degree. This includes all [+V] concepts like TALL, ADOLESCENT, RED and FURNITURE, and also includes OPEN and EMPTY, but it excludes [−V] concepts like ODD NUMBER and GRANDMOTHER.

	−P	+P	
		−PE	**+PE**
+G	tall, not red, dirty	adolescent	red, chair, open, clean
−G	inanimate, odd number	grandmother, bird	n/a

To sum up: the fact that all the examples Kamp and Partee give of [+V−P] concepts are relative adjectives is not a coincidence. If we define a prototype of a concept

C in terms of maximal embodiment of C, it follows that only naturally bounded concepts can have a prototype: for unbounded concepts, no instantiation could ever be *most* C. On the linguistic side, we see this conceptual (un)boundedness reflected in the structure of the degree scale associated with the adjective corresponding to the concept.

2.2 Gradable Nouns

In the literature on gradability and degree, the usual examples are adjectives (and adverbs)—however, it has often been observed that other lexical categories may be gradable as well (e.g. Bolinger 1972; Abney 1987; Doetjes 1997; Kennedy and McNally 2005; Sassoon 2007; Morzycki 2009; de Vries 2010). For example, there exists a class of nouns that behaves exactly like the relative adjectives (modulo independent syntactic differences)—for example, they can be modified by precisely those adjectives that, in adverbial form, modify degree in adjectival predicates:

(7) a. Bill is enormously stupid.
b. Bill is an enormous idiot.
(8) a. Sally is incredibly nerdy.
b. Sally is an incredible nerd.

Moreover, it can be shown that these nouns denote degree predicates by exploiting the observation that such predicates are monotone in the following sense (Heim 2000):

(9) **Monotonicity of degree predicates**
A function f of type $\langle d, et \rangle$ is **monotone** iff
$\forall x \forall d \forall d' [f(d)(x) = 1 \wedge d' < d \rightarrow f(d')(x) = 1]$

This monotonicity is detectable in several ways; for example, it is responsible for the following contrast (Katz 2005; Nouwen 2009):

(10) a. John is surprisingly/unexpectedly/incredibly tall. (degree reading)
b. John is unsurprisingly/expectedly/credibly tall. (no degree reading)

The reasoning here is as follows. As demonstrated by the entailment pattern in (11), evaluative modifiers like *surprising* are downward monotone: they reverse the entailment relations in their scope.

(11) a. Mary read a romance novel ⇒ Mary read a book.
b. It's surprising for Mary to read a romance novel ⇐ It's surprising for Mary to read a book.

So, when a modifier like *surprisingly* is applied to a degree predicate, the direction of monotonicity of the degree predicate is reversed: [[*surprisingly tall*]] is a predicate for which it holds that $\forall x \forall d \forall d' [\mathbf{surprising}(\mathbf{tall}(d')(x)) = 1 \rightarrow \mathbf{surprising}(\mathbf{tall}(d)(x)) \wedge d' < d = 1]$. In words: if it is surprising for John to be tall to a degree d, him being tall to any higher degree $d' > d$ would also be surprising. With a non-downward monotone modifier like *unsurprisingly*, there is no such entailment reversal: if it is unsurprising for John to be tall to a degree d, him being tall to any lower degree $d' < d$ would also be unsurprising. But in fact, by the definition in (9), John *is* tall to all lower degrees d'. This makes being *unsurprisingly tall* is a trivial property: everyone with a height, no matter how small, is *unsurprisingly tall*. Nouwen (2009) assumes that such trivially true interpretations are ruled out for pragmatic reasons; hence, the only available interpretation for a sentence like (10a) can be paraphrased as "It is unsurprising/expected/credible that John is tall".

Morzycki-nouns show the same pattern when modified by an evaluative adjective, which can be accounted for straightforwardly when we assume that these nouns, too, denote monotone degree predicates:

(12) a. Bill is an unbelievable/extraordinary/indescribable idiot. (degree reading)
b. Bill is a believable/ordinary/describable idiot. (no degree reading)

In a way very similar to Nouwen's approach above, Morzycki (2009) and de Vries (2010) account for the contrast between (8) and the following parallel sentences, which cannot be used to express the proposition that Bill's degree of idiocy is very low:

(13) a. #Bill is diminutively stupid.[3] (no degree reading)
b. Bill is a diminutive idiot. (no degree reading)

Given this strong evidence for the claim that nouns like *idiot* and *nerd* (other examples are *fan, psychopath, airhead, goat cheese aficionado, simpleton, loser* and *weirdo*) are linguistically gradable in exactly the sense that adjectives like *tall* are, we may wonder which aspect of their conceptual semantics makes them so. Morzycki speculates that these nouns are gradable because they 'identify a single scale': what makes a *weirdo* a weirdo is just their degree of weirdness, which means that different instances of WEIRDO can easily be ordered on such a scale.

However, while it may well be true that linguistic gradability relies on a single integrated degree scale, the link between this degree scale and a predicate's conceptual properties does not seem as transparent as Morzycki suggests. For instance, there are nouns like *nerd* that are clearly gradable (as (13b) shows), but are conceptually multidimensional. Nerdiness is a cocktail of many different properties and does not correspond to any straightforward ranking (John has taught himself Ancient Greek and suffers from acne. Mary is painfully shy and loves to code. Who is the bigger nerd?). Conversely, what makes a *circle* a circle is just circularity—which provides an extremely transparent way of ranking, for example, a series of drawings along a scale of circle-ness—yet the noun *circle* is not gradable in Morzycki's linguistic sense. Finally, it is unclear why the distinction between one- versus many-dimensional concepts should play a role in the availability of linguistic gradability in the first place, as many multidimensional adjectives are linguistically gradable (e.g. *intelligent* or *healthy*).[4]

[3] I am hashing this sentence because I am not quite sure it is interpretable at all.

[4] (Multi)dimensionality does seem to play a role in nominal gradability in a different way: in a series of papers, Galit Weidman Sassoon argues that multidimensional predicates fall into different

On the other hand, however, the fact that *circle* does not admit degree readings of size and evaluative adjectives may be exactly what we should expect given that CIRCLE is a [+P] concept that can be maximally embodied. The same holds for maximal-value absolute adjectives (e.g. **The door is enormously/unbelievably open*). Perhaps nouns like *circle* correspond to absolute adjectives, being gradable in a similar sense. This is the view espoused by Sassoon (2007): in Sassoon's view, nouns like *circle* or *chair* are gradable because their referents can be more or less typical examples of the concept the noun denotes. This is supported by the grammaticality of sentences like the following:

(14) a. This shape is more an oval than a circle.
b. This object is almost/barely a chair.

In fact, the classes of nouns with which Morzycki and Sassoon concern themselves more or less seem to be in complementary distribution in the same way that relative and absolute adjectives are:

(15) a. Bill is a huge/incredible idiot. (degree reading)
b. This object is a huge/incredible chair. (no degree reading)
(16) a. *?Bill is nearly/barely an idiot.
b. A stool is nearly/barely a chair.[5]

Thus, a possible synthesis of Morzycki's and Sassoon's approaches to nominal gradability, and a connection to our adjectival data, suggests itself: like absolute adjectives, 'Sassoon nouns' are compatible with endpoint modifiers but do not allow size and evaluative degree modification, and like relative adjectives, 'Morzycki nouns' are incompatible with endpoint modifiers but do allow degree readings of size and evaluative modifiers. This suggests that the former class corresponds to [+G+P] concepts (note that this is exactly how Kamp and Partee classify CHAIR), and the latter to [+G−P] ones.

In the remainder of this paper, I will not concern myself much with the conceptual properties of the Sassoon nouns, because Sassoon herself has written on this extensively. Instead, I will focus on empirically defending the claim that Morzycki-nouns denote [+G−P] concepts.

classes depending on the way the values of its dimensions are integrated, and that this classification correlates with a predicate's ability to appear in various degree constructions (e.g. Sassoon 2016; Sassoon and Fadlon 2017).

[5] I am inclined to think that this particular example, as well as (14a), are cases of metalinguistic gradability (cf. McCawley 1988's 'metalinguistic comparison': *Your problems are more financial than legal*), but in later papers, Sassoon has presented more quantificational evidence that at least 'social' nouns (human roles such as *journalist* and artifacts such as *chair*) are relatively OK in various real degree constructions (Sassoon and Fadlon 2017). See also Sect. 4 for more on the possible connections between Sassoon's findings and the approach presented in this paper.

3 Gradable Nouns as [−P] Concepts

I propose that the class of Morzycki-nouns can be defined in prototype-theoretic terms as exactly those nouns that denote [+G−P] concepts: categories whose membership is gradable, but which lack a prototype. For nouns that are directly derived from [+G−P]-denoting adjectives, this does not actually seem a bold proposal. Dutch, for example, has a relatively productive operation of adding the suffix *-erd* to property-denoting adjectives in order to form a noun meaning 'a person with this property': thus, *slim* 'clever' becomes *slimmerd*, *dik* 'fat' becomes *dikkerd*, *gemeen* 'mean' becomes *gemenerd*, and so on. These nouns, like the adjectives, are all linguistically gradable, so there is no reason to assume that they do not inherit their conceptual structure from the adjective. They exemplify Morzycki's aforementioned definition of a 'gradable noun': a noun with a single measurable dimension. But as I already noted, the class of gradable nouns also includes nouns that seem more complex nouns like *nerd, genius*, and perhaps *idiot*, whose interpretation depends on multiple dimensions. Take *nerd*, which is probably the best example. Whether someone can be called a nerd depends on many things: their IQ, computer skills, certain aspects of their look, their knowledge of obscure science fiction movies, and so on. Some of these dimensions may themselves have prototypes. How can we know whether NERD, as a whole, has one?

First, note that for every putative NERD-prototype we come up with, we can, in principle, imagine someone who is even more nerdy: someone whose IQ is ever so slightly higher, whose eyesight is just the tiniest bit worse, whose programming skills extend to just one more programming language. This seems to indicate that the concept denoted by *nerd* cannot be maximally embodied, and hence has no prototype.

Now, consider the concept GENIUS. There are many dimensions one can associate with genius—high IQ, natural talent for a particular art or science, excellence at a very young age, representing a turning point in the history of their field, lasting relevance—and no matter which genius we pick, we can come up with someone who embodies these dimensions more. However, there are certain people we associate with the quintessential genius: 9 out of 10 people will probably mention Albert Einstein when asked to list examples of geniuses, and the name is likely to be on top of most of those lists as well. We can answer the question "Was Albert Einstein a genius?" affirmatively without the least bit of hesitation. All these are measures that are highly correlated with prototypicality (Laurence and Margolis 1999). So does this mean that GENIUS has a prototype?

I propose the ready availability of Albert Einstein as the quintessential genius is related to the difference between prototypicality—the degree to which a particular instance resembles an abstract prototype—and accessibility: the ease with which a particular instance is retrieved from memory (cf. Ashcraft 1978). This difference has been demonstrated perhaps most strikingly in the previously mentioned series of studies by Armstrong et al. (1983). When asked to rate different even numbers for exemplariness of well-defined categories like EVEN NUMBER, subjects gave graded responses (favouring 2 and 4 over 34 and 806, for example), even though they all agreed that no even number could be ‘more’ or ‘less’ even than any other even number. Armstrong et al. speculate that such choices reflect the identification heuristics we employ when categorising various kinds of objects: for example, when deciding whether a natural number is even, most people will check whether it is divisible by 2, a check that is probably quicker to perform for some numbers than for others. So, graded responses do not necessarily tell us anything about the underlying conceptual structure of a category: they may merely tell us that some instances are faster and easier to categorise than others, with resemblance to a prototype only one of many possible heuristics. This is what I will refer to as the *accessibility* of an instance.[6] I speculate that anything that makes a given instance salient facilitates its accessibility; in the case of human nouns like GENIUS this might be factors like: familiarity with someone’s life and/or work, the frequency with which this person is mentioned in the media, or a strong emotional opinion on someone. Thus, Albert Einstein counts as the ‘quintessential’ genius not because he embodies the concept more than anyone else (although he obviously embodies it to a large degree), but because he is extremely well-known to the extent that his name has become synonymous with ‘an intelligent person’. There is no sense in which genius is measured in terms of ‘resemblance to Einstein’ similar to, for example, the way ‘bird-ness’ is measured in terms of resemblance to a prototypical bird.

In short, I expect that ‘typicality’ ratings for GENIUS will correlate strongly with (semi-)independent factors like familiarity and emotional attitude. On the other hand, familiarity has been shown not to affect typicality judgements in concepts with clear prototypes such as BIRD (Malt and Smith 1982); I expect the same will hold for emotional attitude.

In the remainder of this section, I describe 3 different experiments. **Experiment 1** uses a classic typicality rating task (Rosch 1973) in order to elicit typicality ratings for various instances of BIRD and GENIUS. The same participants also rated the same instances on two dimensions that should be independent from typicality: familiarity and emotional attitude. The experiment was designed to substantiate the above claim that graded membership ratings for the concept GENIUS do not primarily reflect degrees of resemblance to an Einstein-like prototype, but are heavily influenced by an instance’s accessibility.

3.1 Experiment 1

3.1.1 Method

Participants

An online questionnaire (made with Google Forms) was sent out to a total of 19 participants, recruited from my personal Facebook friends, who participated for a

[6] Armstrong et al. use the term ‘exemplariness’, but since this might be misconstrued as referring to Exemplar Theory—a competitor to Prototype Theory—I will use a more neutral term.

chance to win one of three five-euro gift certificates. All participants were speakers of Dutch and almost all were college or university graduates.

Procedure

The questionnaire consisted of four subparts (the fourth was part of experiment 2 and will be discussed in the next section). Part 2 was a classic category membership rating task, with instructions inspired by Rosch's seminal typicality rating experiment (Rosch 1973). In this task, participants were asked to rate how well they felt a certain individual represented a given category on a scale from 1 to 7, with 1 representing a perfect example and 7 a really bad one. Part 1 and 3 used the same seven-point scale rating system, but in part 1 subjects were asked to indicate their familiarity with several instantiations of a given concept, while in part 3 they indicated their emotional attitude towards these instantiations.

The questionnaire as a whole was prefaced by a general introduction, explaining that I was interested in questions of categorisation such as "Is an avocado a fruit?" and "Is a skateboard a vehicle?", and the various bases on which people make such decisions. Each individual subtask was prefaced by instructions specific to the task; these instructions were repeated at the top of each page of that subtask, so participants were able to review the instructions at each page of the questionnaire. The instructions (translated from Dutch and slightly shortened) can be found in Table 1.

Stimuli

In each of the 3 parts of the questionnaire, the same two concept names were presented (BIRD and GENIUS), each with 10 instantiations. Instantiations were presented in the form of a picture accompanied by the name and a short description; examples of these descriptions can be found in Table 2. The 10 instances of GENIUS, chosen to represent a wide range of fields, achievements, time periods and nationalities, as well as both genders, were: Leonardo da Vinci, Emily Dickinson, Dmitri Mendeleev, Steve Jobs, Hildegard of Bingen, Marie Skłodowska Curie, Rabindranath Tagore, Albert Einstein, Michael Jackson and Björk. The 10 instances of BIRD, also chosen in order to present a highly diverse list, were: blackbird, ostrich, rock bunting, Egyptian goose, gannet, kiwi, green honeycreeper, emperor penguin, kingfisher and little grebe.

3.1.2 Results and Discussion

Table 3 shows the mean typicality, familiarity and attitude ratings for all instances of GENIUS and BIRD (ordered from high to low typicality).

Following the observations and speculations made by Ashcraft (1978), Armstrong et al. (1983) and Malt and Smith (1982), and my own intuitions about the accessibility factors that might influence people's tendency to identify certain people as 'prototypical' geniuses, I predict that apparent typicality judgements for concepts denoted by gradable nouns (GENIUS, in this case) will be strongly correlated with typicality-independent measures of accessibility (familiarity and emotional attitude, in this case), while typicality ratings for BIRD will be independent from familiarity and attitude: they reflect 'degree of bird-ness' and nothing else.

Table 1 Instructions for experiment 1

Task 1: familiarity ratings
Your task is to rate each bird and each genius on a scale from 1 to 7, indicating your familiarity with the genius or bird species in question. Give an 1 to birds or geniuses that you know very well; give a 7 to birds or geniuses that are completely new to you; and give suitable intermediate ratings to the ones in between.
(At the top of the page containing the GENIUS instances) Study the pictured/described geniuses. Do you recognise both the name and the face of the genius, are you aware of their field and the way they contributed to it, do you know the important biographical details? Then give the genius a 1. Have you never heard of this person before? Give them a 7. If the genius looks/sounds familiar to you but you don't know a lot of details, give them a 4, and so on. Follow your intuition: there's no need to keep mental lists of all the facts you know and your knowledge will not be tested in any way.
(At the top of the page containing the BIRD instances) Study the pictured/described birds. Do you recognise both the name and appearance of the bird, have you often seen it on pictures or in real life, do you know its characteristics and habits (like habitat and food)? Then give the bird a 1. Have you never seen the bird (pictured) before and does its name mean nothing to you, give it a 7. If the bird looks/sounds familiar to you but you don't know a lot of details, give it a 4, and so on. Follow your intuition: there's no need to keep mental lists of all the facts you know and your knowledge will not be tested in any way.
Task 2: typicality
Your task, again, is to rate each pictured individual on a scale from 1 to 7, but in contrast to the previous part it's not about your knowledge of the bird or genius in question, but about the extent to which you feel the pictured individual is a good and representative example of the category in question. If the pictured individual is it a perfectly representative example of what you think it means to be a bird or a genius, give him/her/it an 1. If the pictured individual is a very bad example of the concept—an animal that hardly resembles a bird, a person who you feel isn't much of a genius at all—give him/her/it a 7. A moderately representative bird or genius gets a 4, and so on. Don't think about it too much, just follow your intuition.
(Repeated—just mentioning 'genius' or 'bird', respectively - at the top of both the page containing the GENIUS instances, and the page containing the BIRD instances)
Task 3: emotional attitude
Study the pictured birds and geniuses and indicate to what extent your feelings about them are positive or negative. Does a particular bird or a particular genius make you super happy, give him/her/it a 1. If you feel it's a very unpleasant individual, give him/her/it a 7. Is your emotion neutral, give him/her/it a 4, and so on. It doesn't matter whether you can justify or explain your feelings; just follow your intuition.
(Repeated—just mentioning 'genius' or 'bird', respectively—at the top of both the page containing the GENIUS instances, and the page containing the BIRD instances)

Table 2 Example of the kind of descriptions used in Task 1, 2 and 3 of Experiment 1

ROCK BUNTING: this songbird measures 16 cm and occurs in Asia, Northern Africa and Southern Europe. It builds its nest on or near the ground and feeds on insects and birds.
RABINDRANATH TAGORE (1861–1942): Bengali poet, composer, writer, painter, independence and peace activist and educational reformer. Published his first volume of poetry when he was 16 and his first opera when he was 20. Became the first non-Western winner of the Nobel Prize for literature in 1913 and founded his own university with the prize money. Composed the national anthems of both India and Bangladesh.

Table 3 Mean typicality, familiarity and attitude ratings for all instances of BIRD and GENIUS, ordered from high to low typicality

	Typicality	Familiarity	Attitude
Einstein	1.4	1.5	2.3
Da Vinci	1.5	1.7	2.1
Curie	2.8	3.2	2.5
Mendeleev	3.2	5.6	3.3
Jobs	3.7	1.8	4.4
Jackson	3.9	1.5	3.8
Tagore	4.5	6.6	3.7
Bingen	4.5	5.5	3.5
Björk	4.7	2.9	3.3
Dickinson	4.8	4.5	3.0
Blackbird	1.4	1.8	2.9
Rock bunting	1.8	6.2	2.5
Kingfisher	1.9	2.8	1.8
Honeycreeper	2.1	6.6	2.0
Gannet	2.8	4.6	3.4
Grebe	3.1	6.3	4.0
Eg. goose	3.2	4.8	4.0
Ostrich	5.3	1.5	3.3
Kiwi	5.4	3.1	3.2
Em. penguin	5.7	2.3	1.7

In order to test this prediction, I obtained Spearman correlation values comparing all responses on task 1 to all responses on task 2, and similarly for task 3 and task 2, for both GENIUS and BIRD. Furthermore (following a suggestion from Galit Weidman Sassoon, p.c.) I calculated individual correlation values for each participant and then checked whether the obtained correlations differed significantly between the BIRD and GENIUS tasks by means of a paired-samples t-test.

The results shown that the prediction is borne out. There are moderately strong and highly significant ($\alpha = 0.01$) correlations between 'typicality' and familiarity and especially 'typicality' and attitude for GENIUS, while the correlations for BIRD are negligible and/or insignificant (Table 4). Furthermore, as Table 5 shows, these differences in correlation—when obtained for each individual participant—are also highly significant.

Of course these results only apply to the GENIUS-BIRD comparison, so they are not enough to support any conclusions about the difference between Morzycki-concepts and uncontroversially [+P] concepts in general. What I have hoped to show is that various measures that have been experimentally connected to prototypicality—such as fast categorisation speed, a high chance of being mentioned by many people

Table 4 Spearman's ρ correlation values between typicality (task 2), familiarity (task 1) and attitude (task 3) ratings. Starred values are significant with $p < 0.01$; the non-starred correlation is significant with $p < 0.05$. N = 188

	Typicality and familiarity	Typicality and attitude
BIRD	−0.1858	0.19724*
GENIUS	0.353*	0.58593*

Table 5 The result of applying a paired-samples t-test to individual participants' Spearman correlation values. N = 19

	Mean	Mean		Standard	
	BIRD	GENIUS	Paired t	Error	p
Typicality-familiarity	−0.2647	0.3900	7.5552	0.087	<0.0001
Typicality-attitude	0.1447	0.5777	3.8397	0.113	0.0012

when asked to list instances of a concept, and a high chance to appear at the top of such lists—may also be the result of other accessibility-facilitating factors, so finding such patterns does not necessarily reflect the prototype structure of a given concept.

3.2 Experiment 2

Experiments 2 and 3 were designed to test the hypothesis that the dimensions most prominently associated with gradable noun concepts involve unbounded gradable properties, while the dimensions most prominently associated with non-gradable noun concepts are either non-gradable or bounded. The motivation behind this is twofold.

First, the degree to which an individual embodies a certain complex concept is a function of the number of associated properties that individual has, but also of the degree to which it embodies these properties. Whether or not a concept has a prototype—whether it is possible to maximally embody that concept—therefore depends on whether it is possible to maximally embody the properties of which this concept is composed. For associated dimensions that involve non-gradable, all-or-nothing properties (such as having eyes or being over 18 years old), maximal embodiment simply means satisfying that property. For dimensions involving specific degrees of gradable properties (such as having light blue eyes or weighing 30 kilos), maximal embodiment means having the ideal value. But dimensions that involve an *unbounded* degree of some gradable property (such as being tall, having high blood pressure, or loving opera) cannot be maximally embodied, since it is always possible to imagine some individual who is just a tiny bit taller, has a slightly higher blood pressure, or an even greater fondness of opera. From this it follows that any concept that is composed of such unbounded dimensions cannot be maximally embodied. The following hypothesis seems intuitive to me: the more central the role of unbounded dimensions in determining concept membership, the more central the resulting lack of maximal embodiment will be to our understanding of that concept, and the more likely it is to affect the concept's linguistic behaviour. Given our current assumptions, this means that we expect membership of Morzycki-concepts to be determined largely on the basis of unbounded dimensions, while membership of Sassoon-concepts will be determined largely on the basis of non-gradable dimensions or gradable dimensions with an ideal value.

However, this argument hinges on the implicit assumption that *realistic* degrees do not play any role in determining concept membership. This assumption seems somewhat at odds with Rosch and Mervis (1975)'s claim that our mental representation of a concept (and its prototype) is essentially based on family resemblance calculations performed on actual encountered instantiations of that concept. This might lead us to expect that all concept dimensions are naturally bounded by the limits of reality. For example, the heavier a person is the more they embody the concept FAT, which seems to signify that FAT is unbounded, but there are still limits to how heavy a human being can realistically be. Perhaps these natural limits serve to impose a bound on the concept's dimensions, such that it is possible after all to maximally embody a concept like FAT. In other words, while a person who weighs 800 kilos is certainly fatter than a person who weighs only 250 kilos, the latter might be a more prototypical embodiment of the concept FAT because it has more in common with the actual instances of FAT on the basis of which a prototype for the concept might be calculated. If this is the case, our reasoning behind the claim that certain concepts lack a prototype, and consequently our attempt to link this property to linguistic gradability, collapses. Before we can conclude that membership of gradable noun concepts is determined on the basis of unbounded dimensions, we therefore need to show that unbounded dimensions play a role in our conceptual cognition in the first place, regardless of considerations of realism.[7]

In this section, I describe two experiments (the second one a follow-up on the results of the first) designed to test the hypothesis that Morzycki-nouns are mainly associated with unbounded dimensions, while Sassoon-nouns are mainly associated with bounded ones. Experiment 2 consisted of a dimension-naming task, as in Rosch and Mervis (1975). Rosch and Mervis (1975) observe that the dimensions that are most often listed by people are also the most accurate predictors of prototypicality; the results of Experiment 2 should give us a fairly good impression of the dimensions that are most prominently associated with a concept, and hence of the properties

[7] This question is reminiscent of the one explored in Barsalou (1985). Barsalou notes that the prototype for goal-oriented concepts like DIET FOOD seems to involve non-realistic values: the best example of a DIET FOOD is a product that contains zero calories, despite the fact that zero-calorie food does not exist in real life. Thus, a concept's prototype can have properties that none of its instances have. Barsalou's ideal-based approach to prototypes isn't fully applicable here because I am discussing [−P] concepts, but at least it shows that conceptual structure is not fully determined by the properties of actually encountered instances.

that play the largest role in determining concept membership. Subsequently, Experiment 3 was designed to test whether a new set of participants preferred bounded, unbounded or non-graded ‘all-or-nothing’ interpretations of the dimensions elicited in Experiment 2.

3.2.1 Method

Participants
Experiment 2 was part of the same Google Forms questionnaire that also included Experiment 1, and served to break up the various subtasks of Experiment 1. It had the same 19 participants as Experiment 1.
Procedure
Participants were presented with five nominal concept names at a time (once in between Task 1 and Task 2, once in between Task 2 and Task 3 of the questionnaire), 10 concept names in total. They were asked to take approximately 30 seconds per category name and make a list of properties they associated with members of this category. Each part was preceded by identical instructions, which were closely modeled on the ones used by Rosch and Mervis (1975; their Experiment 1); a translated and slightly abridged version can be found in Table 6.
Stimuli
Five Morzycki-nouns and five Sassoon-nouns were used in Experiment 2, with the classification based on my own intuitions about their compatibility with evaluative

Table 6 Instructions for experiment 2

In this part, you will be shown 5 words for various categories. Your task is to list, for each category, what you think are characteristic properties for a member of this category. Here is an example for the category DOG‘:
Characteristics of a dog: • it barks • it has a long tail • it wags its tail • it reaches approximately to my knees • it makes a very good pet • it can do tricks • it hates cats
This task is NOT about free association: perhaps dogs remind you of your eccentric Great Aunt Margot or of that one time the neighbours’ dog ate your homework, but don’t write ‘Great Aunt Margot’ or ‘homework’. It’s about characteristic properties a typical dog possesses according to you. It does not matter whether you can think of exceptions—not all dogs hate cats, but it’s still a property that I strongly associate with a typical dog and that’s why I included it in the list above.
Guidelines: 3 to 8 characteristics for each category, or what you can think of in approximately 30 seconds. Just write down what naturally occurs to you, don’t force anything—it doesn’t matter whether you write down a lot or a little.

and size modification. The nouns were included in randomised order, but not individually randomised for each participant.

(17) **Sassoon**: BIRD, CAR SALESMAN, VEGETABLE, ITEM OF FURNITURE, STUDENT (all single words in Dutch).

(18) **Morzycki**: NERD, HIPSTER, FRATBOY, ARSEHOLE, NEUROTIC.

Table 7 Results of experiment 2: the five most commonly mentioned dimensions for each concept

BIRD	CAR SALESMAN	VEGETABLE
Has feathers	Smartly dressed	Healthy
Has wings	Smooth talker	Green
Flies	Untrustworthy	Grows on the ground
Lays eggs	Friendly, jovial behaviour	Cooked before eating
Has a beak	Knowledgeable about cars	Eaten for dinner[a]
FURNITURE	STUDENT	
Practical use	Attends college/university	
Used inside the house	Young (early twenties)	
Made of wood	Active social life	
Has legs	Intelligent	
For sitting or putting things on	Lives in a student room/dorm	
NERD	FRATBOY	HIPSTER
Good with computers	Drinks lots of beer	Alternative/eccentric style
Wears glasses	Posh accent[b]	Has a beard
Intelligent	Wears a suit jacket and tie	Wears glasses
Socially awkward	Loud	Obsessed with special, unique and ‘pure’ foods/beer/coffee
Peculiar interests/obsessions	Arrogant	Loves obscure bands
ARSEHOLE	NEUROTIC	
Mean, unsympathetic personality	Nervous, agitated behaviour	
Mean, unsympathetic behaviour	Very detail-oriented	
Is a man	Worries a lot	
Selfish	Has tics and compulsions	
Doesn’t care about others	Insecure	

[a] The Dutch have their daily hot meal at dinnertime

[b] NB: The Dutch word used by participants (*bekakt*) has a significantly more negative connotation than English *posh*—perhaps the best comparison is English *la-di-da*

3.2.2 Results

The results of Experiment 2 were tallied, and the five most commonly mentioned dimensions were identified for each concept. The threshold of 5 dimensions was chosen because most less-commonly mentioned dimensions were often mentioned just once or twice and/or seemed a result of free association (e.g. ‘comedy’ for NERD, ‘San Francisco’ for HIPSTER or ‘free as a bird’ for BIRD). On average, participants listed the following number of dimensions for each concept: ITEM OF FURNITURE 3.2, CAR SALESMAN 3.4, STUDENT 3.5, VEGETABLE 3.8, BIRD 4.9; NEUROTIC 2.6, ARSEHOLE 3.1, HIPSTER 3.5, NERD 4.4, FRAT BOY 4.9.

There were several ties, which I resolved by collapsing some very similar properties. For example, I collapsed several HIPSTER properties like ‘obsessed with ‘pure’ and ‘authentic’ food’, ‘coffee geek’ and ‘loves rare beers’ into one umbrella dimension ‘preoccupied with special, unique and ‘pure’ foods/beer/coffee’, and collapsed ‘perfectionist’, ‘control freak’ and ‘obsessed with tiny details’ into ‘very detail-oriented’ in the case of NEUROTIC.

The five most commonly named dimensions for each concept can be found in Table 7.

3.3 *Experiment 3*

For the motivation behind Experiment 3, see Sect. 3.2.

3.3.1 Method

Participants

Experiment 3 was sent out as a Google Forms questionnaire to 23 people (also Dutch, recruited from my personal Facebook friends, and participating for a chance to win a gift certificate), approximately a week after Experiment 1/2. The two groups of participants overlapped to some extent: between 9 and 15 participants in Experiment 3 had previously filled out the questionnaire that included Experiments 1 and 2.[8]

Procedure

Participants were presented with a list of 50 items (divided into 5 blocks, each on a new page), each based on one of the concept dimensions listed in Table 7. Each of these dimensions was presented in three different ways, and participants were asked to choose the formulation they felt captured the membership criterion most accurately. The questionnaire as a whole was prefaced by a short introduction identical to the introduction to Experiment 1/2, explaining that I was interested in questions

[8] Participants could provide their e-mail addresses in order to compete for the gift certificates; there was an overlap of 9 addresses between the first and second questionnaires, in addition to 6 participants (in total) who did not provide an address.

Table 8 Instructions for experiment 3

You will be shown 50 membership criteria for various concepts, for example "A typical dog weighs around 25–30 kg" or "A typical vehicle has wheels". Each of these criteria is formulated in several different ways. For instance:
A typical vehicle: (a) has wheels (b) has 4 wheels (c) has an even number of wheels (d) has wheels; the more the better
Your task is to choose the formulation that, according to you, best represents how you would apply this particular criterion when you are judging whether something or someone is a member of a given category or not. In the above case, you could choose (a) if you think that having wheels as such is important, not so much the particular number of wheels; (b) if you think a vehicle with 4 wheels (like a car) is clearly a 'better', more typical vehicle than a vehicle with some other number of wheels (like a bicycle or a train), and (c) if you feel that vehicles with an even number of wheels (bicycle, car, train) are more typical than vehicles with an odd number of wheels (unicycle, tricycle, rickshaw). Finally, you could choose option (d) if you think that vehicles with lots of wheels are inherently more typical than vehicles with few wheels. NB: 'better' here means "a better example of the category VEHICLE"—it has nothing to do with a positive value judgement.
If your favourite option isn't listed, choose the one that most resembles it.
The given options are different for each question, but always speak for themselves. It is important to follow your intuition when choosing an answer. There are no right or wrong answers and it doesn't matter one bit whether your answers are rational or consistent.

of categorisation such as "Is an avocado a fruit?" and "Is a skateboard a vehicle?", and the various bases on which people make such decisions. This introduction was followed by an instruction, a translation of which can be found in Table 8. A short summary of the instructions (without the examples) was repeated at the top of each of the 5 pages of the task.

Stimuli

Each of the 50 dimensions obtained in Experiment 2 was presented in three different ways, and participants were asked to choose the formulation they felt captured the membership criterion most accurately. Thus, for example, one dimension associated with NERD would be the wearing of glasses, which was then formulated in three different ways (the labels 'all-or-nothing', 'bounded' and 'unbounded' have been added here for clarity and were not part of the original items):

(19) A real, typical nerd:

a. **All-or-nothing**: wears glasses.
b. **Bounded**: wears strong glasses, but not ridiculously strong ones.
c. **Unbounded**: wears glasses; the stronger the better.

For a complete list of items, see Appendix B.

According to the 'unbounded' option, a higher score on a particular dimension also means a higher concept membership score, regardless of whether this is realistic.

According to the ‘bounded’ option, typical instantiations of a certain concept possess a certain quantifiable property, but only up to some extent. According to the ‘all-or-nothing’ option; the degree to which an instantiation possesses a certain property is not important, and the only relevant distinction is between individuals that have the property and individuals that lack it.

In formulating the bounded options, I deliberately used modifiers like ‘excessively’, ‘ridiculously’ and ‘unrealistically’ in order to present these boundaries as natural and realistic; I surmised that if people picked the unbounded option regardless of its implied excessiveness or ridiculousness, this would indicate that considerations of realism did not play a role in determining concept membership along the dimension in question.

All items followed the pattern exemplified above: a description of the form *A real, typical X:* followed by the 3 options, always in the same order (all-or-nothing, bounded, unbounded). The 50 items of the questionnaire (10 concepts x 5 dimensions) were presented in 5 blocks, each containing one dimension for each concept, in random order.

3.3.2 Results and Discussion

The predictions of the present approach were borne out. For the Sassoon nouns, participants overwhelmingly favoured the bounded and all-or-nothing options; but for the Morzycki nouns, they picked the ‘unbounded’ options significantly more often. The graph in Fig. 1 shows the total proportion of all-or-nothing, bounded and unbounded picks for each of the 10 concepts. To make sure the differences were not purely due to one or two different dimensions, I assigned a final label to each of the 50 dimensions based on the type that was picked most often (thus, a dimension that has more ‘all-or-nothing’ responses than either of the other two types was labelled ‘all-or-nothing’, and so on). The results become even more pronounced, as shown in Fig. 2: the only Sassoon noun with a dimension that was judged unbounded by a majority of participants was *car salesman* (the dimension in question was ‘is a smooth talker’), while all Morzycki nouns had at least one unbounded dimension (and three of them at least 3 out of 5). Finally, I made sure that the differences between Sassoon and Morzycki nouns held within individual participants by comparing the number of all-or-nothing, bounded and unbounded responses for each participant across the two noun categories by means of a paired-samples t-test. The results can be seen in Table 9. As the table shows, all three dimension types showed a significant difference: people picked the all-or-nothing and bounded options significantly more often for the non-gradable nouns, and the unbounded options significantly more often for the gradable nouns.

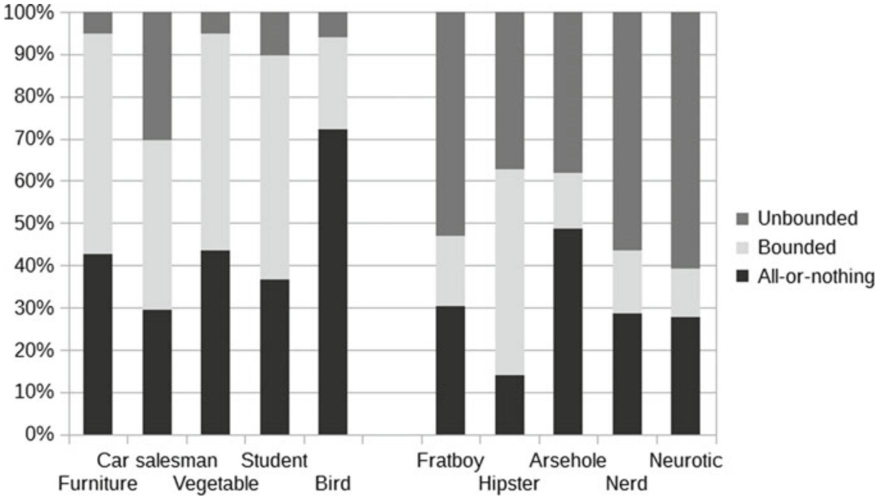


Fig. 1 Total proportion of picks for each of the three dimension types. The Sassoon nouns are on the left, the Morzycki nouns on the right

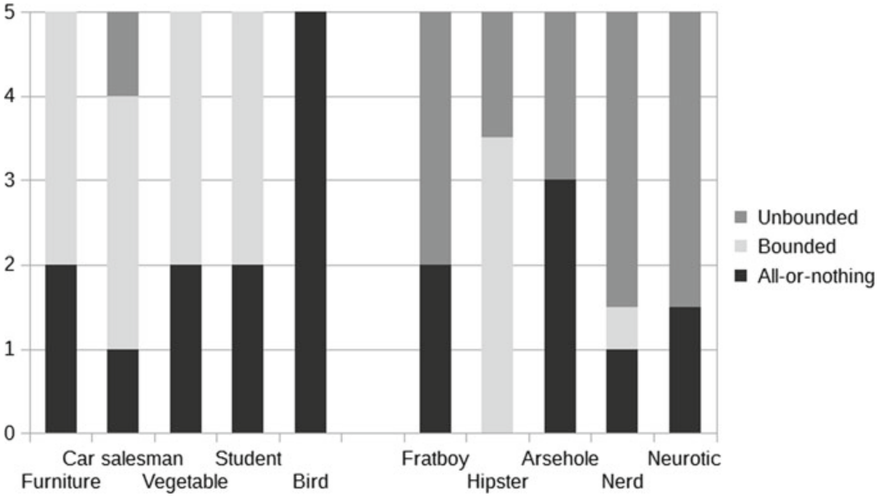


Fig. 2 The number of all-or-nothing, bounded and unbounded dimensions for each concept; the assigned type is the one picked most often by the participants. The Sassoon nouns are on the left, the Morzycki ones on the right

Table 9 Comparison between the number of all-or-nothing, bounded and unbounded picks for the five Morzycki nouns and the five Sassoon nouns, for each individual participant, by means of a paired-samples t-test. N = 23

	Mean no. of picks	Mean no. of picks	Standard		
	Morzycki	Sassoon	Paired t	Error	p
All-or-nothing	7.48	11.22	3.5850	1.043	0.0016
Bounded	5.22	10.91	6.9805	0.816	<0.0001
Unbounded	12.3	2.87	9.221717	1.023	<0.0001

4 Discussion and Conclusions

The results of these small-scale experiments support the hypothesis that the class of 'Morzycki-nouns'—nouns that pattern with relative adjectives in allowing size and evaluative modification, but not endpoint modification, and can be shown to be linguistically gradable using several monotonicity-based tests—correspond to [+G−P] concepts in Kamp and Partee's typology, just like relative adjectives. They also show that considerations of realism do not seem to play a role in determining concept membership for [−P] concepts (supporting an understanding of prototypicality in terms of maximal embodiment and nothing else). Moreover, the results of Experiment 1 suggest that the fact that some [−P] concepts, like GENIUS, intuitively seem to have a prototype may actually reflect independent aspects of instance accessibility.

Let us briefly return to the questions I started out this paper with. In the introduction I tentatively concluded, following many others, that an accurate theory of conceptual cognition probably needs elements from both Prototype Theory and the classical, definition-based theory of concepts. Previous researchers have mainly based this conclusion on philosophical and psychological arguments; I have added a linguistic one. By showing that the various concept classes identified by Kamp and Partee (1995) show distinct linguistic behaviour (and conversely, that various natural language expressions that show distinct linguistic behaviour also fall into different concept classes), I have provided additional support for the psychological reality of such a typology—most importantly, the distinction between concepts that have prototypes and concepts that lack them (or whose prototypes do not determine their extension).

A few problems and questions remain. First, it should be noted (as an anonymous reviewer did) that the Kamp and Partee-inspired argument based on which I claimed that concepts like NERD lack prototypes—it is always possible to imagine someone who embodies the concept more—also holds for concepts like GRANDMOTHER, yet GRANDMOTHER is commonly assumed (also by Kamp and Partee) to have a prototype and does not correspond to a Morzycki-noun. I do not have an immediate answer to this, but I speculate that an answer may be sought in another aspect of the Kamp/Partee typology I did not really go into: the [±PE] distinction between [+P] concepts whose extension is determined by typicality and [+P] concepts whose extension is determined by definitional criteria. Note that it is usually possible to ‘coerce’ [±G+P−PE] concepts like GRANDMOTHER into [+PE] concepts:

(20) I spent all day on the couch. I feel like such a granny.
(21) a. Mary was acting like a big girl.
b. John was acting like a big girl.

The speaker of (20) does not mean that she feels like a woman who has grandchildren; instead, she’s claiming something about her resemblance to the GRANDMOTHER prototype. The contrast in (21) is particularly striking: if we use *big girl* to describe someone who fits the definitional criteria of GIRL, the adjective *big* receives its standard intersective size interpretation: to act like a big girl, for a girl, is to act like a girl who is big. In contrast, if we use the same predicate to describe someone outside of the extension of GIRL, *girl* effectively starts behaving like a Morzycki noun, meaning something like ‘girlish person’. When used this way, the size adjective receives a degree reading: to act like a big girl, *for a boy*, is to act very girlishly (with all its sexist connotations). Do these ‘coerced’ versions of GRANDMOTHER and GIRL have prototypes, and if they do, what would that mean for the present analysis of Morzycki nouns as [−P] concepts? Perhaps future research may show the [−P] hypothesis is too strong in its current form.

Another interesting observation involving [+P−PE] concepts may connect the present approach to the work of Galit Weidman Sassoon on nominal gradability. Intuitively, [−PE] concepts are concepts that somehow ‘exist independently’ from the way they are experienced by humans, while [+PE] concepts are defined by human perception. There is no law of nature deciding which objects count as *furniture* and which do not—*furniture* is a category made up by humans in order to describe something they created. On the other hand, platonic objects such as circles and natural kinds such as grandmothers ‘exist’ and obey certain natural laws regardless of human perception or activity. In other words, there seems to be an intuitive connection between what Sassoon calls ‘social nouns’—human roles and artifacts—and [+PE] concepts on the one hand, and natural kind nouns and [−G−PE] concepts on the other. Recently, (Sassoon and Fadlon 2017) have shown that social nouns are significantly more compatible with various degree constructions than natural kind nouns; following the Kamp/Partee typology, the latter correspond to sharp concepts (or non-gradable concepts, in my adaptation), while the former (necessarily) correspond to vague (or gradable) ones. Sassoon and Fadlon (2017) account for their findings in a different way, claiming that the differences in linguistic gradability drives from the different ways the values on the concept’s dimensions are integrated into a single degree scale, but perhaps there is a way to unify or at least connect the two approaches.

Finally, let me point out a possible connection between emotional attitude and linguistic gradability. Many of the Morzycki-nouns are quite emotionally charged (either positive or negative), while the Sassoon-nouns generally are more neutral. Furthermore, Dutch has a reliable way to come up with new Morzycki-nouns by compounding any noun with an expletive from its impressive collection (such as *kut* ‘cunt’, *klote* ‘bollocks’ or *tering* ‘tubercolosis’)—e.g. *kutklus* ‘detestable job’, *klotehond* ‘awful dog’, *teringwereld* ‘horrible world’. These linguistically gradable compounds are evidently highly emotionally charged as well. I do not have an immediate explanation for this, but I do not think it is a coincidence—note also that the Sassoon noun that behaved most like the Morzycki nouns in Experiment 3 was *car salesman*, which people also tend to have an emotionally charged reaction to (among the properties most often attributed to car salesmen in Experiment 2 was ‘untrustworthy’, and properties in the vein of ‘slick’, ‘smarmy’ and ‘pushy’ also got several mentions). Hopefully, further research will shed more light on the relation of emotional attitude to the linguistic gradability of nouns.

Acknowledgements I’m grateful to Yoad Winter for funding my participation in the 2015 Madrid workshop from his NWO VICI grant no. 277-80-002. I would also like to thank Galit Weidman Sassoon and two anonymous reviewers for their valuable comments and suggestions. Finally, I gratefully acknowledge the following photographers who made their content available for non-commercial free use: Cristiano del Riccio, Georges Biard, Matt Yohe, Tony Wills, Nevit Dilmen, Martien Brand, Wikipedia editor JZ85, Glen Fergus, Mike Pazzani, Samuel Blanc, and Andreas Trepte.

Appendix 1: Experiment 1

All experimental materials in these appendices are translated from the original Dutch. The experiment only used images either in the public domain (PD-1923/PD-US-no notice) or licensed under Creative Commons terms (CC BY/CC BY-NC), but are not included below as they failed to meet the copyright requirements for the present volume.

The items on the two lists below were shown 3 times at different points in the questionnaire; each time, all 10 instantiations of the concept were shown in the same order and on a single page. In each of the 3 parts, subjects rated each item on a 7-point scale (by ticking a box) according to a different criterion: first, familiarity; second, typicality; and third, emotional attitude. Items consisted of a name and a picture with a mouse-over description.

Appendix 2: Experiment 3

The items on the list below were shown in blocks of 10 items per page; within each page, items were randomised differently for each participant. Each item started with the statement “A real, typical X...” followed by 3 possible continuations; first, an All-or-nothing one (A); second, a Bounded one (B); and third, an Unbounded one (U). Participants had to choose exactly 1 of the continuations by ticking a box.

The Dutch construction translated here somewhat inaccurately with ‘but not’ is *maar ook weer niet* (‘but also again not’), which is used to qualify a previous

Birds		
Name	Picture (url)	Mouse-over text
Blackbird	https://en.wikipedia.org/wiki/File:Male_blackbird-b.jpg	This 25-cm-tall songbird occurs mainly in Europe. It's an omnivore with an extensive song repertoire.
Ostrich	https://en.wikipedia.org/wiki/File:Ostrich_Struthio_camelus_Tanzania_3739_cropped_Nevit.jpg	The ostrich lives in Africa and is the biggest bird on earth. It can't fly but it can run very fast. Females lay their eggs in a common nest and take turns brooding.
Rock bunting	https://en.wikipedia.org/wiki/File:Emberiza_cia_Martien_Brand.jpg	This 16-cm-tall songbird occurs in Asia, Northern Africa and Southern Europe. It nests on or close to the ground and feeds on insects and seeds.
Egyptian goose	https://commons.wikimedia.org/wiki/Alopochen_aegyptiaca#/media/File:Alopochen_aegyptiacus_-_Egyptian_goose.JPG	This waterfowl isn't actually a goose but a duck. Originally from Africa; the ones we find in the Netherlands are feralised ornamental birds.
Gannet	https://commons.wikimedia.org/wiki/File:Jan-van-gent.JPG	This large, aerodynamic seabird hunts for fish in the North Sea and Atlantic Ocean. Its legs and wings are weak, so it can only fly in strong winds. Is a stellar diver.
Kiwi	https://commons.wikimedia.org/wiki/File:Tokoeka.jpg	This New-Zealand bird is roughly the size of a chicken, but lays eggs six times the size of a chicken's egg. It can't fly and is mainly nocturnal.
Green honeycreeper	https://www.flickr.com/photos/pazzani/5553987995	A small tropical songbird that occurs in Central and South America. Feeds predominantly on nectar.
Emperor penguin	https://commons.wikimedia.org/wiki/File:Emperor_Penguin_Manchot_empereur.jpg	The biggest penguin; measures up to 120 cm and weighs up to 45 kilos. Subsists mainly on fish. It can't fly, but it can dive up to 500 meters deep and stay underwater for up to 18 min.
Kingfisher	https://commons.wikimedia.org/wiki/File:Common_Kingfisher_Alcedo_atthis.jpg	This 16-cm-tall bird occurs in Europe, Asia and Northern Africa. It dives for fish and nests in steep riverbanks.
Little grebe	https://commons.wikimedia.org/wiki/File:Tachybaptus_ruficollis_ruficollis.jpg	This small, shy water bird is related to the great crested grabe. It occurs in large parts of Europe, Asia and Africa and feeds mainly on water insects and larvae.

utterance by explicitly negating a stronger alternative, e.g. *Ze was boos, maar ook weer niet woedend* 'she was angry, but not exactly furious'. It is somewhat weaker than plain 'but not'; a speaker who utters *maar ook weer niet X* seems to be hedging her commitment to the belief that X is false. As a result, the B continuations involve a slight asymmetry that is mostly lost in the translations below: the boundedness of the property is of secondary importance to possessing it in the first place.

Geniuses		
Name	Picture	Mouse-over text
Leonardo da Vinci	https://commons.wikimedia.org/wiki/File:Possible_Self-Portrait_of_Leonardo_da_Vinci.jpg	(1452-1519): Italian artist, scientist, architect and inventor. Considered one of the best painters ever, with the Mona Lisa as his most famous work. His designs include things as wide-ranging as defence works, musical instruments to unbuildable but ingenious flying machines. Was also a gifted astronomer and physiologist.
Emily Dickinson	https://commons.wikimedia.org/wiki/Emily_Dickinson#/media/File:Emily_Dickinson_daguerreotype_(cropped).jpg	(1830–1886): American poet who lived her life in near-total seclusion. Published only a few poems during her lifetime (which her publisher adjusted to conform to the age's style and taste; her enormous body of work was only discovered after her death. Now considered one of the greatest American poets, whose unconventional style placed her far ahead of her time.
Dmitri Mendeleev	https://en.wikipedia.org/wiki/Dmitri_Mendeleev#/media/File:DIMendeleevCab.jpg	(1834–1907): Russian chemist and inventor of the Periodic Table of Elements, a spatial ordering of all chemical elements based on their properties, which came to him in a dream. It correctly predicted the existence of various elements that hadn't yet been discovered.
Steve Jobs	https://commons.wikimedia.org/wiki/File:Steve_Jobs_Headshot_2010-CROP.jpg	(1955–2011): Entrepreneur and pioneer of the 70 s computer revolution. Saved the struggling Apple company with innovative technology and groundbreaking, iconic design. Was also CEO of Pixar, which under his leadership produced several of the most critically acknowledged animation films ever.
Hildegard von Bingen	https://commons.wikimedia.org/wiki/Hildegard_von_Bingen#/media/File:Hildegard_von_Bingen.jpg	(1098–1179): Benedictine abbess, writer, poet, composer, mystical theologian, scientist and philosopher. Is considered one of the founders of natural history as a scientific field. For her religious poetry, she invented her own script and hundreds of new words. Her music is still regularly performed.
Marie Skło-dowska Curie	https://commons.wikimedia.org/wiki/Maria_Sklodowska-Curie#/media/File:Marie_Curie_1900_-_DIG17379.jpg	(1967–1934):[a] Polish-French physicist and chemist, pioneer in the field of radioactivity (which eventually caused her death). She was the first female professor at the university of Paris, the first woman to win a Nobel prize, the first person to win a second Nobel prize, and the only recipient of two Nobel prizes in different categories.
Rabindranath Tagore	https://commons.wikimedia.org/wiki/Rabindranath_Tagore#/media/File:Rabindranath_Tagore_in_1909.jpg	(1861–1942): Bengali poet, composer, writer, painter, independence and peace activist and education reformer. Published his first volume of poetry when he was 16, and his first opera when he was 20. Was the first non-western winner of the Nobel prize for literature (in 1913) and used the prize money to found his own university. Wrote the national anthems of both India and Bangladesh.
Albert Einstein	https://commons.wikimedia.org/wiki/Category:Portrait_photographs_of_Albert_Einstein\#/media/File:Einstein-formal_portrait-35_(cropped).jpg	(1879–1955): German physicist and founder of the theory of relativity. At 27, in a single year, he published 4 revolutionary physics papers on different topics, written in his spare time next to his day job as an office clerk. Published over 300 scientific articles during his lifetime and won a Nobel prize in 1921.
Michael Jackson	https://commons.wikimedia.org/wiki/File:Michael_Jackson_Cannescropped.jpg	(1958–2009): Eccentric singer, songwriter, dancer and producer known as the 'King of Pop'. Began his singing career as a five-year-old, in the family group The Jackson Five. His album 'Thriller' is the best-sold album of all time; the eponymous music video is considered revolutionary and is the only music video ever to be included in the American national film registry.
Björk	https://www.flickr.com/photos/26377221@N06/2498816562	(1965-): This Icelandic singer and multi-instrumentalist released her debut album when she was 11. Is known for her groundbreaking, avant-gardistic music and unique music videos, on which she collaborates with international film directors, artists and fashion designers. In 2011, she released an album consisting fully of interactive apps, which has since been included in the permanent collection of the MoMA in New York. Has also received multiple acting awards.

[a] Note the typo in these dates (1967 instead of 1867); I do not expect this to have influenced the results in any way

Items: NERD	Page
A real, typical nerd:	1
A has a peculiar hobby/interest that he/she puts a lot of time into and knows a lot about	
B has a peculiar hobby/interest that he/she puts a lot of time into and knows a lot about, but not an unrealistic lot	
U has a peculiar hobby/interest that he/she puts a lot of time into and knows a lot about—the more the better	
A real, typical nerd:	2
A has great computer skills	
B has great computer skills, but not unrealistically great	
U has great computer skills—the greater the skills, the better	
A real, typical nerd:	3
A is socially awkward	
B is socially awkward, but not extremely socially awkward	
U is socially awkward—the more socially awkward, the better	
A real, typical nerd:	4
A wears glasses	
B wears strong glasses, but not ridiculously strong	
U wears strong glasses—the stronger, the better	
A real, typical nerd:	5
A is intelligent	
B is intelligent, but not absurdly intelligent	
U is intelligent—the more intelligent, the better	

Items: FRAT BOY	Page
A real, typical frat boy	1
A behaves in an arrogant manner	
B behaves in an arrogant manner, but not absurdly arrogant	
U behaves in an arrogant manner—the more arrogant, the better	
A real, typical frat boy	2
A wears a suit jacket and tie	
B wears a suit jacket and tie, but not 24/7 U wears a suit jacket and tie—the more often the better, he preferably sleeps in them too	
A real, typical frat boy	3
A has a posh accent	
B has a posh accent, but not extremely posh	
U has a posh accent—the posher, the frattier	
A real, typical frat boy	4
A drinks a lot of beer	
B drinks a lot of beer, but not an unrealistic lot	
U drinks a lot of beer—the more, and the more frequently, the frattier	
A real, typical frat boy	5
A is loud and obnoxious	
B is loud and obnoxious, but not ridiculously loud and obnoxious	
U is loud and obnoxious—the louder and more obnoxious, the better	

Items: HIPSTER	Page
A real, typical hipster	1
A wears glasses	
B wears glasses, but not extremely prominent ones	
U wears glasses—the more prominent, the more hipsterish	
A real, typical hipster	2
A has an alternative, eccentric style	
B has an alternative, eccentric style, but not extremely alternative or eccentric	
U has an alternative, eccentric style—the more alternative/eccentric, the better	
A real, typical hipster	3
A has a long and/or wild beard	
B has a long and/or wild beard, but not extremely long or wild	
U has a long and/or wild bird—the longer/wilder, the more hipsterish	
A real, typical hipster	4
A loves obscure music	
B loves obscure music, but not extremely obscure	
U loves obscure music—the more obscure and unknown, the better	
A real, typical hipster:	5
A is preoccupied with special, unique and ’pure’ food/coffee/beers	
B is preoccupied with special, unique and ’pure’ food/coffee/beers, but not extremely obsessed by them	
U is preoccupied with special, unique and ’pure’ food/coffee/beers—the bigger his/her obsession, the better	

Items: ARSEHOLE	Page
A real, typical arsehole:	1
A cares little about other people’s needs	
B cares little about other people’s needs, but is not completely uncaring	
U cares little about other people’s needs—the less, the better	
A real, typical arsehole	2
A has an unsympathetic personality	
B has an unsympathetic personality, but also not pathologically unsympathetic	
U has an unsympathetic personality—the more unsympathetic, the better	
A real, typical arsehole:	3
A is a man	
B is a man, but not an absurdly masculine testosterone bomb	
U is a man—the more masculine, the better	
A real, typical arsehole	4
A behaves in a rude, mean manner	
B behaves in a rude, mean manner, but not excessively rude or mean	
U behaves in a rude, mean manner—the worse and the more often, the better	
A real, typical arsehole	5
A is selfish	
B is selfish, but doesn’t always think of nobody but himself	
U is selfish—the more selfish, the better	

Items: NEUROTIC	Page
A real, typical neurotic	1
A worries about everything	
B worries about everything, but not inordinately deeply or frequently	
U worries about everything—the deeper and more frequent the worries, the more neurotic	
A real, typical neurotic A has tics and compulsive tendencies B has tics and compulsive tendencies, but moderately U has tics and compulsive tendencies—the more, the more neurotic	2
A real, typical neurotic	3
A behaves in a nervous and agitated manner	
B behaves in a nervous and agitated manner, but not excessively nervous and agitated	
U behaves in a nervous and agitated manner—the more nervous and agitated, the more neurotic	
A real, typical neurotic	4
A is insecure	
B is insecure, but not overly insecure	
U is insecure—the more insecure, the more neurotic	
A real, typical neurotic	5
A is focused on tiny details	
B is focused on tiny details, but not absurdly so	
U is focused on tiny details—the more focused, and the more trivial the details, the better	

Items: BIRD	Page
A real, typical bird:	1
A lays eggs	
B lays eggs, but not extremely many	
U lays eggs—the more the better	
A real, typical bird	2
A flies	
B flies, but not all day long	
U flies—the longer and the more often, the better	
A real, typical bird	3
A has feathers	
B has feathers, but also spots without feathers	
U has feathers all over its body, the more the better	
A real, typical bird	4
A has wings	
B has wings, but not overly prominent ones	
U has wings—the more prominent the better	
A real, typical bird	5
A has a beak	
B has a beak, but not an extremely big or striking one	
U has a beak—the more big/striking, the better	

Items: CAR SALESMAN	Page
A real, typical car salesman	1
A knows a lot about cars	
B knows a lot about cars, but not an extreme lot	
U knows a lot about cars—the more, the better	
A real, typical car salesman	2
A dresses smartly	
B dresses smartly, but not *too* smart	
U dresses smartly—the smarter, the better	
A real, typical car salesman	3
A is a smooth talker	
B is a smooth talker, but not absurdly smooth	
U is a smooth talker—the smoother, the better	
A real, typical car salesman	4
A is untrustworthy	
B is untrustworthy, but not extremely untrustworthy	
U is untrustworthy—the more untrustworthy, the better	
A real, typical car salesman	5
A behaves in a friendly and jovial manner	
B behaves in a friendly and jovial manner, but not overly friendly and jovial	
U behaves in a friendly and jovial manner—the more friendly and jovial, the better	

Items: FURNITURE	Page
A real, typical item of furniture	1
A has a surface for sitting or putting something on	
B has a surface for sitting or putting something on, which is neither too big nor too small	
U has a surface for sitting or putting something on—the bigger the surface the better	
A real, typical item of furniture	2
A is made of wood	
B is made of wood, but also incorporates other materials	
U is made of wood—the more wooden bits it has, the better	
A real, typical item of furniture	3
A is used inside the house	
B is used inside the house, but sometimes outside too	
U is used only inside the house, never outside	
A real, typical item of furniture	4
A serves a practical purpose	
B serves a practical purpose, but also has an aesthetic/artistic component	
U serves a practical purpose—the more practical, the better	
A real, typical item of furniture	5
A has legs	
B has 4 legs	
U has legs—the more legs, and/or the higher their prominence, the better	

Items: VEGETABLE	Page
A real, typical vegetable	1
A is eaten for dinner	
B is eaten dinner, but occasionally as part of other meals too	
U is eaten for dinner and never as part of another meal	
A real, typical vegetable	2
A is healthy	
B is healthy, but not some extremely healthy superfood	
U is healthy—the healthier, the better	
A real, typical vegetable	3
A is green	
B is mainly green, but has some other colours too	
U is green—the greener the better	
A real, typical vegetable	4
A needs to be cooked before you can eat it	
B needs to be cooked before you can eat it, but not too long	
U needs to be cooked before you can eat it—the longer the better	
A real, typical vegetable	5
A grows on the ground (and not on a tree)	
B grows on the ground, but not directly on the ground	
U grows on the ground—the closer to the ground, the better	

Items: STUDENT	Page
A real, typical student	1
A lives in a student room/dorm	
B lives in a student room/dorm, but often returns to his/her parents	
U lives in a student room/dorm and spends as little as possible time at his/her parents'	
A real, typical student	2
A studies at college or university	
B studies at college or university, but isn't focused on his/her studies 24/7	
U studies at college or university—the more time spent on studying, the better	
A real, typical student	3
A is intelligent	
B is intelligent, but not absurdly intelligent	
U is intelligent—the more intelligent, the better	
A real, typical student	4
A is young	
B is young, but not extremely young	
U is young—the younger the better	
A real, typical student	5
A has an active social life	
B has an active social life, but not outrageously active	
U has an active social life—the more active, the better	

References

Abney, S. (1987). *The English noun phrase in its sentential aspect*. Dissertation, MIT.

Armstrong, S., Gleitman, L., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*(3): 263–408.

Ashcraft, M. (1978). Property norms for typical and atypical items from 17 categories: A description and discussion. *Memory & Cognition*, *6*(3), 227–232.

Barsalou, L. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 629–654.

Barsalou, L. (1987). The instability of graded structure: implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101–140). Cambridge: Cambridge University Press.

Bolinger, D. (1972). *Degree words*. The Hague: Mouton.

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

Cresswell, M. (1976). The semantics of degree. In B. Partee (Ed.), *Montague Grammar*. New York: Academic Press.

Cruse, D. A. (1980). Antonyms and gradable complementaries. In D. Kastovsky (Ed.), *Perspektiven der lexikalischen Semantik* (pp. 14–25). Bonn: Bouvier.

Doetjes, J. (1997). *Quantifiers and selection: On the distribution of quantifiying expressions in French, Dutch and English*. Dissertation, Leiden University.

Fodor, J., & Lepore, E. (1996). The red herring and the pet fish: Why concepts still can't be prototypes. *Cognition*, *58*(2), 253–270.

Gärdenfors, P. (2004). Conceptual spaces as a framework for knowledge representation. *Mind and Matter*, *2*(2), 9–27.

Hampton, J., & Jonsson, M. (2012). Typicality and compositionality: The logic of combining vague concepts. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford handbook of compositionality*. Oxford: Oxford University Press.

Heim, I. (2000). Degree operators and scope. In B. Jackson & T. Matthews (Eds.), *Proceedings of Semantics and Linguistic Theory* (Vol. 10, pp. 40–64). Ithaca, NY: Cornell University).

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*(2), 129–191.

Katz, G. (2005). Attitudes towards degrees. In *Proceedings of Sinn und Bedeutung* (Vol. 9, pp. 183–196).

Kennedy, C. (1997). *Projecting the adjective: The syntax and semantics of gradability and comparison*. Dissertation, University of California, Santa Cruz.

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

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

Lassiter, D., & Goodman, N. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In T. Snider (Ed.), *Proceedings of Semantics and Linguistic Theory* (Vol. 23, pp. 587–610).

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings*. Cambridge, MA: MIT Press.

Lehrer, A. (1985). Markedness and antonymy. *Journal of Linguistics*, *21*(2), 397–429.

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

Malt, B., & Smith, E. (1982). The role of familiarity in determining typicality. *Memory & Cognition*, *10*(1), 69–75.

McCawley, J. (1988). *The syntactic phenomena of English*. Chicago: University of Chicago Press.

Morzycki, M. (2009). Degree modification of gradable nouns: Size adjectives and adnominal degree morphemes. *Natural Language Semantics*, *17*(2), 175–203.

Nouwen, R. (2009). Monotone amazement. In *Proceedings of Amsterdam Colloquium* (Vol. 15, pp. 167–172).

Osherson, D., & Smith, E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*(1), 35–58.

Prinz, J. (2012). Regaining composure: A defense of prototype compositionality. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford handbook of compositionality*. Oxford: Oxford University Press.

van Rooij, R. (2011). Vagueness and linguistics. In G. Ronzitti (Ed.), *The vagueness handbook*. Dordrecht: Springer.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. Moore (Ed.), *Cognitive development and the acquisition of language*. New York: Academic Press.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

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

Sassoon, G. W. (2007). *Vagueness, gradability and typicality: A comprehensive semantic analysis*. Dissertation, Tel Aviv University.

Sassoon, G. W. (2012). A slightly modified economy principle: Stable properties have non stable standards. In E. Cohen (Ed.), *Proceedings of the Israel Association of Theoretical Linguistics* (Vol. 27, pp. 163–182). MIT Working Papers in Linguistics.

Sassoon, G. W. (2016). *Multidimensionality in the grammar of gradability*. Unpublished manuscript, Bar Ilan University.

Sassoon, G. W., & Fadlon, J. (2017). The role of dimensions in classification under predicates predicts their status in degree constructions. *Glossa: a journal of general linguistics* *2*(1): 42. 1–40.

von Stechow, A. (1984). Comparing semantic theories of comparison. *Journal of Semantics*, *3*(1–2), 1–77.

Storms, G., Boeck, P. D., Hampton, J., & Mechelen, I. V. (1999). Predicting conjunction typicalities by component typicalities. *Psychonomic Bulletin & Review*, *6*(4), 677–684.

de Vries, H. (2010). *Evaluative degree modification of adjectives and nouns*. MA thesis, Utrecht University.

Yoon, Y. (1996). Total and partial predicates and the weak and strong interpretations. *Natural Language Semantics*, *4*(3), 217–236.

Education as a Source of Vagueness in Criteria and Degree

Steven Verheyen and Gert Storms

Abstract Individual differences in application are considered a hallmark of vague terms. When a term is truly vague there exists a range of applications that are considered permissible by competent users of the language. The divergence in application may be the result of indeterminacy with respect to the conditions for application (vagueness in criteria) and indeterminacy with respect to the extent of application given fixed conditions (vagueness in degree). We propose a formal procedure to determine whether individual application differences result from vagueness in criteria and/or vagueness in degree. The procedure provides an experimental perspective on vagueness in that it involves the comparison of two groups of participants that differ on a variable of interest. The procedure establishes whether the variable systematically affects application of a term. We present a case study in which we compare categorization data from participants who went on to higher education after completing compulsory education and participants who did not. Application of the proposed procedure shows that education systematically affects categorization. Higher education participants tend to apply common terms like VEGETABLES, FURNITURE, and TOOLS more conservatively than compulsory education participants do (vagueness in degree). For terms they are arguably more familiar with, like SCIENCES, they are found to employ different conditions for application (vagueness in criteria). The results demonstrate that part of the permissible variation that is deemed characteristic of vagueness reflects sociolinguistic variation.

S. Verheyen (✉)
Institut Jean Nicod, Département d'études Cognitives, ENS, EHESS,
PSL Research University, CNRS, Pavillon Jardin, 29, rue d'Ulm, 75005 Paris, France
e-mail: steven.verheyen@ens.fr

G. Storms
Laboratory of Experimental Psychology, Faculty of Psychology and Educational Sciences,
KU Leuven, Tiensestraat 102, 3000 Leuven, Belgium
e-mail: gert.storms@kuleuven.be

E. Castroviejo et al. (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure*,
Language, Cognition, and Mind 4, https://doi.org/10.1007/978-3-319-77791-7_6

Keywords Categorization · Individual differences · Threshold theory Permissible variation · Sociolinguistics · Vagueness in criteria · Vagueness in degree

1 Vagueness and Individual Differences

Whether a woman of 1m75 is TALL or chess is a SPORT are questions without single, matter-of-fact answers. TALL and SPORT are vague words, meaning that there is no established demarcation of the instances to which they apply and the instances they do not apply to. Individuals can use these words in different ways without committing an error (Kölbel 2004; Raffman 2014; Wright 1995). This "permissible variation" becomes readily apparent in categorization tasks in which the participants diverge widely regarding the instances they feel these words apply to (Black 1937; McCloskey and Glucksberg 1978). Among psychology undergraduates, the odds in favour of calling a woman of 1m75 TALL are 65:35 (Verheyen et al. 2018) and the odds in favour of considering chess a SPORT are about 50:50 (Verheyen et al. 2010), for instance.

The idea that some of these individual differences in categorization are systematic and can be brought back to properties of the participants, has often been entertained, but seldom demonstrated. Both Barsalou (1993) and Smith and Samuelson (1997) have suggested that in addition to the context language users find themselves in, their individual learning histories, aptitudes, and dispositions influence their categorization behavior. The idea also finds support in the work of Gardner (1953) and Verheyen et al. (2010), who showed that participants display relatively stable categorization patterns across tasks. Which properties of the individuals are responsible for the observed stability? And to what extent can individual differences in categorization be attributed to different participant properties?

Questions like these naturally fit an experimental perspective on vagueness. A comparison of the categorization behaviour of two groups of participants who differ in a property of interest, allows one to determine if the property under investigation systematically affects the participants' word use. The work described in this chapter is an illustration of how one might effectively do that. Inspired by a quote from Chomsky: "*Word meaning is intimately bound up with matters of knowledge and belief*" (1980: 225), we chose to study the effect of education on vagueness[1].

The effect of education on semantic tasks has rarely been explicitly investigated (see Rosenzweig 1964, for a notable exception), unless educational level is used as a proxy for verbal skill (Kuperman et al. 2013). Participants' level of education is regularly included as a control variable in semantic norming studies, though (e.g., De Witte et al. 2015; Loonstra et al. 2001). In these studies the transition from compulsory to higher education regularly yields pronounced effects (Charchat Fichman et al.

[1] To be fair, although Chomsky recognizes that variation is an important part of language, he considers the study of the *shared* knowledge of language users paramount.

2009; Keuleers et al. 2015). For illustrative purposes, we too therefore decided to focus on the categorization differences between participants who went on to complete higher education after compulsory education and those who did not.

In what follows we describe how a property such as the level of education of participants can be brought into a formal account of vagueness. This account takes the form of a statistical model that makes a number of assumptions regarding the manner participants arrive at a categorization decision *and* the kind of differences one might expect to see herein. Indeed, while the vagueness of words comes with variation in their application by different individuals, it is not the case that anything goes. The latter would render communication among language users impossible. An explanation of the permissible individual variation should thus be an intrinsic part of any formal account of vagueness (Black 1937). The proposed statistical model works by characterizing categorization differences between groups as vagueness in criteria or vagueness in degree.

2 Vagueness in Criteria and Degree

Devos (1995, 2003) distinguishes vagueness in criteria and vagueness in degree (see also Alston 1964; Kennedy 2013; Machina 1976). He defines vagueness in criteria as the indeterminacy with respect to (the combination of) the conditions for application of a term. Individuals might employ different criteria for establishing whether an activity is a SPORT or not. While some might emphasize that SPORTS require physical activity, others might require an element of competition. According to the first criterion *hiking*, but not *chess*, is likely to be considered a SPORT, while the reverse holds when the second criterion is employed. Devos (1995, 2003) defines vagueness in degree as the extent to which a term can be applied given that the conditions have been determined. Even when individuals agree on which criterion to employ to establish whether an activity is a SPORT or not, they might still disagree as to whether a particular activity sufficiently meets that criterion. That is, while some might deem both *hiking* and *running* sufficiently demanding, others might feel that only *running* requires sufficient physical activity to be considered a SPORT.

Devos (1995, 2003) argues that vagueness in criteria coincides primarily with nouns and vagueness in degree with adjectives. The rationale behind this is that while for many adjectives there exists a unique criterion that determines application (e.g., height for TALL, price for EXPENSIVE), many criteria can be considered for applying a noun like SPORT (competitiveness, physical activity, ...). This line of reasoning ignores the fact, however, that (a) many adjectives too are multifaceted (Kamp 1975; Klein 1980; Sassoon 2012) and (b) the application of many multifaceted nouns is effectively governed by one-dimensional constructs such as typicality or similarity to the target category (Hampton 1998, 2007; McCloskey and Glucksberg 1978). We therefore see no reason to expect a principled relationship between vagueness in criteria and nouns on the one hand, and vagueness in degree and adjectives on the other.

What Devos (1995, 2003) terms vagueness in degree is addressed in the so-called threshold theory (Hampton 1995, 1998, 2007; Raffman 1994, 1996). The threshold theory provides a psychological account of individual differences in categorization. It does so by positing that categorizers position a threshold that separates members from non-members on a dimension along which the candidate items are organized. The threshold theory thus assumes that categorization is governed by a single latent dimension, which can be composed of one or more (weighted) substantial criteria (see Égré 2017; Keefe 2000, for similar assumptions)[2]. In the work of Raffman (1994, 1996), who is mainly interested in adjectives such as RICH and RED, the dimension reflects a single substantial criterion such as dollar amounts or wavelengths. In the work of Hampton (1995, 1998, 2007), who is primarily concerned with noun categories such as FRUITS, VEHICLES, and SPORTS, the dimension is thought to reflect the items' typicality or similarity to the target category, which can be shown to reflect a weighted combination of several substantial criteria (Dry and Storms 2010; Hampton 1979; Rosch and Mervis 1975; see De Deyne et al. 2014, for evidence that the same holds for multifaceted adjectives). Crucially, however, in both accounts all categorizers are assumed to rely on the same dimension for their categorization and only to differ with respect to the threshold they employ. When categorizers use different thresholds on the same dimension, they are essentially diverging on the *degree* to which the target term applies, given fixed conditions (be it wavelength for RED, or a weighted combination of competitiveness and activity level for SPORTS).

Although vagueness in criteria lay outside its original scope, the threshold theory can easily be extended to encompass it, by allowing not only the threshold, but also the dimension that is relied upon for categorization, to be subject to individual variation. Hampton (2006) already alluded to this possibility when he showed that the weighting of substantial criteria may differ from person to person (see Hampton and Passanisi, 2016; Verheyen and Storms 2013; Zee et al. 2014, for additional empirical support). The resulting dimensions of categorization could differ rather subtly when the same criteria are merely accentuated differently, or more profoundly when they reflect the use of distinct criteria as in the SPORTS example at the beginning of this section. Both are potential instantiations of vagueness in *criteria*; the latter being a special case which amounts to setting to zero the weights of all potential criteria, except the one that is relied on.

2.1 A Formalization of Threshold Theory

Verheyen et al. (2010) provided a formalization of the threshold theory that can be applied to binary categorization tasks, in which categorizers go through a set of candidate items, indicating whether (1) or not (0) these belong to the target category. Like the threshold theory, the model assumes a single dimension along which the

[2] This assumption does preclude the possibility that individuals employ a disjunctive set of criteria, such as "SPORTS should be competitive *or* involve intense physical activity".

candidate items and the categorizers' thresholds are positioned and it is their relative position that determines the answers. Unlike the original theory, the thresholds in the model do not deterministically separate members from non-members.

The model encompasses a free parameter β for each one of the items *i* and a free parameter θ for each one of the categorizers *c*. The β_i values reflect a rank ordering of the items according to the propensity with which they are endorsed (the more often *i* is endorsed as a category member, the higher β_i is). The θ_c values reflect a rank ordering of the categorizers according to the number of items they endorsed (the fewer items *c* endorses, the higher θ_c is). As such, the values of β_i and θ_c can respectively be interpreted as the extent to which item *i* meets the categorization criteria and categorizer *c*'s threshold (Verheyen et al. 2010). The better an item meets the categorization criterion, the further it is positioned on the dimension. The higher the requirements for category membership categorizers' impose, the further their thresholds θ_c are positioned along the dimension.

Both the item positions (β_i) and the thresholds (θ_c) are estimated from the categorization data. As such, the model does not presume any a priori knowledge about the underlying dimension, although it can be subsequently interpreted by inspecting the relative positions of the items indicated by β_i. Items are positioned further along the dimension the better they meet the categorization criteria. If physical activity thus governed the categorization decisions for SPORT, *hiking* would be found to the right of *chess*. If competition governed the decisions, the relative positions of *hiking* and *chess* would be reversed.

According to the model formula in Eq. (1), the more the position of the item (β_i) surpasses the position of the categorizer's threshold (θ_c) along the dimension, the higher the probability that categorizer *c* considers item *i* a category member will be (and vice versa):

$$\Pr\left(Y_{ci} = 1\right) = \frac{e^{\beta_i - \theta_c}}{1 + e^{\beta_i - \theta_c}}. \tag{1}$$

The membership function thus starts of at 0 (clear non-member) for items that fall short of the categorizer's threshold ($\beta_i \lll \theta_c$) and steadily increases until 1 (clear member) for items that clearly surpass it ($\beta_i \ggg \theta_c$). The threshold θ_c thus does not rigorously separate members from non-members into, but rather reflects the point at which the categorizer is indifferent with respect to the category membership decision. The probability that the model assigns to the categorization answer for an item that coincides with the threshold θ_c is 0.5, making the decision effectively a coin toss. By casting them as stochastic variables, the model leaves room for some statistical variation in the categorization answers, which is particularly interesting for explaining intra-individual differences in categorization (see Verheyen et al. 2010, for details); but this does not mean that categorizers are completely free to fill in the meaning of a word. The categorization patterns are constrained by the positions of the items (β_i), which reflect how the items score on the criteria that are relevant for category membership. Like the threshold theory, the model assumes that the grounds for making the categorization decisions are shared by all categorizers. In Eq. (1) the agreement on which items score high/low on the relevant criteria is reflected in a single set of β_i estimates. That is, there is one dimension that governs categorization. As such, the model is only concerned with degree differences in categorization. It can be naturally extended to encompass criteria differences, though.

2.2 A Formalization of Criteria and Degree Differences

The model in Eq. (1) makes establishing whether a participant's group membership (e.g., compulsory vs. higher education) affects vagueness in criteria and degree tangible. Vagueness in degree would show in a meaningful difference in the threshold positioning θ_c of the members of the two groups. This would indicate that one group applies the term under investigation more conservatively than the other group does. Vagueness in criteria would show in a meaningful difference in the item positioning β_i of the two groups. This would indicate that different dimensions govern categorization in the groups. In order to establish whether two groups employ different criteria for categorization, the modeling framework again requires no intuitions about potential criteria. It suffices that the estimated item positions are different for the two groups, as this constitutes evidence that the used criteria are not the same. The nature of these criteria can later be investigated through interpretation of the items' relative positions in the manner described above.

To make statistical inferences regarding the existence of both types of vagueness feasible, the model in Eq. (1) needs to be extended as neither the threshold positions, nor the item positions are group dependent. In addition to indices *c* and *i* to indicate individual categorizers and items, respectively, an index *g* is introduced to make a distinction between groups. In order to establish whether group membership is a source of vagueness in degree, we assume that the thresholds follow a normal distribution with a group-specific mean and variance: $\theta_{cg} \sim N\left(\mu_{\theta_g}, \sigma^2_{\theta_g}\right)$. By constraining μ_{θ_1} to equal 0, μ_{θ_2} can be thought of as the mean threshold difference between the groups (compulsory vs. higher education). If μ_{θ_2} is reliably different from zero, this constitutes evidence for vagueness in degree resulting from education.

To establish whether group membership is a source of vagueness in criteria, we introduce binary latent indicators D_i that signal whether individuals who have the same threshold but are from different groups have a different probability of categorizing item *i* as a category member and therefore require a separate β_i estimate. If D_i equals 0 the model reads:

$$\Pr\left(Y_{cig} = 1 \,|D_i\, = 0\right) = \frac{e^{\beta_i - \theta_{cg}}}{1 + e^{\beta_i - \theta_{cg}}}. \tag{2}$$

β_i has no index *g* here, signaling that the position of item *i* is the same in both groups.

If D_i equals 1 the model reads:

$$\Pr\left(\mathrm{Y}_{cig} = 1 \,|\mathrm{D}_i \;= 1\right) = \frac{\mathrm{e}^{\beta_{ig}-\theta_{cg}}}{1 + \mathrm{e}^{\beta_{ig}-\theta_{cg}}}. \tag{3}$$

β_i does receive an additional index *g* here, signaling that item *i* is positioned differently in the two groups. When the modeling procedure uncovers items with different positions in the two groups, this constitutes evidence for vagueness in criteria resulting from education.

3 Method

In order to demonstrate the modeling framework, we will re-analyze part of the data from Verheyen et al. (2018) in which adult participants who were recruited online completed a categorization task. The re-analysis is restricted to the data from female participants, who make up the majority of the participants sample (55%), since previous research has shown that gender produces both vagueness in degree and vagueness in criteria (Stukken et al. 2013). We want to avoid mistaking gender differences for education differences.

3.1 Participants

The selection of participants is comprised of 1036 adult female participants aged 18—92 (*M* =56.82, *SD* =16.10) from Flanders (Belgium). Since not all participants completed all categories, the actual number of participants per category ranges between 1004 and 1011. Sixty-four percent of the participants indicated to have obtained a diploma beyond secondary education, which is the compulsory level in Flanders. They make up the higher education group.[3]

3.2 Materials

The materials were Dutch translations of the categories and items in Hampton et al. (2006). In what follows we will use the original English terms to refer to them. The materials included eight categories with 24 items each. The majority of these items were borderline items for the target categories, but clear members and non-members

[3] Because the higher educated participants tended to be younger, we also conducted an additional analysis in which we equated the two groups in terms of age. We biased this analysis against the educational effect found in the main analysis by choosing for the higher education group participants with a more practical university college education over participants with a scientific university education whenever possible. This additional analysis yielded similar results, indicating that the results of the main analysis are not due to a confounding of education with age.

were included as well to make the task more natural. The categories will be printed in small capitals (FRUITS, VEGETABLES, FISH, INSECTS, SPORTS, SCIENCES, TOOLS, and FURNITURE) and the items in italic (*avocado*, *garlic*, *shrimp*, *maggot*, *chess*, *economics*, *funnel*, and *piano*, are examples of borderline items for the respective categories). A list of the original materials along with their Dutch translation can be found in the Appendix.

3.3 Procedure

The data were gathered through a web survey. The participants completed a categorization task in which they were asked to indicate for the eight categories whether the 24 candidate items belonged to the category or not. In addition to "yes" and "no" participants could also answer "I don't know the item". It was emphasized that we were interested in participants' personal opinions rather than the answers considered appropriate by the general public or official authorities. The categories were presented on separate pages in a random order. The corresponding items were presented in randomized lists. Participants could proceed at their own pace. The majority of participants completed the survey in less than ten minutes.

3.4 Model Analysis

The models that have been discussed in this paper are all existing models that have been developed in the Item Response Theory (IRT) literature. IRT models tend to be used to infer latent traits from individuals' manifest responses to the items in a questionnaire or test (Hambleton et al. 1991). Verheyen et al. (2010) recognized the potential of IRT models for the study of vagueness when they introduced the model in Eq. (1) as a formalization of the threshold theory. In the IRT literature this model is known as the Rasch model (Rasch 1960). The model we describe in this chapter was introduced in the IRT literature by Frederickx et al. (2010) to investigate group bias in high stakes testing situations.

We analyzed each category's categorization data separately using the extended model. This was done using WinBUGS (Lunn et al. 2000) following the procedures for the Bayesian estimation of the model outlined in Frederickx et al. (2010). These include the specification of the priors for the model parameters. For every analysis five chains were run of 10,000 iterations each, with a burn-in sample of 1,000. To determine whether compulsory education (group 1) and higher education (group 2) participants employ different criteria for categorization, we investigate whether there are items for which the posterior probability of indicator $\mathrm{D_i}$ exceeds .5, indicating that the item is positioned differently in the two groups. To determine whether the two education groups differ regarding the degree they feel the target categories apply, we investigate the mean threshold difference between them. We deem a difference in degree reliable if the 95% credibility interval for μ_{θ_2} does not include 0.

3.5 Predictions

This is an exploratory study that is first and foremost intended to be an illustration of a modeling framework to characterize group differences in categorization as degree or criteria differences. The study was not designed with the intent of investigating the effect of education level on categorization. The choice for education level as the group variable was a matter of convenience, as we had this information available from a large scale study where this was not the variable of interest. We did not entertain any a priori hypotheses as to the extent to which we would observe criteria and degree differences resulting from education differences, as to our knowledge, the effect of education level on semantic categorization has not been investigated yet. We did deem education level a promising group variable for an illustration, though, as it seemed probable that individual differences in knowledge would affect how terms are applied.

4 Results

4.1 Vagueness in Criteria

We hardly found any evidence for vagueness in criteria. For FRUITS, VEGETABLES, INSECTS, SPORTS, and TOOLS there were no items for which the indicator D_i exceeded .5. That is, in the higher education group the candidate items were positioned the same as in the compulsory education group. For FISH and FURNITURE one item was positioned differently. Higher educated participants considered *sea horses* more representative FISH than compulsory education participants did (higher β_i in the higher education group). They also considered *shelves* more representative FURNITURE than compulsory education participants did.

For SCIENCES the picture looked completely different. There were twelve items (50%) for which D_i exceeded 0.5. *Astrology*, *philosophy*, *palm reading*, *literature*, and *psychology* were considered less representative of SCIENCES in the higher education group than in the compulsory education group (lower β_i in the higher education group). The items *geography*, *geometry*, *meteorology*, *mineralogy*, *chemistry*, *dentistry*, and *nutrition* were considered more representative by higher education participants than by compulsory education participants (higher β_i in the higher education group).

4.2 Vagueness in Degree

For FRUITS and SPORTS the mean threshold difference could not reliably be discerned from zero, indicating that there is no reliable difference in the number of items endorsed by the two groups. A reliable positive threshold difference was found for VEGETABLES, INSECTS, FURNITURE, FISH, and TOOLS. Higher education participants used a higher threshold than the compulsory education participants did, resulting in the endorsement of fewer items as category members by the former group. For the category of SCIENCES, we found the opposite pattern: the mean threshold difference was reliably negative, indicating that higher education participants tended to endorse more items as SCIENCES than compulsory education participants did.

4.3 Criteria-Degree Interplay

The findings regarding group differences in degree should be interpreted in light of the findings regarding vagueness in criteria. To aid this interpretation we have included a figure that depicts categorization patterns for the three combinations of vagueness in criteria and degree we identified through the model analyses: (i) absence of vagueness in criteria and degree, (ii) vagueness in degree but not criteria, (iii) vagueness in both criteria and degree. The three combinations are exemplified by the results for FRUITS, INSECTS, and SCIENCES, respectively.

All three panels in Fig. 1 show for 24 candidate items the proportion of participants endorsing the items as a category member. Each panel contains two graphs, one for the compulsory education participants (black circles) and one for the higher education participants (gray squares). The items are organized along the horizontal axis in increasing order of endorsement according to the compulsory education group. The item on the far left is thus the one least endorsed by the compulsory education participants, while the item on the right is the one most endorsed by these participants.

For FRUITS (left panel) the analyses yielded no reliable differences between the two education groups. The absence of vagueness in degree and criteria clearly shows in the categorization proportions as well: the black and gray curves almost completely overlap. This pattern of categorization is exemplary for SPORTS as well.

For INSECTS (middle panel) the analyses yielded a reliable threshold difference, but no differently positioned items. The absence of vagueness in criteria shows in that the categorization curves of the education groups take the same shape: the items are ordered in the same manner in both groups. The vagueness in degree shows in the displacement of the two curves: the gray squares are systematically lower than the black circles, indicating that the higher educated participants were more strict in categorizing items as category members. This pattern of categorization is exemplary for the majority of the studied categories (VEGETABLES, FURNITURE, FISH, TOOLS).

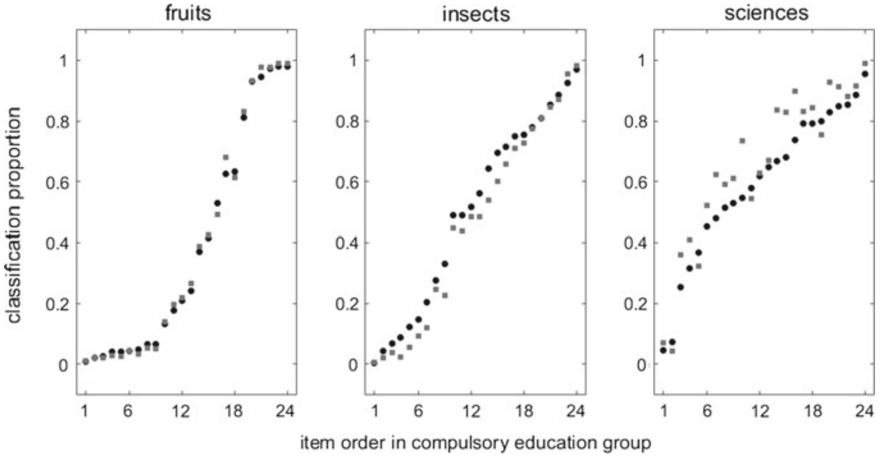


Fig. 1 Categorization proportions for FRUITS (left), INSECTS (middle), and SCIENCES (right) of the compulsory education group (black circles) and the higher education group (gray squares). Items are ordered along the horizontal axis according to their categorization rank in the compulsory education group

For SCIENCES (right panel), we observe that the categorization proportions of the higher education group tend to be higher (instead of lower) with respect to those of the compulsory education group. This is an indication of the vagueness in degree the analyses established. The additional vagueness in criteria shows in that the two curves no longer have the same shape: The order of the categorization proportions in the two groups is not the same. It is not just the case that the curve of one group is shifted with respect to the other group, as we saw for INSECTS. The nature of the group difference is different for individual items (those for which D_i exceeded .5) with the compulsory – higher education categorization divide being greater for some than for others and in some instances even showing the reverse pattern. This was the case for items 2 (*palm reading*), 5 (*literature*), 11 (*philosophy*), and 19 (*astrology*). They were considered less representative category members by the higher education group, along with *psychology* (item 12) for which the categorization proportion difference is smaller than one would expect in light of the group threshold difference. For other items such as item 7 (*nutrition*), 10 (*dentistry*), and 16 (*mineralogy*) the categorization difference is larger than expected based on the threshold difference solely.

4.4 Discussion

We observed pronounced differences in semantic categorization between the group of compulsory education participants and the group of higher education participants. The model analyses qualified the categorization differences more often as degree differences (with higher education participants endorsing fewer items than compulsory education participants), than as criteria differences. From this observation one should conclude that education can give rise to both these types of vagueness, but not that degree differences are more prevalent than criteria differences. Which combination of degree and criteria differences emerges, appears to be dependent upon the stimulus materials, and the employed materials were not selected with the education difference in mind. The current study does offer a number of interesting hypotheses regarding the origin of the degree and criteria differences between the education groups, which can be tested in future research with materials tailored to these questions (see Conclusions section for details).

5 Conclusions

The purpose of this chapter was to introduce a procedure that allows group differences in categorization to be identified as criteria and/or degree differences. As such, the procedure yields a number of insights regarding the nature and the sources of vagueness. We established that for noun categories, both vagueness in criteria and vagueness in degree are likely in play at any given moment. From the early work by McCloskey and Glucksberg (1978) and the subsequent replication of their work by Verheyen et al. (2010) we already knew that individuals diverge in the degree to which they consider items members of noun categories. Verheyen and Storms (2013), on the other hand, identified latent groups of participants who use different criteria for categorizing items in categories such as FISH, SPORTS, and SCIENCES. The novelty of the current procedure lies in its focus on vagueness in degree and vagueness in criteria *simultaneously*, and in its ability to relate them to external information about the participants, such as their education. In the current application it allowed us to show that higher education participants tend to apply common terms like VEGETABLE, FURNITURE, and TOOL more conservatively than compulsory education participants do (vagueness in degree) and that for terms they are arguably more familiar with, like SCIENCES, they employ different conditions for application (vagueness in criteria).

The procedure opens up the possibility to investigate which other properties of the participants systematically affect vagueness. In other work along these lines, we have so far identified a number of factors with ties to degree and criteria differences. We found degree differences in the application of adjectives that one can also apply to oneself to be related to one's own standing on the relevant dimension. Subjects' height and weight requirements for applying TALL and HEAVY, for instance, correlate positively with their personal measurements (Verheyen et al. 2018). Gender results in categorization patterns that are opposite to those we observed for education. Using a selection of categories that were expected to yield gender differences in categorization (e.g., CLOTHING, PROFESSIONS, SPORTS, TOYS) few degree differences were observed, while most categories yielded criteria differences (Stukken et al. 2013). Age gives rise to very intricate patterns of categorization. Both young children (<13 years old) and older adults (>62 years old) are found to overextend common noun categories compared to young adults (degree difference; Verheyen et al. 2011a; Verheyen et al. 2018). Older and young adults also differ in the criteria they use for categorization (White et al. 2018). For the categorization of storage containers, for instance, older adults rely more on “classic” materials such as glass or cardboard, whereas younger adults emphasize relatively “new” materials such as plastics. Using the proposed modeling framework, it is straightforward to extend this line of study to the comparison of different contexts, cultures, language groups, and even regional varieties, and/or to look into categorization differences at the level of individual items instead of entire categories, as we have done.

Ultimately, this research program will allow us to determine to what extent vagueness, as shown in inter-individual application differences, can be accounted for in terms of a limited number of external participant properties. As such, the program could be characterized as researching vagueness from a sociological or psychological perspective, depending on the nature of the property of interest that is investigated. The complete resolution of vagueness is not within reach of this program, however, as intra-individual application differences (Hampton and Passanisi 2016; McCloskey and Glucksberg 1978) cannot be explained in these terms. Our findings do raise the fundamental question of whether sociolinguistic variation of the kind uncovered in this study should be part of a theory of vagueness proper, or whether such a theory should only address the variation that is left after these external influences have been partialled out[4]. A choice for the latter might lead to inter-individual application differences being struck as hallmarks of vagueness, as it is impossible to ever ascertain whether all relevant participant properties have been taken into account.

The current project falls short when it comes to explaining where the effects of education on categorization come from. The purpose of this project was to establish *whether* education gives rise to particular categorization differences. Establishing *how* education gives rise to these differences, is more of an endeavor for sociolinguists and differential psychologists and is out of the scope of this chapter. Below we will nevertheless offer some suggestions as to the origins of the education effects we established, with the primary purpose of indicating how they could be formally tested within the framework we have proposed.

One explanation for the observation that compulsory education participants have broader categories than higher education participants do, might be the latter group’s higher lexical familiarity (familiarity with printed words). Individuals that score high on lexical familiarity, have been found to display a higher rejection rate of non-category members that are semantically related to the category (Lewellen et al. 1993), reminiscent of the difference in degree we established. According to this explanation, educational level would be a proxy for verbal skill (see Kuperman and Van Dyke 2013, for support of this argument). An alternative explanation for the difference in degree could also be attempted in terms of personality characteristics that correlate with education. Education level correlates positively with both

[4] Conversely, one might ask whether we are not construing sociolinguistics too broadly if that might encompass the possession of specific forms of non-linguistic knowledge, which might explain the differences we found (see below).

openness and conscientiousness (Denissen et al. 2008), but academic performance is more strongly related to conscientiousness than to openness (Rocklin 1994; see also Paunonen and Ashton 2001). The observation that higher educated participants are more conscientious than compulsory education participants are, could be used to explain the degree differences observed in the majority of categories. According to this reasoning, higher education participants would ultimately reject more semantic foils because their deliberations are more thorough and deliberate than those of the compulsory education participants are. However, based on the relationship between education and openness, one could have predicted the opposite pattern as well. Higher education participants would then include more items in their categories than compulsory education participants do because they are more imaginative and creative and can therefore more easily come up with conditions in which an item fulfills the requirements for category membership.

For the pronounced differences in criteria for the category of SCIENCES, one can similarly come up with several explanations: (i) higher education participants differ in familiarity with the different disciplines on the basis of information they acquired directly in higher education courses or indirectly through contact with fellow students, (ii) higher education participants have been explicitly instructed about the nature of SCIENCES, (iii) higher education participants have been exposed to the manner in which higher education institutes are organized. Evidence for the second explanation can be found in our data in that higher education participants make a stronger distinction between pseudosciences like *astrology* and *palm reading* and prototypical natural sciences like *chemistry*, *meteorology*, and *mineralogy*. Evidence for the third explanation can be found in the clustering of *dentistry* and *nutrition* with medicine among higher education participants. All three explanations offered for the vagueness in criteria with respect to the SCIENCES category are reminiscent of categorization differences between experts and novices (e.g., Chi et al. 1981).

We did not entertain any a priori predictions as to how education might give rise to categorization differences. For demonstrative purposes we re-analyzed an existing data set, which was not gathered with the purpose of explaining education differences. If one were interested in the effects of lexical familiarity, conscientiousness, or openness on degree differences in categorization, it would be straightforward to collect this information from the participants and apply the procedure outlined in this chapter after dichotomizing these external variables (as is commonly done in experimental approaches). Alternatively, if one would like to honor the continuous nature of these variables, one could extend the model hierarchically to evaluate their relationship with the estimated thresholds (for a demonstration see Verheyen et al. 2011a). To interpret the criteria differences, one could regress the item positions onto the criteria under consideration and see which of these impact the item positions in the two groups differently. This is the approach taken by Verheyen and Storms (2013) and Verheyen et al. (2015). The regression could also be made part of the modeling by extending the model hierarchically, as in Verheyen et al. (2011b). Ideally, if one were to have specific hypotheses about potential criteria, one would compile the item set in a way that optimally allows one to disentangle the various criteria under consideration. Manipulations like these would make the proposed procedure an even more valuable experimental perspective on vagueness.

Acknowledgements Part of this research was conducted while SV was a postdoctoral fellow at the Research Foundation-Flanders (FWO). At the time of writing, SV was funded by ANR project *TriLogMean* (ANR-14-CE30-0010). SV also acknowledges grants ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL* for research conducted at the Department of Cognitive Studies of ENS in Paris. SV and GS thank Elisabeth Droeshout for her help with the data collection.

Appendix, Part 1: Categories and Items from Hampton, Dubois, and Yeh (2006) Along with Their Dutch Translation Used in the Current Study

FRUITS	FRUIT	VEGETABLES	GROENTEN	FISH	VISSEN	INSECTS	INSECTEN
acorn	eikel	apple	appel	alligator	krokodil	amoeba	amoebe
almond	amandel	artichoke	artisjok	catfish	zeewolf	ant	mier
avocado	avocado	asparagus	asperge	clam	mossel	bacterium	bacterie
banana	banaan	bamboo shoot	bamboescheut	crab	krab	bat	vleermuis
carrot	wortel	bread	brood	eel	paling	caterpillar	rups
coconut	kokosnoot	celery	selder	frog	kikker	centipede	duizendpoot
cucumber	komkommer	cereal	graan	goldfish	goudvis	dust mite	mijt
date	dadel	chili pepper	peper	gull	meeuw	earthworm	regenworm
eggplant	aubergine	cloves	kruidnagel	jellyfish	kwal	grashopper	sprinkhaan
ginger	gember	dandelion	paardenbloem	lobster	kreeft	hamster	hamster
mint	munt	garlic	look	oyster	oester	head lice	luis
mushroom	champignon	lettuce	sla	plankton	plankton	leech	bloedzuiger
olive	olijf	milk	melk	salmon	zalm	lizard	hagedis
onion	ajuin	parsley	peterselie	sardine	sardine	maggot	made
orange	sinaasappel	peanut	pinda	sea horse	zeepaardje	mosquito	mug
pine cone	denneappel	pineapple	ananas	seal	zeerob	moth	mot
pomegranate	granaatappel	potato	aardappel	shark	haai	scorpion	schorpioen
pumpkin	pompoen	rice	rijst	shrimp	garnaal	silkworm	zijderups
rhubarb	rabarber	sage	salie	sponge	spons	snail	slak
strawberry	aardbei	seaweed	zeewier	squid	inktvis	spider	spin
sugar beet	suikerbiet	soybean	soja	starfish	zeester	tapeworm	lintworm
tomato	tomaat	spinach	spinazie	tadpole	kikkervisje	tarantula	tarantula
walnut	walnoot	turnip	raap	trout	forel	termite	termiet
watermelon	watermeloen	watercress	waterkers	whale	walvis	wasp	wesp

Appendix, Part 2: Categories and Items from Hampton et al. (2006) Along with Their Dutch Translation Used in the Current Study

SPORTS	SPORTEN	SCIENCES	WETENSCHAPPEN	TOOLS	WERKTUIGEN	FURNITURE	MEUBELS
aerobics	aerobics	advertising	advertising	axe	bijl	ashtray	asbak
ballroom dancing	salondansen	agriculture	landbouw	broom	bezem	bed	bed
billiards	biljart	archaeology	archeologie	calculator	rekenmachine	book	boek
bridge	bridgen	architecture	architectuur	dictionary	woordenboek	bookends	boekensteun
bullfighting	stierengevecht	astrology	astrologie	funnel	trechter	bucket	emmer
chess	schaken	astronomy	sterrenkunde	hammer	hamer	chair	stoel
conversation	praten	chemistry	scheikunde	key	sleutel	curtains	gordijnen
croquet	croquet	criminology	criminologie	pen	balpen	cushion	poef
crosswords	kruiswoordpuzzelen	dentistry	tandheelkunde	photograph	foto	desk	bureau
darts	darts	economics	economie	pitchfork	hooivork	dishwasher	vaatwasmachine
fishing	vissen	geography	geografie	rake	hark	door mat	mat
frisbee	frisbee	geometry	geometrie	scalpel	scalpel	lamp	lamp
hiking	trektocht	literature	literatuur	scissors	schaar	painting	schilderij
hunting	jagen	mathematics	wiskunde	screw	schroef	piano	piano
jogging	joggen	medicine	geneeskunde	screwdriver	schroevendraaier	pillow	kussen
kite flying	vliegeren	meteorology	meteorologie	sewing needle	naald	plate	schotel
mountaineering	bergbeklimmen	mineralogy	mineralogie	shovel	schop	refrigerator	koelkast
picnicking	picknicken	nutrition	voedingsleer	stone	steen	rug	vloerkleed
skiing	skiën	palm reading	handlezen	string	koord	shelf	schap
surfing	surfen	pharmacy	farmacie	toothbrush	tandenborstel	suitcase	aktentas
swimming	zwemmen	philosophy	filosofie	tractor	tractor	table	tafel
tennis	tennis	psychology	psychologie	trunk	koffer	telephone	telefoon
weightlifting	gewichtheffen	religious studies	godsdienstleer	umbrella	paraplu	television	televisie
wrestling	worstelen	sociology	sociologie	varnish	vernis	waste basket	vuilbak

References

Alston, W. P. (1964). *Philosophy of language*. Englewood Cliffs: Prentice-Hall.

Barsalou, L. W. (1993). Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A. F. Collins, S. E. Gathercole, M. A. Conway, & P. E. Morris (Eds.), *Theories of memories* (pp. 29–101). East Sussex, UK: Lawrence Erlbaum Associates.

Black, M. (1937). Vagueness: An exercise in logical analysis. *Philosophy of Science, 4*(4), 427–455.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121–152.

Chomsky, N. (1980). *Rules and representations*. Oxford, UK: Basil Blackwell.

De Deyne, S., Voorspoels, W., Verheyen, S., Navarro, D. J., & Storms, G. (2014). Accounting for graded structure in adjective categories with valence-based opposition relationships. *Language and Cognitive Processes, 29*(5), 568–583.

Denissen, J. J. A., Geenen, R., van Aken, M. A. G., Gosling, S. D., & Potter, J. (2008). Development and validation of a Dutch translation of the Big Five Inventory (BFI). *Journal of Personality Assessment, 90*(2), 152–157.

Devos, F. (1995). Still fuzzy after all these years. A linguistic evaluation of the fuzzy set approach to semantic vagueness. *Quaderni di Semantica, 16*(1), 47–82.

Devos, F. (2003). Semantic vagueness and lexical polyvalence. *Studia Linguistica, 57*(3), 121–141.

De Witte, E., Satoer, D., Robert, E., Colle, H., Verheyen, S., Visch-Brink, E., et al. (2015). The Dutch linguistic intraoperative protocol: A valid linguistic approach to awake brain surgery. *Brain and Language, 140,* 35–48.

Dry, M. J., & Storms, G. (2010). Features of graded category structure. *Acta Psychologica, 133*(3), 244–255.

Égré, P. (2017). Vague judgment: A probabilistic account. *Synthese, 194*(10), 3837–3865.

Charchat Fichman, H., Santos Fernandez, C., Alves Lourenço, R., Martins de Paiva Paradela, E., Carthery-Goulart, M. T., & Caramelli, P. (2009). Age and educational level effects on the performance of normal elderly on category verbal fluency tasks. *Dementia & Neuropsychologia, 3*(1), 49–54.

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement, 47*(4), 432–457.

Gardner, R. W. (1953). Cognitive styles in categorizing behavior. *Journal of Personality, 22*(2), 214–233.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Beverly Hills, CA: Sage.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18*(4), 441–461.

Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language, 34*(5), 686–708.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65*(2–3), 137–165.

Hampton, J. A. (2006). Concepts as prototypes. *The Psychology of Learning and Motivation: Advances in Research and Theory, 46,* 79–113.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science, 31*(3), 355–384.

Hampton, J. A., Dubois, D., & Yeh, W. (2006). The effects of pragmatic context on classification in natural categories. *Memory & Cognition, 34*(7), 1431–1443.

Hampton, J. A., & Passanisi, A. (2016). When intensions don't map onto extensions: Individual differences in conceptualization. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 42*(4), 505–523.

Kamp, H. (1975). Two theories about adjectives. In E. L. Keenan (Ed.), *Formal semantics of natural language* (pp. 123–155). Cambridge, UK: Cambridge University Press.

Keefe, R. (2000). *Theories of vagueness*. Cambridge, UK: University Press.

Kennedy, C. (2013). Two sources of subjectivity: Qualitative assessment and dimensional uncertainty. *Inquiry, 56*(2–3), 258–277.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology, 68*(8), 1665–1692.

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and Philosophy, 4*(1), 1–45.

Kölbel, M. (2004). Faultless disagreement. *Proceedings of the Aristotelian Society* (Vol. 104, pp. 53–73).

Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of work recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance, 39*(3), 802–823.

Lewellen, M. J., Goldinger, S. D., Pisoni, D. B., & Greene, B. G. (1993). Lexical familiarity and processing efficiency: Individual differences in naming, lexical decision, and semantic categorization. *Journal of Experimental Psychology: General, 122,* 316–330.

Loonstra, A. S., Tarlow, A. R., & Sellers, A. H. (2001). COWAT metanorms across age, education, and gender. *Applied Neuropsychology, 8*(3), 161–166.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS: A Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*(4), 325–337.

Machina, K. F. (1976). Truth, belief and vagueness. *Journal of Philosophical Logic, 5*(1), 47–78.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition, 6*(4), 462–472.

Paunonen, S. V., & Ashton, M. C. (2001). Big five predictors of academic achievement. *Journal of Research in Personality, 35*(1), 78–90.

Raffman, D. (1994). Vagueness without paradox. *Philosophical Review, 103*(1), 41–74.

Raffman, D. (1996). Vagueness and context-relativity. *Philosophical Studies, 81*(2–3) 175–192.

Raffman, D. (2014). *Unruly words*. Oxford, UK: Oxford University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rocklin, T. (1994). Relation between typical intellectual engagement and openness: Comment on Goff and Ackerman (1992). *Journal of Educational Psychology, 86*(1), 145–149.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573–605.

Rosenzweig, M. R. (1964). Word associations of French workmen: Comparisons with associations of French students and American workmen and students. *Journal of Verbal Learning and Verbal Behavior, 3*(1), 57–69.

Sassoon, G. W. (2012). A typology of multidimensional adjectives. *Journal of Semantics, 30*(3), 335–380.

Smith, L. B., & Samuelson, L. K. (1997). Perceiving and remembering: Category stability, variability and development. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 161–195). East Sussex, UK: Psychology Press.

Stukken, L., Verheyen, S., & Storms, G. (2013). Representation and criterion differences between men and women in semantic categorization. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 3474–3479). Austin, TX: Cognitive Science Society.

Verheyen, S., Ameel, E., & Storms, G. (2011a). Overextensions that extend into adolescence: Insights from a threshold model of categorization. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2000–2005). Austin, TX: Cognitive Science Society.

Verheyen, S., De Deyne, S., Dry, M. J., & Storms, G. (2011b). Uncovering contrast categories in categorization with a probabilistic threshold model. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 37*(6), 1515–1531.

Verheyen, S., Dewil, S., & Égré, P. (2018). Subjectivity in gradable adjectives: The case of *tall* and *heavy*. *Mind & Language*. In Press.

Verheyen, S., Droeshout, E., & Storms, G. (2018). *Age-related degree and criteria differences in semantic categorization*. Ms. University of Leuven and École Normale Supérieure.

Verheyen, S., Hampton, J. A., & Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the threshold model. *Acta Psychologica, 135*(2), 216–225.

Verheyen, S., & Storms, G. (2013). A mixture approach to vagueness and ambiguity. *PLoS ONE, 8*(5), e63507.

Verheyen, S., Voorspoels, W., & Storms, G. (2015). Inferring choice criteria with mixture IRT models: A demonstration using ad hoc and goal-derived categories. *Judgment and Decision Making, 10*(1), 97–114.

White, A., Storms, G., Malt, B. C., & Verheyen, S. (2018). Mind the generation gap: Differences between young and old in everyday lexical categories. *Journal of Memory and Language, 98,* 12–25.

Wright, C. (1995). The epistemic conception of vagueness. *The Southern Journal of Philosophy, 33*(S1), 133–159.

Zee, J., Storms, G., & Verheyen, S. (2014). Violations of the local independence assumption in categorization. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1670–1675). Austin, TX: Cognitive Science Society.

Intensification, Gradability and Social Perception: The Case of *totally*

Andrea Beltrama

Abstract The notion of *social meaning* has been widely investigated in sociolinguistic research (Eckert, Journal of Sociolinguistics, 12(4):453–76, 2008); yet, it is rarely considered in experimental semantics, mainly due to the assumption that this type of meaning is relatively independent from the semantic properties of its carrier. Following a recent strand of inquiry (Acton and Potts, Journal of Sociolinguistics, 18(1):3–31, 2014; Glass, Selected papers from NWAV 43, 2015; Jeong and Potts, Proceedings of SALT, 26, 1–22, 2016), this paper contributes to filling this gap by exploring the role of semantic and pragmatic factors in determining the salience of the social meaning of the intensifier *totally*. Relying on a social perception task, it is shown that listeners perceive the social meaning of this expression—measured in terms of Solidarity and Status attributes—as particularly prominent in situations in which the morpheme combines with a commitment scale provided by the pragmatics, as opposed to when it combines with a scale lexically supplied by the subsequent predicate. This evidence suggests that listeners keep track of semantic information when making social evaluations about speech, pointing to social perception as a novel methodology for research in experimental semantics.

Keywords Intensification · Social meaning · Adjectives · Social perception
Variation · Markedness

1 Introduction: What is Social Meaning?

Scholars in semantics and pragmatics have focused on *meaning* as the conventional content associated with linguistic forms, enriched with inferences drawn from the linguistic and non-linguistic context. In other domains of linguistics, however, the notion of meaning has been investigated under a completely different light. In particular, sociolinguists and linguistic anthropologists pursue the investigation of *social meaning* (Eckert 1989), that is, the cloud of socio-psychological qualities that expressions convey about language users, which typically range from demographic traits

A. Beltrama (✉)
University of Konstanz, Egg, Universitaetsstrasse 10, 78464 Konstanz, Germany
e-mail: andrea.beltrama@uni-konstanz.de

E. Castroviejo et al. (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure*,
Language, Cognition, and Mind 4, https://doi.org/10.1007/978-3-319-77791-7_7

(e.g., gender, age) to more local, idiosyncratic categories (e.g., "Jocks", "Burnouts", "Yuppies" and similar. See Agha 2005; Podesva 2007 for further discussion). A typical example of social meaning is represented by the use of raised diphthongs in the island of Martha's Vineyard, investigated in a seminal study by Labov (1963). As tourism from the mainland came to undermine fishing as the main source of income in the local economy, fishermen from the Vineyard began to centralize the pronunciation of diphthongs /ay/ and /aw/ to a much greater extent than any other occupational group on the island, thus turning a generic geographical marker of the island's dialect into a resource to convey more specific ideological meanings such as "locality", "loyalty", and resistance against the looming socio-economic transformations.

Despite the common label, semantic and social content have typically been seen as pertaining of independent domains. Even though they can be both seen as *bits* of content that attach to linguistic forms, a number of empirical observations justify this divide. First, the two types of meaning do not attach to the same units: sounds, for example are devoid of semantic meaning, and yet often carry a rich cloud of social meanings (e.g., Martha's Vineyard or the association of full releases of /t/ with attributes like "articulate", "prissy", "educated". See Campbell-Kibler (2007) for further discussion). Second, semantic and social meaning have a different semiotic status. While the former is conventionally associated with linguistic forms, the latter is only indirectly *indexed* by them (Silverstein 2003), emerging as more contingent and perspective dependent. Third, while semantic meaning is relatively fixed within a speech community and impermeable to the influence of extra-linguistic factors, social meaning is deeply affected by the broader social, cultural and ideological context, as the discussion above made clear.

At the same time the fact that social meaning has a more fluid nature and is heavily affected by non-linguistic factors does not mean that it lacks systematicity, or that it is blind to the language internal properties of its carriers. Quite the contrary, studies focusing on different phenomena and methodologies have pointed to a principled interaction between the perception of social meaning and language structure and processing. In particular, it has been shown that listeners keep track of fine-grained acoustic or syntactic properties when constructing social evaluations about language users (Squires 2013; Staum Casasanto 2008; Bender 2000); that social meaning plays an important role in speech perception (e.g., Niedzielski 1999; Campbell-Kibler 2010; D'Onofrio 2015); and that the social meanings conveyed by phonological variables—e.g., the association between monophthongized diphthongs and the Southerns states of the US—survive negation and other environments in which at-issue meanings are normally suppressed, making a case for grouping social meaning together with other types of non-at-issue meaning traditionally investigated in semantics and pragmatics (Smith et al. 2010). In this paper, I aim to expand on these investigations to explore whether the perception of social meaning is constrained, or at least affected, by fine-grained semantic and pragmatic properties of linguistic expressions. Specifically, I focus on the following question: can the compositional mechanism whereby an expression is interpreted affect the expression's suitability to serve as a vehicle of social meaning?

2 *Totally*: A Promising Case Study

Intensifiers, and more broadly modifiers that target scalar dimensions, emerge as a promising test case for this question. On the one hand, they tend to be used more frequently by speakers with specific socio-demographic characteristics (e.g., young and female: Ito and Tagliamonte 2003; Tagliamonte 2008 among others), emerging as a powerful linguistic resource to convey social meanings. Crucially, sociolinguists have been extremely careful to point out that the social meaning of a linguistic form neither boils down to, nor is it directly caused by such correlations between frequency of use and such demographic traits (Ochs 1992; Eckert 2008, 2012 among others). Quite the contrary, the relationship between them is a dialectic one: social meanings of a form feed off of the form's variability across the demographic space; but patterns of use in turn, are constantly reinforced and retransformed by the social meanings that speakers assign to the form.[1] The takeaway is that, while the nexus between patterns of use identity categories and social meanings is a complex one, the life of social meanings is crucially tied to the presence of socially-conditioned linguistic variation, making intensification a good testing ground for the investigation of this type of content.

On the other hand, intensifiers present considerable variation on the semantic front as well, as they appear to be able to contribute their meaning through different compositional mechanisms, including: direct binding of the degree argument supplied by a gradable predicate (Heim 2000; Kennedy and McNally 2005 among others); manipulation of scales grounded in the contextual parameters of interpretation of the expression (Anderson 2013; Beltrama and Bochnak 2015; McNabb 2012); modification of gradable epistemic/emotive attitudes held by the speaker towards the propositional content (Giannakidou and Yoon 2011; Morzycki 2011; Bochnak and Csipak 2014). The empirical richness of these expressions on both the compositional semantic and the sociolinguistic front opens up the possibility of investigating whether a principled relationship links these two dimensions. The present paper explores this issue by focusing on the intensifier *totally*.

2.1 Totally *and Social Meaning: A Preliminary Look*

Let us begin by considering the following example accompanying the picture of a cap whose bill has been designed to resemble Donald Duck's beak.[2]

(1) I *totally* had this hat as a child …The bill *totally* quacks when you squeeze it.

Even a cursory look is sufficient to observe that the use of *totally* in this particular context comes with a somewhat rich social meaning. First, it conveys a flavor of marked informality and reduced social distance, suggesting that the interlocutors are

[1] I am grateful to an anonymous reviewer for bringing this issue to my attention.

[2] https://instagram.com/p/zEZEQQqYPY/.

close to one another, share a set of norms or values and easily agree on the content of the conversation. Besides these effects, the intensifier additionally conveys a set of social attributes about the social identity of its typical users, which track macro-social categories—e.g., young age—as well as more specific personae and social types. This emerges in the following excerpt from the website Urban Dictionary,[3] a popular repository of social stereotypes that can be used to have a preliminary grasp on the social significance of specific linguistic expressions in the sociocultural context of North America. While these commentaries are not sufficient to have an exhaustive map of the social content conveyed by a linguistic variable—they merely reflect those attributes that are stereotypical enought to undergo explicit circulation in a community—they already point to a rather rich constellation of social attributes, highlighting *totally* as a highly salient social meaning carrier.

1. It's a word used by ditzy young girls that means definitely or for sure.
2. Valley Girl Speak that means "Of course!"
3. A word used by girly girls, poppers, and rich spoiled little brats.
4. A word used for emphasis. Makes you sound kinda cheerleaderish when you use it.

2.2 **Totally:** ***Semantic/pragmatic Meaning***

On the semantic and pragmatic front, *totally* likewise presents a rich empirical picture. On a general level, the intensifier combines with a bounded scale and requires that the scalar maximum on such a scale be reached.[4] It is precisely in the way in which this scale is supplied that variation enters the picture. In standard cases, the scale is provided by the following predicate as in (2): both *full* and *agree* come with a bounded ordering hardwired in their lexical meaning, providing *totally* with an argument to operate on. I refer to these cases as *lexical totally*.

(2) a. The bus is *totally* full. Lexical
b. She *totally* agrees with me. Lexical

In other cases (in (3)), though, *totally* combines with predicates that do *not* supply a scale operating on the *commitment* that the speaker has towards the proposition (McCready and Kaufmann 2013; Irwin 2014). I refer to these cases as *speaker-oriented totally*.

[3] http://www.urbandictionary.com/define.php?term=totally.

[4] Authors have put forward different proposal to model this meaning—-see Kennedy and McNally (2005) for a degree-based approach and Toledo and Sassoon (2011), Sassoon and Zevakhina (2012) for a non-degree-based one among others. The formalization of the contribution of the modifier in this use is orthogonal to the aims of the current paper, and I will therefore remain agnostic as to whether a degree-based or a non-degree based approach is to be preferred.

(3) a. You should *totally* click on that link! It's awesome.[5] Speaker-oriented
b. A dude *totally* walked off a train, threw his shit down & camped out.[6] Speaker-oriented

Despite sharing reference to maximality, the speaker-oriented usage of *totally* is empirically distinct from the lexical one. First, because it does not combine with a lexical scale it cannot be replaced by modifiers like *partially* and *almost* (in (4a)). Second, it contributes its meaning at the non at-issue level, as shown by the fact that it resists being embedded under questions or negators and it cannot be challenged independently from the rest of the propositional content.[7]

(4) a. * You should *partially/almost* click on that link! It's awesome.
b. *You shouldn't *totally* click on that link.
c. *Should he totally click on that link?
d. She should *totally* click on that link!
B: # **No!** She should click on that link, but you're not committed to saying that!

As far as the exact nature of the contribution of *totally* is concerned, I have proposed in my previous work an analysis of the intensifier as a Common Ground managing operator, whereby the speaker expresses the belief that there should be no option other than adding the anchor proposition to the Common Ground. More precisely, adopting the conversational model of Farkas and Bruce (2010) I have argued that *totally* signals that, in view of the speaker's conversational goals, all worlds in possible Common Grounds projected by the assertion (i.e., its Projected Set) should be worlds in which the proposition is true. On this view, speaker-oriented *totally* operates as a universal quantifier over sets of worlds. As such, while it does not target a scale in the traditional sense it still targets a gradient domain of sort, all the while retaining a common semantic core with the maximizing function of the lexical version Beltrama (2018).

An important corollary of this proposal is that, while the intensifier is generally paraphrased with epistemically flavored adverbs like *definitely* or *certainly*.[8] *Totally* is crucially different from these operators, which are instead grounded in private individual certainty of the speaker towards the truth of the proposition. While subtle this difference is empirically substantiated by the lack of interchangeability between *totally* and these adverbs in certain contexts. Let us consider, for example contexts in which the intensifier is used out of the blue such as (3b), reproduced below as (5a).

[5] https://www.facebook.com/TheBiscuitGames/posts/488916347870627 accessed on June 5th 2015.

[6] http://bartidiothalloffame.com/dude-totally-walked-off-a-train-threw-his-shit-down-camped-out-embaracado-station/.

[7] From this perspective it shows compositional properties similar to other expressions that specify the attitude of the speaker such as expressives (Potts 2005b), certain evidentials (Faller 2002; Murray 2014; Rett and Murray 2013); and other speaker-oriented adverbs, (see Ernst 2009).

[8] The OED added a dedicated entry in 2005: "In weakened use as an intensifier: (modifying an adjective) very, extremely; (modifying a verb) definitely, absolutely."

(5) a. A dude *totally* walked off a train, threw his shit down & camped out.

b. # A dude *certainly/definitely* walked off a train, threw his shit down & camped out.

The content of the sentence is inherently bizarre as it describes a fact that strongly deviates from our background assumptions about the world—e.g. people do not normally camp on train platforms. By pushing for the addition of the proposition to the Common Ground, the presence of *totally* serves as a strategy for the speaker to preemptively address the potential skepticism of the listener, who would have good reasons to question the update on the ground of its low plausibility. On the other hand, without a previous discourse move that openly raises an issue as to whether the proposition is actually true pure epistemic operators like *certainly* and *definitely* sound remarkably less natural than *totally*.

Finally, note that, while it is normally straightforward to distinguish between the speaker-oriented and the lexical version of the intensifier, the boundary between the two uses is less clearcut in particular contexts. This is observed, for instance in occurrences of the intensifier with *extreme* adjectives (Morzycki 2012)—e.g., *awesome amazing*. These adjectives, while gradable do not lexicalize a bounded scale. As such, the presence of *totally* in their proximity is predicted to instantiate the speaker-oriented version of the intensifier. Yet, we observe that, when *totally* modifies these adjectives, it is considerably less deviant than standard cases of speaker-oriented *totally* when we run the diagnostics discussed above.[9] While I will remain agnostic throughout the paper as to what semantic/pragmatic factors are behind this empirical observation, the somewhat murkier status of speaker-oriented *totally* with extreme adjectives will be important to one of the hypotheses that will be laid out with respect to the mapping between the intensifier's semantic and social meaning.[10]

(6) a. Bob is totally awesome

b. ? Bob is {*not totally/almost totally/completely/entirely*} awesome.

c. ? Bob is *almost totally* awesome.

d. ? Is Bob *totally* awesome?

e. ? Bob is *completely/entirely* awesome.

3 From Semantic to Social Meaning: Hypotheses

In light of the discussion above the flexibility of *totally* in terms of the type of targeted scale provides a window into the relationship between mechanisms of semantic

[9] The symbol ? indicates a minor degree of deviance.

[10] A possible departure point for an explanation could be rooted in the fact that extreme adjectives themselves pattern somewhat in between relative and absolute ones, as extensively discussed by Morzycki (2012). By referring to properties with an inherently high degree for example they could make it easier for the listener to coerce their open scale into a bounded one as suggested by Paradis (2000).

composition and social meaning, raising the following question: does *totally*'s suitability to convey social information about language users change depending on whether the intensifier targets a lexical or a speaker-oriented scale? Before making a hypothesis about the relationship between these two components, it is first necessary to consider what proposals have been made in the literature to capture the relationship between social meaning and linguistic features on a broader level. I first review the extant literature in this area, focusing on the role of markedness as a bridge between linguistic expressions and the salience of social meaning and then proceed to formulate two hypotheses on the behavior of *totally*.

3.1 Linguistic Constraints on Social Meaning: The Role of Markedness

3.1.1 Frequentistic Markedness

Since Wolfram (1969) various studies pointed out a positive correlation between markedness and the *salience* of social meaning, observing that, by virtue of their heightened noticeability, marked variants are better designed for conveying social meaning than their unmarked counterparts. Concerning the exact characterization of the notion of noticeability, most investigators link it to the violation of frequentistic expectations that is associated with the use of marked forms, which therefore stand out as particularly surprising for the hearer.[11] In a foundational study, Bender (2000)' shows that "copula deletion" in African American English is perceived as more strongly associated with African American ethnic identity in environments in which this particular variant is least frequent—eg. before an NP. Conversely, the perceived intensity of the social meaning decreases in the environments in which copula deletion is more frequent, hence less marked (i.e., before auxiliary verbs), unveiling a principled connection between syntactic environments, frequency of use and the salience of the relevant social meaning. The contrast is exemplified in (in 7), where Ø represents the absence of an overt auxiliary.

(7) a. She Ø a nurse.
b. I don't think John Ø gonna make it.

Similar arguments have been provided for the social meanings carried by phonological variables. Podesva (2011) for example shows that rising contours in declarative sentences, by virtue of being considerably less frequent than falling ones, emerge as a suitable resource for doctors to convey a variety of social meanings, including concern and attentiveness towards the patient; conversely, Jeong and Potts (2016) show that questions asked with falling intonation, the least frequent tune for this

[11] Campbell-Kibler (2007) for example suggests that "it is likely that those variants which depart more strongly or unexpectedly from a listener's customary experience are more apt to be noticed and assigned meaning than those which differ only slightly".

speech act, convey an especially rich package of social information. By the same token, Callier (2013) provides evidence that creaky voice in mid-phrasal position, a linguistic context where it is less frequent, is perceived more negatively than in phrase-final position. Finally, Grinsell and Thomas (2012) show that the modal *finna* in African American English is a particularly prominent index of ethnic identity when occurring with inanimate subjects and atelic predicates, that is, in a linguistic environment to which it has extended only recently, and in which the expression is still highly infrequent. Taken together, these studies constitute an important step towards understanding of how social meaning is linguistically constrained. At the same time these investigations focus on a very specific type of linguistic factor. By reducing the language-internal properties of an expression to its frequentistic distribution, they cannot assess whether more inherent features of linguistic forms—e.g., those pertaining to their semantics, syntax or pragmatics—also contribute to determine the suitability to serve as a carrier of social meaning.

3.1.2 Pragmatic Markedness

Such issues could not be addressed in the investigations discussed above for a simple reason: these studies are concerned with units with either no independent syntactic/semantic structure—e.g. sounds—or with a basic meaning unambiguously shared across different variants—e.g. copula deletion—making frequency the only linguistic dimension along which the variants at stake significantly vary. In order to verify whether social meaning can also be constrained by non-frequentistic types of markedness, one has to focus on cases of variation in which the variants do have independent and fully fleshed semantic and pragmatic content, thus providing another layer of linguistic structure that can impact social meaning salience. Two studies, in particular, have undertaken this direction.

Building on an observation by Lakoff (1974), Acton and Potts (2014) argue that demonstratives like *this* and *that* index a sense of "emotional closeness between speaker and hearer" (p. 351) when compared to run-of-the-mill determiners like *your* or *the*. The contrast is exemplified by the following minimal pair:

(8) a. That left front tire is pretty worn. Lakoff (1974: ex. 32)
b. Your left front tire is pretty worn. Lakoff (1974: ex. 33)

In a similar fashion, Glass (2015) contrasts deontic modals like *got/have to* which merely expresses obligation in light of a set of circumstances/body of law to *need to* which additionally conveys that the obligation is good for the hearer's well-being, arguing that the latter conveys an additional social component of care or presumptuousness. Both studies echo previous sociolinguistic literature in suggesting that the notion of markedness remains the guiding principle of the mapping between linguistic forms and social meaning. Yet, they show that the correlation between the salience of social meaning and the markedness of the variant need not be framed in purely frequentistic terms, but can also be grounded in basic pragmatic principles.

More specifically, both demonstratives and *need* emerge as marked with respect to functionally similar competitors that vie for the same slot, and yet provide a simpler semantic contribution: *the/your* for demonstratives; and *have to/got* for *need*. As such, both forms exemplify Horn (1984)'s principle of the *division of pragmatic labor* according to which, if two forms have the same referential content and different degrees of complexity, the more complex tends to be restricted to non-stereotypical situations, in which the communicated content extends beyond the bare literal meaning of the expression. Within this view, demonstratives are even more marked in contexts in which a determiner could have been left out altogether, thus emerging as completely unnecessary for referential purposes. Proper names provide a clear example of this.[12]

(9) That Henry Kissinger sure knows his way around Hollywood!

In such contexts, the proper noun already identifies a unique referent. As such, the presence of the demonstrative is especially redundant, thus acquiring especially high potential to become socially meaningful through the contrast with the semantically equivalent demonstrative-free alternative that could have been used instead. The authors observe that it is in these contexts that demonstratives are more frequently used by politicians (and, in particular, Sarah Palin) as a stylistic resource to foster a sense of proximity with the listener, indicating that, once again, markedness is functional to the operation of highlighting social meaning.

3.2 **Totally** ***and markedness asymmetries***

The emerging picture is one in which linguistic markedness—both in its frequentistic and pragmatic notion—provides a non-social criterion that can set apart suitable and less suitable linguistic carriers of social content, illuminating how the circulation of social meaning can be parasitic on forces that are endemic to the linguistic system, and not just grounded in the socio-cultural ideological landscape. It now becomes possible to consider the specific case of *totally* with a focus on the following question: can the semantic variation that characterizes the intensifier allow us to make prediction concerning the social salience of the different uses of *totally*? As discussed above the two basic variants of the morpheme differ in terms of the dimension that they target: lexical *totally* quantifies over the degrees supplied by the lexical meaning of a bounded predicate; speaker-oriented *totally* quantifies over a scale of pragmatic commitment that is grounded in the speaker's attitude towards the content of the assertion. I argue that this distinction at the semantic level does indeed correspond to a markedness asymmetry between the two uses of *totally* thus leading us to make a clear prediction about what we should expect in terms of social meaning salience.

[12] The example only holds for languages like English, where proper names do not require a determiner. The same social effects are not predicted to hold, instead, for languages that grammatically require the presence of a determiner in this context, such as Greek, even though, to my knowledge the prediction has not been tested scientifically thus far.

On the one hand, lexical *totally* modifies a property within the propositional content, restricting the interpretation of the modified predicate in a non-trivial fashion. Let us consider the example below:

(10) John's personality is different from Katie's personality.

(11) John's personality is *totally* different from Katie's personality.

Totally crucially increases the informativity of the utterance changing the truth conditions of the proposition. While (10) is satisfied whenever the two personalities are at least slightly different from one another, (11) is only satisfied in a scenario in which the two personalities have no overlapping whatsoever. While in a less obvious fashion, the same applies to situations in which lexical *totally* occurs next to maximum standard adjectives such as *full*.

(12) The glass is *full*.

(13) The glass is *totally* full.

While different proposals have been put forward to model the meaning of the intensifier in this environment, what is common across them is that *totally* affects at the very least the *extension* of the modified predicate thus affecting the propositional content of the utterance in a non-trivial fashion. Under certain accounts (Sassoon and Zevakhina 2012; Toledo and Sassoon 2011), *totally* has been analyzed as an operator that *widens* the comparison class of the predicate. As such, it shifts upwards the standard that we use to determine whether the adjective holds true or not, strengthening the interpretation and affecting its truth conditions. Under other accounts (Kennedy and McNally 2005; Kennedy 2007), *totally* has been claimed not to change the truth conditions of the predicate at least in a strict sense. On this view, *full* already encodes maximality when occurring in its positive form. Yet, even under such accounts the modification by *totally* nevertheless makes the interpretation of the predicate more restrictive. By excluding those "close-enough" cases that, as part of *full*'s pragmatic halo (Lasersohn 1999), would count as true in the positive form, the intensifier crucially changes the extension of the predicate thus bringing about a significant effect on the informativity of the utterance even if it does not technically change its truth-conditions.

The same does not apply to speaker-oriented *totally*. First, this version of the intensifier does not affect the propositional content, as shown by the fact that it operates on an independent compositional tier (see Sect. 2.2). Second, the contribution of speaker-oriented *totally* is already part of the sincerity conditions of every assertion. Barring obviously defective contexts of communication, the assertion of a proposition is in fact by default accompanied by the commitment of adding *p* to the Common Ground.[13] Under this view, speaker-oriented *totally* appears to express a pragmatic move that already underlies the speech act that it modifies. As such, the very same

[13] In a more general sense all cooperative interlocutors are working towards the goal of enriching the amount of mutual knowledge coordinating their moves to maximize the number of propositions that they mutually accept as true (Stalnaker 1978).

message could have been conveyed by an utterance withouth *totally* resulting in the minimal pair below:

(14) A: Is your name Emily?
 a. B: Yes, it's *totally* Emily.
 b. B: Yes, it's Emily.

The contrast between (14a) and (14b) exemplifies a case of markedness based on the division of pragmatic labor, where (14a) is an utterance that, *ceteris paribus* could have been made in a simpler way. As such, the use of *totally* with speaker-oriented scales emerges as a inherently salient: the morpheme adds to the complexity of the utterance while not making any additional contribution to what would have been conveyed without its presence. The seemingly redundant of *totally* in this context thus creates the conditions for the emergence of "extra" meanings. It imbues the presence of the intensifier with special social and pragmatic significance highlighting the social content that it contributes.

3.3 **Totally** *Scale Type and Social Meaning Salience: Hypotheses*

In light of this discussion, speaker-oriented *totally* emerges as a more suitable linguistic resource to convey the social identity of its users than lexical *totally* leading us to predict a correlation between the salience of the intensifier's social meaning and the availability of a lexical scale in the linguistic context. More specifically, I hypothesize that the intensifier should be more likely to be interpreted as a social marker when it occurs in contexts that make no bounded scale lexically available thus making a speaker-oriented interpretation the only possible one. Conversely, the social meaning should be less salient when *totally* combines with a bounded gradable predicate and thus can receive a lexical interpretation.

(15) **Hypothesis 1**: *Totally* is more likely to be interpreted as a carrier of social meanings when it targets a speaker-oriented rather than a lexical scale.

If this hypothesis is confirmed, the question emerges as to whether the salience of social meaning reflects the gradience of the distinction between the two semantic variants of *totally* discussed in the end of Sect. 2.2. If this is the case I hypothesize that, with extreme adjectives, the social meaning of *totally* should have intermediate salience between the lexical and the speaker-oriented use given the fact that, while a bounded lexical scale is not available it can be easily coerced.

(16) **Hypothesis 2**: The social meaning of *totally* should be most salient for clear cases of speaker-oriented *totally*; least salient for clear cases of lexical *totally*; and intermediate with *extreme adjectives*.

I test these hypotheses via a social perception experiment.

4 The Experiment

4.1 Methods

Experimental methods have long been used to investigate language attitudes in social psychology. An especially popular technique in particular, has been the *matched guise* task, first introduced by Lambert et al. (1960) (see Campbell-Kibler 2007 for an overview of the literature). This particular design consists of the collection and measurement of the reactions and attitudes of listeners towards instances of language use manipulated by the researcher to test the effect of a particular independent variable. Despite their popularity in other fields, it is not until the last ten years that these methods have been systematically applied to test sociolinguistically-related questions (see Campbell-Kibler 2010; Drager 2013 for further details). A crucial assumption of this method is that social evaluation is a proxy into the social meaning of the variable as it allows us to have access to "what social information listeners can extract from the speech of particular speakers, and which linguistic cues they rely on to do so" (Campbell-Kibler 2010). This method has two important advantages for our purposes. First, it provides the opportunity of manipulating the type of scale targeted by *totally* in different sentences while leaving the rest of the proposition unchanged, allowing us to isolate scale type as the only changing factor across conditions. Second, by providing a way to measure the intensity of social meaning in terms of a series of evaluative scales, it allows us to detect at a fine-grained level how the perception of the social meaning changes as a function of the semantic/pragmatic features of *totally*. As such, it represents a viable methodology to test questions about the linguistic factors that constrain the perception of social meaning, just like the one addressed in this paper.[14]

4.1.1 Building Test Scales

As the first step, I conducted a preliminary study to construct the evaluation scales to be used to measure social meaning in the actual experiment. The study was designed with the software Qualtrics and subsequently circulated on Amazon Mechanical Turk. 60 subjects, who self-declared to be native speakers of American English and between 18 and 35 years old, were recruited and paid $ 0.50 for participating. First, each subject saw in written a sentence containing an instance of lexical *totally* and one of speaker-oriented *totally*. The order in which the two instances was randomized, so that each subject saw either an instance of lexical or speaker-oriented *totally* first.

[14] An obvious disadvantage of this methodology, by contrast, is that it is less ecologically faithful than other techniques for data collection (e.g., ethnography). In particular, it has been suggested by sociolinguists that social meaning is a complex semiotic entity that cannot be separated from the other linguistic and non-linguistic *practices* through which humans interact and make sense of the world (Eckert 2000). As such, investigating it through the lens of a set of attributes that rate speech samples in isolation obviously comes with a price in terms of empirical simplification.

For each sentence each subject was asked to provide four adjectives to describe its imagined speaker by filling out blank spaces on a computer screen. Based on the most recurring adjectives in the responses, a total of eight social attributes describing the listener were selected as particularly salient in connection with the use of the intensifier: four of them are predicted to be positively affected, while four of them are instead predicted to be negatively affected by the presence of *totally*. I label these sets of dimensions Solidarity and Status attributes respectively, using them as the dimensions of social evaluation to tap into the social meaning of *totally*.

- **Solidarity**: Friendliness, Coolness, Outgoingness, Excitability
- **Status**: Articulateness, Maturity, Intelligence Seriousness[15]

For clarification purposes, it is important to observe that adopting the Solidarity and Status categories is primarily motivated by the convenience of having a conventional term that uniquely identifies each class of social evaluation scales, while at the same time connecting with the labels commonly used in the literature on social perception studies and language attitudes (Lambert et al. 1960; Campbell-Kibler 2010 among others). Thus, while most of the scales can indeed be seen as loosely related to either social proximity (i.e. solidarity) or social distance (i.e., status), I do not intend to make a specific commitment to claiming that each of the attributes related to these notions in a strict sense.

4.1.2 Stimuli

Two factors were crossed in a 3 × 4 design. The first factor manipulates the semantic variant of *totally* along the lexical vs speaker-oriented axis of variation by presenting the intensifier in combination with three distinct classes of adjectives. To cue lexical *totally* the intensifier was used next to *(maximum standard) absolute adjectives* (Kennedy and McNally 2005), which lexicalize a bounded scale as part of their lexical meaning (e.g., “bald”). To cue the speaker-oriented reading, instead, open-scale *relative* adjectives (e.g., “tall”), which offer a commitment scale as the only possible target for the intensifier. In addition, *extreme adjectives* (e.g., “awesome”) were used as an intermediate case between the two other categories. I predict that *totally* affects the social perception of the speaker of the sentence in the following way (Table 1). In the other factor, the type of modifier accompanying the adjective came in four different conditions: the target intensifier, *totally*; two control intensifiers, *really* and *completely* and the positive non-intensified form. On the one hand, *completely* contrary to *totally* is exclusively able to target lexical scales. As such, it cannot modify speaker-oriented scales, resulting in ungrammaticality when used with an

[15] Note that, for building the scales predicted to be negatively affected by the intensifier, I took into consideration both adjectives referring to a high degree of their antonym and adjectives negating the quality itself. For example the decision to adopt “Intelligent” as a dimension negatively impacted by *totally* was motivated by subjects entering both “unintelligent” and “dumb” as descriptors of the speaker in the pilot.

Table 1 Critical conditions and predictions

Adjective type	Bounded scale availability	Markedness of totally	Social meaning salience
Absolute	✓	Low	Low
Extreme	≈	Medium	Medium
Relative	No	High	High

open-scale adjective. On the other hand, *really* has a less selective semantics than *totally*. It does not require the availability of an upper-bounded scale but, as discussed in the semantics literature can modify any type of scalar predicate (McNabb 2012; Constantinescu 2011). Since all the adjectives used in the experiment are indeed scalar, the intensifier should always operate at the lexical level, showing no semantic difference across the adjective types. In light of these properties, I predict that, if an effect of the semantic type of *totally* is observed on the social meaning, the same effect should not be observed on the two control intensifiers. Finally, as I discuss below, the positive form serves as a baseline condition to assess the contribution of each intensifier to the social meaning. Having this contrast is necessary to filter out any effect on social meaning that is contributed by other elements in the sentence such as the adjectives themselves. 12 items, each with a different set of adjectives, were crossed in a Latin Square Design.[16] Table 2 provides a full paradigm for an item across all conditions.

4.1.3 Procedure and Statistical Analysis

Every subject saw a total of 12 written sentences, one sentence for each condition.[17] Each sentence was followed by a series of questions aimed at assessing solidarity-based and non-solidarity-based traits of social meaning discussed above. They were presented in the form of a 1-6 Likert scale where 1 indicated the minimum value and 6 the maximum value. Subjects were explicitly instructed to answer the questions following their instincts and to be very honest and straightforward, even if they felt compelled to provide a particularly negative judgments of the speaker. A full list of the questions, together with the possible answers, is reported below.

(17) **Sentence**: I just met the new boss. He’s totally bald.

1. How **articulate** does the speaker sound? 1 ……6
2. How **mature** does the speaker sound? 1 ……6
3. How **intelligent** does the speaker sound? 1 ……6
4. How **serious** does the speaker sound? 1 ……6

[16] See Appendix for full set of experimental items.

[17] Due to the high number questions following each item, no fillers were used so as to avoid overwhelming subjects throughout the study and help them stay focused at all times.

Table 2 A full item

Adj type	Int type	Sentence
Absolute	Totally	I just met the new boss. He's *totally* **bald**
Extreme	Totally	I just met the new boss. He's *totally* **awesome**
Relative	Totally	I just met the new boss. He's *totally* **tall**
Absolute	Ø	I just met the new boss. He's **bald**
Extreme	Ø	I just met the new boss. He's **awesome**
Relative	Ø	I just met the new boss. He's **tall**
Absolute	Completely	I just met the new boss. He's *completely* **bald**
Extreme	Completely	I just met the new boss. He's *completely* **awesome**
Relative	Completely	I just met the new boss. He's *completely* **tall**
Absolute	Really	I just met the new boss. He's *really* **bald**
Extreme	Really	I just met the new boss. He's *really* **awesome**
Relative	Really	I just met the new boss. He's *really* **tall**

5. How **friendly** does the speaker sound? 1 ……6
6. How **outgoing** does the speaker sound? 1 ……6
7. How **cool** does the speaker sound? 1 ……6
8. How **excitable** does the speaker sound? 1 ……6

The study was created with Qualtrics and carried out online. 36 self-declared native speakers of American English, age 18–35, were recruited on Amazon Mechanical Turk and compensated $2 for their participation. For statistical analysis, mixed-effects models were ran for each attribute with the R statistical package *lmer4* (Bates and Walker 2015). The fixed effect predictors included Adjective and Intensifier and their interactions, and the random effects included at least random intercepts for subjects and items. When a higher-level main effect or interaction was significant, I followed up with planned paired-comparisons between the relevant conditions. In light of the experimental questions, I am especially interested in comparing each intensifier with the base form of the adjective. This would allow me to assess if, and how, each intensifier affects the social meaning for each of the adjective types.

Table 3 Mixed effect model summary for Solidarity attributes

Factor	Excitable		Outgoing		Friendly		Cool	
	F-value	p-value	F-value	p-value	F-value	p-value	F-value	p-value
Intensifier	4.7	<0.001	1.4	–	0.6	–	3.5	<0.05
Adjective	11.0	<0.0001	7.3	<0.0001	1.4	–	6.1	<0.001
Adj:Int	3.4	<0.05	2.3	<0.05	2.5	<0.05	2.3	<0.05

(18) Planned comparisons:

a. {Totally/Really/Completely} Rel Adj versus Bare Rel Adj

b. {Totally/Really/Completely} Ext Adj versus Bare Ext Adj

c. {Totally/Really/Completely} Abs Adj versus Bare Abs Adj.

4.2 Results

For both Solidarity and Status attributes I report the summary of the main effect and interactions in a dedicated table (Tables 3 and 5).[18] I then report the results of the planned comparisons in a separate table for Solidarity and Status attributes.

4.2.1 Solidarity

Table 3 reports the summary of the mixed effects models for the Solidarity attributes. For all attributes, an interaction between Intensifier and Adjective was found, reflecting the fact that *totally* with relative adjectives is perceived as higher in solidarity. In addition, a main effect of Adjective was found for Excitable Outgoing and Cool. Finally, a main effect of Intensifier was found for Excitable and Cool.

I now focus on the specific contrasts between intensified forms and the base forms, which allow us to gauge the effect of *totally completely* and *really* in the different linguistic environments in which they were tested. Table 4 reports the differences between the perception of the sentence with the intensifier and the perception of the sentence with the base form for the corresponding adjective type. Results for *totally* are in bold face. Other significant contrasts between intensifier and base form are indicated with *, with threshold for significance set at $p < .05$.

For all attributes, *totally* with relative adjectives was perceived as significantly higher than the corresponding base forms. No significant contrasts are found for *totally* with extreme adjectives or absolute adjectives. With the latter, however, *totally*

[18] Whether it is desirable to generate *p* values for fixed effect models has been widely discussed recently within the R community. For reporting purposes, the *p* values were generated with the function summary(aov(model)).

Table 4 Perception for Solidarity attributes: differentials

Attribute	Relative				Extreme				Absolute			
	Base	Tot	Com	Rea	Base	Tot	Com	Rea	Base	Tot	Com	Rea
Exc	3.51	*+**0.61**	+0.25	+0.01	3.80	+0.34	+0.42	−0.08	3.19	+0.54	−0.05	−0.08
Out	3.65	*+**0.74**	+0.26	+0.26	4.34	−0.09	+0.14	−0.48	3.80	+0.05	−0.39	+0.05
Fri	3.68	*+**0.65**	+0.37	+0.34	4.20	−0.26	−0.03	−0.23	3.94	+0.00	−0.44	+0.17
Cool	3.02	*+**0.85**	+0.06	−0.02	3.45	−0.03	+0.26	−0.40	2.97	+0.17	−0.18	+0.00
Avg	3.47	*+**0.72**	+0.23	+0.14	3.95	−0.01	+0.20	−0.30	3.47	+0.19	−0.26	+0.04

Table 5 Mixed effet model summary for status attributes

Factor	Articulate		Mature		Intelligent		Serious	
	F-value	p-value	F-value	p-value	F-value	p-value	F-value	p-value
Intensifier	6.0	<0.001	10.0	<0.0001	8.8	<0.01	10.9	<0.001
Adjective	1.6	–	3.7	<0.05	4.3	<0.05	3.4	<0.05
Adj:Int	3.1	<0.01	1.8	–	2.0	–	1.3	–

displays a trend to raise the solidarity perception, which is particularly evident with Excitability. Concerning the other intensifiers, no systematic contrast is observed that holds across all the attributes. It can be observed, though, that *completely* with absolute adjectives tend to lower the perception of solidarity.

4.2.2 Status

Table 5 reports the summary of the mixed effects models for the Status attributes. For all attributes, a main effect of Intensifier was found, with *totally* being associated with lower Status perception than the other conditions. A main effect of Adjective was found for Mature Intelligent and Serious, with absolute adjectives being rated higher than extreme and relative ones. Finally, an interaction between Intensifier and Adjective is found for Articulate.

As was done for Solidarity attributes, I now focus on the specific contrasts between intensified forms and the base forms. Table 6 reports the differences between the perception of the sentence with the intensifier and the perception of the sentence with the base form for the corresponding adjective type. Results for *totally* are in bold face. Significant contrasts between intensifier and base form are indicated with *.

For all attributes, *totally* with relative adjectives and with extreme adjectives is perceived as significantly lower than the corresponding base forms. No significant contrasts are found for *totally* with absolute adjectives, even though *totally* displays a trend to decrease the perception with these predicates as well. Concerning the other intensifiers, no significant contrast is observed across all the attributes. Yet, we observe that *completely* with relative adjectives displays a marked trend to derease the perception with relative adjectives, with effects that near significance (all $ps < 0.1$). At the same time we note that *completely* with absolute adjectives displays a trend to raise the status perception, featuring an effect that goes in the opposite direction to the one observed for the other adjective types. No effect is observed for *really*.

Table 6 Perception scores for status attributes: differentials

Att	Relative				Extreme				Absolute			
	Base	Tot	Com	Rea	Base	Tot	Com	Rea	Base	Tot	Com	Rea
Art	3.68	*−**0.87**	−0.54	+0.23	3.91	*−**0.86**	−0.26	−0.14	3.47	+0.03	+0.55	−0.02
Mat	3.68	*−**0.93**	−0.54	+0.11	4.05	*−**1.20**	+0.43	−0.50	3.77	−0.42	+0.31	−0.03
Int	3.60	*−**0.84**	−0.37	+0.34	4.00	*−**1.03**	−0.35	−0.37	3.77	−0.19	+0.17	+0.08
Ser	4.22	*−**1.01**	−0.55	−0.14	4.25	*−**1.15**	−0.31	−0.12	4.22	−0.31	+0.00	+0.00
Avg	3.80	*−**0.90**	−0.50	+0.13	4.05	*−**1.08**	−0.37	−0.28	3.81	−0.23	+0.26	+0.01

4.3 Discussion

4.3.1 *Totally*

The current study aims to investigate how the social perception of *totally* is affected by variations in the semantic properties of the intensifier across different linguistic contexts. Two hypotheses were tested. First, I predicted that instances in which *totally* targets a speaker-oriented scale should be more likely to be interpreted as carriers of social meaning than cases of lexical *totally*. The prediction is confirmed for all the attributes used in the experiment: when *totally* occurs next to an unbounded adjective an environment in which only a speaker-oriented reading is licensed, listeners perceive the intensifier as a salient marker of social identity along eight different dimensions. By contrast, when *totally* occurs next to an absolute adjective and a lexical reading is possible the intensifier does not significantly impact the social evaluation of the sentence.

The second hypothesis tested whether the salience of the social meaning would reflect the continuum in the distinction between lexical and speaker-oriented uses, predicting that the social meaning should have intermediate intensity with extreme adjectives. This prediction, however, is not borne out, as we observe that for none of the tested dimensions a continuum along these lines emerges. Quite the contrary, the social perception of *totally* in this environment is polarized across different dimensions of evaluation. Concerning Solidarity, *totally* has no effect, leaving the perception unchanged from the positive unintensified form. Concerning Status, the effect of *totally* is instead comparable to—in fact, even stronger than—the one observed for relative adjectives. A possible explanation of this result could be that Extreme adjectives, by virtue of referring to properties that are already instantiated to a very high degree tend to come with a considerable emotive charge. As such, they turn out to feature a remarkably high value on their own along Solidarity attributes—as Table 5 shows, the average Solidarity difference between the bare forms of these adjectives and the bare form of relative and absolute adjectives is 0.48, a much wider gap than the one observed for Status attributes—thus masking the independent contribution of *totally*.[19]

The emerging picture is one in which, by and large the (lack of) availability of a lexical scale correlates with the social salience of *totally* in a given context, suggesting a connection between the semantic and the social components of the meaning conveyed by the intensifier. At the same time the data from the experiment do not present conclusive evidence as to whether the social meaning of *totally* is a gradient phenomenon. In other words, is it the case that only the speaker-oriented use has a distinctive independent social meaning, while the lexical use lacks it altogether? Or is it the case that *totally* has the same underlying social meaning across both variants,

[19] At the same time it must be noted that the Solidarity mean ratings of the positive form of extreme adjectives is still near the middle of the scale rather than being skewed towards the top. As such, it would be hasty to explain the lack of Solidarity effects on *totally* in terms of a ceiling effect of the bare forms. I am grateful to an anonymous reviewer for directing my attention to this observation.

which is just more salient in the speaker-oriented use and less salient in the lexical one? Possible evidence in favor of the former alternative is that we fail to observe the presence of cases with intermediate social salience as extreme adjectives could have been. On the other hand, evidence in support of the latter is that lexical *totally* features a trend that mirrors the effect observed for its speaker-oriented counterpart, even though the effect is not large enough to reach statistical significance. This suggests that, even in unmarked environments, the intensifier might still be associated with a similar social evaluation, although of much lower perceptual salience. Note however, that the trend observed on the evaluation of lexical *totally* could also be due to a different reason, which further complicates the current picture. In particular, it might be the case that certain subjects gave a speaker-oriented interpretation also to occurrences of *totally* with absolute adjectives, which are indeed in principle ambiguous between the two uses. For example a sentence like "John is *totally* bald" could be taken to either mean that John has zero hair left, or that the speaker is maximally committed to asserting the proposition that John is bald, with the two readings being possibly truth-conditionally distinct. While it is reasonable to expect that, in the absence of further information about the broader discourse context, the lexical interpretation should have been considerably more easily accessible to the subjects than the speaker-oriented one the availability of the latter leaves open the possibility that the weak social effects observed on absolute adjectives could be due to some subjects assigning *totally* the marked interpretation in this environment as well.[20] In sum, while the body of evidence provided by the experiment points to a connection between the semantic and social meaning of *totally* it is not sufficient to make a case either in favor or against the idea that such a social meaning is gradient in nature. I thus leave the exploration of this question to further research.

4.3.2 *Completely* and *Really*

Concerning the effect of the control intensifiers, no systematic pattern emerges. As predicted, *really* has a minor impact on all the evaluation scales and presents no significant difference across the tested adjective types. Concerning *completely* we also observe that the intensifier does not change the perception of the sentence with the positive form in a systematic way. This, at the very least, suggests that the effect observed for *totally* with relative adjectives is not due to a mismatch in scalar structure or to the perception of the construction as ungrammatical. If that were the case we should expect to observe the same effect on *completely* which however we don't. At the same time it is worth observing that *completely* closely approximates the effect of *totally* on relative adjectives, nearing statistical significance especially with respect to Status attributes. This finding raises the question as to why *completely* displays

[20] As an anonymous reviewer suggests, a possible way to further explore the factor(s) driving the social perception of *totally* with absolute adjectives would be to use intonation to disambiguate between the lexical and the speaker-oriented reading, and verify how this impacts the social perception of the intensifier.

a trend that is not featured by *really*. I propose two alternative explanations, which can be explored in further research. One possibility is that the effects of scale type on *completely* are grounded in the ungrammaticality of the combination, rather than to the particular semantic properties of the expression. Saying "completely tall", in other words, amounts to saying something that is located outside of the grammatical knowledge of the speakers, and then evokes whichever social features are associated with a "default other" who does not fully master the grammar of English.[21] The other possibility is that *completely* is also on the way of grammaticalizing a speaker-oriented meaning similar to the one of *totally*. As such, it begins to display the same markedness effects of *totally* even though it is not deep enough in the grammaticalization trajectory to trigger such effects as consistently as *totally*.[22] This hypothesis would fit in well with the observation that the shift from the lexical to the speaker-oriented domain is rather common for maximizers across languages (see Hoeksema 2011; Tribushinina and Janssen 2011 on Dutch *helmaal*).

5 Taking Stock

In this Sect. 1 take stock of the experimental results from a broader angle returning to the original question that informed this article: How can the social meaning of an expression be constrained by its semantic/pragmatic properties?

5.1 Scalarity and Social Meaning Salience

The experimental findings indicate that the salience of the social meaning associated with *totally* co-varies with the semantic/pragmatic properties of the intensifier. While the presence of speaker-oriented *totally* significantly impacts the social perception of a sentence the presence of lexical *totally* has a much weaker effect. To explain this result, I have argued that speaker-oriented *totally* is a suitable candidate to convey social meaning in virtue of its status as a marked variant. By pragmatically evoking a simpler, semantically equivalent alternative utterance that could have been used in its substitution, this use of the intensifier is naturally equipped to strike the listener's attention as a noticeable linguistic choice. As such, on a par with what has been observed for other socially meaningful expressions, it is associated with a language-internal mechanism that makes it apt to be assigned "extra" meanings besides its regular semantic/pragmatic ones, including those pertaining to the social dimension. On the other hand, as a consequence of its semantics, lexical *totally* does not sufficiently stand out in terms of markedness. It operates within the propositional content

[21] I thank an anonymous NWAV 44 reviewer for suggesting this explanation.

[22] I thank E. Allyn Smith and Tim Leffel for suggesting, separately and (almost) simultaneously, this explanation.

of the utterance failing to invoke the contrast with a simpler alternative. As such, this version of *totally* does not have the inherent salience that marked expressions carry, failing to draw the listener's attention in the way in which its speaker-oriented counterpart does. The emerging picture is one in which, through the mediation of markedness, fine-grained semantic properties like the different types of scales targeted by an intensifier can affect the perception of social meaning. This, in turn, provides preliminary evidence that, when making social evaluations about linguistic form, listeners keep track of the semantic/pragmatic properties of these forms, suggesting that these two types of meaning, while empirically distinct, are also connected in a principled fashion. While a broader empirical basis is necessary to further test this claim, it is interesting to observe that, at first glance intensifiers with similar semantic/pragmatic flexibility to the one shown by *totally* seems to display also a comparable association with social meaning. A particularly relevant example is *so* which can modify gradable predicates (in (19b)) as well as pragmatic attitudes related to speaker commitment (in (19a, see Zwicky 2011; Irwin 2014; Potts 2005a for further discussion).

(19) a. We are *so* going to lose game tonight.
b. John is *so* tall.

In this sense it is quite revealing that the attitudinal variant of *so* has indeed been informally described as a salient carrier of social meanings in comparison to its lexical counterpart, as suggested by labels such as "Generation X so" (Zwicky 2011) or "Drama so" (Irwin 2014). Such an association provides encouraging, if provisional, evidence that the socio-semantic properties of *totally* might be shared with other intensifiers, thus highlighting the domain of scalarity and gradability as a highly promising venue to continue research on this topic (see Beltrama (2016) for evidence from the Italian suffix *-issimo*).

5.2 *Lingering Questions*

As the systematic investigation of the interface between semantic and social meaning has just begun, a number of questions remain open to further investigation. With respect to the particular phenomenon of intensification, I would like to point out two.

5.2.1 Why *That* Social-Meaning?

First, while we have an understanding of why *totally* becomes associated with *some* social meaning, are we also in the position of explaining why it is associated with *that* particular social meaning just by looking at its semantic and pragmatic profile? In other words, why does it emerge as an index of high solidarity and low status? Providing a complete answer solely on the basis of the linguistic properties appears to be an ambitious task. It is well known that the outcome of any enregisterment process is heavily driven by extra-linguistic ideological and historical factors. As Agha (2005) suggests, the social recognition of linguistic features as indexes of speakers qualities is the result of a continuous process of circulation, renegotiation and reanalysis, which cannot be pre-determined by the sheer linguistic features of these forms. Yet, the question remains as to whether such features, besides rendering certain expressions a more or less suitable site for the emergence of social meaning, can also have any effects on the particular type of indexical content that becomes associated with them. I have suggested elsewhere that a possible route to cast light on this issue would be to consider more carefully the attitude conveyed by *totally* in its speaker-oriented use Beltrama (2018). More specifically, there could be a qualitative connection between the solidarity boosting effect of *totally* at the social level and the interpersonal convergence that is associated with the act of emphasizing commitment to adding a proposition to the Common Ground. In other words, given the intersubjective nature of the commitment targeted by *totally* the use of the morpheme could serve as a pragmatic tool to foster agreement and convergence between the interlocutors, thus resulting in the association of the users of *totally* with social qualities that highlight inclusiveness and proximity at the social level. Under this view, the commitment to involving the interlocutor in the construction of the Common Ground percolates up to the more durable categories of social identity, contributing to indexing users of *totally* as kind of persons that are likewise committed to fostering inclusion and proximity at the social level. If this were true it would then be possible to posit a *constitutive* relationship between *totally* and some of the social attributes, opening up another dimension of interaction between semantic and social content (see Ochs 1992; Moore and Podesva 2009; Acton and Potts 2014; Glass 2015 for similar proposals).

5.2.2 The Role of Diachronic Innovations

Second, I would like to briefly elaborate on a possible objection that could be moved to the proposed conclusion of the study: What if social indexicality is not foregrounded by the semantic properties of the variable per se but, more simply, by the fact that speaker-oriented *totally* emerged at a later diachronic state than lexical *totally* and is therefore more easily associated with the social features of typical language innovators? Under this view, the association with different degrees of social meaning salience would just be an accident of language change relatively independent from the pragmatics. A way to respond to this observation would be the following: if recency were the only driving factor, we should expect very little difference between speaker-oriented *totally* and *totally* with extreme adjectives.[23] And yet, as shown

[23] A search on the Corpus of Historic American English (COHA, Davies 2010) shows that, while lexical *totally* has been around since the beginning of the 20th century (and, incidentally, also well before), the intensifier in the other two contexts emerged fairly recently, and almost simultaneously. While the attestation of the first occurrence with extreme adjectives predates the first attestation of speaker-oriented *totally* by 20 years, the very low number of occurrences of both contexts in the corpus suggests some caution in taking such a 20 year gap as significant.

in the experiment, the two variants do not behave in the same way. This suggests that, while the current study does not preclude the possibility that recency effets played a role in shaping the social meaning, an account entirely based on diachronic innovation would need to be supported by stronger evidence.

To provide a more definitive response it would be possible to run a follow up study that exclusively focuses on speaker-oriented *totally*. The following contrast suggests a potentially promising environment to test the hypothesis.

(20) **John**: I can't remember if Luke got married at 25. Doubt about *p*
Mark: Yes, he totally got married at 25.

(21) a. A man totally got off the train, threw his shit down and camped out.
b. Iowa senator totally thinks you should be drug tested for child support payments.

In (20), the incongruence between the interlocutor's and the speaker's view provides an explicit justification for the act of stressing commitment to adding the proposition to the Common Ground on the part of John. By contrast, other contexts, for example (21), present no such clue: here *totally* is used with assertions that describe objective facts and do not address doubts or questions from the interlocutors. As such, the use of *totally* in contexts like (21) appears to be even more marked than the use in contexts like (20): the act of stressing commitment is not openly called for by the discourse structure but stems from the outlandish/surprising content of the assertion. If markedness is the crucial factor driving the salience of the social meaning, marked cases of speaker-oriented *totally* should therefore be more socially meaningful than unmarked ones, providing a neat empirical ground to argue against the idea that the social meaning differences between lexical and speaker-oriented *totally* are entirely driven by a historical contingency. Evidence supporting this idea is discussed in Beltrama (2016).

6 Conclusion

While representing a preliminary step, the current study opens up a novel area of research on the study of meaning, highlighting the interface between social and semantic content as a ripe and largely uncharted, domain of investigation. This line of research, if adequately developed, carries two important implications. On the theoretical level, it can lead us to adopt a more comprehensive view of linguistic meaning, in which social meaning is seen as a *bona fide* type of content to be investigated side by side with the logical and pragmatic properties of expressions. On a methodological level, it points to social perception studies as a promising technique to explore the behavioral correlates of semantic and pragmatic features, expanding the toolbox for the experimental investigation of meaning.

Acknowledgements I would like to thank Ivano Caponigro, Annette D'Onofrio, Penny Eckert, Itamar Francez, Susan Gal, Emily Hanink, Anastasia Giannakidou, Julian Grove Sunwoo Jeong, Chris Kennedy, Dan Lassiter, Tim Leffel, Alda Mari, Costas Nakassis, Rob Podesva, Chris Potts, Teresa Pratt, Michael Silverstein, Laura Staum Casasanto, E. Allyn Smith, Ming Xiang and the other linguists at the University of Chicago and Stanford University who provided insightful comments on different parts of this project. I would also like to thank the audience from the ExpGrad Workshop in Madrid, and in particular Elena Castroviejo, Louise McNally and Galit Sassoon for organizing the conference and including my contribution in this volume. All errors and oversights are my own.

Appendix: Experimental Materials

1. Someone found a bottle of wine on the street. It was {totally/really/very/∅} {big/gigantic/full}.
2. The drive from New York to Chicago is {totally/really/very/∅} {long/awful/flat}.
3. Compared to Atlanta, Portland is {totally/really/very/∅} {small/astonishing/quiet}.
4. I just met the new boss. He's {totally/really/very/∅} {tall/awesome/bald}.
5. I met John's brother. He's {totally/really/very/∅} {young/gigantic/different from him}.
6. We jump in it and ...the water was {totally/really/very/∅} {cold/freezing/frozen}.
7. Traveling on the 4th July weekend is {totally/really/very/∅} {pricey/great/ unaffordable}.
8. Dad finally found a picture of his wedding, but it's {totally/really/very/∅} {small/ridiculous/blurry}.
9. The ice cover on the lake is {totally/really/very/∅} {thin/massive/safe} right now.
10. Take a look at this story. It's {totally/really/very/∅} {deep/amazing/absurd}.
11. Biking from school to the train station is {totally/really/very/∅} {fast/creepy/safe}.
12. The walk home from here is {totally/really/very/∅} {short/gorgeous/straight}.

References

Acton, E., & Potts, C. (2014). That straight talk. Sarah Palin and the sociolinguistcs of demonstratives. *Journal of Sociolinguistics*, *18*(1), 3–31.

Agha, A. (2005). Voice, footing, enregisterment. *Journal of Linguistic Anthropology*, *15*, 38–59.

Anderson, C. (2013). Inherent and coerced gradability across categories: Manipulating pragmatic halos with *sorta*. In T. Snider (Ed.), *Proceedings of Semantics and Linguistic Theory* (Vol. 23, pp. 81–96).

Bates, D., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Beltrama, A. (2016). *Bridging the gap: Intensifiers between semantic and social meaning*. Dissertation, University of Chicago.

Beltrama, A. (2018). *Totally* between subjectivity and discourse. Exploring the pragmatic side of intensification. *Journal of Semantics*. ffx021. https://doi.org/10.1093/semant/ffx021

Beltrama, A., & Bochnak, M. R. (2015). Intensification without degrees cross-linguistically. *Natural Language and Linguistic Theory*, *33*(3), 843–879.

Bender, E. (2000). *Syntactic variation and linguistic competence: The case of AAVE copula absence*. Dissertation, Stanford University.

Bochnak, R., & Csipak, E. (2014). A new metalinguistic degree morpheme. In T. Snider, S. D'Antonio, & M. Weigand (Eds.), *Proceedings of Semantics and Linguistic Theory* (Vol. 24, pp. 432–452).

Callier, P. (2013). *Linguistic context and the social meaning of voice quality variation*. Dissertation, Georgetown University.

Campbell-Kibler, K. (2007). Accent, (ing) and the social logic of listener perceptions. *American Speech*, *82*(1), 32–84.

Campbell-Kibler, K. (2010). New directions in sociolinguistic cognition. In *University of Pennsylvania Working Papers in Linguistics* (Vol. 15.2, pp. 31–39).

Constantinescu, C. (2011). *Gradability in the nominal domain*. Dissertation, Leiden University.

Davies, M. (2010–). The corpus of contemporary American English: 450 million words, 1990–2012. http://corpus.byu.edu/coca/.

D'Onofrio, A. (2015). Persona-based information shapes linguistic perception: Valley girls and california vowels. *Journal of Sociolinguistics*, *19*(2), 241–256.

Drager, K. (2013). Experimental methods in sociolinguistics. In J. Holmes & K. Hazen (Eds.), *Research methods in sociolinguistics: A practical guide* (pp. 58–73). Malden, MA: Wiley-Blackwell.

Eckert, P. (1989). *Jocks and burnouts: Social identity in the high school*. New York: Teachers College Press.

Eckert, P. (2000). *Language variation as social practice*. Oxford: Blackwell.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453–76.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of variation. *Annual Review of Anthropology*, *41*, 87–100.

Ernst, T. (2009). Speaker-oriented adverbs. *Natural Language and Linguistic Theory*, *27*(3), 497–544.

Faller, M. (2002). *Semantics and pragmatics of evidentials in cuzco quechua*. Dissertation, Stanford University.

Farkas, D., & Bruce, K. B. (2010). On reacting to assertions and polar questions. *Journal of Semantics*, *27*(1), 81–118.

Giannakidou, A., & Yoon, S. (2011). The subjective mode of comparison: Metalinguistic comparatives in Greek and Korean. *Natural Language and Linguistic Theory*, *29*, 621–655.

Glass, L. (2015). Need to vs. have to and got to: Four socio-pragmatic corpus studies. In *Selected papers from New Ways of Analyzing Variation 43* (Vol. 21.2, pp. 79–88).

Grinsell, T., & Thomas, J. (2012). *Finna* as a socially meaningful modal in African American English. Talk presented at the 48th meeting of the Chicago Linguistic Society. University of Chicago, Chicago.

Heim, I. (2000). Degree operators and scope. In B. Jackson & T. Matthews (Eds.), *Proceedings of Semantics and Linguistic Theory* (Vol. 10, pp. 40–64).

Hoeksema, J. (2011). Discourse scalarity: The case of Dutch *helemaal*. *Journal of Pragmatics*, *43*(11), 2810–2825.

Horn, L. (1984). Towards a new taxonomy of pragmatic inference: Q-based and r-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11–42). Washington: Georgetown University Press.

Irwin, P. (2014). So [totally] speaker-oriented: An analysis of "Drama SO". In R. Zanuttini & L. R. Horn (Eds.), *Microsyntactic Variation in North American English* (pp. 29–70). Oxford: Oxford University Press.

Ito, R., & Tagliamonte, S. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in english intensifiers. *Language in Society*, *32*(2), 257–279.

Jeong, S., & Potts, C. (2016). Intonational sentence-type conventions for perlocutionary effects: An experimental investigation. In M. Moroney, C.-R. Little, J. Collard, & D. Burgdorf (Eds.), *Proceedings of Semantics and Linguistic Theory* (Vol. 26, pp. 1–22).

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

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

Labov, W. (1963). The social motivation of a sound change. *Word*, *18*, 1–42.

Lakoff, R. (1974). Remarks on 'this' and 'that'. In *Proceeding of the Chicago Linguistic Society* (Vol. 10, pp. 345–356). Chicago, IL: Chicago Linguistic Society.

Lambert, W. E., Hodgson, R. E., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken language. *Journal of Abnormal and Social Psychology*, *60*(1), 44–51.

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

McCready, E., & Kaufmann, M. (2013 November 29). *Maximum intensity*. Paper presented at the Semantics Workshop, Keio University.

McNabb, Y. (2012). Cross-categorial modification of properties in Hebrew and English. In A. Chereches (Ed.), *Proceedings of Semantics and Linguistic Theory* (Vol. 22, pp. 365–382).

Moore, E., & Podesva, R. J. (2009). Style, indexicality, and the social meaning of tag questions. *Language in Society*, *38*(4), 447–485.

Morzycki, M. (2011). Metalinguistic comparison in an alternative semantics for imprecision. *Natural Language Semantics*, *19*(1), 39–86.

Morzycki, M. (2012). Adjectival extremeness: Degree modification and contextually restricted scales. *Natural Language and Linguistic Theory*, *30*(2), 567–609.

Murray, S. E. (2014). Varieties of update. *Semantics and Pragmatics*, *7*(2), 1–53.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Social Psychology (Special Edition)*, *18*(1), 62–85.

Ochs, E. (1992). Indexing gender. In A. Duranti & C. Goodwin (Eds.), *Rethinking context: Language as an interactive Phenomenon* (pp. 335–358). New York, Cambridge: Cambridge University Press.

Paradis, C. (2000). It's well weird. Degree modifiers of adjectives revisited: The nineties. In J. M. K. Pages (Ed.), *Corpora galore: Analyses and techniques in describing English* (pp. 147–160). Amsterdam & Atlanta: Rodopi.

Podesva, R. J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics*, *11*(4), 478–504.

Podesva, R. J. (2011). Salience and the social meaning of declarative contours: Three case studies of gay professionals. *Journal of English Linguistics*, *39*(3), 233–264.

Potts, C. (2005a). Lexicalized intonational meaning. In S. Kawahara (Ed.), *UMOP 30: Papers on prosody* (pp. 129–146). Amherst, MA: GLSA.

Potts, C. (2005b). *The logic of conventional implicature*. Oxford: Oxford University Press.

Rett, J., & Murray, S. E. (2013). A semantic account of mirative evidentials. In T. Snider (Ed.), *Proceedings from Semantics and Linguistic Theory* (Vol. 23, pp. 453–472). Ithaca, NY: CLC Publications.

Sassoon, G.W., & Zevakhina, N. (2012). Granularity shifting: Experimental evidence from degree modifiers. In A. Chereches (Ed.), *Proceedings of Semantics and Linguistic Theory* (Vol. 22, pp. 226–246). Ithaca, NY: CLC Publications.

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language and Communication*, *23*(3–4), 193–229.

Smith, E. A., Hall, K. C., & Munson, B. (2010). Bringing semantics to sociophonetics: Social variables and secondary entailments. *Laboratory Phonology*, *1*(1), 121–155.

Squires, L. (2013). It don't go both ways. limited bidirectionality in sociolinguistic perception. *Journal of Scoiolinguistics*, *17*(2), 200–237.

Stalnaker, R. (1978). Assertion. In *Syntax and semantics* (Vol. 9). New York: Academic Press.

Staum Casasanto, L. (2008). *Experimental investigations of sociolinguistic knowledge*. Dissertation, Stanford University.

Tagliamonte, S. A. (2008). So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language and Linguistics*, *12*(2), 361–394.

Toledo, A., & Sassoon, G. (2011). Absolute vs. relative adjectives—Variance within vs. between individuals. In N. Ashton, A. Chereches, & D. Lutz (Eds.), *Proceedings of Semantics and Linguistic Theory* (Vol. 21, pp. 135–154), Ithaca, NY: CLC Publications.

Tribushinina, E., & Janssen, T. (2011). Re-conceptualizing scale boundaries: The case of Dutch *helemaal*. *Journal of Pragmatics*, *43*(7), 2043–2056.

Wolfram, W. (1969). *A sociolinguistic description of Detroit Negro speech*. Center for Applied Linguistics (Washington).

Zwicky, A. (2011). Gen X So. http://arnoldzwicky.org/2011/11/14/genx-so/.

Perceived Informativity and Referential Effects of Contrast in Adjectivally Modified NPs

Helena Aparicio, Christopher Kennedy and Ming Xiang

Abstract Referential Effects of Contrast (RECs) involving reference resolution of adjectivally modified NPs (e.g., *the tall glass*) have been attributed to pragmatic reasoning based on the informativity of modification (Sedivy et al. Cognition, 71(2):109–147, 1999; Sedivy, Journal of Psycholinguistic Research, 32(1):3–23, 2003; Sedivy, Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions, MIT Press, Cambridge, MA, pp. 345–364, 2004, a.o.). Recently, it has been claimed that informativity alone cannot account for all the attested interactions between adjectival meaning and context and that factors related to efficiency in the search of a referent also play an important role (Rubio-Fernández, Frontiers in Psychology, 7(153), 2016). Building on Aparicio et al. (Proceedings of Semantics and Linguistic Theory, vol. 25, 2015), this paper demonstrates that perceived informativity plays an important role in RECs, but lexical semantic properties of different adjective classes are also relevant. We present results from a Visual World eye-tracking study which shows that adjective classes differ in whether they introduce RECs, and results from an offline judgment task which show that this difference correlates to some extent with the perceived informativity of members of these classes. Color adjectives, relative adjectives and maximum standard absolute adjectives were rated as overinformative when used as modifiers in the absence of contrast, and gave rise to RECs; minimum standard absolute adjectives were not rated as overinformative when used as modifiers in the absence of contrast, and did not give rise to RECs. Taken together, our results show that perceived informativity plays an important role in RECs. We also discuss

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-77791-7_8) contains supplementary material, which is available to authorized users.

H. Aparicio (✉)
Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: haparici@mit.edu

C. Kennedy · M. Xiang
University of Chicago, Chicago, IL, USA
e-mail: ck@uchicago.edu

M. Xiang
e-mail: mxiang@uchicago.edu

E. Castroviejo et al. (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure*,
Language, Cognition, and Mind 4, https://doi.org/10.1007/978-3-319-77791-7_8

additional differences between the adjective classes which suggest that differences in lexical semantics can further contribute to differences in RECs.

Keywords Gradable adjectives · Context-sensitivity · Informativity · Visual world · Referential effects of contrast · Online processing

1 Perceived Informativity and Referential Effects of Contrast

There exists ample evidence that listeners process linguistic input incrementally (Crain and Steedman 1985; Altmann and Steedman 1988; Eberhard et al. 1995, among many others), and that pragmatic information pertaining to different sources is quickly integrated during online processing (Hanna et al. 2003; Hanna and Tanenhaus 2004; Grodner and Sedivy 2011). For instance, in a classic eye-tracking study, Tanenhaus et al. (1995) showed that contextual visual information, introduced by the manipulation of the visual display, was immediately adopted by the listeners to guide their online parsing decisions. This experimental paradigm, which later came to be known as the Visual World (VW) paradigm, has proven especially sensitive in detecting effects of context during online processing. In VW eye-tracking experiments, participants' eye-movements are tracked as they look at arrays of objects while listening to an auditory instruction that typically requires them to visually identify an object in the display in order to perform the experimental task. Eye-movements are a particularly good measure of language processing in reference-resolution tasks because eye-fixations reflect with millisecond granularity what objects in the visual context are being considered as potential referents of the linguistic input (Cooper 1974; Eberhard et al. 1995; Tanenhaus et al. 1995; Pyykkönen-Klauck and Crocker 2016). Therefore, eye-movement patterns can be used to make inferences about whether and at what point of linguistic processing the information of the visual context becomes relevant.

Within Visual World studies, a hallmark of this rapid online integration of pragmatic information comes from *Referential Effects of Contrast* (henceforth RECs). The effect was initially reported by Sedivy et al. (1999) in a study investigating how properties of the visual context influenced the processing of NPs containing an attributive prenominal adjective like *tall*. In the experiment, participants heard instructions such us *'Pick up the tall glass'* while looking at displays of four objects. Two conditions were tested. A **Contrast** condition supported a contrasting interpretation of the adjective by including, alongside the target object (e.g., a tall glass), a contrast object that could be described by the noun but not the adjective in the instruction (e.g., a short glass). In the second condition, the **No-Contrast** condition, the contrasting object was substituted with a distractor, i.e. an object that could not be described either by the head noun or the modifier in the instruction. All trials contained a competitor object that presented a higher degree of the property in the

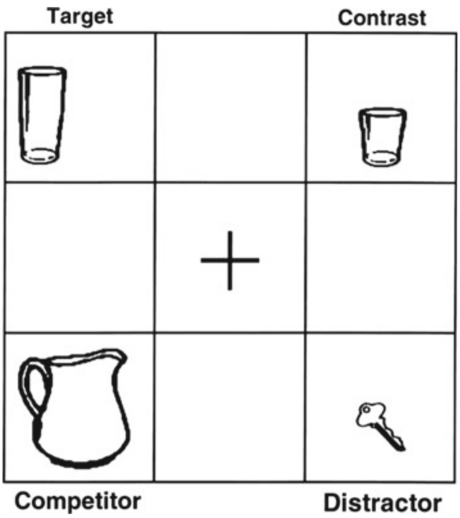


Fig. 1 Experimental set up used by Sedivy et al. (1999) (Contrast Condition)

instruction when compared to the target, but could not be felicitously described by the adjective (e.g., a pitcher that was taller than the glass, but was itself not tall for a pitcher, see Fig. 1).

The main finding of the experiment was that participants' fixations converged on the target faster in the Contrast condition than they did in the No-Contrast condition. Crucially, in the Contrast condition participants zoomed into the target object at a point in which the head noun had not yet been processed. Therefore, this decision was performed at a time in which the linguistic instruction was still ambiguous between the two objects that could be described by the adjective in the instruction (i.e., the target and the competitor), suggesting that the presence of the contrasting object was used very early.

Despite the fact that RECs have been consistently replicated with adjectivally modified NPs (Sedivy et al. 1999; Sedivy 2004; Weber et al. 2006; Grodner and Sedivy 2011; Wolter et al. 2011; Aparicio et al. 2015; Leffel et al. 2016), the exact mechanisms underlying these effects are not fully understood, and it remains an open question whether all the RECs reported in the literature are born equal (cf. Sedivy 2003, 2004). The crucial difference between the Contrast condition and the No-Contrast condition is that in the former, the visual display includes objects that contrast only with respect to the information provided by a noun modifier, not with respect to the information provided by the head noun; while in the latter all objects in the display contrast with respect to the information provided by the noun. This makes the use of a modifier non-contrastive or "redundant," since the head noun alone suffices to distinguish the intended referent from the other objects in the display. A referential contrast is observed when visual target identification takes place significantly faster in the Contrast condition compared to the No-Contrast condition. Such effects receive a natural pragmatic explanation in terms of the interaction of the Gricean Maxims of Quantity and Manner (Grice 1975). Since a definite description with a restrictive modifier is both more complex and more informative than a corresponding description without a modifier, a speaker's use of a modified form provides an indication that she intends to refer to an object that contrasts relative to the modifier but not the noun, which in turn facilitates referential fixation in the Contrast condition but not in the No-Contrast condition.

A naive version of the Gricean account of RECs would lead to the expectation that (cooperative) uses of modifiers should be restricted to contexts involving contrast; i.e., contexts in which the modifier is not redundant, in the sense described above. However, there is evidence that speakers frequently use modifiers in referential NPs even in the absence of contrast (Pechmann 1989; Nadig and Sedivy 2002; Sedivy 2003; Maes et al. 2004; Sedivy 2004; Koolen et al. 2011). Certain patterns seem to emerge in the use of such apparently redundant adjectives. Experimental production tasks have consistently shown that color adjectives are more likely to be used redundantly than other classes of adjectives like dimensional or material adjectives (Pechmann 1989; Belke and Meyer 2002; Nadig and Sedivy 2002; Sedivy 2004). Several factors have been found to be good predictors of when a speaker is more likely to use a redundant adjective. For instance, color adjectives that denote a stereotypical property of the object (e.g., a *yellow* banana) are less likely to be used redundantly (Sedivy 2003), while atypical color adjectives are more likely to be used redundantly (Westerbeek et al. 2015). A second factor affecting the production of redundant adjectives in referential communication tasks is the amount of variation present in the visual scene. Speakers are more likely to utter an overspecified description when the visual scene contains color variability, i.e. the visual display is polychrome, than when it does not, i.e., the visual display is monochrome (Koolen et al. 2013; Rubio-Fernández 2016).

The fact that speakers not only often choose to include overspecified adjectives as part of their utterances, but also do so in systematic ways is unexpected in the context of the naive Gricean view, in which all redundant adjectives are suboptimal from an informativity point of view. Rubio-Fernández (2016) suggests that overspecification should be recast in terms of efficiency rather than informativity, as modifiers may facilitate target identification by helping the hearer optimize the visual search of the target object (see Paraboni et al. 2007; Arts et al. 2011 for similar claims). In this respect, efficiency can be regarded as a pragmatic cooperative phenomenon. Assuming that hearers are sensitive to the systematicities in the production patterns of redundant adjectives, different adjective classes could in principle be associated with different expectations regarding the probability that a given adjective will be used contrastively. This is relevant for VW experiments such as the ones discussed above, as it leads to a more nuanced prediction than the naive Gricean view, namely that only those adjective classes for which a redundant adjective is *perceived* as providing too much information in the context should give rise to such effects, i.e. there should be a correlation between perceived overinformativity and strength of referential contrast. The resulting picture, like the naive Gricean one, remains rooted in reasoning about (over-)informativity of a complex form, but allows for variation in classes of modifiers based on the extent to which they are independently perceived as over-informative or not.

To test this hypothesis we conducted two experiments to explore the relation between RECs and perceived informativity. In Experiment 1 (Sect. 2), we extend a prior study of RECs in so-called "relative" versus "absolute" adjectives by Aparicio et al. (2015) to the class of "minimum standard" absolute adjectives. We show that minimum standard absolute adjectives fail to trigger RECs, in contrast to the relative and maximum standard absolute adjectives analyzed by Aparicio et al., as well as to color adjective controls. In Experiment 2 (Sect. 3), we compare all four classes of adjectives for perceived informativity, and show that minimum standard adjectival modifiers differ from all the other classes of adjectival modifiers in not being perceived as overinformative in the absence of contextual support for contrastive interpretations, in support of the perceived informativity-based view of RECs described above. However, among the other three classes of adjectives, we also found that the magnitude of the perceived (over)informativeness does not completely map to the size of the RECs reported in Aparicio et al. (2015). We conclude with discussion of the role that lexical semantic factors may play in driving perceived informativity and variable RECs.

2 Experiment 1: Variable RECs Across Adjective Classes

In a VW study modeled after Sedivy et al.'s (1999) design, Aparicio et al. (2015) examined RECs in definite descriptions containing modifiers from three classes of adjectives: relative adjectives, maximum standard absolute adjectives and color adjectives. (For general discussion of these adjectives and their semantic and pragmatic properties, see Unger 1975; Pinkal 1995; Rotstein and Winter 2004; Kennedy and McNally 2005; Kennedy 2007; McNally 2011.) Aparicio et al.'s decision to examine these adjectives was based on an interest in the potential role that different kinds of context dependence play in the interpretation of adjectives generally, and in the generation of RECs in particular. Relative adjectives (RelAs) such as *big, small, tall* and *short* are inherently context-sensitive, because their "threshold" for application can change across contexts. For example, the threshold for determining what individuals fall in the extension of the predicate '*tall*' will be significantly higher in a discussion about basketball players (who tend to be taller than average) than in a discussion about jockeys (who tend to be shorter than average). The set of objects or individuals used to determine the threshold of relative adjectives, e.g. basketball players versus jockeys, is usually referred to as the comparison class, and is one of the parameters that plays a role in fixing the extension of a relative adjective in context.

Maximum standard absolute adjectives (MaxAAs) like *full, empty, straight* and *flat* manifest a different type of context dependence. Unlike RelAs, MaxAAs have context *independent* uses that are true of an object just in case it manifests a maximal degree of the relevant property. In such an use, '*empty*' is true of a cookie jar, for example, just in case it contains no cookies at all. MaxAAs also have uses that tolerate deviation from a maximal degree, however: in many contexts, a cookie

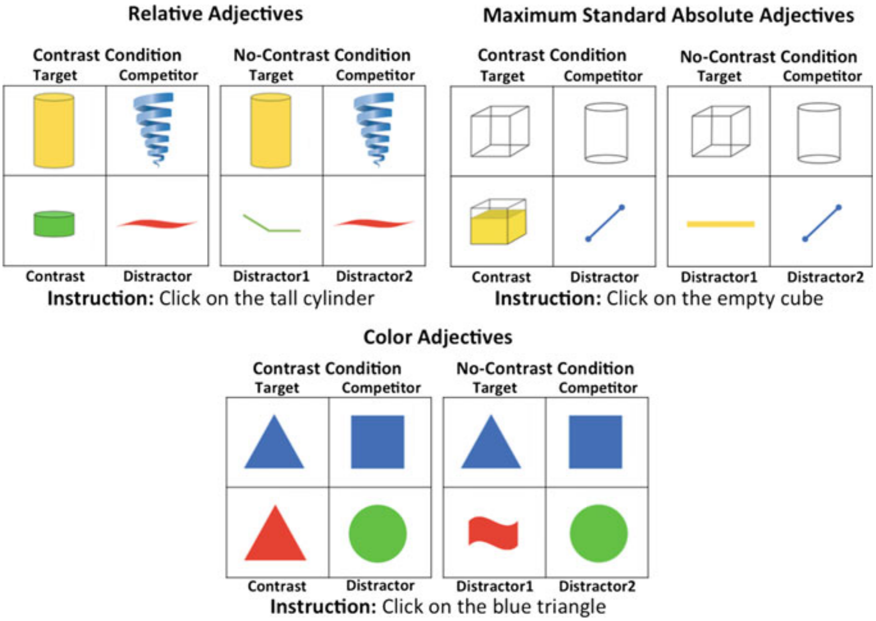


Fig. 2 Item Examples (Aparicio et al. 2015)

jar containing just one or two cookies could be felicitiously described as empty (especially if the goal is to get someone to fill it again). A question of current research is whether such uses of MaxAAs arise from the same semantic principles that regulate context dependent interpretations of RelAs, or whether they involve a pragmatic phenomenon of “imprecise” uses of expressions with context invariant denotations (see e.g., Sassoon and Toledo 2011; Lassiter and Goodman 2013, 2017; Qing and Franke 2014; Leffel et al. 2016).

Although the study in Aparicio et al. (2015) did not address this question directly, it provided a baseline examination of the processing of RelAs versus precisely interpreted MaxAAs used as modifiers in definite descriptions, with color adjectives (ColAs) as a control.[1] Following Sedivy et al. (1999), two critical kinds of visual displays were tested, illustrated in Fig. 2. In the Contrast condition, the visual display contains: (1) a TARGET object (e.g., a tall cylinder) that participants are requested to click on; (2) a COMPETITOR that shares the target property but presents a different shape (e.g., a tall spiral); (3) a CONTRAST object that belongs to the same comparison class as the target, but could not be described by the adjective in the instruction (e.g., a short cylinder); and (4) a DISTRACTOR object that could not be described by the adjective in the instruction, nor does it belong to the same comparison class (e.g., a

[1] Although color adjectives are both context dependent and vague, they are sensitive to different kinds of contextual parameters from RelAs and MaxAAs. See Rothschild and Segal (2009), Kennedy and McNally (2010), Clapp (2012) for discussion.

wavy line). The No-Contrast condition was created by substituting the contrasting object with a second distractor. With the exception of color-adjective trials, none of the shapes in the visual array shared color. Aparicio et al. found that all three adjective types displayed RECs, though there were differences in the time-course of the effects: for ColAs and RelAs, RECs appeared before information about the head noun was available to participants. However, in the case of MaxAAs the REC was delayed and did not obtain until the noun window. This led the authors to conclude that lexical processing can also play an important role in further shaping RECs, a point to which we return in Sect. 4.

Our experiment extends the Aparicio et al. design to a second class of absolute adjectives: "minimum standard" absolute adjectives (MinAAs) such as *bent, spotted, bumpy* and *striped*. Like MaxAAs, MinAAs have context invariant uses, but unlike MaxAAs, they merely require their arguments to have greater than a minimum degree of the relevant property. A *bent rod*, for example, is a rod with some degree of bend; and a *spotted shirt* is a shirt with some number of spots.[2] Our goal in examining MinAAs was both to fill out the empirical picture of RECs in relative versus absolute adjectives that was only partially provided in the Aparicio et al. study, and to identify potential differences in REC effects among natural classes of adjectival modifiers.

2.1 Design

Following Aparicio et al. (2015), we used geometric shapes to construct the visual stimuli with the goal of controlling for potential effects of world-knowledge about

[2] Several linguistic tests diagnose whether an absolute adjective makes use of a maximum versus minimum versus relative standard. For instance, Kennedy (2007) points out that these three classes give rise to different entailment patterns when used in comparatives. In comparatives of the form *X is more A than Y*, MinAAs entail that X is A (i); MaxAAs entail that B is not A (ii); and (unmarked) RelAs entail neither that X is (not) A nor that Y is (not) A (iii).

(i) a. The red towel is wetter than the blue towel. ⇒
b. The red towel is wet.

(ii) a. The red towel is drier than the blue towel. ⇒
b. The blue towel is not dry.

(iii) a. The red towel is bigger than the blue towel. ⇒
b. The red towel is (not) big.
c. The blue towel is (not) big.

The distribution of modifiers like *slightly* and *completely* are also often described as tests for MinAA and MaxAA status, respectively, but strictly speaking, these modifiers test for minimum and maximum scalar endpoints, respectively, which are independent of—though generally correlated with—maximum and minimum standards.

Table 1 Adjective-Noun pairs tested in Experiment 1

Min. St. Absolute Adjective	Noun
Bent	Line
Bumpy	Square/triangle
Curved	Line
Open	Circle
Spotted	Square/circle
Striped	Square/triangle

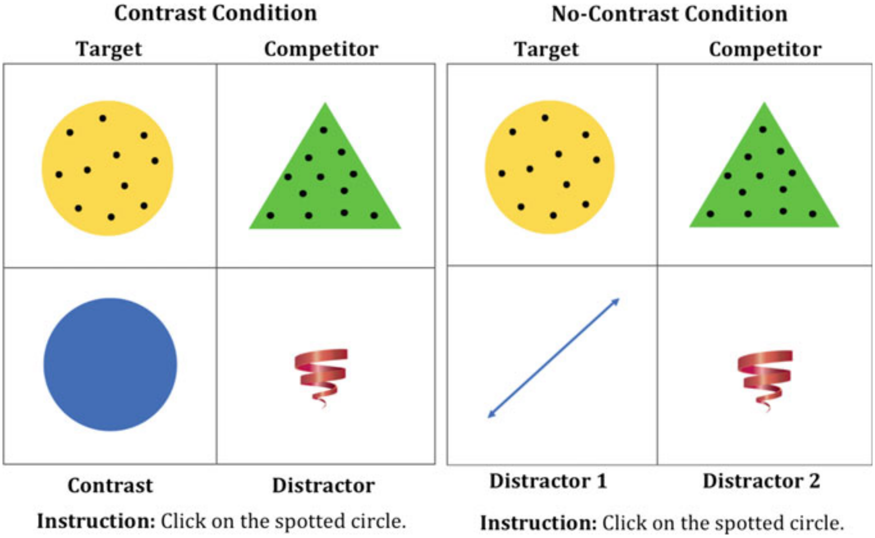


Fig. 3 Item example for eExperiment 1

artifacts on adjective interpretation. Six MinAAs were included in one experiment, which are listed in Table 1.

Two conditions were tested (see Fig. 3). In the Contrast condition, the visual display contains: (1) a TARGET object (e.g., a spotted circle) that participants are requested to click on; (2) a COMPETITOR that shares the target property but presents a different shape (e.g., a spotted triangle); (3) a CONTRAST object that belongs to the same comparison class as the target, but could not be described by the adjective in the instruction (e.g., a circle with no spots); and (4) a DISTRACTOR object that could not be described by the adjective in the instruction, nor does it belong to the same comparison class (e.g., a short spiral). The No-Contrast condition was created by substituting the contrasting object with a second distractor. None of the shapes in the visual array shared color.

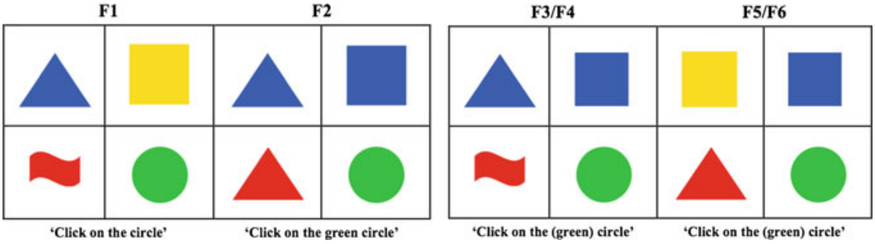

F1
F2
F3/F4
F5/F6
'Click on the circle'
'Click on the green circle'
'Click on the (green) circle'
'Click on the (green) circle'

Fig. 4 Fillers (Experiment 1)

Ten experimental items were constructed.[3] Conditions were distributed in two lists using a Latin Square design. Both the order of the trials within each list and the position of the four pictures within each trial were randomized. Each list was complemented with 60 filler trials. All adjectives used in filler trials were color adjectives (*red*, *green*, *yellow* and *blue*), and pictures always consisted of 2D shapes with plain colors.

As in Aparicio et al. (2015), six different types of fillers (10 trials per type) were constructed (see Fig. 4). In the first type (F1), none of the figures shares shape or color and the instruction does not contain a modifier. In the second type of filler (F2), the visual display is equivalent to the Contrast condition in the color-adjective trials. However, these filler trials differ from the Contrast condition in that the auditory instruction targets the distractor. In the third type of filler (F3), none of the objects share shape, although two of the pictures share color. The instruction contains a modifier but it does not target any of the two shapes that share color. The fourth type of filler (F4) only differs from F3 in that the instruction does not include a modifier. In the fifth type of filler (F5), none of the figures in the visual array shares color. However, two of the shapes belong to the same comparison class. The instruction contains a modifier and targets one of the two pictures that does not share shape with any of the other pictures in the visual array. Finally, the sixth type of filler (F6) is like F5, except that the instruction does not make use of a color adjective.

2.2 *Materials*

2.2.1 Visual Stimuli

Pictures used in experimental trials as targets, contrasts and competitors (a total of 29 pictures) were normed in a series of three description-picture matching studies on Mechanical Turk. The purpose of the norming studies was to standardize the interpretation preferences of the visual stimuli within and across adjective types.

[3] See supplementary materials to this chapter for a full list of the experimental items used in Experiment 1.

More specifically, the norming studies ensured that all target and competitor objects were recognized to satisfy the relevant adjectival property, whereas contrast objects (used in the Contrast condition) were recognized to NOT instantiate the relevant adjectival property. Due to space constraints, we do not report further details about the results of the three norming studies here. In addition, 18 more images were used as distractors. Whenever possible, distractors were drawn from the pool of objects that had been used as target, competitor or contrast in other trials.

2.2.2 Auditory Stimuli

Auditory stimuli were recorded in a sound booth by a female native speaker of English. For each recording, the onsets and offset of the adjective were measured in order to determine the mean duration of the three groups of adjectives tested. The mean duration of the adjective for all trials was 503 ms (SD = 76.09). None of the adjectives bore pitch accent or rising tone.

2.3 Apparatus

Eye movements were recorded with a Tobii T60 Eye-tracker sampling at 60 Hz. Viewing was binocular and both eyes were tracked, although analyses were performed on data belonging to the right eye exclusively.

2.4 Procedure

Participants saw a visual display with four pictures. Their eye movements were tracked while listening to instructions such as ‘*Click on the spotted circle*’. Participants were instructed to click on the picture that they thought fitted the description in the auditory instruction best. Only clicks that took place after the offset of the auditory instruction triggered the next trial. There was a 2 second long preview window between the onset of the visual display and the onset of the auditory instruction. Before each trial, a fixation cross appeared in the middle of the screen. A red box framing the cross appeared when participants fixated on it. Participants were instructed to click on the cross when the red box appeared in order to proceed to the next trial. This was done so that eye movements to the four objects could be measured from a default position that was equidistant to the four pictures in the display. At the beginning of the experiment, participants had four practice trials to help them become familiar with the task.

2.5 Participants

Participants were fifty-one undergraduate and graduate students at the University of Chicago (34 females, M = 20.7, range 18–34). All participants were native speakers of American English. Undergraduate students did the experiment to fulfill a research awareness requirement for a linguistics course. Graduate students were paid $10. All participants had normal or corrected to normal vision. Subjects were excluded from data analysis if they met at least one of the following two criteria: (1) track loss for a given subject was higher than 40%; and (2) before the head noun became available, a subject did minimal scanning of any part of the display (i.e., when the aggregated proportion of fixations to the four pictures in the display was $<10\%$ of the total recorded fixations, probably because the subject was only fixating on the fixation cross in the center of the screen). The latter criterion intends to exclude participants who were passively waiting for the head noun information before processing the instruction. The application of these two criteria resulted in the exclusion of 11 subjects. The results reported in the following section correspond to data from 40 participants between the ages of 18–34 (26 females, M = 20.57).

2.6 Results

Analyses were performed on two consecutive windows (W1 and W2) of 150 ms starting from the onset of the adjective, such that the right boundary of W2 coincided with the onset of the head noun (set at 703 ms after offsetting the adjective window by 200 ms to adjust for the time required to plan and launch an eye-movement). A third window (W3) of 150 ms starting at the onset of the head noun was also analyzed. W1 and W2 contain fixations reflecting the processing of the adjective, whereas W3 contains fixations reflecting the processing of the head noun. Analyses were run on the aggregated proportion of fixations in each of the three windows (see Fig. 5). One adjective-noun combination was removed from the data analysis, since the stimuli was found to not appropriately represent the adjectival property.

Figure 5 contains the proportions of fixations to each of the four objects in the visual display for each condition. Eye fixations to the target and the competitor objects were analyzed. In order to determine whether target versus competitor disambiguation occurred faster in the Contrast than in the No-Contrast condition, a two-way ANOVA using Object Type (target vs. competitor) and Condition (Contrast vs. No-Contrast) as factors was run in each window. Results did not reveal any significant main effect of Condition in any of the time windows examined (all $Fs(1, 39) > 0.5$, $ps > 0.1$). W1 and W2 did not show a significant main effect of Object Type ($Fs(1, 39) > 1.88$, $ps > 0.1$). Even tough the main effect of Object Type reached significance in W3 ($F(1, 39) = 4.12$, $p < 0.05$), pair comparisons between target versus competitor for the Contrast and No-Contrast conditions separately did not yield any significant results ($ps > 0.1$). No interactions between Object Type and

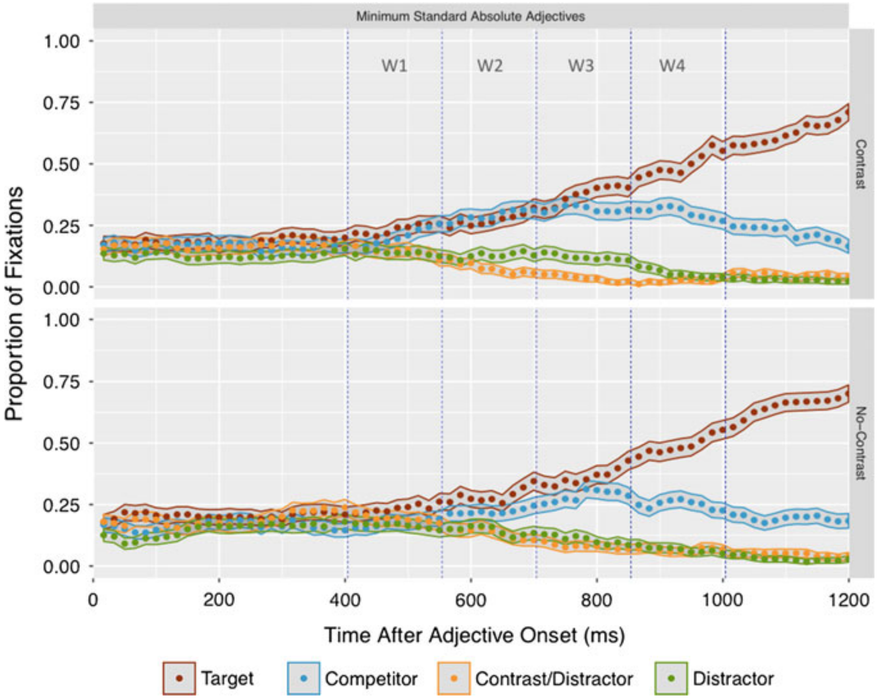


Fig. 5 Proportions of fixations to each of the four objects in the display over time starting at the adjective onset for each adjective type. The vertical dashed blue lines mark the boundaries of the four windows defined for data analysis, with the noun onset coinciding with the right boundary of W2 (703 ms from the onset of the adjective)

Condition (all $Fs(1, 39) > 0.01$, $ps > 0.3$) were observed in any of the three windows. To verify whether there were any RECs in even later time windows, a fourth 150 ms window (W4) spanning from 853-1003 ms was examined. As in W3, a two-way ANOVA showed a main effect of Object Type ($F(1, 39) = 31.00$, $p < 0.00001$), but no significant main effect of Condition ($F(1, 39) = 1.31$, $p > 0.2$), or Object Type x Condition interaction ($F(1, 39) = 0.47$, $p > 0.4$) was observed. A one-way ANOVA with Object Type as factor revealed a significant difference between the two levels for both the Contrast ($F(1, 39) = 12.59$, $p < 0.002$) and the No-Contrast condition ($F(1, 39) = 26.43$, $p < 0.00001$) such that participants fixated significantly more on the target object than the competitor object.

In addition to the ANOVA analysis reported above, a second analysis using logistic mixed effects models was also performed. The goal of this analysis was to determine whether there were significant differences in the rate at which the proportions of fixations to the target objects in the Contrast and the No-Contrast conditions increased as a function of time. Figure 6 plots the proportion of fixations over time to the target objects in the two conditions tested. The existence of a significant difference, such that the target object in the Contrast Condition received a higher proportion of looks

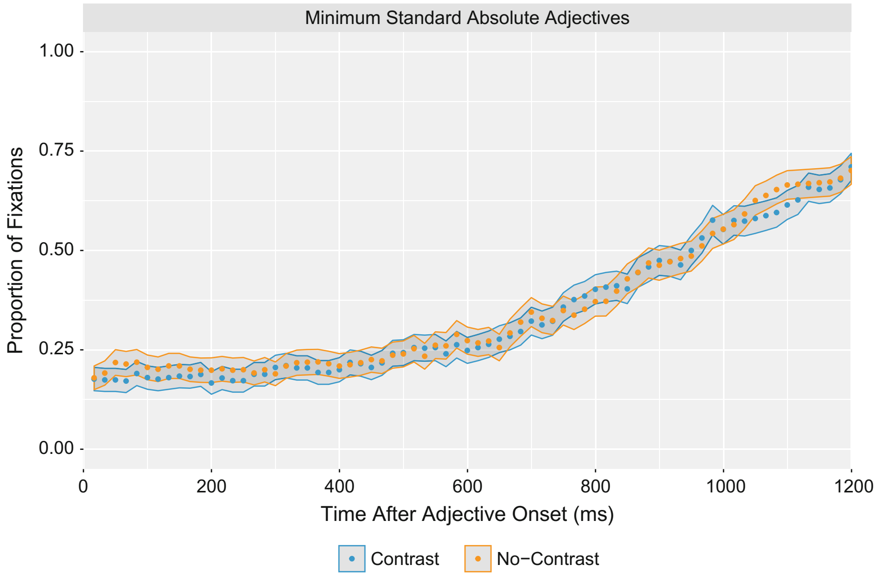


Fig. 6 Proportions of fixations over time to the target objects in the Contrast and the No-Contrast condition. The plotted window starts at the adjective onset and spans for 1200 ms

earlier than the target object in the No-Contrast condition would be indicative of a REC. A window spanning from the onset of the adjective to the end of W3 (853 ms) was defined for data analysis. The factors CONDITION and TIMEPOINT were included as main effects, with SUBJECTS and ITEMS factored in as random effects.

As in the previous analysis, no significant interaction between CONDITION: TIMEPOINT was found ($\beta = -0.0004208$, $p > 0.1$), confirming that MinAAs did not trigger RECs.

2.7 *Discussion*

Our results clearly show that MinAAs do not give rise to RECs, since target versus competitor disambiguation times did not differ significantly across conditions. The same results were achieved when the proportions of looks to the target objects in the Contrast and the No-Contrast conditions were compared. Therefore, information about the visual context was not used by participants during the adjective window to make predictions about potential referents at a point in which the linguistic instruction was ambiguous given the visual context. Rather participants only relied on the linguistic information available to them to narrow down the set of potential referents in the visual display as the auditory instruction unfolded. The current results contrast with the findings reported by Aparicio et al. (2015), who found RECs for each of the three adjectives tested, i.e. RelAs, ColAs and MaxAAs. Taken together, these two sets of results show that not all prenominal adjectives are equally context-sensitive, even when there is contextual support for a contrastive interpretation.

A important question is whether all the differences in the availability and properties of the observed RECs result from pragmatic reasoning—as modulated by the informativity considerations discussed in Sect. 1 regarding the use of overspecified prenominal adjectives—or whether RECs are also affected by grammatical factors related to the lexical-semantic properties of each adjective class. Experiment 2 seeks to address this question by quantifying how informative each of these adjective classes are perceived to be when used restrictively versus redundantly.

3 Experiment 2: Perceived Informativity

Experiment 2 addresses the question of whether all the adjective types tested by Aparicio et al. (2015) and the current eye-tracking experiment (see Sect. 2) are perceived as equally informative when the display contains a contrastive object (Contrast condition), compared to displays that do not (No-Contrast condition). With this goal in mind, Experiment 2 consisted of an offline judgement task, where participants were instructed to rate whether the instructions used in the eye-tracking experiments provide a sufficient amount of information to confidently identify the target object in the relevant visual display.

If the online eye-tracking effects reported by Aparicio et al. (2015), as well as the results reported above for Experiment 1, are shaped by differences in the perceived informativity, we predict the following patterns of results for Experiment 2: First, since MinAAs are the only type of adjective that do no give rise to RECs, we don't expect to find any differences in perceived informativity between the Contrast and the No-Contrast conditions. All other adjectives should show a significant difference between these two conditions such that the No-Contrast condition is perceived as more overinformative than the Contrast condition. Second, based on the timing of the RECs observed for each adjective type, we would expect that the magnitude of the overspecification penalty should be greater for MaxAAs than for ColAs and RelAs.

The same lists and adjectives (RelAs = 9, MaxAAs = 4, MinAAs = 6, ColAs = 4) used in the eye-tracking studies were tested with a total of 60 experimental items (20 containing RelAs, 10 containing MaxAAs, 10 containing MinAAs and 20 containing ColAs). Conditions were distributed in two lists using a Latin Square design. Both the order of the trials within each list and the position of the four pictures within each trial were randomized (see Fig. 7). The same 60 filler trials used in Experiment 1 were included (see Sect. 2).

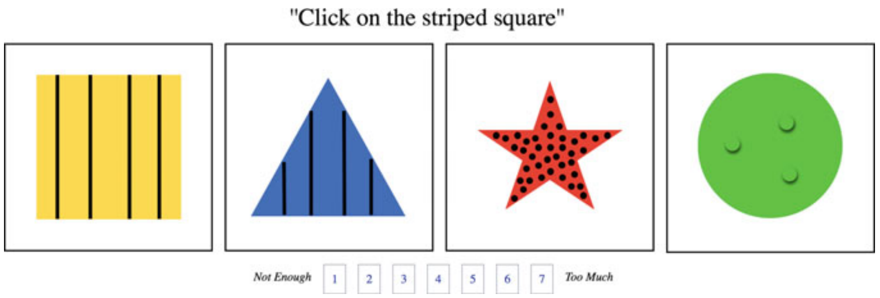


Fig. 7 Item example for Experiment 2

3.1 Methods

3.1.1 Materials

Stimuli consisted of the same visual displays used by Aparicio et al. (2015), a total of 100, plus the 20 visual displays tested in the eye-tracking experiment reported in Sect. 2. The auditory instructions used in both eye-tracking experiments were transcribed and accompanied the visual displays.[4]

3.1.2 Procedure

Participants saw displays of four pictures on a computer screen coupled with a written statement such as '*Click on the striped square*'. For each of the displays, participants were instructed to rate whether the instruction provided a sufficient amount of information to identify the right target. Judgments were indicated on a 1–7 scale, where 1 corresponded to *'Not enough information'* and 7 corresponded to *'Too much information'*. At the beginning of the experiment, participants had three practice trials to help them become familiar with the task.

3.1.3 Participants

Participants were 32 native speakers of English between the ages of 18–35 (12 females; meanage = 30) recruited through the website Amazon Mechanical Turk. Three subjects were removed from data analysis because they were not between 18–35 leaving a total of 29 (10 females; meanage = 29). All participants were payed $3.

[4] See supplementary materials to this chapter for a full list of the experimental items used in Experiment 2.

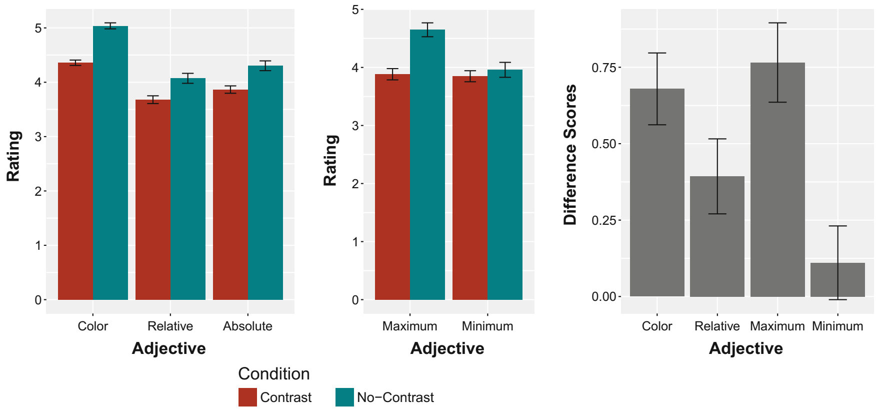


Fig. 8 **"Left"** Rating means for color, relative and absolute adjectives; **"Central"** Rating means for maximum and minimum standard absolute adjectives; **"Right"** Difference scores between the Contrast and the No-Contrast condition for each adjective type

3.2 Results

Means were obtained for all adjective types. Visual inspection of the left plot in Fig. 8 reveals that the No-Contrast condition received higher ratings compared to the Contrast condition for ColAs, RelAs and absolute adjectives (AAs). For the class of AAs, data from MaxAAs and MinAAs were combined. The ratings in the Contrast condition were used as the baseline comparison against the ratings in the No-Contrast condition, as the former represents ratings pertaining to the condition containing the optimal amount of information, since target identification would not be possible in the absence of the adjective. Paired t-tests confirm that the differences between the two conditions were statistically significant (ColAs: $t(28) = -5.78$, $p < 0.0001$; RelAs: $t(28) = -3.20$, $p < 0.01$; AAs: adjectives $t(28) = -3.85$, $p < 0.001$). However, closer inspection to the two subclasses of AAs (central plot, Fig. 8) shows that the difference between conditions observed for AAs is mostly driven by MaxAAs, which present the higher ratings in the No-Contrast condition. A paired t-test confirmed that this difference was highly significant ($t(28) = -5.89$, $p < 0.0001$). MinAAs, on the other hand, showed a non-significant difference across conditions ($t(28) = -0.91$, $p > 0.3$).

A 2-way ANOVA using Adjective Type and Condition as factors was run on the three classes of adjectives that showed significant differences between the two conditions, i.e., ColAs, RelAs and MaxAAs. A significant interaction for Adjective Type x Condition was detected ($F(2, 56) = 7.64$, $p < 0.008$), showing that the magnitude of the effect was different across the three adjective types. In order to further explore this interaction, a 2-way ANOVA was run in three different subsets of the data. The interaction remained significant for the subset containing RelAs and MaxAAs ($F(1, 28) = 10.70$, $p < 0.002$), and the subset containing RelAs and

ColAs ($F(1, 28) = 13.10$, $p < 0.001$), while it did not reach significance for the data subset containing only ColAs and MaxAAs ($F(1, 28) = 0.7$, $p > 0.4$). This suggests that the magnitude of the effect was comparable for ColAs and MaxAAs (see the right panel of Fig. 8 containing the difference scores obtained by subtracting the Contrast condition from the No-Contrast condition for each adjective type), and that the ADJECTIVE TYPE x CONDITION interaction detected for the full data set was driven by differences between ColAs and MaxAAs on the one hand and RelAs on the other.

3.3 Discussion

For ColAs, RelAs and MaxAAs, the No-Contrast condition received significantly higher ratings than the Contrast condition. This means that participants perceived a difference between the optimally informative baseline in the Contrast condition and the No-Contrast condition, which they judged to contain more information than necessary. Interestingly, no parallel effect was found for MinAAs, suggesting that participants did not perceive differences between the degree of informativity of the two conditions tested. Our results also revealed that the magnitude of the effect of perceived informativity was not the same for ColAs, RelAs and MaxAAs. The results from the 2-way ANOVA interaction and the t-tests indicate that the effect was bigger for ColAs and MaxAAs than it was for RelAs, while no significant difference in perceived informativity was found between ColAs and MaxAAs. The main conclusion that can be extracted from these results is that perceived informativity is indeed modulated by adjective class. In the general discussion (Sect. 4), we address the relationship between perceived informativity and RECs.

4 General Discussion

Out of the four adjective classes tested in Experiment 1 and in Aparicio et al.'s (2015) study, we were able to detect RECs for ColAs, RelAs and MaxAAs. However, MinAAs failed to display a REC, as target versus competitor disambiguation took place in the same time window, i.e. W4, for both the Contrast and the No-Contrast condition (see Fig. 9). An important finding of Aparicio et al.'s (2015) is that there exist non-trivial timing differences in the RECs of ColAs and RelAs on the one hand, and MaxAAs on the other. For the former, the effect took place in W2, during the adjective window, whereas for the latter the effect did not occur until W3, a window that already reflects processing of the head noun. In the case of ColAs and RelAs, participants committed to the target object at a point in which the linguistic input was still ambiguous between two objects in the visual display (i.e., target and competitor), whereas for MaxAAs, target identification was facilitated in the Contrast condition, but was nevertheless significantly delayed, as participants did

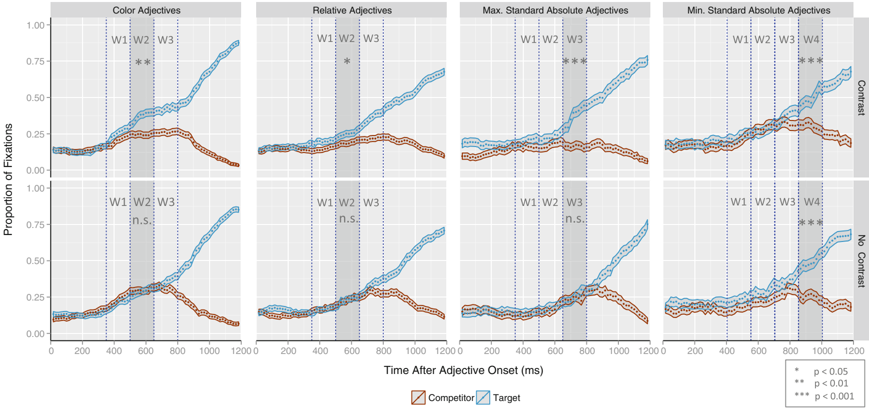


Fig. 9 Proportions of fixations to target versus competitor over time starting at the adjective onset. Data belonging to ColAs, RelAs and MaxAAs are reproduced from Aparicio et al. (2015). All windows are 150 ms long. For each adjective, the right boundary of W2 coincides with the onset of the head noun. The grayed time windows correspond to the first window in which a significant difference was found

not discriminate between target and competitor until information about the head-noun was available to them.

Experiment 2 also revealed important asymmetries in the effect of perceived informativity across adjective types. MinAAs were the only class of adjectives that did not display differences in perceived informativity between the Contrast and the No-Contrast condition. Interestingly, MinAAs were also the only adjective class that did not give rise to RECs. However, ColAs, RelAs and MaxAAs did show an overspecification penalty, as indicated by the significantly higher ratings obtained for these three adjective classes in the No-Contrast condition, which was not compatible with a contrastive interpretation of the adjective.

Taken together, the previous results reported by Aparicio et al. (2015), as well as the results from Experiment 1 and 2 suggest that informativity is an important factor in RECs, as shown by the relation between RECs and the offline measure of perceived informativity: adjectives that showed an overspecification penalty (ColAs, RelAs and MaxAAs) also gave rise to RECs, whereas adjectives that did not show and overspecification penalty (MinAAs) did not display RECs. However, the timing differences observed in the RECs of ColAs, RelAs and MaxAAs could not be uniquely attributed to the overspecification penalties detected by Experiment 2 for these three types of adjectives. As discussed above, the magnitude of the perceived (over)informativeness was different across the three adjective types with RelAs showing a significantly smaller effect compared to ColAs and MaxAAs, for which the size of the effect was comparable. If perceived informativity was the only source of RECs we would expect ColAs and MaxAAs to pattern alike with respect to the timing of their RECs, showing earlier effects compared to RelAs. However, this is not what Aparicio et al.'s (2015) results show, with MaxAAs being delayed with respect to ColAs and RelAs. We therefore conclude, that informativity cannot be the only factor driving RECs.

Based on these results, we would like to suggest that there exist at least two non-mutually exclusive sources of the RECs. The first one pertains to perceived informativity considerations related to quantity and manner-based pragmatic reasoning about referential contrast triggered by the mention of the prenominal adjective. Second, RECs are also modulated by differences in lexical processing incurred by distinct lexically encoded types of context-dependence. The differences in the timing of the REC of RelAs and MaxAAs can be explained in this way. While relative adjectives like *tall* resort to context in order to fix the value of their *semantic* threshold (typically computed with respect to a contextually salient comparison class), MaxAAs like *empty* have been argued to only interact with context in order to fix a *pragmatic* threshold of imprecision (Kennedy 2007; Syrett et al. 2009; van Rooij 2011; Burnett 2014; Qing and Franke 2014; Leffel et al. 2016). If lexical context-sensitivity is an important component of the timing resolution of RECs, it is conceivable that RelAs could trigger RECs with a different time course from MaxAAs. But the exact mechanism that relates context-sensitivity to the time course of RECs still remains a question for future research. Another question that remains to be explored is whether the early REC attested for ColAs also results from facilitated lexical processing (though see Aparicio et al. (2015) for an argument against this view). In principle, the adjectival threshold of ColAs is not assumed to depend on a contextually salient comparison class for its resolution (Kennedy and McNally 2005). This may mean that other high level perceptual factors such as the visual saliency of color might underlie the timing resolution of the REC for ColAs.

Given the abundance of results showing that speakers have a greater tendency to use ColAs redundantly than any other class of adjectives (see Pechmann 1989; Belke and Meyer 2002; Nadig and Sedivy 2002; Sedivy 2004, among many others), it is somehow unexpected that Experiment 2 showed such a clear penalty for overspecified uses of ColAs. If hearers are sensitive to the probabilities of use of overspecified adjectives, ColAs would be expected to give rise to the lowest overspecification penalty among all the adjectives tested in Experiment 2. It is possible that the nature of the stimuli used in our experiment had an effect on how overinformative ColAs were perceived to be. In a production experiment, Rubio-Fernández (2015) shows that the rates of overspecification of ColAs vary depending on the nature of the object. Rubio-Fernández found lower rates of color overspecification with geometric shapes in polychrome displays than in displays containing garments, a type of object for which color is a more central feature. A final important issue is the question of why MinAAs did not show differences in perceived informativity in the two conditions tested. At this point, we do not have an explanation for the lack of sensitivity to the visual context displayed by this adjective class. Further research will have to determine why this class of adjective does not seem to be associated with an expectation of contrastive use.

5 Conclusion

The experiments presented in this paper had the goal of determining whether informativity-based reasoning about the use of a prenominal modifier is the sole driver of Referential Effects of Contrast involving adjectivally modified NPs. By examining four different classes of adjectives, we have shown that perceiving the use of a particular class of adjective as overinformative when used redundantly is related to whether such adjective class should give rise to a REC. However, while pragmatic reasoning is an important source of these effects, it cannot alone account for the variety of attested patterns of RECs. We conclude that lexical semantic factors determining how context-sensitive a given adjective class is further contributes to the temporal resolution of such effects.

Acknowledgements We would like to thank Aaron Hill, Jackson Lee, Gabriel Aparicio and Katie Franich for their assistance during various stages of the eye-tracking experiment reported in this paper, as well as Michelle Namkoong and Sonia Juan Rubio for help in the creation of part of the visual stimuli. Finally, we would also like to thank the audiences at the workshop on *Gradability, Scale Structure, and Vagueness: Experimental Perspectives*, and the 22nd AMLaP Conference. This project was supported by a NSF grant (BCS 1227144) to C. Kennedy and M. Xiang.

References

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*(3), 191–238.

Aparicio, H., Xiang, M., & Kennedy, C. (2015). Processing gradable adjectives in context: A visual world study. In S. D'Antonio, M. Moroney, & C. R. Little (Eds.), *Proceedings of Semantics and Linguistic Theory*. Publisher: Linguistic Society of America and Cornell Linguistics. CircleUrl: http://journals.linguisticsociety.org/proceedings/index.php/SALT/issue/view/132.

Arts, A., Maes, A., Noorman, K., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*(1), 361–374.

Belke, E., & Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different" decisions. *European Journal of Cognitive Psychology*, *14*(2), 237–266.

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

Clapp, L. (2012). Indexical color predicates: Truth conditional semantics vs. truth conditional pragmatics. *Canadian Journal of Philosophy*, *42*(2), 71–100.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107.

Crain, S., & Steedman, M. (1985). On not being led up the garden path: the use of context by the psychological parser. In D. Dowty, L. Kartunnen, & A. Zwicky (Eds.), *Natural Language Parsing*. Cambridge, MA: Cambridge University Press.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*(6), 409–436.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts*. New York: Academic Press.

Grodner, D., & Sedivy, J. C. (2011). The effect of speaker-specific information on pragmatic inferences. In N. Pearlmutter & E. Gibson (Eds.), *The processing and acquisition of reference* (pp. 239–272). Cambridge, MA: MIT Press.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, *28*(1), 105–115.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*, 43–61.

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

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

Kennedy, C., & McNally, L. (2010). Color, context, and compositionality. *Synthese*, *174*(1), 79–98.

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231–3250.

Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The effect of scene variatio on the redundant use of color in definite reference. *Cognitive Science*, *37*(2), 395–411.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In T. Snider (Ed.), *Proceedings of Semantics and Linguistic Theory* (Vol. 23, pp. 587–610). Ithaca, NY: CLC.

Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a Bayesian model of interpretation. *194*(10), 3801–3836.

Leffel, T., Xiang, M., & Kennedy, C. (2016). Imprecision is pragmatic: Evidence from referential processing. In M. Moroney, C. R. Little, J. Collard, & D. Burgdorf (Eds.), *Proceedings of Semantics and Linguistic Theory* (Vol. 26). Publisher: Linguistic Society of America and Cornell Linguistics. CircleUrl: https://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/26.836/3688.

Maes, A., Arts, A., & Noordman, L. (2004). Reference management in instructive discourse. *Discourse Process*, *37*(2), 117–144.

McNally, L. (2011). The relative role of property type and scale structure in explaining the bahavior of gradable adjectives. In R. Nouwen, R. van Rooij, U. Sauerland, & H.-C. Schmitz (Eds.), *Vagueness in communication*. Lecture Notes in Computer Science (Vol. 6517, pp. 151–168). Berlin, Heidelberg: Springer.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking contraints in children's on-line reference resolution. *Psychological Science*, *13*(4), 329–336.

Paraboni, I., van Deemter, K., & Masthoff, J. (2007). Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, *33*(2), 229–254.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110.

Pinkal, M. (1995). *Logic and lexicon: The semantics of the indefinite*. Dordrecht: Kluwer.

Pyykkönen-Klauck, P., & Crocker, M. W. (2016). Attention and eye movement metrics in visual world eye tracking. In P. Knoeferle, P. Pyykkönen-Klauck, & M. W. Crocker (Eds.), *Visually situated language comprehension* (pp. 67–82). Amsterdam: John Benjamins.

Qing, C., & Franke, M. (2014). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In T. Snider, S. D'Antonio, & M. Weigand (Eds.), *Proceedings of Semantics and Linguistic Theory* (Vol. 24, pp. 23–41). Ithaca, NY: CLC.

Rothschild, D., & Segal, G. (2009). Indexical predicates. *Mind and Language*, *24*(4), 467–493.

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

Rubio-Fernández, P. (2015). *Redundancy is efficient–and effective, too*. Paper presented at the XI Conference on Architectures and Mechanisms for Language Processing (AMLaP).

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology*, *7*(153).

Sassoon, G. W., & Toledo, A. (2011). *Absolute and relative adjectives and their comparison classes*. Unpublished manuscript. Amsterdam university and Utrecht university.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23.

Sedivy, J. C. (2004). Evaluating explanations for referential context effects: Evidence for Gricean meachanisms in online language interpretation. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 345–364). Cambridge, MA: MIT Press.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Syrett, K., Kennedy, C., & Lidz, J. (2009). Meaning and context in children's understanding of gradable adjectives. *Journal of Semantics*, *27*(1), 1–35.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Unger, P. (1975). *Ignorance: A case for scepticism*. Oxford: Clarendon Press.

van Rooij, R. (2011). Vagueness and linguistics. In G. Ronzitti (Ed.), *Vagueness: A guide* (Chap. 6, pp. 123–179). Dordrecht: Springer.

Weber, A., Braun, B., & Crocker, M. W. (2006). Finding referencts in timr: Eye-tracking evidence for the role of contrastive accents. *Language and Speech*, *49*(3), 367–392.

Westerbeek, H., Koolen, R., & Maes, A. (2015). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in Psychology*, *6*(935).

Wolter, L., Skovbroten Gorman, K., & Tanenhaus, M. K. (2011). Scalar reference, contrast and discourse: Separating effects of linguistic discourse from availability of the referent. *Journal of Memory and Language*, *65*(3), 299–317.

Modified Fractions, Granularity and Scale Structure

Chris Cummins

Abstract Pragmatic enrichments arising from the use of modified fractions have been little studied, but offer interesting insights into the subtleties of scale structure and granularity. In this chapter I present some new experimental data on the interpretation of these expressions. I argue that these data suggest that modified fractions, like modified integers, give rise to pragmatic enrichments which are conditioned by scale granularity, but that we need to refine the notion of granularity somewhat to extend it to this domain. There is also evidence for enrichments that are not easily captured in classical quantity implicature terms, but which I suggest could be explained by appeal to typicality effects.

Keywords Fractions · Number · Modified numerals · Granularity · Implicature
Typicality · Scales

1 Introduction

Expressions of numerical quantity have been the focus of much study in experimental semantics and pragmatics. In many cases, this research is inspired by the realisation that there is an incompatibility between the "expected" meaning of expressions, based on mathematical considerations, and their communicative meaning in linguistic interactions. Consider, for example, (1)–(3).

(1) Most of the American population is female. (Solt 2016)
(2) A hexagon has at most 10 sides. (Nouwen 2010)
(3) London has more than 1000 inhabitants. (Cummins et al. 2012)

In each case, we can offer a plausible account of the semantics of the expressions by appeal to mathematical considerations. We could argue that "most X is Y" means

C. Cummins (✉)
Linguistics and English Language, University of Edinburgh, Edinburgh, UK
e-mail: c.r.cummins@gmail.com

E. Castroviejo et al. (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure*,
Language, Cognition, and Mind 4, https://doi.org/10.1007/978-3-319-77791-7_9

that the quantity of X that is Y exceeds the quantity of X that is not Y; that "at most 10 X are Y" means that the quantity of entities that are both X and Y does not exceed 10; and that "more than 1000 X are Y" means that the quantity of entities that are both X and Y exceeds 1000.

Under analyses of this kind, (1)–(3) are true. However, these examples are not unanimously judged as true by actual users of language, who consider them anomalous. As in (1), "most" is not considered to be felicitous when referring to values just a little above 50%. As in (2), "at most *n*" is not considered to be felicitous when referring to values clearly and invariably below *n*. As in (3), "more than *n*" is not considered to be felicitous when used out of the blue and referring to values far in excess of *n*.

In these cases, and many others besides, there are broadly two possible approaches to explaining the lack of felicity. One is to argue that the "mathematical" intuition about the semantic analyses is wrong, and that in fact these expressions have a more complex semantics (Geurts and Nouwen 2007; Solt 2016). The other is to argue that the anomalies are pragmatic, and arise for principled reasons that have nothing to do with the semantics per se (Cummins and Katsos 2010).

Of course, it is quite possible in principle that both explanations are correct, i.e. that the semantic analyses are more complex than initially supposed, but these meanings are also subject to pragmatic enrichment. None of the specific accounts offered appears to rely on both semantic and pragmatic effects—perhaps it would not be parsimonious to do so—but nevertheless it seems to be common ground that pragmatic enrichments are widespread in the domain of number.

Most strikingly, the interpretation of numerals itself has been widely argued to rely upon pragmatic factors. Unmodified numerals can convey either a "punctual" or a "lower-bound" meaning—that is, (4) can be interpreted as equivalent in meaning either to (5) or (6). In this case, our intuition might be that (5) is strongly preferred, but (6) is still available, for instance in a context such as (7).

(4) John has three A-levels.
(5) John has exactly three A-levels.
(6) John has at least three A-levels.
(7) You need three A-levels to be considered for the job. Is John eligible?

In one class of accounts, the semantics of numerals is lower-bounding and the punctual interpretation arises because of a quantity implicature (Gazdar 1979; Levinson 1983; Horn 1989): broadly, if the speaker of (4) knew that John had four A-levels, she would typically say so. On other accounts (e.g. Carston 1998; Geurts 2006; Breheny 2008) the semantics of numerals is punctual or underspecified, and pragmatic inference is required either to select a meaning or to obtain the lower-bound reading where required.

Other categories of numerical expression also appear to give rise to pragmatic enrichments, but the problem of determining which alternatives are in play is a more complex one. For instance, as discussed by Fox and Hackl (2006), (8) does not implicate (9); however, Cummins et al. (2012) show that items such as (10) are widely judged to convey meanings such as (11) (but not (12)).

(8) John has more than four children.
(9) John does not have more than five children.
(10) There's room for more than 80 people.
(11) There is not room for more than 100 people.
(12) There is not room for more than 81 people.

One potential explanation for this difference (Cummins 2012) is that the availability of the alternatives depends upon the properties of the numbers involved as well as on the information content of the sentence. On this account, (8) is only felicitous if the speaker is ignorant as to the truth or falsity of informationally stronger alternatives, or the precise issue of whether John has four or more children is currently under discussion. In either case, the stronger assertion ("…more than five…") is out of court as an alternative, and hence the implicature (9) fails to arise. By contrast, (10) could be felicitous even for a knowledgeable speaker, as it is a convenient approximation that uses a round number (putatively accessible at a lower cognitive cost than non-round numbers; cf. Krifka 2002). The corresponding sentence with "…more than 100…" would be a viable alternative, but that with "…more than 81…" would not, as this uses a costlier non-round number.

Whether or not this particular explanation is along the right lines, it seems inevitable that we have to consider the distinct properties of different numbers in order to understand their pragmatic behaviour in full. With respect to the issue of unmodified numerals and their meaning, it may be possible simply to construe the number line as a homogeneous sequence of equally-spaced scale points. However, research on the psychology of number itself (Dehaene 1997; Butterworth 1999) clearly indicates that our cognitive representation of integers is much more structured than this: some numbers (10, 20, 50, 100, …) are major reference points, while others (7, 13, 101, …) are not. And the existence of this structure is known to have linguistic consequences: round numbers are more widely used than non-round numbers (Jansen and Pollmann 2001), and round numbers are capable of being used to express approximate values (Krifka 2002).

2 Expressing Fractional Quantities

There are several ways in which we can quantify the size of the subdivisions of a whole.[1] We can do this using scalar quantity expressions (a few/little, some, many/much, most, all), percentages, fractions, or derived expressions such as "two

[1] As an anonymous reviewer pointed out, we can also use fractions to quantify over parts of things that are not obviously characterised as "wholes"—for instance, "half a kilogram of…". The examples in this chapter all concern proportions of a finite quantity, and consequently involve proper fractions (those that lie between 0 and 1). Given that most uses of fractions for quantities above 1 also involve proper fractions, in combination with integers—we usually say "two and a quarter" rather than "nine quarters"—I would expect the observations here to apply to the broader class of fractional expressions of quantity.

out of every five". And these latter numerically-based categories of expression can themselves be further modified by expressions such as "more/less than", "at least/most", "up to", "about", and so on.

One immediate question that arises is how the availability of these distinct means of expression bears upon their perceived meaning. Returning to example (1), repeated below, we can see that this could be expressed in various other ways, such as (13)–(15).

(1) Most of the American population is female.
(13) More than 50% of the American population is female.
(14) More than half of the American population is female.
(15) More than one out of every two Americans is female.

(15) seems potentially anomalous—perhaps "more than one" is initially interpreted as "at least two"—but (13) and (14) appear to be valid alternatives to (1). We might then ask whether these three options are semantically equivalent, and if they are, whether this has pragmatic consequences (for instance, whether one option is marked and thus gives rise to some form of markedness implicature). This question is explored in detail, for "most" versus "more than half", by Solt (2016).

2.1 Inferences from Modified Fractions

A further question of interest, both for pragmatics and for the nature of the interface between number and language, is how the various different expressions of fractional quantity relate to one another. Is it the case, for instance, that the use of one modified fraction implicates the falsity of another? Consider (16) and (17).

(16) More than one-fifth of the participants were literature students.
(17) More than two-fifths of the participants were literature students.

It seems reasonable to conjecture that the use of (16) by an informed and cooperative speaker could be held to implicate the falsity of (17). But this is not obvious on theoretical grounds. Note that—to recall again Fox and Hackl's (2006) observation—(18) does not implicate the falsity of (19).

(18) More than one of the participants was a literature student.
(19) More than two of the participants were literature students.

Setting aside the question of precisely why this is so, it is conceivable that (16) could pattern with (18), if we consider (16) to be effectively quantifying over the number of "fifths of the participants that were literature students". Thus, one immediate question is whether the numerators of modified fractions behave like modified numerals, for pragmatic purposes. If so, then at least some modified fractions will fail to give rise to quantity implicatures that would theoretically be predicted, while others may give rise to a restricted class of implicatures, negating only a subset of the informationally stronger alternatives. Concretely, for instance, we would expect "more than two-fifths" potentially to implicate "not more than three-fifths" on this account, but not to implicate "not more than half", as a half does not correspond to a whole number of fifths.

Considering the whole class of proper fractions, it is clear that we will need some way to constrain the set of alternatives that are to be considered in the calculation of quantity implicatures. Given any proper fraction, we can identify proper fractions that are arbitrarily close to it (in either direction): for a fraction p/q, consider for instance the set {3p/2q, 4p/3q, 5p/4q, …}. If "more than p/q" were to implicate the falsity of the corresponding expression with any other member of this set, it would, in the limit, convey "not more than p/q", which is clearly absurd. In practice we could argue that these quantities are not all calculable by speaker and hearer, and not easily expressible, and for one of these reasons the problematic implicatures are ruled out of court. However, it is not straightforward to draw the line between what can and cannot be inferred on any principled grounds.

In suggesting (17) as a potentially consequential candidate alternative to (16), I am implicitly acknowledging the potential relevance of granularity considerations, in the sense of Krifka (2009), in determining which alternatives are pragmatically active. The notion of granularity, traceable to Curtin (1995), captures the fact that measured quantity can be reported at various different levels of precision, the levels differing specifically in the density of representation points. For instance, in the domain of time reporting, "hours" form a coarse-grained scale, with "minutes" forming a finer-grained scale, and "quarters" (units of 15 minutes) constituting an intermediate scale. (16) and (17) could be argued to be matched in granularity, as they are both expressions at the "fifths" level.

Cummins et al. (2012) argue that granularity is relevant to scalar implicatures, and specifically that modified coarse-grained numerals do not implicate the falsity of modified finer-grained alternatives. Empirically, it is an open question whether the same claim holds for fractions, but in principle the same arguments should apply: the alternatives of finer granularity enable the speaker to formulate more informative expressions, but these expressions incur a greater cognitive cost, both for the speaker and the hearer. Consequently, the speaker's failure to use a more informative finer-grained alternative can be explained away as being due to cost considerations, rather than being interpreted as a signal that the speaker is not in a position to commit to the more informative assertion that would have arisen. Applying this reasoning to fractions, we might expect that the use of a coarse-grained modified fraction will not give rise to pragmatic enrichments based on the existence of finer-grained alternatives. However, we might expect the use of such an expression to give rise to pragmatic enrichments related to equally (or more) coarse-grained alternatives.

2.2 *The Granularity of Fractions, and Its Consequences*

Given the definition of granularity, it would be natural to suppose that fractions' fineness of granularity increases with the increasing size of the denominator. If so, the claim articulated above can be reformulated as follows: fractions with small denominators will not give rise to implicatures concerning alternatives with larger denominators. For instance, the Cummins et al. (2012) argument would seem to suggest that (20) should not implicate the falsity of (21).

(20) More than three-quarters of the participants were literature students.
(21) More than nine-tenths of the participants were literature students.

However, there are problems with interpreting the notion of granularity in such an intuitive way for the case of fractions. Notably, it raises a potential conflict with Krifka's (2009) observations about the construction of granularity scales. He makes two observations: that scales are optimal (in expressive power) if their scale points are distributed in a systematic way (for instance, equidistant, or logarithmically distributed), and that "scales of different granularity levels should align".

The status of Krifka's (2009) latter observation, about the alignment of scales, is not made entirely clear. It could be read as a desideratum in order for granularity scales to be easily usable, or it could be read as a requirement in order for two scales to coexist on the same underlying domain. All the examples that Krifka discusses involve scales that align in this way, but we can readily imagine candidate pairs of scales that do not: for instance, we might count eggs in sixes or twelves, at one granularity level, and in hundreds, at another level. The hundreds are not all scale points on the sixes scale. For that matter, we might measure distance in miles, at a coarse granularity level, or metres, at a fine granularity level, in which case the scale points will (almost) never precisely coincide.[2]

The former observation, that scale points should be sensibly distributed, makes tacit appeal to the idea that we wish to be able to describe the quantities that we want to talk about efficiently in terms of scale points. If the scale points—with which convenient expressions are associated—are clustered unevenly and do not span the full range of values that we wish to discuss, they are less helpful to us. The appropriate distribution of scale points clearly depends on the distribution of the values that we wish to discuss. For instance, a five-point rating scale with the options <*OK*, *good*, *very good*, *excellent*, *superb*> would be helpful if most of the things we want to rate are good to a greater or lesser extent, but unhelpful in permitting us to distinguish between things that are variously bad. In the case of expressions of proportion, we might reasonably suppose that we would like to be able to discuss all values across the range (0, 1) with similar levels of acuity.[3]

If we consider proper fractions with a single, fixed denominator q, this criterion is satisfied, as they are uniformly distributed between zero and one. If we add to this

[2] The International Yard and Pound Agreement of 1959 defines the yard as exactly 0.9144 m, so in fact a mile is exactly 1609.344 m and 125 miles is thus 201,168 m. This latter distance is the first point at which the scale points for mile and metre coincide.

[3] This assumption may not always be tenable: if we are talking predominantly about rare events, we might find it more useful to be able to distinguish events occurring with probability 0.001 and probability 0.01 than to be able to distinguish events with probability 0.5 and probability 0.6. However, a simple system of fractions with small denominators would perform very poorly according to this criterion too.

system a further set of fractions with a denominator that is a multiple of q, both of Krifka's conditions are met. However, if we add to the system a set of fractions of a different denominator that is not a multiple (or factor) of q, both criteria will be violated: the scale points will not be evenly distributed over the range (0, 1), nor will the scale points align.

To take a concrete example, if our scales are based around halves, quarters and eighths, there will be seven equally-distributed scale points between 0 and 1 on the "eighths" scale, three of which are also scale points on the "quarters" scale; one of these is also a scale point on the "halves" scale. If our scales are quarters and thirds, there are three scale points on the quarters scale and two scale points on the thirds scale between 0 and 1, none of which coincide. Consecutive scale points in this case are unevenly spaced: the gaps between them are 1/4, 1/12, 1/6, 1/6, 1/12 and 1/4.

To illustrate the potential limitations of such a system, imagine that speaker and hearer are committed to using a system that relied upon thirds and quarters, that each possible expression within this system is equally costly to use, and that the speaker is known to be fully knowledgeable and cooperative. A description "more than a quarter" would be highly informative in such a system: it would convey that the value in question lay between a quarter and a third. By contrast, "more than three-quarters" would be much less informative: it would only convey that the value lay in the (three times larger) range between three-quarters and 1. This underscores the point that the efficiency of scales depends upon systematic distribution of the scale points. Unless it is particularly important for some reason that values in the middle of the range (0, 1) are especially easy to describe economically and accurately, this arrangement is inefficient.

Intuitively, it seems clear that some denominators that would cause problems in such a system—e.g. sevenths—are seldom used. However, it seems very plausible that speakers could use a system that employed both quarters and tenths, or halves and thirds, despite such a system violating both Krifka's (2009) criteria.[4] If so, this would suggest either that granularity is not an appropriate construct for capturing alternatives in the domain of fractions, or that the notion of granularity must be generalised somewhat from Krifka's definition in order to be applied here.

If we were to allow the notion of granularity to be elaborated or generalised to treat the domain of fractions, it is natural to consider whether we should make similar arrangements for other systems of quantity. The scale structure of fractions presumably reflects some kind of cognitive preference on the part of humans. Jansen and Pollmann (2001: 200) conjecture that "doubling and halving (sometimes followed by halving again) are basic means to manipulate quantities", which predicts a central role for halves and quarters in the organisation of the system. Among other researchers on the topic, Sigurd (1988) argues for the relevance of the base system to numerical cognition, and if applied to the case of fractions, this suggests that tenths and hundredths should also have some kind of conceptual primacy. Taking these considerations simultaneously into account, we might also predict a role for fifths and

[4] For instance, as pointed out by an anonymous reviewer, recipes can rely on thirds and quarters in combination, with quantities such as "1/3 cup" and "1/4 cup" being simultaneously salient.

twentieths in the system. However, if fifths are represented as “pairs of tenths”, we might expect them to be cognitively less accessible than tenths, which runs counter to the prediction that we would make about fifths and tenths based on standard considerations of granularity (i.e. that fifths are coarser-grained and therefore more accessible than tenths).

Moreover, although it seems plausible that the operation of halving is more cognitively salient than any other operation of division, it seems perfectly feasible to conceptualise entities such as thirds by direct division. If the operation of dividing by three is a great deal more complex than that of dividing by two, we might expect thirds to be less accessible than quarters, again running counter to a straightforward granularity-based account.

The above discussion has entirely concerned the denominators of fractions, but we might also expect the complexity of the numerators to bear upon how fractions are treated by speakers. We could think of a fraction p/q as being represented via a series of stages, in which the whole is first divided into q equal parts (perhaps via a series of distinct operations of division) and then collections of p of these parts are considered. If fractions are indeed conceptualised in this way, we would expect unit fractions—those of the form 1/q, for integer q—to be preferred over other fractions with the same denominator. If an individual’s system of fractions effectively comprises a large number of unit fractions, plus a few full sets of fractions with special denominators such as two, three and ten, this system will have a particularly uneven distribution of scale points: specifically, many scale points will be clustered relatively near zero. Such a system would be justified if it is particularly important to be able to distinguish small proportions from one another with a high resolution.

Could the granularity of fractions have any implications for other expressions of quantity? There is clearly a potential interplay, as discussed above in the case of the decimal system: if the preference for a particular kind of scale structure arises in for number in general, we could reasonably expect that to carry over to the construction of fractions. The existence of the decimal system makes powers of ten especially important in representing and manipulating expressions of numerical quantity, and that might also influence our preferences as to how we use fractions. Non-decimal systems might also play a part: for instance, the way clock time is represented might encourage us to use the operations of dividing by four and dividing by 60, and practice with these operations might promote use of the relevant fractions.

In the other direction, if we have internalised a particular system of fractions, presumably we will be inclined to apply this when dealing with quantities that—unlike clock time—do not have pre-established points of division. If, for argument’s sake, eighths are more salient than thirds, we would expect to find eighths being used preferentially as a way of partitioning up quantity in a novel domain. Moreover, if it were to transpire that (for instance) three-eighths is a more salient fraction than seven-eighths, we might be more likely to talk about units of three-eighths in a novel domain than about units of seven-eighths.

2.3 Testing the Predictions by Appeal to Implicature

Granularity has been argued to have several implications for the way in which quantity expressions are understood. Krifka (2002) argues that coarse-grained numerals attract approximate interpretations, while fine-grained numerals are restricted to precise interpretations. In a situation in which 98 people are present, (22) can be judged as true, although interestingly (23) cannot, even when 99 people are present.

(22) There were 100 people there.
(23) There were more than 100 people there.

Conversely, (24) is false if exactly 100 people are present, and (25) is true only on an existential reading, which is hard to access in this case.

(24) There were 102 people there.
(25) There were 98 people there.

We would expect the same to apply to fractions, but with a possible caveat: namely that all the fractions widely used are coarse-grained enough to attract some kind of approximative reading. Supposing that the US population were precisely 320 million, we would not normally expect (26) and (27) to refer respectively to exactly 160 million people and exactly 64 million people.

(26) Half of the US population is clustered in just 146 counties.
(27) By 2050, one fifth of the US population will be aged 65 or over.

However, to the extent that fractions get interpreted as approximations, this is not a unique attribute: numbers used in continuous measurement behave in much the same way. The speaker of (28) is not understood to be referring to a distance of precisely 400 cm, let alone 4000.0 mm. It is fine-grained cardinal quantities that are the exception, in their apparent preference for exact readings.

(28) The passage into the Mound of the Hostages is four metres in length.

Building upon Krifka's observations, we could attempt to quantify the salience of particular fractions by examining the size of the regions for which they are acceptable approximations (with or without the explicit use of a hedge such as "about" or "around")—that is, the diameter of their pragmatic halos, in the sense of Lasersohn (1999). A potential issue to address in this case would be dealing with overlapping regions. For instance, if the quantity under discussion is 29%, successively less accurate approximations for this would include "two sevenths", "three tenths", "one quarter" and "one third", and each of these could be presented bare or with a hedging modifier. It is not entirely clear on theoretical grounds whether these should all be acceptable, or whether there is some implied trade-off in which the acceptability of one term is associated with a decrease in acceptability for the others. This might hamper our ability to use data pertaining to the "approximative power" of a fraction as a measure of its salience.

In this chapter I adopt a different approach: I ask participants for range interpretations of modified fractions. If these fractions are able to convey quantity implicatures, as argued above, then the ranges obtained will depend upon the presence of salient alternatives. Acceptance that a range could extend beyond a particular value will thus indicate that that value it is not considered to be a sufficiently salient alternative to mandate its usage, even in cases where doing so would yield expressions that were semantically true. So, for instance, if "more than one quarter" is understood to correspond to the range 25–50%, that indicates that "a half" (=two quarters) is a salient alternative to "one quarter", but also that "one third", "two fifths" etc. are not.

A potential challenge in adopting this approach is that the format in which participants are instructed to give their answers might influence their interpretation. For instance, requiring participants to respond in terms of fractions might represent a confound—simpler fractions would presumably be privileged in the responses. In the following, the use of percentages was adopted to give participants more flexibility in their response: however, it must be acknowledged that this also has its limitations, as it might confer an advantage on fractions that are expressible in precise percentage terms (for instance, promoting the use of tenths).[5]

3 Experiments: Pragmatic Bounds for Modified Fractions

This section reports a series of short experiments designed to explore whether modified fractions give rise to pragmatic bounds of the type posited in the above discussion. This preliminary research serves partly as a proof of concept and partly as a first attempt to map out some of the terrain, by establishing the pragmatic relations that are judged to hold between different modified fractions within the system.

Four versions of the stimuli were created and administered separately. Details of the specific stimuli are given below. The general procedure was the same for all sets of stimuli, and was as follows.

3.1 Method

The experiments were conducted using the Amazon Mechanical Turk platform. In each version, participants were presented with the following cover story and instructions:

A market research company has conducted a detailed survey on a large group of people, and has written up the results. For instance, "More than 50% of the participants are female", "Less than 20% of the participants own two cars", and so on.

[5] An anonymous reviewer noted that this approach also still relies upon a high level of general numeracy on the part of the participants, with respect to their ability to interpret fractions. Given the relatively small number of outright errors with simple fractions in the following experiments, I would argue that this turned out not to constitute a major concern; however, caution is clearly necessary in interpreting the participants' pragmatic behaviour with respect to fractions that did elicit a lot of errors (e.g. "more than 6/7" in experiment 1 below).

You're now going to read some expressions that have been used to summarise the results from the survey. For each one, please state the range of possible values, in percent, that you think the expression means.

For example, if the expression is "about half", you might say that that means between 45 and 55%, or between 40 and 60%, etc.

There are no 'correct' answers: we're interested in knowing what you think.

The experimental items consisted of a series of modified fractions presented in a pseudorandom order. In each case, the modifier "more than" or "less than" was used, and the fraction was presented in word form rather than numerals (e.g. "more than three fifths"). The entire list was presented on a single page.

Responses were coded as "literal" if they reflected the semantic bound without any implicature (for example, interpreting "more than four fifths" as corresponding to the range 80–100%), "pragmatic" if they reflected an inferred bound (for example, interpreting "more than four fifths" as corresponding to the range 80–90%), and "error" if the responses failed to respect the semantics of the expression (for example, interpreting "more than four fifths" as corresponding to the range 60–80%). Where present, pragmatic bounds were recorded and analysed.

3.2 Participants

For each version of the experiment, 20 participants were recruited from US locations. Each participant was paid $0.50 for participation.

3.3 Experiment 1

Version 1 of the experiment was directed primarily towards establishing whether modified fractions towards the edges of the (0, 1) range attracted literal interpretations, and secondarily towards establishing whether modified fractions towards the middle of the (0, 1) range attracted pragmatic readings that were conditioned by the presence of "one half" in the system.

3.3.1 Materials

The following 15 items were pseudorandomised in order and presented in the context of the cover story shown above:

- less than one third/one quarter/one fifth/one sixth/one seventh/one eighth
- more than two thirds/three quarters/four fifths/five sixths/six sevenths/seven eighths
- more than two fifths/three sevenths/four ninths.

Table 1 Results of experiment 1

Item	Error	Literal	Pragmatic	Pragmatic lower/upper bounds
Less than 1/3	1	10	9	5, 5, 10, 10, 11, 15, 25, 25, 26
Less than 1/4	2	9	9	5, 10, 10, 10, 12, 15, 15, 19, 20
Less than 1/5	2	12	6	5, 10, 10, 10, 11, 15
Less than 1/6	1	11	8	5, 5, 5, 5, 5, 9, 10, 12
Less than 1/7	3	12	5	5, 5, 10, 10, 10
Less than 1/8	6	10	4	5, 5, 8, 10
More than 3/4	0	8	12	80, 80, 84, 85, 85, 85, 90, 90, 90, 90, 90, 95
More than 4/5	3	11	6	89, 90, 90, 90, 95, 95
More than 5/6	10	4	6	89, 90, 90, 95, 95, 97
More than 6/7	13	2	5	90, 90, 90, 92
More than 7/8	4	9	7	90, 90, 90, 93, 95, 95, 98
More than 2/5	9	10	21	40, 45, 45, 45, 47, 47, 50, 50, 50, 55, 55, 55, 55, 60, 60, 60, 60, 60, 60, 60, 60
More than 3/7	13	3	4	49, 49, 50, 55
More than 4/9	8	5	7	49, 49, 55, 55, 60, 60, 65
Total	75	116	109	

This corresponds to a "40–40" response, which could reflect error; here I charitably assume that the respondent interprets this expression as ruling out anything as high as 41%

Due to a coding error, "more than two thirds" was omitted and "more than two fifths" was repeated within the experiment as administered.

3.3.2 Results

The numbers of error, literal and pragmatic responses for each item, along with the pragmatic responses given, are presented in the table (Table 1).

3.3.3 Discussion

Generally, participants appear to have been competent with the task: although the overall error rate was 75/300 (=25%), the majority of these errors arose in cases involving relatively little-used fractions that are not straightforward to convert into percentages. I will not attempt to interpret the pragmatics of error responses as we cannot assume the participants' competence with respect to those items.

Of the 225 semantically correct responses, 109 (=48.4%) exhibited some form of pragmatic narrowing, which coheres with the prediction that implicatures are available from modified fractions. Most, although not all, of the pragmatic bounds offered by participants correspond to potentially salient alternative fractions. Of the

41 pragmatic responses to the "less than" items, 13 refer to 5% (1/20) and 14–10% (1/10), and of the 36 pragmatic responses to the corresponding "more than" items, 16 refer to 90% (9/10)—in fact, 18 refer to this alternative fraction if we consider 89% also to represent a bound for "less than 9/10". In the case of "more than 2/5", we see some responses reflecting the presence of "one half" as an alternative, but with more responses based on the next point on the fifths scale (3/5, i.e. 60%).

In addition to these responses, there are some that cannot be easily understood in terms of alternative fractions. For instance, 55% is attested as a pragmatic upper bound for "more than 3/7" and "more than 2/5". As a fraction, this could be interpreted as corresponding to either 11/20 or 5/9, neither of which is predicted to be salient (although when 55% occurs as a pragmatic upper bound for "more than 4/9", it seems natural to attribute that to the salience of 5/9 as an alternative). There are several possibilities as to how these bounds are arising: perhaps the presence of "more than 4/9" in the experiment has made ninths atypically salient as alternatives, or perhaps the value given reflects an impressionistic range interpretation (something more akin to the typicality effects postulated by Geurts and van Tiel (2013)) or a compromise between two possible interpretations. However, with this small sample, it is also highly plausible that the returned value represents an error, with the participant intending to give a percentage value that corresponded to a more salient fraction. We return to this question later.

3.4 Experiment 2

Version 2 of the experiment aimed to test whether the repeated use of terms on a specific scale, e.g. fifths, would cause alternatives drawn from the same scale to become more salient, or whether the presence of coarser-grained alternatives from other scales would condition the pragmatic readings that were obtained.

3.4.1 Materials

The following 14 items were presented in the context of the same cover story as in experiment 1:

- more/less than one quarter/a half/three quarters
- more than one fifth/two fifths
- less than three fifths/four fifths
- more than one tenth/seven tenths
- less than three tenths/nine tenths.

The order of presentation was fixed. Participants first saw the items involving quarters and halves, then the items involving fifths, and finally the items involving tenths. Within each subset of items, the order was pseudorandomised.

Table 2 Results of experiment 2

Item	Error	Literal	Pragmatic	Pragmatic lower/upper bounds
More than 1/4	1	5	14	29, 29, 30, 30, 30, 32, 32, 35, 35, 40, 49, 49, 49, 50
Less than 1/4	0	15	5	15, 15, 15, 20, 20
More than 1/2	0	9	11	55, 60, 65, 65, 65, 70, 74, 74, 74, 74, 75
Less than 1/2	0	9	11	25, 25, 26, 26, 26, 32, 34, 39, 40, 40, 41
More than 3/4	0	14	6	79, 80, 84, 85, 85, 90
Less than 3/4	0	7	13	51, 51, 51, 51, 56, 60, 60, 60, 65, 65, 67, 70, 71
More than 1/5	2	7	11	23, 24, 24, 24, 24, 24, 25, 25, 25, 30, 39
More than 2/5	1	8	11	45, 45, 48, 49, 49, 49, 49, 49, 50, 59, 65
Less than 3/5	4	7	9	49, 50, 51, 51, 51, 53, 55, 55, 56
Less than 4/5	1	8	11	50, 55, 60, 61, 68, 71, 74, 76, 76, 76, 76
More than 1/10	2	8	10	14, 15, 19, 19, 19, 19, 19, 20, 20, 24
Less than 3/10	1	8	11	20, 20, 20, 21, 21, 25, 25, 26, 26, 26, 26
More than 7/10	4	7	9	74, 74, 75, 75, 79, 79, 79, 79, 90
Less than 9/10	2	7	11	50, 79, 81, 81, 81, 81, 81, 81, 81, 86, 87
Total	18	119	143	

3.4.2 Results

The numbers of error, literal and pragmatic responses for each item, along with the pragmatic responses given, are presented in the table (Table 2).

3.4.3 Discussion

Participants generally appeared to find this version of the task easier, and returned a much lower error rate (18/280 = 6.4%). Just over half the total responses reflected pragmatic bounds.

The absence of finer-grained fractions than quarters from the initial part of the item list seems to have made a difference to the interpretation of "less than one quarter" and "more than three quarters". In experiment 1, these attracted 9 literal to 9 pragmatic responses and 8 literal to 12 pragmatic responses, respectively. In this version, they attracted respectively 15 literal to 5 pragmatic responses and 14 literal to 6 pragmatic responses. Although neither of these differences reaches significance under Fisher's exact test, the numerical difference is suggestive that participants in version 2 of the experiment are inclined to draw implicatures based on the next quarter, when presented with a series of stimuli involving quarters. However, this is not a hard and fast rule: for instance, most of the responses for "more than one quarter" reflect a tighter pragmatic bound than that provided by "not more than a half".

It appears that making quarters salient, as scale points, has had some influence on the interpretation of subsequent expressions with fifths. "More than one fifth" attracted nine responses which could be interpreted as relating to the inference "not more than one quarter". Similarly, "more than two fifths" attracted seven responses which could be interpreted as relating to the inference "not more than half (=two quarters)", and 9 out of 11 pragmatic responses for this item involved a bound of 50% or lower. This appears to contrast with responses for this item in experiment 1, where the majority of pragmatic responses involved a bound above 50%.

Finally, fifths having been presented, the expressions with tenths gave rise to predictable interpretations: the majority of pragmatic responses made reference to the next scale point on the tenths scale (which in these cases is also a point on the fifths scale). In the case of "less than three tenths" and "more than seven tenths", responses were split between referring to the next tenths scale point and referring to the next quarter.

These preliminary results suggest that the overall picture is complex. Generally, repeated reference to a scale appears to make its scale points more salient, which is evident in the interpretation of subsequent items referring both to that scale and finer-grained scales. However, even under these circumstances, there is no guarantee that participants will choose to draw implicatures based on separate coarse-grained scales—e.g. referring to quarters when cognising about values expressed in tenths—and may instead draw weaker inferences based on the current term's scale-mates.

3.5 Experiment 3

Version 3 of the experiment was designed to test whether the repeated use of a less salient scale would exert any effect on the interpretation of subsequent expressions, either using that scale or using a different scale.

3.5.1 Materials

The following 16 items were presented in the context of the same cover story as in experiment 1:

- more/less than one sixth/one third/two thirds/five sixths
- more/less than one tenth/three tenths/seven tenths/nine tenths.

The order of items was manipulated such that those involving sixths or thirds were presented first, then those with tenths. The order of presentation was pseudo-randomised within each group of items.

Table 3 Results of experiment 3

Item	Error	Literal	Pragmatic	Pragmatic lower/upper bounds
More than 1/6	7	2	11	19, 20, 20, 20, 25, 25, 25, 30, 30, 40, 50
More than 1/3	7	3	10	40, 40, 40, 40, 40, 45, 49, 50, 50, 67
More than 2/3	6	5	9	70, 74, 74, 74, 75, 75, 75, 80, 90
More than 5/6	10	4	6	90, 90, 90, 90, 95, 95
Less than 1/6	7	6	7	5, 5, 7, 10, 10, 12, 12
Less than 1/3	9	4	7	10, 15, 20, 25, 25, 26, 26
Less than 2/3	6	4	10	40, 40, 44, 50, 55, 55, 55, 55, 60, 60
Less than 5/6	14	2	4	60, 70, 78, 80
More than 1/10	3	6	11	15, 15, 15, 15, 15, 16, 20, 20, 20, 20, 25
More than 3/10	6	6	8	35, 39, 39, 40, 40, 40, 45, 45
More than 7/10	3	6	11	74, 75, 78, 79, 79, 79, 80, 80, 80, 90, 90
More than 9/10	5	11	4	92, 95, 97, 98
Less than 1/10	1	15	4	4, 5, 5, 7
Less than 3/10	5	6	9	10, 20, 20, 20, 20, 21, 25, 26, 26
Less than 7/10	4	6	10	55, 60, 60, 60, 60, 60, 61, 65, 65, 67
Less than 9/10	3	5	12	10, 70, 70, 70, 79, 80, 80, 81, 81, 85, 85, 85
Total	96	91	133	

3.5.2 Results

The numbers of error, literal and pragmatic responses for each item, along with the pragmatic responses given, are presented in the table (Table 3).

3.5.3 Discussion

Unlike the case of quarters (experiment 2), the repeated use of sixths does not appear to elicit many pragmatic bounds that refer to thirds and sixths. Within the sixths scale itself, participants who derive pragmatic bounds tend to prefer more informative bounds, often referring to tenths. When then presented with expressions with tenths, participants do not tend to infer bounds referring to thirds or sixths, even in cases where these would be more informative than the bounds actually inferred. For instance, “less than seven tenths” attracts a modal lower bound of 60%, rather than 67%; all of the eight pragmatic upper bounds offered for “more than three tenths” exceed 34%. This suggests that the pragmatic influence of thirds and sixths is relatively weak in the participants’ systems of fractions, compared to the influence of tenths. In addition, there are high error rates in the conditions involving “five sixths” in particular, perhaps suggesting that participants in this experiment had difficulty in evaluating this in percentage terms. As noted earlier, this particular experimental setup, relying on percentage responses, may be promoting the use of tenths—and perhaps disadvantaging the use of thirds and sixths—to an atypical extent.

3.6 Experiment 4

Version 4 of the experiment was intended as a control for version 3, reversing the order of presentation, to see whether influence could spread between scales in the opposite direction.

3.6.1 Materials

The following 16 items were presented in the context of the same cover story as in experiment 1:

- more/less than one tenth/three tenths/seven tenths/nine tenths
- more/less than one sixth/one third/two thirds/five sixths.

The order of items was manipulated such that those involving tenths were presented first, then those with sixths and thirds. The order of presentation was again pseudorandomised within each group of items.

3.6.2 Results

One participant failed to finish the task, and a further participant gave decimal responses without a clear system, so their results are omitted. The numbers of error, literal and pragmatic responses for each item, along with the pragmatic responses given, are presented in the table (Table 4).

3.6.3 Discussion

The pattern of responses in this experiment closely mirrors that of experiment 3, suggesting that the order of presentation makes relatively little difference for these items. Again, there is little evidence of thirds and sixths being used as pragmatically relevant alternatives. By contrast, tenths continue to be pragmatically relevant when dealing with terms on the thirds/sixths scale, but there is no indication that the prior mention of tenths has promoted inferences involving tenths. In this case, terms involving “five sixths” did not appear to present any particular difficulty to the participants.

Table 4 Results of experiment 4

Item	Error	Literal	Pragmatic	Pragmatic lower/upper bounds
More than 1/10	3	3	12	14, 15, 15, 15, 19, 19, 19, 20, 20, 20, 20, 40
More than 3/10	3	3	12	35, 36, 39, 39, 39, 40, 40, 40, 40, 50, 55, 65
More than 7/10	1	4	13	74, 75, 75, 77, 79, 79, 79, 79, 80, 80, 80, 90, 90
More than 9/10	0	14	4	95, 95, 95, 98
Less than 1/10	0	11	7	5, 5, 5, 6, 6, 6, 7
Less than 3/10	6	4	8	5, 10, 17, 20, 20, 21, 25, 25
Less than 7/10	3	4	11	20, 45, 60, 60, 60, 61, 61, 65, 66, 66, 67
Less than 9/10	1	4	13	50, 55, 75, 76, 80, 80, 80, 80, 80, 81, 81, 82, 86
More than 1/6	5	3	10	19, 21, 25, 25, 25, 30, 31, 32, 35, 35
More than 1/3	6	3	9	39, 45, 45, 48, 49, 50, 50, 65, 65
More than 2/3	5	6	7	72, 74, 75, 75, 75, 80, 85
More than 5/6	4	7	7	89, 90, 90, 90, 92, 95, 95
Less than 1/6	5	8	5	8, 10, 11, 12, 14
Less than 1/3	3	6	9	5, 15, 19, 20, 25, 25, 25, 26, 29
Less than 2/3	5	4	9	34, 34, 50, 51, 51, 55, 55, 60, 60
Less than 5/6	6	4	8	70, 70, 70, 75, 75, 75, 76, 78
Total	56	88	144	

3.7 General Discussion

The results from these small experiments strongly suggest that people are inclined to interpret modified fractions in a pragmatically restricted way, and that a lot of the readings they obtain are predictable on the basis of a quantity implicature analysis that considers other salient fractions as alternatives. More specifically, the results indicate that quarters and tenths are especially pragmatically relevant alternatives, under such an analysis, while the presence of other potential scale points (such as thirds) does not give rise to any striking pragmatic effects. The presence of literal responses, especially in cases where the pragmatically stronger alternatives are not obvious, suggests that participants have not felt obliged to give pragmatic responses under these experimental conditions.

There is also potential evidence here against an account of fractions in which we consider them simply to be quantifying over parts—for instance, taking "more than two-fifths" to mean "more than two of the fifths". This approach would explain why some participants obtain interpretations involving the negation of the next point on the scale corresponding to this denominator (in this case, "not more than three-fifths"). However, it fails to predict interpretations also attested in these data which seem to rely upon scale points from other scales. Some participants appear to interpret "more than two-fifths" as implicating "not more than half" (and similarly for other expressions under test). To explain this purely in terms of quantifying over parts, we would need to read this as an instance of "more than two (fifths)" implicating "not more than two-and-a-half (fifths)", an enrichment which is not generally predicted to be available. Thus I take these results to cast doubt on the usefulness of an analysis of fractions in which the nominator is treated as though it were a free-standing numeral. Having said that, the availability of the weaker ("not more than three-fifths") implicature does suggest that the denominator actually present in the numeral is privileged in some weaker sense, and that the use of a particular fraction at least heightens the salience of alternative expressions that share this denominator.

Generally, if these experimental results are an accurate reflection of the reality, we can observe that the concept of granularity requires some modification to be applied to the case of fractions. As discussed earlier, it appears that major and minor scale points are not necessarily aligned in this domain; nor is it clear that coarse-grained scales (defined in terms of the distance between successive representation points) are necessarily less cognitively costly to work with than finer-grained alternative scales. This raises the broader question of whether Krifka's (2009) criteria for granularity scales should be seen as hard-and-fast rules or merely generalisations that admit potential exceptions.

That said, we should exercise some caution in interpreting the experimental results presented here as evidence that tenths, in particular, are necessarily salient scale points. Recall that participants were asked to respond using percentages: this system draws attention to the possibility of divisibility by ten, and could be argued to make the tenths scale points more salient than would otherwise be the case (because they correspond to round numbers in percentage terms). Similarly, this methodology could be argued to privilege quarters and fifths, whose scale points are expressed as round integer percentages, over thirds, sixths, sevenths etc., whose scale points are not.

Even with this caveat, the experimental findings support a view of the fraction system—as represented by hearers—that is more complex than would be supposed based on a naïve application of granularity criteria. Participants are able to access readings of modified fractions that appear to rely on alternatives of the same or coarser granularity, but also sometimes able to access readings that rely on alternatives of finer granularity.

The availability of such alternatives is for instance, crucial in obtaining restricted readings for expressions such as "more than (a) half", as was achieved by a majority of participants in experiment 2. It is striking that these participants all favoured interpretations in which "more than a half" conveyed maximally 75% (in most cases, considerably less). Speculatively, we might note a potential point of contact here with the literature on "most" (e.g. Solt 2016). As discussed earlier, a crucial observation is that "most" is not judged felicitous when referring to quantities that are just over 50%, as in (1), repeated below.

(1) Most of the American population is female.

If it is the case that "more than (a) half" attracts a narrower interpretation than would be predicted on its semantics, this might suggest that "more than (a) half" is a particularly good competitor to "most" for values within its pragmatically typical range (i.e. a little over 50%). This in turn might suggest that the distribution of the meanings or interpretations of “most” should skew higher. Of course, one could counter that “most” would then run into competition from other alternatives (“more than two thirds”, “more than three quarters”)—but it is perhaps reasonable to suppose that they are not such salient options as “more than a half”, and consequently that “most” is more likely to be the optimal expression when we are dealing with values in that somewhat higher range (assuming that the speaker does not wish to commit to a precise value).

Finally, it is interesting to note that a substantial minority of the responses elicited appear to reflect a pragmatic enrichment that does not appear to correspond to a specific and highly salient alternative—for instance, “more than two fifths” being judged to convey a value of 40–47% (two responses in experiment 1). Apart from error on the part of the participants, there are several possible reasons for this. It could be that the participants are giving an impressionistic response based on something like a typicality effect associated with the expression that was used (for instance, having a notion that a 7% range is somehow “about right” for this expression). Alternatively, their response may reflect some kind of compromise between competing possible enrichments, based on different alternatives. A third possibility is that the participants are indeed drawing quantity implicatures based on salient alternatives, but that these are not the kinds of alternatives that have been considered in this chapter—for instance, participants might think that something above 47% would have been better described as “about half”. In order fully to understand the possibility of readings of this latter type, we would need to explore the domain of expressions of proportion in much greater generality. However, it is perhaps reasonable to assume that the availability of such alternatives will be attenuated in experiments which consistently omit these alternatives, as in this case (although note that “about half” was presented as an example in the cover story). We might therefore hope that suitably designed experiments will enable us fairly straightforwardly to control for any interference effects from alternative expressions from outside the domain of interest.

4 Conclusion

The relatively understudied domain of fractions appears to exhibit a complex structure which can be seen as a reflection of cognitive preferences about divisibility and salience. Modified fractions give rise to pragmatic enrichments that resemble quantity implicatures, and these indicate the presence of salient alternatives within the system. In the experiments presented here, quarters and tenths are shown to constitute especially salient scale points, and can potentially be seen as the “coarse-grained” representation points in the domain of fractions, although the normal rules of granularity do not straightforwardly apply here. Future work will aim to map the salience of fractions more thoroughly, and also take into account their relation to other expressions of proportional quantity.

References

Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics, 25*(2), 93–139.

Butterworth, B. (1999). *The mathematical brain*. London: Macmillan.

Carston, R. (1998). Informativeness, relevance, and scalar implicature. In R. Carston & S. Uchida (Eds.), *Relevance theory: Applications and implications* (pp. 179–236). Amsterdam: Benjamins.

Cummins, C. (2012). Modelling implicatures from modified numerals. *Lingua, 132,* 103–114.

Cummins, C., & Katsos, N. (2010). Comparative and superlative quantifiers: Pragmatic effects of comparison type. *Journal of Semantics, 27*(3), 271–305.

Cummins, C., Sauerland, U., & Solt, S. (2012). Granularity and scalar implicature in numerical expressions. *Linguistics and Philosophy, 35*(2), 135–169.

Curtin, P. (1995). *Prolegomena to a theory of granularity*. MA thesis, University of Texas at Austin.

Dehaene, S. (1997). *The number sense*. New York: Oxford University Press.

Fox, D., & Hackl, M. (2006). The universal density of measurement. *Linguistics and Philosophy, 29*(5), 537–586.

Gazdar, G. (1979). *Pragmatics: Implicature, presupposition and logical form*. New York: Academic Press.

Geurts, B. (2006). Take 'five': The meaning and use of a number word. In S. Vogeleer & L. Tasmowski (Eds.), *Non-definiteness and plurality* (pp. 311–330). Amsterdam: John Benjamins.

Geurts, B., Nouwen, R. (2007). 'At least' et al.: The semantics of scalar modifiers. *Language, 83*(3), 533–559.

Geurts, B., & Van Tiel, B. (2013). Embedded scalars. *Semantics and Pragmatics, 6*(9), 1–37.

Horn, L. R. (1989). *A natural history of negation*. Chicago: University of Chicago Press.

Jansen, C. J. M., & Pollmann, M. M. W. (2001). On round numbers: Pragmatic aspects of numerical expressions. *Journal of Quantitative Linguistics, 8*(3), 187–201.

Krifka, M. (2002). Be brief and vague! And how Bidirectional Optimality Theory allows for verbosity and precision. In D. Restle & D. Zaefferer (Eds.), *Sounds and systems: Studies in structure and change, a Festschrift for Theo Vennemann* (pp. 439–458). Berlin: Mouton de Gruyter.

Krifka, M. (2009). Approximate interpretations of number words: A case for strategic communication. In E. Hinrichs & J. Nerbonne (Eds.), *Theory and evidence in semantics* (pp. 109–132). Stanford CA: CSLI Publications.

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

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Nouwen, R. (2010). Two kinds of modified numerals. *Semantics and Pragmatics, 3*(3), 1–41.

Sigurd, B. (1988). Round numbers. *Language in Society, 17*(2), 243–252.

Solt, S. (2016). On quantification and measurement: The case of 'most' and 'more than half'. *Language, 92*(1), 65–100.

Decomposition and Processing of Negative Adjectival Comparatives

Daniel Tucker, Barbara Tomaszewicz and Alexis Wellwood

Abstract Recent proposals in the semantics literature hold that the negative comparative *less* and negative adjectives like *short* in English are morphosyntactically complex, unlike their positive counterparts *more* and *tall*. For instance, the negative adjective *short* might decompose into LITTLE TALL (Rullmann, Dissertation, 1995; Heim, Proceedings of Semantics and Linguistic Theory, vol. 16, 2006, Proceedings of Sinn und Bedeutung, vol. 12, 2008; Büring, Proceedings of Semantics and Linguistic Theory, vol. 17, 2007). Positing a silent LITTLE as part of adjectives like *short* correctly predicts that they are semantically opposite to *tall*; we seek evidence for this decomposition in language understanding in English and Polish. Our visual verification tasks compare processing of positive and negative comparatives with *taller* and *shorter* against that of less symbolically-rich mathematical statements, $A > B$, $B < A$. We find that both language and math statements generally lead to monotonic increases in processing load along with the number of negative symbols (as predicted for language by e.g. Clark and Chase, Cognitive Psychology, 3:472–517, 1972). Our study is the first to examine the processing of the gradable predicates *tall* and *short* cross-linguistically, as well as in contrast to extensionally-equivalent, and putatively non-linguistic stimuli (cf. Deschamps et al, Cognition, 143:115–128, 2015 with quantificational determiners).

D. Tucker (✉)
Northwestern University, 2016 Sheridan Road, Evanston, IL 60208, USA
e-mail: danieltucker2017@u.northwestern.edu

B. Tomaszewicz
Universität zu Köln, Albertus-Magnus-Platz, 50923 Köln, Germany
e-mail: btomasze@uni-koeln.de

B. Tomaszewicz
Instytut Filologii Angielskiej, Uniwersytet Wrocławski, Kuźnicza 22, 50-138
Wrocław, Poland

A. Wellwood
University of Southern California, 3709 Trousdale Parkway,
Los Angeles, CA 90089, USA
e-mail: wellwood@usc.edu

E. Castroviejo et al. (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure*,
Language, Cognition, and Mind 4, https://doi.org/10.1007/978-3-319-77791-7_10

Keywords Experimental semantics · Gradable adjectives · Negation
Comparatives · Morphology · Psycholinguistics · Symbolic processing
Decomposition

1 Introduction

How does formal semantics relate to language understanding? And, how can linguistic processing bear on questions about the atoms of compositional interpretation? Recent proposals in the literature on superlatives (Hackl 2009; Szabolcsi 2012), negative comparatives (Rullmann 1995; Büring 2007; cp. Heim 2008), and positive comparatives (Solt 2015; Wellwood 2012, 2015) have highlighted the compositional role of units below the word level. With negative comparatives, much recent debate has centered on whether forms like *shorter* decompose into LITTLE-TALL plus - ER. We look for evidence of such decomposition in processing, by investigating the time it takes to judge sentences containing *taller* and *shorter* as true or false of simple pictures.

The results of early cognitive psychology studies (Just and Carpenter 1971; Clark and Chase 1972; Trabasso et al. 1971; Clark et al. 1973, *inter alia*) report longer processing times for 'negative' statements vis-à-vis their positive analogues. These effects have been found both for sentences with overt sentential negation (e.g. *The dots are not red* vs. *The dots are red*), as well as sentences featuring 'linguistic negation' (e.g. *Few of the dots are red* vs. *Many of the dots are red*; *A minority of the dots are red* vs. *A majority of the dots are red*; cf. Klima 1964). Throughout this early literature, 'negative' features were consistently found to impact the time it took to process a sentence.

We test for these effects with *taller* (positive) and *shorter* (negative). If negative 'features' are specifically linguistic, then it is possible that such an asymmetry might not be observed with the processing of mathematical statements like $A > B$ and $A < B$. Deschamps et al. (2015) tested a similar hypothesis in their study of *more/less than half* and *many/few*, contrasting processing of those expressions with that of extensionally-equivalent, quasi-algebraic inequalities. They found that the sentences with relevantly negative quantifiers in English took longer to process than the corresponding ones with positive quantifiers, but no such asymmetry was observed for the analogous math statements.

This paper contributes to early results in comprehending negation, but links the processing of negative sentences directly to how the meanings of these sentences are characterized in contemporary formal semantics. Like Deschamps et al. (2015), we examine the effects of polarity on processing linguistic and non-linguistic statements; unlike those authors, we examine the possibility of an additional effect of 'congruence'—whether a statement is true or false of a picture (Just and Carpenter 1971; Trabasso et al. 1971). Congruence played an important role in the construction of early cognitive models of sentence-picture verification with negative statements, and so can support a finer-grained picture of the underlying cognitive processes involved in these tasks.

Our investigation is broadly compatible with research conducted under the banner of the Interface Transparency Thesis, offered to precisify a representational role for formal semantics within the broader project of cognitive science (Lidz et al. 2011). The idea is that cognition, by default, carries out procedures that align with the operations specified in the semantic representation of a sentence. If a thesis like this is correct, investigations of processing will be a useful tool for understanding the nature of speakers' semantic representations in general, in addition to paving the way for tests that mediate between specific representational proposals.

In what follows, we first discuss the recent proposals for decomposition in negative adjectival comparatives in order to motivate our processing studies (Sect. 2.1). Next, we recall both early and recent results investigating the processing of 'implicit' negation in cognitive psychology and in linguistics (Sect. 2.2). Then, we present the results of a sentence-to-picture verification task in English (Sect. 3) and in Polish (Sect. 4). To preview, our results provide support for the decompositional analysis of forms like *shorter* in both languages. Section 5 concludes.

2 Background and Motivation

Positive gradable adjectives like *tall* are morphemes—they are not amenable to further morphological analysis. However, Büring's (2007) theory decomposes negative gradable adjectives like *short* into two parts, glossed LITTLE and TALL (cf. Heim 2008). Evidence for decomposition is seen explicitly on the surface in some languages; in Hixkaryana, the antonym of an adjective like *long* is formed by two pieces, i.e. *kawo-hra*, which Bobaljik (2012) glosses as 'long-not'. Our research brings to bear a new kind of evidence for these questions through an examination of gradable adjectives like *tall* and *short* in English and Polish, seeking a different kind of evidence for decomposition in sentence processing.

In this section, we motivate our experimental project: Sect. 2.1 reviews the decompositional approach in semantics, and Sect. 2.2 discusses relevant contemporary and classic literature that informs our linking hypotheses.

2.1 Morphosyntax and Semantics of Shorter

In the contemporary degree semantics tradition, *tall* is analyzed as involving a relation between individuals and their heights, and a sentence like (1a) is interpreted as a comparison between those heights. 'Heights' are formalized as degrees or sets of degrees, and gradable adjectives like *tall* as relations between individuals and those degrees (Cresswell 1976; Heim 1985, 2001; Kennedy 1999, among many others).

The question for this section is: how does the analysis of comparatives with *tall* relate to those with *short*, as in (1b)?

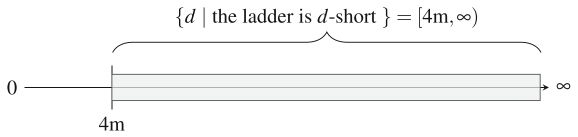
(1) a. Al is taller than Bill is.
b. Bill is shorter than Al is.

(1a) and (1b) stand in a mutual entailment relationship: competent speakers of English intuitively infer that if (1a) is true, (1b) is guaranteed to be true, and vice versa. Is this entailment relation due to their shared *forms*, or something else? On the traditional view, speakers' intuitive awareness of this relationship is not a matter of logic, per se: if both *tall* and *short* are atomic, then their dual nature isn't syntactically 'visible'. Kennedy captures the mutual entailment relation by way of something like a meaning postulate: where S is a scale, pos_S is a positive adjective associated with S and neg_S is its antonym, $pos_S(x) > pos_S(y) \Leftrightarrow neg_S(y) > neg_S(x)$ (Kennedy 2001, p. 56).

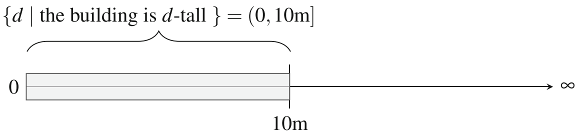
Büring's (2007) decompositional approach, in contrast, supports an analytic relationship between (1a) and (1b). His analysis begins by considering Kennedy's (2001) explanation of the oddity of (2), which is argued to follow from the hypothesis that *tall* and *short* relate individuals to incommensurable sorts of degrees, positive and negative. More formally, the measure function expressed by the negative antonym, SHORT, maps the entity referred to by *the ladder* to a set of degrees like that in (3), while TALL maps *the building* to a set of degrees like that in (4).[1] What Heim (2008) calls *Kennedy's constraint* is that - ER cannot compare positive and negative degrees.

(2) ? The ladder is shorter than the building is tall. ?HEIGHT

(3)


$\{d \mid \text{the ladder is } d\text{-short}\} = [4\text{m}, \infty)$
0
∞
4m

(4)


$\{d \mid \text{the building is } d\text{-tall}\} = (0, 10\text{m}]$
0
∞
10m

Büring points out that, as given, Kennedy's explanation for (2) incorrectly predicts that (5) should be odd as well. Since, as Kennedy suggests, a negative adjective like *short* introduces a negative set of degrees, and a positive adjective like *wide* introduces a positive set of degrees, (5) should also be anomalous.

(5) The ladder is shorter than the building is wide. LENGTH

[1] Note that Kennedy's analysis differs from Rullmann's in that Rullmann had the negative antonym 'flip' what was otherwise a positively-oriented scale (i.e. reverse the ordering relations). In contrast, Kennedy (and subsequent authors presupposing his ontology) proposes that negative antonyms introduce sets of degrees that extend from a point d to infinity, the complement of the set introduced by the positive antonym (see especially Kennedy 2001, p. 55, examples (46) and (48), for discussion).

Büring suggests that decomposition is critical to understanding this pattern. By decomposing *short* into the pieces TALL and LITTLE (where LITTLE TALL is semantically equivalent to Kennedy's SHORT), he is able to argue that the component LITTLE is also shared with the decomposed form of *less* (i.e. LITTLE - ER; Heim 2006). This raises the potential for (1b) to be analyzed as ambiguous between two structures, one containing the bundling [LITTLE - ER] TALL and the other - ER [LITTLE TALL]. (5) would be interpretable on the first bundling as a less-than relation between the positive degrees introduced by TALL and WIDE. It would not be interpretable on the other bundling, since that would express a greater-than relation between the negative degrees introduced by LITTLE TALL and the positive degrees introduced by WIDE, which is barred by Kennedy's constraint.

This analysis can account for the contrast between (2) and (5) as follows. In principle, there could be two bracketings for (2), but either would be problematic. On the bundling - ER [LITTLE TALL] for *shorter*, (2) would express a greater-than comparison between positive TALL and negative LITTLE TALL, barred by Kennedy's constraint. If *shorter* were bundling [LITTLE - ER] TALL, (2) would express a less-than comparison between two instances of positive TALL. This last structure is, presumably, barred by an independent rule or preference that the second of a pair of identical adjectives delete in the *than*-clause of a comparative (cf. Bresnan 1973).

In addition to accounting for (2) and (5), Büring's account extends to cases of ambiguity with *less high* and *lower* that are not evidenced by comparatives with their antonym *higher*, (6a)–(6c) (Seuren 1973; Rullmann 1995). (6a) describes a helicopter flying some degree higher than the maximal height a plane can safely fly, while both (6b) and (6c) can describe a helicopter flying some degree lower than the maximal height a plane can safely fly, or some degree lower than the minimal height a plane can safely fly. This pattern is predicted if LITTLE is able to Quantifier Raise (Lakoff 1970; May 1977; Heim and Kratzer 1998, *inter alia*) in the *than*-clause higher or lower than *can*. (See also Rullmann 1995 for relevant data involving NPI licensing.)

(6) a. The helicopter was flying higher than a plane can fly. NOT AMBIGUOUS
b. The helicopter was flying less high than a plane can fly. AMBIGUOUS
c. The helicopter was flying lower than a plane can fly. AMBIGUOUS

Though promising, such an account faces challenges. As Heim (2008) points out, an account like Büring's would seem to predict that adjectives with *less* should always be substitutable with their negative antonym and *-er* without a change in meaning. So far this prediction is not correct in the general case. Heim shows that, while (7a) can be judged true if Polly's speed may, but needn't, exceed Larry's (perhaps because she has more time to get to her destination), (7b) cannot be read this way: (7b) only has the reading where whatever speed Polly drives, it *has to* be less than Larry's.

(7) a. Polly needs to drive less fast than Larry needs to drive. AMBIGUOUS
b. Polly needs to drive more slowly than Larry needs to drive. NOT AMBIGUOUS

Nonetheless, rolling-back the decompositional analysis for *short* entirely would, as Heim notes, have trouble explaining contrasts like that between (2) and (5). In light of this and other data, Heim posits that there are in fact two distinct LITTLEs, a scopally-mobile one for the decomposition of *less*, and a scopally-immobile one for the decomposition of *short*. One question that potentially arises for this part of her proposal is why the sentences in (8) 'feel different'; if (8a) has an instance of a covert LITTLE, and (8b) results from LITTLE morphologically exerting itself on the adjective, why does (8b) seem more difficult to understand than (8a)?[2]

(8) a. The ladder is shorter than the doorway is wide.
b. ? The ladder is shorter than the doorway is narrow.

Distinguishing the finer details of these proposals is not our focus. Rather, we assume that the linguistic evidence amassing in favor of a decompositional analysis of *shorter* is strong, at least strong enough to warrant further investigation. Our interest is in the fact that decompositional proposals can be seen to make explicit predictions about sentence comprehension.

2.2 Relating Language and Vision

How can the decompositional approach be tested in processing? In what follows, we draw a link with research in classic and contemporary research concerning how semantic representations might make contact with extralinguistic cognition. Of primary interest is early research on the processing of different types of 'linguistic negation', as well as recent results targeting similar questions. Ultimately, we suggest that decompositional approaches explicitly predict that negative adjectival comparatives should take longer to judge true or false than positive comparatives.

Beginning with the cognitive psychology literature, many proposals in the late 60 s and early 70 s were made as to what sorts of processing mechanisms would need to be deployed when people considered the truth or falsity of a sentence against a picture. While this literature is broad, we can draw some important conclusions from it. The first is that positive statements are more readily processed than negative (polarity effects), and that it is easier to verify a statement when it is true of its accompanying scene than when it is false (congruence effects).

A core assumption from this early work is that "perceptual events are interpreted" (Clark and Chase 1972), specifically into a sort of propositional format. One motivation for this idea is the simplicity that it affords to understanding how, ultimately, a sentence meaning and a representation of a picture can be compared. If sentence meanings and perceptual events are encoded in a common representational format, the comparison can simply be one of identity—not merely truth-conditional identity,

[2] Possibly more importantly, Beck (2013) has found some slipperiness in the judgments of speakers for the relevant scope data. Thus, so far it seems that the evaluation of decompositional analyses from the perspective of semantic theory should not yet hang on the data in (7).

though this ultimately plays a role—specifically, *identity of representation*. We will be more explicit about this shortly.

Separately from the representational assumptions, models of sentence-picture matching were designed to account for the response latencies of judgments in extremely simple tasks.[3] Typically, this type of task would involve a participant reading a sentence, considering a picture, and indicating whether they understand the sentence to be true or false of the picture. Two importantly different types of tasks were found to make different demands on the participant, and the models were designed to make the right predictions accordingly: the Sentence-to-Picture verification task and the Picture-to-Sentence verification task, which differ only in whether the picture or the sentence is presented first. We focus on the first type of task, since it will be most relevant for our own experiments (though see Sect. 3.4).

On the "Sentence-First Model" (Clark and Chase 1972), the process of comparing a sentence with a picture proceeds in four stages, summarized in (9). Stage 1 involves linguistic decoding/encoding, and Stage 2 involves nonlinguistic perceptual/conceptual processing that eventuates in a representation given in the same general format as the sentence. This general format is thought to be important for comparison to proceed at Stage 3, which might also involve *transformations* of a given representation before the final check for identity. At Stage 4, participants record their judgment, typically using a button press.

(9) **"Sentence-First" processing stages** (Clark and Chase 1972)

i. **Stage 1**: form a mental representation of the sentence
ii. **Stage 2**: form a mental representation of the picture
iii. **Stage 3**: compare the two representations
iv. **Stage 4**: produce a response

Stage 3 is thus crucial. In this model, it involves checking whether two representations 'mean' the same thing, where 'meaning the same' is cashed out in terms of representational identity (Clark 1969b calls this the 'principle of congruence'). However, it would be overly simplistic to assume that this amounts merely to truth-conditional equivalence, or mere representational equivalence based on the initial representation of the sentence or picture. Checking for mere truth-conditional equivalence would predict that evaluating *A is above B* and *B is below A* should take the same amount of time in the same contexts. However, studies have repeatedly shown that there is a cost to sentences with *below* compared to *above*. On the other hand, merely checking whether the two representations match would be overly restrictive: comparing linguistic BELOW(A, B) and visual ABOVE(B, A) should then be judged as 'false', which would be incorrect.

Thus, according to Clark and Chase (1972, p. 478), "Stage 3 must be endowed with a series of comparison operations, each checking for the identity of the subparts

[3] The most explicit overview of the methodology and models is given by Clark and Chase (1972), who cite Clark (1970), Trabasso et al. (1971) as important precursors, as well as an extensive list of even earlier results that informed their view.

of the two representations, and each adding to the computation of *true* and *false*". There are many different ways, in the modern era of computational analogies in semantics research, to conceptualize such 'comparison operations' (e.g., reduction to a canonical form, comparison of evaluation consequences, etc.); we will attempt to remain at a fairly informal level here.

So what parameters affect the latency of a participant's judgment, and how? Clark and Chase (1972) posit a number of parameters, each of which additively contributes (citing Sternberg 1969) to the total response time. The parameters relevant to the present study are summarized in (10). A cost of $+a$ should be observed for evaluating sentences with the 'marked' or 'negative' member of a pair of linguistic opposites (per the hit observed for *below*). And, a cost of $+b$ should be observed for the operations required to determine that the linguistic and visual encodings mismatch (the time for performing operations at Stage 3, i.e. *falsification*). In previous work, these two factors did not interact (Clark and Chase 1972, p. 487). Finally, there is an overall and independent cost of t_0 for the time to plan and execute the response.

(10) **Parameters affecting response latency**

i. a-cost of 'linguistic negation'; *Below* time

ii. b-cost of comparison operations; *Falsification* time

iii. t_0-'wastebasket parameter'; *Base* time

Somewhat differently methodologically from these early studies are the recent papers in the Interface Transparency suite (Pietroski et al. 2009; Lidz et al. 2011). These studies all made use of the Sentence-to-Picture verification task, but limited the viewing time for the picture to 150 or 200 ms, whereas the classic studies tended to give participants essentially as much time with the picture as was necessary to make the judgment. With a restricted viewing time, it was assumed that participants' response latencies reflect operations over the initial representation of the scene in memory.

More recently, Deschamps et al. (2015) tested similar hypotheses but with different linguistic stimuli, and a different experimental set-up. They investigated polarity contrasts with the quantifiers *more/less* and *many/few* versus quasi-mathematical expressions in a verification task that required numerical estimation and comparison. We also test the processing of math expressions against expressions in natural language (English and Polish), asking whether the 'simpler' math expression leads to different effects. Our study differs in that we test comparative adjectives, provide a shorter viewing time for the picture (theirs was 2500–2800 ms), and we include tests for congruence effects.[4]

[4] A further difference is that Deschamps et al. (2015) presented their linguistic statements auditorily.

3 Experiment 1: English Sentence-Picture Matching

We test the predictions of decompositional analyses of *shorter*, which posit that the semantic representation of sentences containing this form are strictly more complex than (and in fact contain) the representation of equivalent sentences with *taller*. In light of the early and recent results indicating that the marked member of a positive-negative pair induces additional processing cost, we expected *shorter* should take longer to process than *taller*. We contrast this processing with that of prima facie 'simpler' mathematical statements like '$A > B$' and '$A < B$'.

3.1 Design and Participants

We designed a sentence-to-picture verification task in a two-2 × 2 design according to linguistic and non-linguistic statements. In our task, participants were presented with a statement, followed by a picture, and asked to judge whether the statement accurately described the picture. Each of our two-2 × 2 sub-designs corresponded to the 'language' that the statement was presented in, either English or Math.

For each of the English and Math sub-designs, we manipulated POLARITY (positive, negative) and CONGRUENCE (congruent, incongruent). As can be seen in Table 1, we considered the expressions that corresponded to a greater-than comparison as 'positive', and those which corresponded to a less-than comparison as 'negative'. Thus, the factor POLARITY varied whether the statement was positive (*taller than*, >) or negative (*shorter than*, <), for a total of 8 statements. The factor CONGRUENCE varied whether the statement was true of the paired picture or not, corresponding to the congruent and incongruent conditions, respectively.

Stimuli. We created 20 pictures featuring two lines marked A and B. The shorter line always appeared in one of two sizes (24 or 42 pixels, with a 160 pixel distance in between), and the longer line differed from the shorter by one of five different length ratios (0.5, 0.75, 0.833, 0.875, 0.9). Figure 1 shows a subset of these visual stimuli: a ratio difference of 0.5 for an "A wins" picture (a) and a "B wins" picture (b); and a ratio difference of 0.75 for an "A wins" picture (c) and a "B wins" picture (d). In half of the pictures, the longer line was labeled 'A' and the shorter line was labeled 'B'; in the other half of the pictures, the shorter line was labeled 'A' and the longer

Table 1 English and Math statements used in Experiment 1

	English	Math
Positive	*A is taller than B*, *B is taller than A*	$A > B$, $B > A$
Negative	*A is shorter than B*, *B is shorter than A*	$A < B$, $B < A$

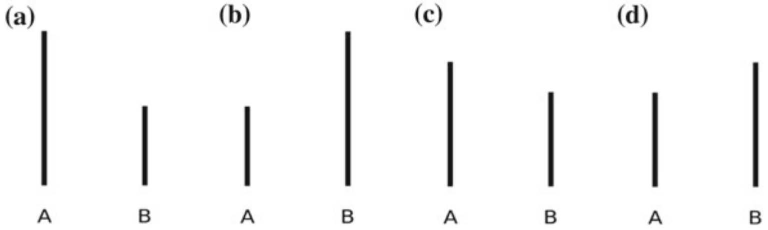


Fig. 1 Sample picture stimuli used in Experiment 1

line was labeled 'B'. Each of these pictures was paired with each of the 8 statements in Table 1. Every possible sentence-picture pair delivered a total of 160 trials.

Procedure. The experiment was designed using jsPsych, a JavaScript library for creating behavioral experiments in a web browser (de Leeuw 2015). After consenting to participate, participants were presented with instructions for the experiment (see below). Following this, participants completed the 160 trials,[5] each of which was structured as follows. At the start of the trial, a statement was presented in the center of the screen, along with an indication that the statement would remain visible until the participant pressed the spacebar. After pressing the spacebar, a center-oriented fixation cross appeared for 200 ms, followed by a display of the picture for 200 ms. 200 ms after the display of the picture, a center-oriented '?' appeared, along with an indication to press 'f' if the statement matched the picture, or 'j' otherwise. Participants had a maximum of 5 s to record their judgment. Trials were organized into 4 blocks, each defined by one combination of linguistic/non-linguistic statements and line order (A first vs. B first). The order of presentation of the blocks and of the trials within the blocks was completely randomized.

Instructions to Participants. The exact instructions given to participants were as below. As we were primarily interested in the timing of the response to our stimuli, we explicitly indicated that participants should attempt to make their judgment as quickly as possible.

> Welcome to the experiment!
>
> There are 160 trials in this experiment. Each trial will consist of a statement, an image, and your response. The statement may be in a natural or mathematical language. You will have as much time as you wish to view the statement, and then press spacebar to see the image. The image will be shown for only 1/5 of a second. Immediately afterwards, your task is to judge whether the statement accurately describes the image.
>
> If the statement accurately describes the image, press the letter **f** on the keyboard.
>
> If the statement does not accurately describe the image, press the letter **j** on the keyboard.
>
> Please make this judgment as quickly as possible. The experiment will automatically advance to the next trial after 5 seconds of no response. The whole experiment should take no longer than 15 minutes to complete.
>
> Ready? Press spacebar to begin the experiment.

[5] No filler task items were used in this experiment or in the second experiment reported below.

Participants. We recruited 15 participants through a Human Intelligence Task (HIT) posted on Amazon's Mechanical Turk. We restricted eligibility to native speakers of English living in the United States who had completed at least 1000 HITs on Mechanical Turk with a HIT approval rate of at least 99%. Participants were compensated $2.50 for participating, and took an average of 13.5 minto complete the HIT. No Mechanical Turk master workers were recruited for this study.

3.2 Predictions

We assume the decompositional analysis of English negative comparatives in line with Büring (2007), the 'simple' hypothesis about math statements, and combine these assumptions with the predictions of the Sentence-First model of Clark and Chase (1972). In what follows, we discuss the predictions for English and math statements separately, and in turn.

Linguistic Stimuli (English). On the decompositional analysis, the semantic representation of a positive comparative is contained within the representation of a negative comparative. Abstracting away from many details, a proposal like Büring's can be summarized as in (11). The major operand of the semantic representation is ER, which specifies a greater-than relation between two quantities. These quantities are provided by TALL(A) and TALL(B) in (11a), and by an operation over such quantities (e.g. complementation) provided by LITTLE, (11b).

(11) a. ⟦A is taller than B.⟧ = ER(TALL(A),TALL(B))

b. ⟦A is shorter than B.⟧ = ER(LITTLE(TALL(A)), LITTLE(TALL(B)))

In light of the early cognitive psychology literature, we expected that the added presence of LITTLE should correspond to an increase in processing load: processing (11b) requires to processing something like (11b) in addition to the contributions of the two instances of LITTLE. Such additional processing steps should correspond to an increase in RTs. Furthermore, we expect an additional cost of evaluating the the semantic representation in situations where it is false of the scene—when the two are *incongruent*.

On the simplest version of the Sentence-First model, these two effects—of polarity and congruence—are expected to be additive to RT: both negativity in the sentence and falsity of the sentence given scene induce independent processing costs. Thus we predicted the fastest RTs in the positive congruent condition, and the slowest in the negative incongruent condition. The expected results can be depicted as in Fig. 2.[6]

[6] Indeed, this is the pattern found by Clark and Chase (1972), when participants evaluated the sentences *A is above B* and *A is below B* in a Sentence-Picture verification task. However, Trabasso et al. (1971) reported an interaction between polarity and congruence, in which RTs were greater for negatives in incongruent situations, yet greater for positives in congruent situations. These results, however, were found in a Picture-Sentence verification task where the contrast in negativity was sentential negation, e.g.: *The patch is/isn't orange*.

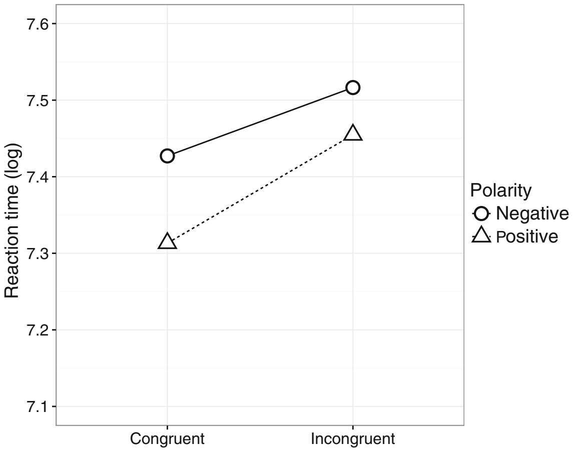


Fig. 2 Predicted main effects of polarity and congruence for natural language, given the decompositional analysis of forms like *shorter* and the Sentence-First model of Clark and Chase (1972)

What about the predictions for accuracy? Clark and Chase (1972) report overall error rates of 9.7% in their task using *above* and *below*, but that these were unequally distributed between the 'positive' conditions with *above*, and the 'negative' conditions with *below*. They report that, in general, higher error rates were observed in conditions where 'more mental operations' needed to be carried out. We thus expected overall error rates to be similar in our task: broadly, higher RTs should pattern with higher error rates.

Non-linguistic Stimuli (Math). Our expectations for Math statements are somewhat less clear. On the one hand, Deschamps et al. (2015) report no effect of polarity on processing quasi-algebraic inequalities, in contrast to English sentences with *many/few* and *more/less*. Such an expectation aligns with the 'simple' hypothesis that statements like $A > B$ and $A < B$ are essentially non-linguistic, and representationally transparent (i.e. non-decompositional), and so should be processed differently than linguistic statements.

However, we might expect an effect of congruence here—whether the statement matches the scene. Clark and Chase's (1972) characterization of congruence effects was that they were essentially an independent consequence of comparing two mismatching representations. In light of this, we do not expect such effects to apply only to linguistic statements. This amounts to the expectation that incongruent situations will lead to increased RTs for processing Math statements.[7]

3.3 Analyses and Exclusions

We report the results of linear and logistic mixed effects model comparisons with maximal random effects structures (i.e. including random intercepts and slopes by

[7] As noted above, congruence effects were not discussed by Deschamps et al. (2015).

subject and item; best generalization for LMEMs, Barr et al. 2013). For all analyses, we used an orthogonal contrast coding scheme that assigned values of −0.5 and 0.5 to each level of POLARITY and CONGRUENCE, respectively. The significance levels (p-values) that we report are derived from comparison of the maximal model in each case, against the same model minus the relevant parameter.

Analyses for RT measures were conducted on the log-transformed RT data to respect the normality assumptions of linear mixed effects models (Gelman and Hill 2007). We plot the log-transformed RT measure, and report both the results in both logRT and milliseconds (ms) for readability. Analyses for response accuracy were summarized by participant by condition and are reported as mean percent correct.

Of the 2400 datapoints we collected, 45 were excluded (1.9%) for either a missed response (i.e., the participant failed to respond within the 5 s time window), or because the response time was greater than three standard deviations from that participant's mean RT. Each main effect reported in the next section was based on an average of 585 observations per condition, while each interaction was based on an average of 295 observations per condition.

All analyses were conducted using R's *lme4* package (Bates et al. 2014).

3.4 Results: RTs

We conducted two separate linear mixed effects model comparisons on the log-transformed RT data. The results for both English and Math are presented in Fig. 3.

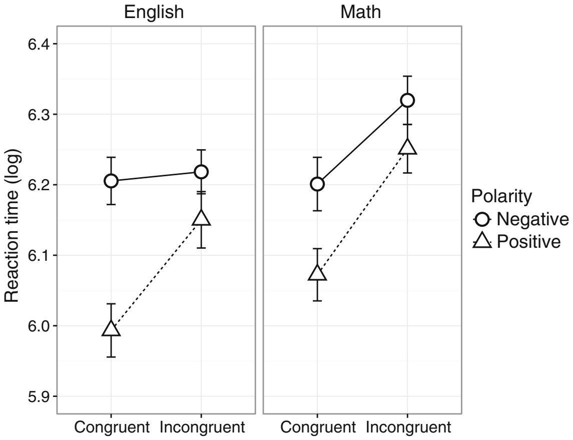


Fig. 3 Mean log RTs and SEs by POLARITY and CONGRUENCE for the linguistic (English) and non-linguistic (Math) sub-experiments of Experiment 1

3.4.1 Linguistic Conditions (English)

Participants took longer to evaluate sentences with *shorter* than with *taller*. This was reflected in a robust main effect of POLARITY (means: negative 6.21, positive 6.07, $\beta = -0.14$, $SE = 0.03$, $\chi^2 = 12.56$, $p < 0.001$) in the predicted direction: RTs in the negative conditions were longer than in the positive conditions (means, in ms: negative 634.12 ms, positive 586.70 ms).

Additionally, participants took longer to reject false statements than to accept true statements. This was reflected in a marginal main effect of CONGRUENCE (means: congruent 6.32, incongruent 6.39, $\beta = -0.09$, $SE = 0.05$, $\chi^2 = 3.34$, $p = 0.067$), in accord with our predictions: a statement's truth or falsity with respect to its accompanying picture had a non-trivial impact on associated RTs (means, in ms: congruent 593.10 ms, incongruent 627.53 ms).

Moreover, accepting true sentences with *taller* was much faster than could be accounted for with just the main effect of congruence. This was reflected in an interaction between POLARITY and CONGRUENCE ($\beta = -0.13$, $SE = 0.07$, $\chi^2 = 3.89$, $p = 0.048$). RTs in the positive congruent condition were shorter than in the negative congruent condition (means: negative 6.21, positive 5.99; means, in ms: negative 636.99 ms, positive 549.36 ms), while there was little difference between the negative incongruent condition and the positive incongruent condition (means: negative 6.22, positive 6.15; means, in ms: negative 631.21 ms, positive 623.90 ms).

3.4.2 Non-linguistic Conditions (Math)

Participants took longer to evaluate Math statements with $<$ than with $>$. This was reflected in a strong main effect of POLARITY (means: negative 6.26, positive 6.16, $\beta = -0.10$, $SE = 0.04$, $\chi^2 = 6.40$, $p = 0.01$), in which reaction times in the negative conditions were substantially longer than for the positive conditions (means, in ms: negative 678.06 ms, positive 610.35 ms). These results stand in contrast to Deschamps et al. (2015), who report no asymmetry in the evaluation of positive and negative Math statements.

Participants also took longer to reject Math statements that didn't match the picture than to reject those that did. This was reflected in a strong main effect of CONGRUENCE (means: congruent 6.14, incongruent 6.29, $\beta = -0.15$, $SE = 0.04$, $\chi^2 = 9.45$, $p = 0.002$): the incongruent conditions took longer to evaluate than the congruent conditions (means, in ms: congruent 602.63 ms, incongruent 685.76 ms). This congruence effect was expected as reflecting a general cost of rejecting false statements.

All of the effects of congruence were accounted for in the main effects, in contrast to our results for English. That is, there was no interaction between POLARITY and CONGRUENCE ($\beta = -0.07$, $SE = 0.11$, $\chi^2 = 0.40$, $p > 0.5$). RTs in the positive congruent condition were faster than in the negative congruent condition (means: negative 6.20, positive 6.07; means, in ms: negative 636.63 ms, positive 568.75 ms). Similarly, RTs in the positive incongruent condition were faster than in the negative

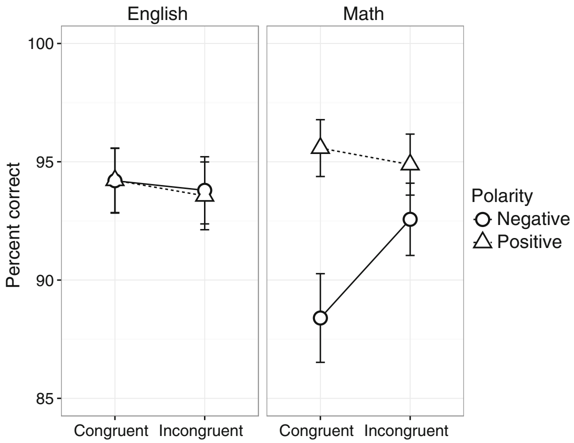


Fig. 4 Mean subject accuracy and SE by POLARITY and CONGRUENCE for the linguistic (English) and non-linguistic (Math) sub-experiments of Experiment 1

incongruent condition (means: negative 6.32, positive 6.25; means, in ms: negative 719.07 ms, positive 652.12 ms).

3.5 Results: Accuracy

To assess response accuracy (a binary variable), we conducted model comparisons over mixed effects logistic regressions. The results are presented graphically in Fig. 4, with accuracy plotted in terms of the percentage of correct responses summarized by participant in each condition.

3.5.1 Linguistic Conditions (English)

Our participants' accuracy was not any worse for sentences with *shorter* than for those with *taller*. This was reflected in the lack of effect of POLARITY on mean response accuracy (means: negative 94.0%, positive 93.9%, $\beta = 0.02$, $SE = 0.31$, $\chi^2 < 0.01$, $p > 0.9$). This result is unexpected in light of the early cognitive psychology literature, which found an inverse correlation between reaction time and response accuracy.

Participants were no less accurate at rejecting false statements than at accepting true statements. That is, we found no effect of CONGRUENCE on mean response accuracy (means: congruent 94.2%, incongruent 93.7%, $\beta = 0.22$, $SE = 0.24$, $\chi^2 = 0.80$, $p = 0.4$): a statement's veracity with respect to its accompanying picture made little difference.

Analyses revealed no interaction was found between POLARITY and CONGRUENCE ($\beta = -0.05$, $SE = 0.48$, $\chi^2 = 0.01$, $p > 0.9$); there was no difference in mean response accuracy in the negative versus positive congruent conditions (means:

negative 94.2%, positive 94.2%). Such was also the case in the negative and positive incongruent conditions (means: negative 93.4%, positive 93.6%).

3.5.2 Non-linguistic Conditions (Math)

In contrast to the results for English, participants were less accurate at evaluating sentences with $<$ than with $>$. This was revealed in a marginal main effect of POLARITY (means: negative 90.5%, positive 95.2%, $\beta = 0.78$, $SE = 0.36$, $\chi^2 = 3.83$, $p = 0.05$): average response accuracy was lower for the negative conditions than for the positive conditions.

Similar to the results for English, participants were as accurate at rejecting false statements as accepting true statements. That is, we found no main effect of CONGRUENCE on mean response accuracy (means: congruent 92.0%, incongruent 93.7%, $\beta = -0.22$, $SE = 0.28$, $\chi^2 = 0.57$, $p > 0.5$): whether the sentence was true of the picture made little difference to average response accuracy.

Finally, no interaction was found between POLARITY and CONGRUENCE ($\beta = 0.78$, $SE = 0.50$, $\chi^2 = 2.31$, $p = 0.1$); accuracy was lower in the negative congruent condition than in the positive congruent condition (means: negative 88.4%, positive 95.6%); such was also the case for the negative and positive incongruent conditions (means: negative 92.6%, positive 94.9%).

3.6 Discussion

In Experiment 1, we found that sentences with *shorter* took longer to process than sentences with *taller*, supporting the decompositional analysis on which *shorter* is strictly more representationally complex than *taller*. Furthermore, evaluating false statements took longer than evaluating true statements (in both English and Math). These results are in line with the earlier results for *above* and *below* and other pairs reported for previous Sentence-to-Picture matching tasks (cf. Clark and Chase 1972). We also found an interaction effect that was not observed in earlier works.

In the Math sub-experiment, we found that statements with $<$ took longer to process than statements with $>$, and that statements which were false of the accompanying picture took longer to process than statements that were true of the picture. In this sub-experiment, we found no interaction effect, suggesting that these results provided a better match to the predictions of the Sentence-First model proposed by Clark and Chase (1972). This is not what a simple hypothesis about how math statements are processed would predict, and it contrasts to the findings of Deschamps et al. (2015), who found that processing quasi-algebraic inequalities was qualitatively

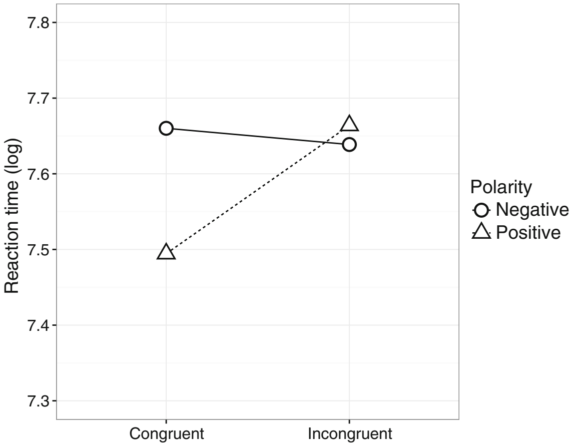


Fig. 5 Results of Clark and Chase's (1972) Picture-to-Sentence verification task with *above* (positive) and *below* (negative), modeled after the presentation in Clark et al. (1973)

different than the processing of natural language. We do not have a good explanation for why their results differ from ours.[8]

We found one major difference between the English and Math sub-experiments, which was the interaction between polarity and congruence. Evaluating true and false statements with *shorter* took roughly the same amount of time, however evaluating true statements with *taller* was much faster than evaluating false statements with *taller*. We did not find a corresponding effect in the Math sub-experiment. What could explain this difference?[9] One possibility, again considering the discussion in Clark and Chase (1972), is that our speeded task involves a different sort of processing for English than for their Math correspondents.

One line of inquiry is suggested by considering the results that those authors found testing sentences with *above* and *below* in a Picture-to-Sentence verification task, as in Fig. 5. On the surface, the "Sentence-First" processing model in (9) and a "Picture-First" model should not look all that different; Stage 1 in a Picture-First model would involve forming a representation of the picture, and Stage 2 forming a representation of the sentence, as opposed to vice versa. Yet, Clark and Chase (1972) crucially assumed that, absent a linguistic cue, there was a default, positive encoding

[8] It is possible that our participants understood the math statements in terms of natural language translations like *A is greater/less than B*, which lead to the language-like effects. The quasi-algebraic expressions tested in Deschamps et al. (2015) consisted of blue and yellow squares on both sides of the $>$ and $<$ operators. Such representations might be less likely to be translated into natural language than $A > B$ and $A < B$, potentially accounting for the differences between our study and theirs.

[9] An anonymous reviewer notes that we so far have not directly compared these two sub-experiments, and so haven't shown that they are statistically different from one another. Conducting a post-hoc LMEM comparison over the combined data from the English and Math sub-experiments, we found no main effect for the contrast-coded factor LANGUAGE (English vs. Math), nor any interactions with that factor. Subsequently, in the text, we focus on the qualitative difference that can be seen in Fig. 3, and which was borne out in the independent 2 × 2 analyses.

of a scene; when there was a linguistic cue, sentence encoding could impact picture encoding.

That is, the Sentence-First model assumes that the representation of the sentence formed during Stage 1 impacts how the picture is encoded during Stage 2. Clark and Chase assumed that given a sentence specified with *above*, the picture will be encoded in terms of the matching ABOVE relation, and given a sentence specified with *below*, the picture will be encoded with the matching BELOW relation. Increased latencies for polarity are seen to arise due to the negative feature on *below* (*a*—Below time), and for congruence due to the mismatching subjects (*b*—Falsification time). An instance of this type of processing is shown schematically in (12).

(12) **"Sentence-First" processing for a negative incongruent trial, Clark and Chase (1972)**

a. **Stage 1**: Read *A is below B* ⇒ BELOW(A, B) $+a$

b. **Stage 2**: See picture of A above B ⇒ BELOW(B, A)

c. **Stage 3**: Are A and B in the same position in the relation? ⇒ No $+b$

d. **Stage 4**: Respond with button press $+t_0$

The major surface difference between this model and the "Picture-First" model is the latter's assumption of a default, positive encoding of the picture at Stage 1, which is then checked against whatever the sentence encoding is. The default encoding in Clark and Chase's experiment is specified in terms of ABOVE. In cases where the sentence and the picture encodings don't immediately match (i.e. whenever it is not the case that the encoding of the picture is ABOVE(A, B) and the sentence is *A is above B*), one will have to transform the sentence to put it in a format that the comparison operations can understand.

Importantly, our Experiment 1 differs from these earlier studies in that we imposed a 200 ms viewing time for the picture, a threshold more often imposed in contemporary experiments (see Sect. 2.2). It is possible that the 'preference' to encode visual scenes in positive terms manifests as a necessity under this kind of time pressure given that it takes approximately 200 ms to initiate a regular saccade movement in response to an unexpected stimulus—with an expected stimulus, peripheral vision may be sufficient (Carpenter 1977; Allopenna et al. 1998).

Thus, there may be a way of thinking about the processing demands imposed in the present task which is relevant to predicting the differences between the English and Math sub-experiments. Suppose that the scene is always encoded positively under the 200 ms time pressure, and that encoding a statement negatively always imposes its own cost (x). Assume that there is an additional 'check' imposed for matching English with the picture on whether the subject of the sentence corresponds to the first position of the (positive) relation encoded by vision (y).[10] Along with the cost of congruence (z), the sum of the processing costs would be as in (13) for one type of trial.

[10] Clark and Chase (1972) point to studies by Huttenlocher (1969) and Clark (1969a, b) for evidence that the 'theme/rheme' distinction is important in these tasks, which is reflected in the specific type of comparison operation that Clark and Chase posit for Stage 3, shown in (12).

(13) **English negative incongruent trial**

a. language: ER(LITTLE(TALL(A)), LITTLE(TALL(B)))

b. vision: ER(TALL(A),TALL(B))

c. unmarked English form? NO. $+x$

d. subjects match? YES.

e. congruent representations? NO. $+z$

f. Respond with button press $+t_0$

If one applies this reasoning to each of the conditions of the English sub-experiment, we might predict the relative magnitudes of the effects that we found (positive congruent t_0, positive incongruent $t_0 + y + z$, negative congruent $t_0 + x + y$, negative incongruent $t_0 + x + z$). In contrast, if the Math task imposes no such 'check' on whether the 'subject' of the statement corresponds to the first position of the (positively-encoded) visual representation, we might predict the relative magnitudes we found there as well (positive congruent t_0, positive incongruent $t_0 + z$, negative congruent $t_0 + x$, negative incongruent $t_0 + x + z$).

Of course, this is a post-hoc analysis, and it remains unclear why checking the match for 'subject' would differ between English and Math statements in this task (apart from the fact that statements like $A > B$ might not necessitate a notion of 'subject'). However, we do observe a clear difference between English and Math, and it is possible that probing the effects of this type of task on processing with different types of statements could provide new insight into how statements are matched with pictures, and why this might differ across 'languages'.

Regardless, limiting participants to 200 ms appears to have had an important effect on the task demands, at least in the case of sentence-to-picture matching with natural language. We have suggested that, in this case, participants could be relying on a bias to positively-encode a scene in order to actually perform the task under these pressures. In the next experiment, we design a very similar task, but do not impose such stringent restrictions on how long participants have to view the scene.

With respect to response accuracy, it is unclear why the error rates did not pattern with response latencies for *taller* and *shorter* as they did in work on *above* and *below*. As an anonymous reviewer suggests, this could be due to a ceiling effect in the English sub-experiment, since accuracy rates there were very high across the board. In the math sub-experiment, however, accuracy patterned with the RT data: accuracy was lower for statements with $<$ than with $>$, suggesting greater difficulty with the more 'negative' of the pair. However, why the English and Math sub-experiments should differ in this respect is a matter we leave for future research.

4 Experiment 2: Polish Sentence-Picture Matching

We test the predictions of decompositional analyses of negative comparatives in Polish, contrasting this with the processing of math statements. Polish *wyższe* and *niższe* comparatives, (14), have a very similar underlying syntactic structure to those of English (Pancheva 2006), and thus will provide an interesting test for the cross-linguistic robustness of the decompositional proposals so far offered explicitly only for English.

(14) a. A jest wyż-sze niż B
A is tall-er than B

b. A jest niż-sze niż B
A is short-er than B

4.1 Design and Procedure

We conducted a sentence-to-picture verification task like Experiment 1, in a two-2 × 2 design according to linguistic (Polish) and non-linguistic (Math) statements. As before, participants were presented with the statement, followed by a picture, and asked to judge whether the statement matched the picture. Experiment 2 differed in that it was conducted while participants' eyes were tracked. However, because the eye-movement data is not relevant for our reaction time hypotheses, we do not investigate that data here.

As in Experiment 1, for each of the Polish and Math sub-designs we manipulated POLARITY (positive, negative) and CONGRUENCE (congruent, incongruent). The 8 statements we tested, sorted by the levels of these factors, are shown in Table 2. Participants were presented with images very much like those presented above in Fig. 1, except that the two lines were spaced further apart (see below). Unlike in Experiment 1, we allowed participants up to 4 s to view the image; this viewing time is more consonant with that employed in the early cognitive psychology studies.

Stimuli. 20 pictures featuring two lines marked A and B were paired with the 8 statements in Table 2 for a total of 80 pairings. This is half of the number of pairings featured in Experiment 1 in order to keep the time required to complete the experiment under 20 min. Each of the 8 conditions was presented with 10 of the 20 pictures in an alternating fashion (counterbalancing how many images with line lengths of 28 pixels and 42 pixels were presented, and in how many of them the line labeled A or the line labeled B was longer). As in Experiment 1, the shorter line appeared in one of two sizes (24 or 42 pixels), and differed from the longer line by one of five

Table 2 Polish and Math statements used in Experiment 2

	Polish	Math
Positive	*A jest wyższe niż B*, *B jest wyższe niż A*	$A > B$, $B > A$
Negative	*A jest niższe niż B*, *B jest niższe niż A*	$A < B$, $B < A$

different length ratios (0.5, 0.75, 0.833, 0.875, 0.9). The distance between the two lines was 700 pixels, substantially larger than in Experiment 1. This was in order to facilitate tracking of participants' eye-movements during picture verification, and to prevent encoding the scene solely using peripheral vision. As noted above, we only report the behavioral results in this paper.

Procedure. The experiment was run on a Windows PC connected to the SR Research EyeLink 1000 Plus eye-tracker. The participants were first presented with the printout of the instructions and the experimenter answered any questions. Participants then saw the same instructions on the screen followed by a practice session with trial structure parallel to the experimental trials but with different statements and pictures. In a trial, a statement was presented until button press, followed by a picture displayed for up to 4s. Participants pressed the left arrow key on the response box when they decided that the picture matched the sentence, and the right arrow key when they decided that it did not. Accuracy was encouraged by an auditory signal in case of a wrong response. When a response was recorded, or no response was made during the 4 s window, the picture disappeared and a new trial began. Each sentence display and each picture display was preceded by a 1 s pause followed by a fixation point (during which drift-correction was performed). Trials were organized into 4 blocks, each defined by one combination of linguistic/non-linguistic statements and POLARITY. Block order and trial order within blocks was pseudo-randomized (no more than three trials of the same type consecutively). After each block participants were able to take a short break. The experiment took approximately 20 min to complete.

Instructions to Participants. The Polish instructions given to participants are presented below translated into English. Unlike in the previous study, we explicitly indicated that participants should attempt to make their judgment as accurately (as opposed to as quickly) as possible.

> Welcome and thank you for your participation in our experiment!
>
> Your task is to read short sentences and mathematical expressions and to decide if they match the pictures. The accuracy of your responses matters. Before the experiment, there will be a practice session, where you will see some examples.
>
> We begin with the process of CALIBRATION: You will see a black point. Look at its white center. The point will be appearing in a different locations. Track the point.
>
> When calibration is successful, we begin the practice session. First, you will see a cross. Look at its center. Next, the text will appear. When you read it, press the bottom button. You will then see a point. Look at its center. Now the picture will appear. Decide whether the picture matches the text.
>
> If it does, press the LEFT button for YES.
>
> If it doesn't, press the RIGHT button for NO.
>
> Do the same for the following pairs of text and pictures.
>
> The accuracy of your responses matters. If you make a mistake, you cannot go back or repeat.
>
> Are you ready for calibration and practice session? Press the bottom button to begin.

Participants. 32 participants were recruited from the University of Wrołcaw, Poland. One participant was excluded due to calibration failures. Participants received a payment equivalent to 9 EUR for participation.

4.2 Predictions

We assumed the decompositional analysis as extended to Polish negative comparatives, and that simple Math statements would be processed similarly. We again test the predictions of the Sentence-First model of Clark and Chase (1972), and discuss these separately for the Polish and Math sub-experiments.

Linguistic Stimuli (Polish). In Experiment 1, the pattern of RTs for English was different than that predicted by the Sentence-First model. We speculated in Sect. 2.2 that this was due to the 200 ms time window in which participants had to view the picture, which seems to have selectively impacted the processing of natural language. A 200 ms constraint on the presentation time prevents the initiation of a saccades in response to an unexpected stimulus (Carpenter 1977; Allopenna et al. 1998), and thus may have contributed to this pattern. With a 4 s window for viewing the picture, the predictions of Clark and Chase's (1972) model of Sentence-to-Picture verification should unambiguously apply. Moreover, since the Polish comparative sentences in Table 2 are compatible with the same syntactic and semantic analyses as their English counterparts in Table 1, any such effects can be attributed to decomposition. Thus, we predicted a main effect of POLARITY, with longer RTs corresponding to the processing of the two instances of LITTLE. We also predicted a main effect of CONGRUENCE—whether the statement was true of the picture. As in the earlier studies reported in the literature, we expect only additive effects of these two factors (i.e., no interaction).

Non-linguistic Stimuli (Math). If the processing of Math statements in Experiment 1 was reflective of the processing of such statements regardless of the viewing time for the picture, we expected to replicate the main effects of POLARITY and CONGRUENCE. If the restricted viewing time did have an impact, the predictions here are less clear.

4.3 Analyses and Exclusions

Similar to Experiment 1, all results we report reflect mixed effects model comparisons with maximal random effects structure. Out of 2480 observations collected for this experiment, 14 observations were excluded as missed responses (approximately 0.57% of the data). As with the previous experiment, results for RT measures are plotted and reported in log space, but also reported in milliseconds (ms) in the prose for ease of interpretation.

4.4 Results: RTs

We report the results of linear mixed effects regressions on the Polish and Math RT data. The results are presented graphically in Fig. 6.

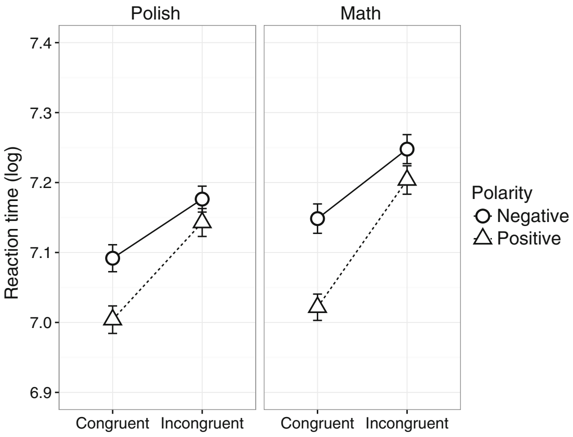


Fig. 6 Mean log RTs and SEs by POLARITY and CONGRUENCE for the linguistic (Polish) and non-linguistic (Math) sub-experiments of Experiment 2

4.4.1 Linguistic Conditions (Polish)

Participants took longer to process Polish negative comparatives than positive comparatives. This was reflected in a main effect of POLARITY (means: negative 7.14, positive 7.09, $\beta = -0.06$, $SE = 0.02$, $\chi^2 = 7.48$, $p < 0.01$) in the predicted direction: the negative conditions took longer to evaluate overall (means, in ms: negative 1384.83 ms, positive 1340.55 ms).

Participants also took longer to reject false statements than to accept true statements. This was reflected in a robust main effect of CONGRUENCE (means: congruent 7.06, incongruent 7.17, $\beta = -0.11$, $SE = 0.02$, $\chi^2 = 18.17$, $p < 0.01$), in accord with our predictions: a statement's truth or falsity with respect to its accompanying picture made a substantial difference to verification response latency (means, in ms: congruent 1279.27 ms, incongruent 1446.11 ms).

These two effects were only additive in our data. There was no interaction of POLARITY and CONGRUENCE ($\beta = -0.08$, $SE = 0.09$, $\chi^2 = 0.77$, $p > 0.1$); RTs in the positive congruent condition were somewhat faster than in the negative congruent condition (means: negative 7.10, positive 7.01; means, in ms: negative 1317.54 ms, positive 1241.36 ms), while a smaller difference in the same direction held in the incongruent conditions (means: negative 7.18, positive 7.16; means, in ms: negative 1451.46 ms, positive 1440.71 ms).

4.4.2 Non-linguistic Stimuli (Math)

Participants took longer to process statements with $<$ than with $>$. This was reflected in a main effect of POLARITY (means: negative 7.22, positive 7.12, $\beta = -0.09$, $SE = 0.02$, $\chi^2 = 23.23$, $p < 0.01$): the negative conditions took longer to process than the positive conditions (means, in ms: negative 1524.02 ms, positive 1377.45 ms).

This result is similar to the results for Math in our Experiment 1, and again stand in contrast to the results reported in Deschamps et al. (2015).

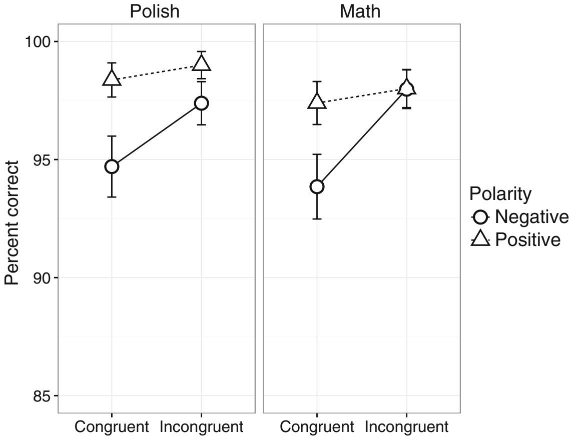
Again, participants took longer to reject false statements than to accept true statements. This was reflected in a main effect of CONGRUENCE (means: congruent 7.09, incongruent 7.25, $\beta = -0.16$, $SE = 0.02$, $\chi^2 = 71.40$, $p < 0.01$): the incongruent conditions had longer associated RTs than the congruent conditions (means, in ms: congruent 1328.17 ms, incongruent 1573.53 ms). This result replicates Experiment 1, and was predicted if there is a general cost for judging statements to be false.

As in the Experiment 1 Math sub-experiment, these effects were only additive. That is, we found no interaction between these two factors ($\beta = -0.05$, $SE = 0.09$, $\chi^2 = 0.32$, $p > 0.1$). RTs in the positive congruent condition were marginally faster than in the negative congruent condition (means: negative 7.15, positive 7.03; means, in ms: negative 1405.18 ms, positive 1251.17 ms); a similar pattern was observed in the incongruent conditions (means: negative 7.29, positive 7.22; means, in ms: negative 1645.22 ms, positive 1501.38 ms).

4.5 Results: Accuracy

Similar to Experiment 1, we assessed response accuracy via mixed effects logistic regression model comparisons. The results for mean response accuracy summarized by participant by condition for both linguistic and non-linguistic statements are presented in Fig. 7.

Fig. 7 Mean subject accuracy and SE by POLARITY and CONGRUENCE for the linguistic (Polish) and non-linguistic (Math) sub-experiments of Experiment 2



4.5.1 Linguistic Conditions (Polish)

Unlike in the Experiment 1 English sub-experiment, participants made more errors on the negative comparatives than the positive comparatives. This was reflected in a main effect of POLARITY (means: negative 96.1%, positive 98.7%, $\beta = 1.12$, $SE = 0.46$, $\chi^2 = 5.32$, $p = 0.02$), in which accuracy was lower in the negative conditions than in the positive conditions.

However, participants were equally accurate at rejecting false statements and accepting true statements. That is, we found no effect of CONGRUENCE here (means: congruent 96.6%, incongruent 98.2%, $\beta = -0.63$, $SE = 0.48$, $\chi^2 = 1.64$, $p > 0.1$): a statement's veracity with respect to its accompanying picture made little difference to response accuracy.

No interaction was found between POLARITY and CONGRUENCE ($\beta = 0.28$, $SE = 0.95$, $\chi^2 = 0.09$, $p > 0.1$); accuracy was lower in the negative congruent conditions than in the positive congruent conditions (means: negative 94.8%, positive 98.4%). The same was observed in the negative and positive incongruent conditions (means: negative 97.4%, positive 99.0%).

4.5.2 Non-linguistic Conditions (Math)

The apparent difference in POLARITY seen in Fig. 7 did not reach statistical significance here, unlike in the Polish sub-experiment. This was revealed in the lack of a main effect of POLARITY (means: negative 95.8%, positive 97.7%, $\beta = 0.56$, $SE = 0.38$, $\chi^2 = 2.25$, $p > 0.1$): there was little difference in accuracy between the negative and positive conditions.

Similarly, participants were no more or less accurate at rejecting false statements than at accepting false statements. That is, there was no main effect of CONGRUENCE on accuracy (means: congruent 95.6%, incongruent 97.9%, $\beta = -0.68$, $SE = 0.41$, $\chi^2 = 2.65$, $p > 0.1$). Whether the statement matched the picture made little difference to average response accuracy.

Finally, these two factors did not interact. No interaction was found between POLARITY and CONGRUENCE ($\beta = 0.80$, $SE = 0.88$, $\chi^2 = 0.83$, $p > 0.1$); accuracy was lower in the negative congruent condition than in the positive congruent condition (means: negative 93.8%, positive 97.4%), which was also observed in the negative and positive incongruent conditions (means: negative 97.7%, positive 98.0%).

4.6 Discussion

In Experiment 2, we found that negative Polish statements took longer to process than their positive counterparts, and rejecting a false statement took longer than accepting a true statement, regardless of whether the statement was provided in Polish or in a quasi-algebraic inequality. These results replicate the major results of Experiment 1, and support the viability of a decompositional analysis of negative comparatives in languages like Polish.

Unlike in the previous natural language sub-experiment, we found no interaction between POLARITY and CONGRUENCE in the Polish data. The effects of these factors appeared to be independent and additive, as predicted by the Sentence-First model of Clark and Chase (1972). In Experiment 2, participants were able to view the picture for up to 4 s, unlike the 200 ms time window imposed in Experiment 1, and moreover participants were encouraged to focus on accuracy over speed. These design parameters were more consistent with the original testing conditions for the Sentence-First model, so this result is perhaps not surprising. Future research should investigate why imposing a shorter viewing window leads to a different pattern of behavior.

The results tended in the same direction in the Math sub-experiment of Experiment 2. We found main effects of both POLARITY and CONGRUENCE, with no interaction between these factors. These results provide further evidence against the simple hypothesis concerning how Math statements might be processed.

In terms of accuracy, we found a main effect of POLARITY, but no effect of CONGRUENCE, and no interaction between these factors—both for Polish and Math. With respect to negation, accuracy appeared to pattern inversely with response latency, in line with Clark and Chase (1972) and Trabasso et al. (1971). This finding differed from Experiment 1 for English, which did not show this pattern; again, this lack of difference appears to be due to the differing amounts of time participants had to view the picture.

The results of these two experiments are thus consistent with the model of the Sentence-to-Picture verification task of Clark and Chase (1972). Perhaps surprisingly, they are met both in a task using natural language (Polish) and putatively non-linguistic statements (Math). Unlike Deschamps et al's (2015) finding, it appears that putatively non-linguistic stimuli can be processed in a highly similar fashion to linguistic stimuli, perhaps suggesting some sort of translation at test.

5 General Discussion

We have considered the processing of positive and negative adjectival comparatives in English and Polish, in contrast to analogous quasi-mathematical statements. We drew an explicit link between the decompositional analysis proposed by Büring (2007) and the additive effects on response latencies in simple tasks presented primarily in Clark and Chase (1972). Our results are predicted by decompositional analyses of *shorter* versus *taller*, given the linking hypotheses we have assumed. If the posited decomposed forms are reflective of speakers' semantic representations of positive and negative comparatives, and if comparing those representations to visual inputs involves additional symbolic manipulation whenever the representations mismatch, the predictions (and our results) follow straightforwardly.

One of the specific costs of processing *shorter/niższe* comparatives over *taller/wyższe* comparatives can be seen to reflect the cost of computing (shown here for A, but equally well for B) LITTLE(TALL(A)) over TALL(A). However, this cost is reflected at the point of making the judgment, *after* participants view the scene. How can we conceptualize this pattern? If we assume that scenes are essentially represented in positive terms (contra Clark and Chase 1972), we can suppose that the cost reflects one of comparing *representations in canonical (positive) form.*

(15) a. Language: *A is shorter than B.* $\overset{lg}{\Rightarrow}$ ER(LITTLE(TALL(A)), LITTLE(TALL(B)))

b. Vision: (a line marked A is longer than a line marked B) $\overset{vis}{\Rightarrow}$ ER(TALL(A), TALL(B))

c. Language representation in canonical form? No.

ER(LITTLE(TALL(A)), LITTLE(TALL(B))) $\overset{lg}{\Rightarrow}$ ER(TALL(B), TALL(A)) '*Below* time'

d. Representations match? No. '*Falsification* time'

Such a picture crucially involves the assumption that perceptual events are interpreted into a kind of representation that is 'written' in the same format as the semantic representations of sentences (Clark et al. 1973; Carpenter 1974). If semantic representations have a propositional format, then it must be possible to interpret the things we see into a similar format to feed later comparison operations. It has been suggested that the manipulation of such symbolic representations is reflected in the time it takes to initiate a response given visual input (Just and Carpenter 1971), as well as by the pattern and duration of eye fixations during tasks involving visual input (Just 1974). This paper has supported this pattern for temporal duration; the question of eye movements will be of particular interest for future research.

The results of Experiment 1 for English raised some questions about this model. There, we saw that the pattern of response latencies for English were different from those for Math, and different again from the predictions of the "Sentence-First" model. In Sect. 3.6, we speculated that these results reflected an additional cost imposed under the time pressure, specifically involving checking whether the entities in the two positions of the linguistically- and visually-derived ER relations were the same. Incorporating such a cost into the processing model helped to explain the pattern we observed for English in that experiment, but it does not explain why it was not observed for Math under the same conditions. We leave the development of these ideas for future research.

We also observed that the predictions of the 'simple' hypothesis for how math statements would be processed was not borne out. In two experiments with different task demands, and very different populations of speakers, we consistently observed patterns of response latency and accuracy that matched the predictions of the Sentence-First model for matching natural language and pictures. A possible hypothesis about why this pattern was observed, briefly entertained in Sect. 3.6, is that participants may have been 'translating' the math statements into their natural language during the initial processing of the statement. If so, we might expect that participants would spend more time on the statement screen for math statements than for sentences. To explore this idea, we conducted a post-hoc independent samples t-test on the length of time participants spent reading the statement before pressing spacebar to advance to the picture in Experiment 1.[11] This comparison was not significant ($t(1199) = 0.075$, $p = 0.9$). Thus, if math statements required an up-front additional cost for translation, it was not reflected in reading times.

This study leaves open the question of whether our data *could not* have been accounted for by positing a non-decompositional analysis in the first place, for instance that posited by Kennedy (2001). Given the other assumptions we made, such a view would hearken back to the early cognitive psychology literature, in which what was responsible for the additional processing cost of negation was some sort of linguistic 'negative feature'—in this case a negative lexical meaning. While such an approach could be made compatible with our findings, it would do so at the cost of transparency at the language-cognition interface. On the decompositional approach, the mapping from syntax to conceptualization is uniform, and the same, for English, Polish, and Hixkaryana (see Sect. 2): explicit constituents of the syntactic representation are related to explicit operations in processing. On the alternative view, this mapping is transparent in Hixkaryana, but requires a detour through the lexicon in English and Polish.

It could thus be particularly fertile to investigate links between processing and linguistic typology. It is well-known that negation is 'special' in language, and one vexing question that has yet to be resolved is *why* it is so special. We have begun to suggest a view on which negative forms are 'non-canonical', while those of other cognitive systems may be 'canonical'—at least in the mental language in which these representations are compared with those derived from other information sources. Comparing or interfacing representations across domains may require transformations of non-canonical representations into canonical ones, which is costly. If linguistic forms are furthermore required to be 'transparent' to non-linguistic cognition, then it seems that negative elements will be very costly indeed.

Two areas stand out as ripe for further investigation along these lines.

For one, Horn (1972) discusses the fact that of the four corners of the Aristotelian Square of Opposition, only three are found as distinct lexical items across languages: an existential (*some*), a universal (*all*), and a negative existential (*none*) (cf. Roelandt 2016). The fourth member of the set—a quantificational determiner equivalent in meaning to the phrasal form *not all* in English–never appears as a lexical item. Beyond the quantificational determiners, Bobaljik (2012) notes that no language, in the over 300 languages that he surveyed, features a synthetic comparative of inferiority. This gap is exemplified in the following examples, with the unattested meanings incorporating elements of the Büring and Heim semantics for *less*.

(16) a. Mary is *smart-er* than Bill. ⇒ $\text{ER}(\text{SMART}(M), \text{SMART}(B))$ ATTESTED

b. * Bill is *smart-le* than Mary. ⇒ $\text{ER}(\text{LITTLE}(\text{SMART}(B)), \text{LITTLE}(\text{SMART}(M)))$ UNATTESTED

Why should these typological gaps exist? Perhaps they reflect constraints on 'how much meaning' can be bundled into a single morpheme (Dunbar and Wellwood 2016), or on 'how much non-transparency' is permitted at the language-cognition

[11] These data were not collected for Experiment 2 due to a programming error.

interface. More broadly, the requirement for a transparent interface might require that the smallest meaningful pieces are alignable in a regular way with representations and processes in non-linguistic cognition. Thus, evidence from processing could provide new insight into what those pieces are, by looking at the kinds, and amount, of information recruited during linguistic understanding.

Acknowledgements We are grateful to Rebecca Way for creating the diagrams illustrating the Kennedy semantics from Sect. 2, and assisting with programming Experiment 1. We also extend our gratitude to Joanna Blaszczak, Andreas Brocher, Johannes Gerwien, Naomi Kamoen, and Maria Mos for their involvement with coding Experiment 2. Finally, we extend a special acknowledgment to Yaman Özakın for his work on creating the picture stimuli used in both experiments. The work on Experiment 2 was supported by the Polish National Science Center (NCN) grant OPUS 5 HS2 (DEC-2013/09/B/HS2/02763).

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7.

Beck, S. (2013). Lucinda driving too fast again-the scalar properties of ambiguous *than*-clauses. *Journal of Semantics*, *30*(1), 1–63.

Bobaljik, J. D. (2012). *Universals in comparative morphology: Suppletion, superlatives, and the structure of words*. Boston, MA: MIT Press.

Bresnan, J. (1973). Syntax of the comparative clause construction in English. *Linguistic Inquiry*, *4*(3), 275–343.

Büring, D. (2007). Cross-polar nomalies. In T. Friedman & M. Gibson (Eds.), *Proceedings of Semantics and Linguistic Theory* (Vol. 17, pp. 37–52).

Carpenter, P. A. (1974). On the comprehension, storage and retrieval of comparative sentences. *Journal of Verbal Learning and Verbal Behavior*, *13*(4), 401–411.

Carpenter, R. H. S. (1977). *Movements of the eyes*. London, UK: Pion Ltd.

Clark, H. H. (1969a). The influence of language in solving three term series problems. *Journal of Experimental Psychology*, *82*(2), 205–215.

Clark, H. H. (1969b). Linguistic processes in deductive reasoning. *Psychological Review*, *76*(4), 387–404.

Clark, H. H. (1970). *How we understand negation*. Paper presented at COBRE workshop on cognitive organization and psychological processes, Huntington Beach, CA.

Clark, H. H., Carpenter, P. A., & Just, M. A. (1973). On the meeting of semantics and perception. In W. Chase (Ed.), *Visual information processing* (pp. 311–381). New York, NY: Academic Press.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*(3), 472–517.

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

Deschamps, I., Agmon, G., Lewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition*, *143*, 115–128.

Dunbar, E., & Wellwood, A. (2016). Addressing the 'two interface' problem: The case of comparatives and superlatives. *Glossa: A Journal of General Linguistics*, 1(1), 5.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Hackl, M. (2009). On the grammar and processing of proportional quantifers: Most versus more than half. *Natural Language Semantics*, *17*(1), 63–98.

Heim, I. (1985). *Notes on comparatives and related matters*. Unpublished manuscript, University of Texas, Austin.

Heim, I. (2001). Degree operators and scope. In C. Fery & W. Sternefeld (Eds.), *Audiatur Vox Sapientiae. A Festschrift for Arnim von Stechow* (pp. 214–239). Berlin: Akademie Verlag.

Heim, I. (2006). Little. In M. Gibson & J. Howell (Eds.), *Proceedings of Semantics and Linguistic Theory* (Vol. 16, pp. 35–58), Ithaca, NY: Cornell University.

Heim, I. (2008). Decomposing antonyms? In A. Grønn (Ed.), *Proceedings of Sinn und Bedeutung* (Vol. 12, pp. 212–225). Oslo: ILOS.

Heim, I., & Kratzer, A. (1998). *Semantics in Generative Grammar*. Malden, MA: Blackwell.

Horn, L. (1972). *On the semantic properties of the logical operators in English*. Bloomington, IN: Indiana University Linguistics Club.

Huttenlocher, J. (1969). *Imaginal processes in reasoning*. Paper presented at the XIX International Congress of Psychology, London, UK.

Just, M. A. (1974). Comprehending quantified sentences: The relation between sentence-picture and semantic memory verification. *Cognitive Psychology*, *6*(2), 216–236.

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, *10*(3), 244–253.

Kennedy, C. (1999). *Projecting the adjective: The syntax and semantics of gradability and comparison*. New York: Garland.

Kennedy, C. (2001). Polar opposition and the ontology of 'degrees'. *Linguistics and Philosophy*, *24*(1), 33–70.

Klima, E. S. (1964). Negation in English. In J. A. Fodor & J. J. Katz (Eds.), *The structure of language: Readings in the philosophy of language* (pp. 246–323). Prentice Hall.

Lakoff, G. (1970). Linguistics and natural logic. *Synthese*, *22*(1–2), 151–271.

de Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

Lidz, J., Halberda, J., Pietroski, P., & Hunter, T. (2011). Interface transparency and the psychosemantics of *most*. *Natural Language Semantics*, *6*(3), 227–256.

May, R. (1977). *The grammar of quantification*. Dissertation, Massachusetts Institute of Technology.

Pancheva, R. (2006). Phrasal and clausal comparatives in Slavic. In *Formal approaches to Slavic linguistics* (Vol. 14, pp. 236–257).

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of *most*: Semantics, numerosity, and psychology. *Mind & Language*, *24*(5), 554–585.

Roelandt, K. (2016). *Most or the art of compositionality: Dutch 'de/het meeste' at the syntax-semantics interface*. Dissertation, KU Leuven.

Rullmann, H. (1995). *Maximality in the semantics of wh-constructions*. Dissertation, University of Massachusetts, Amherst, MA.

Seuren, P. A. M. (1973). The comparative. In F. Kiefer & N. Ruwet (Eds.), *Generative Grammar in Europe* (pp. 528–564). Dordrecht: D. Reidel Publishing Company.

Solt, S. (2015). Q-adjectives and the semantics of quantity. *Journal of Semantics*, *32*(2), 221–273.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276–315.

Szabolcsi, A. (2012). Compositionality without word boundaries: *(the) more* and *(the) most*. In *Proceedings of Semantics and Linguistic Theory* (Vol. 22, pp. 1–25). Ithaca, NY: CLC Publications.

Trabasso, T., Rollins, H., & Shaughnessy, E. (1971). Storage and verification stages in processing concepts. *Cognitive Psychology*, *2*(3), 239–289.

Wellwood, A. (2012). Back to basics: *More* is always *much-er*. In E. Chemla, V. Homer, & G. Winterstein (Eds.), *Proceedings of Sinn und Bedeutung* (Vol. 17). Paris: ENS.

Wellwood, A. (2015). On the semantics of comparison across categories. *Linguistics and Philosophy*, *38*(1), 67–101.

Cumulative Comparison: Experimental Evidence for Degree Cumulation

Rick Nouwen and Jakub Dotlačil

Abstract In this paper we address the question whether it makes sense to assume that the domain of degrees, as used in degree semantics, consists not just of atoms, but also of *degree pluralities*. A number of recent works have adopted that assumption, most explicitly (Fitzgibbons et al. Plural Superlatives and Distributivity, Proceedings of Semantics and Linguistic Theory, Vol. 18, 2008; Beck, The Art and Craft of Semantics: A Festschrift for Irene Heim, MITWPL, Vol. 70: 91–115, 2014; Dotlačil and Nouwen, Natural Language Semantics, 1–34, 2016). In this paper, we provide experimental evidence for degree pluralities by showing that comparatives may express cumulative relations between degrees.

Keywords Comparatives · Plurality · Plural degrees · Degree semantics
Cumulativity · Quantification · Scope

1 Adjectives and Plurality

Adjectives can be distributive or collective. For instance, "tall" is related to the *height* of an individual and is, as such, intrinsically concerned with atoms only: there is no such thing as the height of John and Mary as a group. Other adjectives are different. For instance, the adjective "compatible" is inapplicable to atoms, since degrees of compatibility can only be assigned to groups of entities. This situation is familiar from the semantics of non-adjectival plurality. Predicates like "be a team" are collective in the same sense as "compatible" is, while predicates like "being wounded" are distributive in the same sense as "tall".

R. Nouwen (✉)
Talen, Literatuur & Communicatie, Utrecht University, Utrecht, The Netherlands
e-mail: R.W.F.Nouwen@uu.nl

J. Dotlačil
Institute for Logic, Language & Computation, University of Amsterdam,
Amsterdam, The Netherlands
e-mail: j.dotlacil@gmail.com

E. Castroviejo et al. (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure*,
Language, Cognition, and Mind 4, https://doi.org/10.1007/978-3-319-77791-7_11

A dominant way of thinking about this distinction is that collective predicates have no atomic individuals in their extension, while distributive predicate have nothing but atoms in their extension. One clear advantage of this is that it enables us to account for why collective predicates are never compatible with singular arguments, while distributive ones *are* compatible with plural arguments, as illustrated in (1) and (2).

(1) *John is a good team.

(2) John and Mary are wounded.

To account for (2), all we need to assume is that, at least for plural cases like this, the extension of the predicate is closed under so-called *sum formation*: the operation that forms pluralities $a \sqcup b$ out of atoms a and b. The operator that enforces this kind of closure is often notated as * (after, Link 1983). What accounts for the contrast between (1) and (2) is the fact that * can create pluralities out of atoms, but it cannot create atoms out of pluralities. In other words, for a predicate with a non-empty extension, the extension of $^*P$ will always contain non-atomic entities, while it is not guaranteed to contain atomic ones.

Things are no different for adjectives. It would be natural to assume that "tall" expresses a relation between atomic entities and degrees and that "compatible" expresses a relation between non-atomic entities and degrees. Once the degree slot has been saturated, we will have a distributive predicate for the case of "tall" (as in "being tall" or "being two meters tall") and a collective one in the case of "compatible".[1]

In other words, considerations of plurality appear to have to do with how predicates map to the domain of entities and in particular to the complexity of the members of their extension. It appears then that plurality does not play a role on the side of degrees. Comparison, for instance, is a purely atomic relation, even if the adjective involved is collective. Take (3) as a case in point.

(3) John and Mary are more compatible than Peter and Sue.

What is at stake in (3) is whether the degree of compatibility of the sum of John and Mary exceeds the degree of compatibility of the sum of Peter and Sue. Clearly, even though the components of the comparative are plural, the degrees are not. The same point can be made for (4).

[1] Whereas "tall" lacks collective readings and "compatible" lacks distributive readings, so-called additive adjectives allow both. For instance, (i) either means that the totality of boxes is heavy or that each of them individually exceeds some standard of weight.

(i) The boxes are heavy.

This observation can be straightforwardly taken into account by assuming that the underlying measure function (weight) makes sense for both atoms and plurals in the sense that $\mu(\alpha \sqcup \beta) = \mu(\alpha) + \mu(\beta)$. A sentence like (i) is now ambiguous between a distributive one in which the weight of each atomic box is at stake or a collective one in which the added up weight of all the boxes is taken into account.

(4) John and Mary are taller than Peter.

Here, the predicate "to be taller than Peter" is a distributive predicate, which can be closed under sum formation using the * operator. In this way, it can apply to plural arguments even though both the relation expressed by "tall" and the comparison relation expressed by the comparative only have atoms in their extension.

So far, then, we have seen that where adjectives interact with pluralities, there is no evidence for the need for plural degrees. What is plural in all the above cases is an *e*-type argument of an adjectival or a comparative relation. Recently, however, there have been a number of proposals that assume the existence of non-atomic degrees. Fitzgibbons et al. (2008), for instance, analyse sentences where a superlative predicate is combined with a plural subject (*John and Paul are the tallest students*) as involving a fully pluralised adjective, i.e. a relation between potentially plural entities and potentially plural degrees. Very recently, two other accounts of plural degrees were developed, both aiming to improve on existing analyses of comparatives: Beck (2014) and Dotlačil and Nouwen (2016). In the remainder of this paper, we will present experimental evidence for such proposals. Focusing on our own account, we will start by summarising the plural degree approach to comparatives.

2 Plural Degrees

At the heart of the plural degree approaches to comparatives lies a classical puzzle of the semantics of comparatives, namely how to account for the interpretation of comparatives that have quantified *than* clauses (von Stechow 1984; Larson 1988; Heim 2006). The best paraphrase for a sentence like (5) is one in which the quantifier takes scope outside of the *than* clause, as in (6).[2]

(5) John is taller than every girl is.

(6) Every girl x is such that John is taller than x.

The same intuition holds for differentials: (8) is a very good analysis of (7), since it correctly predicts that (7) entails that every girl has the same height.

(7) John is exactly 2 inches taller than every girl.

(8) Every girl x is such that John is exactly 2 inches taller than x.

The issue is that *than* clauses are islands, cf. the ungrammatical status of the following example from Larson (1988):

[2] Clausal comparatives like (5) are perhaps not entirely natural to all native speakers, presumably given the (simpler) phrasal alternative *John is taller than every girl.* However, our theory is a theory of *clausal* comparatives and so it is important that we only consider that class of comparatives. Note that in our experiment, below, we escape the potential unnaturalness of sentences like (5) by turning to subcomparatives (*the table is longer than the door is wide*). There are of course no phrasal paraphrases of subcomparatives and so these sentences are entirely natural.

(9) *I wonder which door the table is longer than ___ is wide.

This fact makes (6) and (8) useless as blue-prints for the semantic structure of comparatives and differentials, for they would require an island violation. Since Schwarzschild and Wilkinson (2002), semanticists have standardly observed this restriction and developed several accounts that derive the correct interpretation without the island violation.

In line with this tradition, in Dotlačil and Nouwen (2016) we also claim that the wide scope of the quantifier is an illusion. On our account, what rather happens is that the *than* clause denotes a plural degree, namely the sum of degrees containing (nothing but) the heights of every girl. The sentence is then interpreted distributively in the sense that for each atom in that sum, John's height has to exceed that atom.

Before we say a little bit more about the framework that facilitates such an analysis, we zoom out a bit. If works like this or Beck (2014) or Fitzgibbons et al. (2008) are on the right track, then it suggests that the domain of degrees is no different from the domain of entities: both contain plural individuals and the relations we build on top of them are interpreted with respect to the same mechanisms, in particular, as we will see below, distributivity and cumulativity. In this paper, we follow this intuition. If degree plurality is like entity plurality, then we expect to see effects of plurality beyond the phenomena for which we designed the plural framework. That is, we should find evidence for plural interpretation beyond the simple comparatives in (5) and (7).

2.1 A Framework for Plural Degree Semantics

We take degrees to be discrete, atomic entities that are ordered by some ordering $>$. The plural degree semantics we developed in Dotlačil and Nouwen (2016) subscribes to the assumption that atomic degrees may combine to form sums. So, on top of the set of atomic degrees, there is also a set of non-atomic degrees, built from these atoms. If d and d' correspond to two different heights, then $d \sqcup d'$ is the collection that contains nothing but these degrees. Since we take degrees to be discrete, $d \sqcup d'$ equals d only if $d = d'$.

Above, we discussed how distributive $\langle e, t\rangle$-type predicates (i.e. predicates which only have atomic entities in their extension) can take plural arguments by closing the extension under sum formation using the * operator. The interpretation of * for a predicate P is as follows:

(10) a. $P \subseteq {}^* P$
b. If $\alpha \in {}^* P$ and $\beta \in {}^* P$, then also $\alpha \sqcup \beta \in {}^* P$.
c. Nothing else is in $^* P$.

To get a feel of what this definition does, let us briefly explain how, for instance, $^* \{a, b\}$ equals $\{a, b, a \sqcup b\}$. (10-a) states that $\{a, b\} \subseteq {}^* \{a, b\}$. The next condition states that $a \sqcup b \in {}^* \{a, b\}$. (10-c) adds that no other element is in $^* \{a, b\}$. In sum,

$\{a, b\} \subseteq {}^{*}\{a, b\}$. For any atomic predicate P, the result is that $^{*}P$ is only true of a plurality if P is true of each of the atoms of that plurality.

In the literature, a parallel operator exists for $\langle e, \langle e, t\rangle\rangle$-type relations (Krifka 1989; Sternefeld 1998; Beck and Sauerland 2000), often written as **. This is a generalisation of the * operator for sets of pairs of entities instead of just for sets of entities. For R a set of pairs:

(11) $^{**}R$ is the smallest superset of R such that if $\langle \alpha, \beta\rangle \in {}^{**}R$ and $\langle \alpha', \beta'\rangle \in {}^{**}R$ then also $\langle \alpha \sqcup \alpha', \beta \sqcup \beta'\rangle \in {}^{**}R$.

For example, $^{**}\{\langle a_1, b_1\rangle, \langle a_2, b_2\rangle\}$ equals $\{\langle a_1, b_1\rangle, \langle a_2, b_2\rangle, \langle a_1 \sqcup a_2, b_1 \sqcup b_2\rangle\}$. The effect on an originally atomic relation R is that two pluralities A and B stand in the $^{**}R$ relation if for each atom x in A there is at least one atom y in B such xRy and for each atom y in B there is at least one atom x in A such that xRy. This means that interpretative effect of ** is a cumulative reading (Scha 1981). For instance, when (12) is interpreted as (13), it yields the truth-conditions in (14). This makes the sentence true in a situation in which one boy carried two of the boxes and the other one carried the remaining boxes. In such a situation, the distributive reading is false.

(12) The two boys carried the four boxes.

(13) [the two boys [[** carried] the four boxes]]

(14) Each of the two boys carried some of the four boxes and each of the four boxes was carried by (at least) one of the two boys.

The operations * and ** suffice to account for *distributive*, *collective* and *cumulative* readings of (12). Applying the * operator to the predicate [*carried* [*the four boxes*]] allows it to take a plural subject. The resulting reading is compatible with both a distributive and a collective understanding of the sentence (depending on whether or not the extension of the predicate already contained plurality—i.e. boys jointly carrying boxes—or not). On the collective reading, neither of the boys carried the four boxes by themselves, they only did so collectively. Note that this is different from the cumulative reading, which does not entail that any collective carrying took place.[3]

The resulting framework is a minimal plural semantics for (predicates over) the domain of entities. For the case of degrees, we now assume that: (i) the domain of degrees contains both atoms and sums, just like the domain of entities; (ii) predicates and relations that involve degrees can be interpreted using the * and ** operators; (iii) the degree comparison relation $>$ is a relation between atomic degrees.

[3] There may be reasons to think that the distributive and collective understanding are proper *readings*, in which case one needs to posit a distributivity operator that quantifies over atoms (see Lasersohn 1998 for discussion). To keep things simple, we will remain agnostic with respect to this issue, which is orthogonal to our focus below.

2.2 Quantified Than-Clauses as Degree Pluralities

This framework can now be used to solve the puzzle of quantified *than* clauses. The idea is that *than* clauses denote potentially plural degrees, using the interpretation scheme in (15). (See Dotlačil and Nouwen 2016 for details of an underlying compositional semantics, and Beck 2014 for an alternative.)

(15) [than Q/DP is tall] = the smallest degree plurality that contains the height of Q/DP

For a DP like *Mary*, this scheme is going to return the smallest plurality that contains the height of Mary, which is simply the atomic degree Mary's height. For a QP like *every girl*, this scheme is going to return the smallest plurality that contains the height of every girl: girl$_1$'s height$\sqcup \ldots \sqcup$girl$_n$'s height.[4]

If the *than* clause denotes a non-atomic degree, it is in principle incompatible with the comparative semantics, since, as we said above, only atomic degrees are ordered and, so, degree comparison is comparison of atoms only. This means that in order to interpret a comparative with a *than* clause containing a quantificational element, we need to pluralise the comparison relation. For the case of *John is taller than every girls is*, we get:

(16) John's height $^{**}>$ girl$_1$'s height$\sqcup \ldots \sqcup$girl$_n$'s height

The relation $^{**}>$ is true of pluralities A and B if and only if each atom in A exceeds some atom in B and each atom in B is exceeded by some atom in A. If A itself is atomic, this simply boils down to this atom exceeding each atom in B, and so, for (16) to be true, John's height has to exceed all the atoms in the plurality of girl heights, which entails him being taller than the tallest girl. In other words, what in (5)–(8) seemed like a distributive quantifier taking wide scope is really the distribution over atoms in a plurality stemming from the need to pluralise an atomic relation that got given a non-atomic argument.

2.3 A Predicted Effect: Cumulative Comparison

We are assuming that if degrees can be plural then all the interpretation mechanisms we observe for the domain of entities should in principle also be available for degrees. We see no reason to assume a watered-down version of plural

[4] In Beck (2014), *than* clauses with *every* denotes plural degrees by virtue of the fact that *every* DP can have group readings. This is problematic because *than* clauses with *each* DPs yield the same readings as those with *every* DPs, but it is well-known that *each* DPs do not have group readings (Beck 2014, pp. 101–102). While the paraphrase in (15) may suggest that our proposal suffers from the same problem, we should hasten to add that (15) is only a very rough paraphrase of our proposal. In Dotlačil and Nouwen (2016) we compositionally derive plural degree denoting *than* clauses with both *each* and *every* DPs, based on their *distributive* semantics.

semantics for degrees, for instance where degree pluralities exist but relation cumulativity over degree relations does not. The account we sketched for quantified *than* clauses already suggests that this assumption is on the right track. This kind of view, however, also predicts that we should be able to observe further effects of plurality. In particular, the availability of ** accounts for cumulative readings for sentences like (12) and, so, we would expect to see true cumulative readings for $^{**}>$. The interpretation (16) of *John is taller than every girl is* is not evidence for that, since that interpretation is equivalent to the distributive reading we get by pluralising a derived predicate $\lambda d.John'sheight > d$ and applying it to the plurality denoted by the *than* clause.

(17) $[^*(\lambda d.\text{John's height} > d)](\text{girl}_1\text{'s height}\sqcup \ldots \sqcup \text{girl}_n\text{'s})$

In order to find true cumulative readings, we need two plural arguments. The literature contains at least one influential example of where we might find such a reading.

(18) The frigates were faster than the carriers. (Scha and Stallard 1988)

One possible interpretation of (18) is one in which there were groups of ships and in each group the frigates in that group were faster than the carriers in that group. On that reading, the subject distributivity reading is false, since there may be carriers that were faster than one or more frigates, as long as they were not in the same group.

In order to account for this reading, it is natural to resort to **. But for cases like (18), one need not assume that such an operator functions in the domain of degrees. Indeed, Scha and Stallard (1988), Schwarzschild (1996) and Matushansky and Ruys (2006) all analyse (18) as a cumulative relation between *entities*. That is, since (18) is a phrasal comparative, we can analyse it as a relation between entities (here, the frigates and the carriers) and so we can cumulate that relation using **.

This means that examples like (18) are not evidence for a cumulative interpretation of the degree relation $>$, but one could think that its clausal counterpart (19) is.

(19) The frigates were faster than the carriers were.

Clearly, (19) shares with (18) the same cumulative-like reading. However, in order to analyse (19) as a relation between entities, we would need to move out the subject of the *than* clause.[5] This is because clausal comparatives cannot be understood as relations between entities, given that one of the ‘comparees’ is a clause. To turn it into a relation over entities, we would somehow need to abstract over the subject in that clause, something we assume not to be a viable option, given that it would constitute an island violation. This suggests, then, that perhaps (19) does not involve a cumulative relation between entities, but one between degrees. Still, as we explain in Dotlačil and Nouwen (2016) in more detail, (19) is still not definitive proof that cumulative comparison exists. This is because we could arrive at exactly the same truth-conditions using distributivity and dependency. As Winter (2000)

[5] In fact, that would not suffice to gain a relation. See Dotlačil and Nouwen (2016).

shows, cumulative readings are often indistinguishable from distributive readings. For (19), that reading would be along the lines of (20).

(20) The frigates EACH$_i$ were faster than [the carriers]$_i$ were.

The idea is that the definite *the carriers* is interpreted as being dependent on the frigates. All one needs to assume is that distributivity can *bind* definites, something we need anyway to account for examples like (21), which has one reading in which each boy thinks that *he* is the tallest, instead of attributing the contradictory thought to him that all the boys are the tallest.

(21) The boys each think they are the tallest.

This means that if we want to show that cumulative comparison exists we need to use examples with two features: (i) we have to avoid phrasal comparatives, like (18), and use clausal comparatives instead, since phrasal comparatives may be understood as cumulative relations between entities, not degrees; (ii) we need to exclude the option of cumulative-like truth-conditions arising through dependent interpretation. We can accomplish the latter by resorting to distributive quantifiers. Consider for instance a minimal variation on (21): (22).

(22) The boys each think each of them is the tallest.

Whereas (21) has a reading in which *they* depends on the distributive quantification over boys, (22) lacks such a reading. The reason is that if *them* in (22) is interpreted dependently, it will refer to single boys and this renders the distributive quantification by *each* inappropriate, cf:

(23) *Each of John is sick.

Using this for our quest to find cumulative comparison, we arrive at examples like (24): This is an example of a clausal comparative, where there is no option of the subject of the comparative clause to depend on distribution over the matrix subject.

(24) The frigates were faster than each of the carriers were.

Intuitions are admittedly murky, here, and there are several complications: not least of all the fact that the distributive reading tends to be more readily available than the cumulative one, even already for the much simpler (19). For this reason, we turn to an experimental setting, in which we probe the truth-conditions participants assign to sentences of the shape in (24).

3 The Experiment

We tested interpretations of comparatives in a simple verification task. The goal was to find to what extent cumulative readings of clausal comparatives are accepted and how the level of acceptance compares to other readings one might associate with clausal comparatives. The experiment was run in Dutch.

3.1 Experimental Setup

In the experiment, participants were first given a cover story which told of a fictional study that compared people's ability to write (by hand) and type (on a keyboard) in a wide array of different circumstances. For each trial, this study recorded the writing and typing speeds of the participants in the cover story. Each stimulus of our experiment consisted of a fictional graph from the fictional study, depicting the typing and writing speed of three participants for a single trial. Figure 1 shows an example of such a graph. (The original stimuli were in Dutch and contained colours instead of shading.) Here, the speeds of three (fictional) participants (p.1, p.2 and p.3) are displayed. Shaded bars indicated the speed of their handwriting, non-shaded bars the speed of their typing in the trial.

Graphs like these were displayed with sentences that were supposed to provide a true statement about the trial in question. Participants in our experiment had to decide whether the statement was indeed correct.

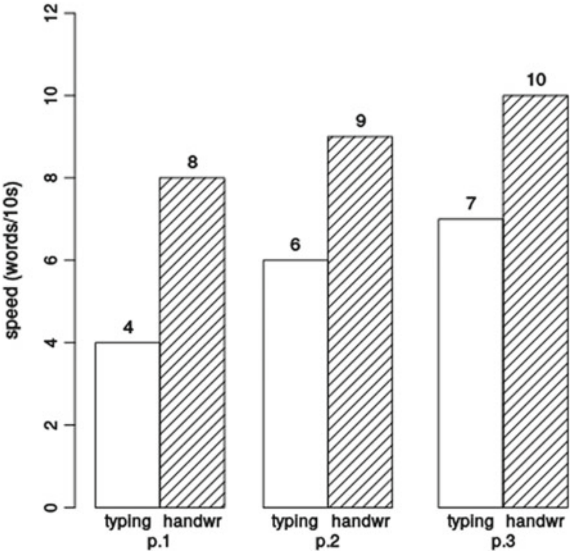


Fig. 1 An example plot used in the experimental stimuli

There were two types of test items appearing with graphs. In the test items, the *than* clause included a distributive universal quantifier (glossed as dist[6]), (25-a), or a plural definite anaphor, (25-b).

(25) a. De deelnemers typten sneller dan elk van hen schreef.
The participants typed faster than dist of them wrote
UNIVERSAL

b. De deelnemers typten sneller dan ze schreven.
The participants typed faster than they wrote
PLDEF

The test items had a verb in the *than* clause and consequently, they had to be treated as clausal comparatives.

Each barplot graphically summarized six data points representing the typing and writing speed of the three participants, as illustrated in Fig. 1. For ease of exposition, we will represent the graphs used in stimuli by enlisting the typing speed/writing speed pairs of the three participants. For instance, the shorthand for Fig. 1 is $\langle 4-8, 6-9, 7-10\rangle$.

It depended on the available readings whether a sentence was compatible with its accompanying barplot or not. We focused on two readings, the distributive and the cumulative reading. These are represented by the propositions in (26) and (27), respectively.

(26) Each of the three recorded typing speeds exceeds each of the recorded writing speeds. distributive reading

(27) For each of the three recorded typing speeds there exists a recording writing speed that is slower and for each of the three recorded writing speeds there exists a typing speed that is faster. cumulative reading

There were 5 tested scenarios for the experimental items. They are summarized in the following table:

Name	Example	Distr. reading	Cumul. reading
Dist	$\langle 8-5, 10-6, 12-3\rangle$	True	True
Cumul1	$\langle 6-5, 10-7, 12-3\rangle$	False on 1 account	True
Cumul2	$\langle 8-6, 6-5, 5-4\rangle$	False on 2 accounts	True
Noreading1	$\langle 4-8, 9-5, 7-6\rangle$	False	False on 1 account
Noreading2	$\langle 7-8, 9-5, 2-3\rangle$	False	False on 2 accounts

[6] We gloss it as such to avoid the issue of deciding whether *elk* in Dutch corresponds more closely to the distributive quantifier *every* or to the distributive quantifier *each*. Syntactically, it behaves like *each*: it appears in partitive constructions and can function as a floating quantifier. But semantically, it express distributivity but it does not seem to force the differentiation condition associated with *each* (Tunstall 1998; Brasoveanu and Dotlačil 2015).

To illustrate the idea behind this setup let us go through the examples. First of all, the example given for *dist* clearly verifies (26), since $8 > 5$, $8 > 6$, $8 > 3$, $10 > 5$, $10 > 6$, $10 > 3$, $12 > 5$, $12 > 6$ and $12 > 3$. Since in our setup, the distributive reading entails the cumulative one, (27) is true too. In the case of *cumul1*, the distributive reading is false. This is because participant 2 wrote faster than participant 1 typed: $7 > 6$. The cumulative reading is still true though, since $6 > 5$, $10 > 7$ and $12 > 3$. That is, the fact that for each participant it was the case that the typing speed exceeded the writing speed satisfies the requirements for the cumulative reading as stated in (27). This requirement also holds in the case of *cumul2*, but here the distributive reading is false on two accounts. Firstly, the typing speed of participant 2 does not exceed the writing speed of participant 1 and the typing speed of participant 3 does not exceed the writing speed of either participant 1 or 2.

In the cases of *noreading1* and *noreading2* both the cumulative and the distributive readings are false. We distinguish two cases here. In *noreading1*, two participants satisfied the cumulative relation imposed by the comparative, while one participant violated it. In the example in the table above, the problematic participant is participant 1 since he typed slower than he wrote (4 vs. 8). In *noreading2*, two participants violated the cumulative relation imposed by the comparative (in the example above, these are participant 1[7] and participant 3). For this reason, the cumulative reading is false in *noreading1* on one account (participant 1) and false in *noreading2* on two accounts (participants 1 and 3).

3.2 Predictions

The theory of Dotlačil and Nouwen (2016) predicts the following. For the test sentence without the distributive quantifier, i.e. the PLDEF item, both the distributive and the cumulative reading should be available. The former should be the default interpretation, derivable via pluralisation of the matrix predicate. The latter is available in two distinct ways: (i) via the cumulative operator, **; (ii) via the distributive operator in tandem with a dependent interpretation of the pronoun. For the test sentence with the distributive quantifier, the UNIVERSAL item, it should also be the case that both readings are available. However, now the option of arriving at the cumulative interpretation via a dependent analysis of the pronoun is excluded. If for some reason inserting * is preferred over inserting **, then we would furthermore predict higher rates of acceptance for *dist* than for *cumul1* and *cumul2*.

[7] Why participant 1? According to the definition in (27), participant 1 should not be problematic since he types faster than some of the other participants write, and he writes slower than one of the other participants type. However, following Schwarzschild (1996), among many others, we assume that cumulative relations are also sensitive to context, which determines which typing/writing speeds are compared. In the case at hand, the context requires that each participant's typing is compared to the writing of the same participant. Since participant 1 types slower than he writes, he presents a case violating the cumulative relation.

Table 1 Predictions in terms of proportion of responses in which participants respond that the sentence correctly describes the graph for the UNIVERSAL item

	dist	cumul1	cumul2	noreading1	noreading2
D&N16	top	top/high	top/high	bottom	bottom
No ** over degrees	top	bottom	bottom	bottom	bottom
Leniency	top	lower	lower still	even lower	bottom

Theories that do not have the option of interpreting the comparison relation cumulatively would potentially make the same prediction as Dotlačil and Nouwen (2016) for the PLDEF item, since in the absence of cumulativity, *cumul1* and *cumul2* are still compatible with a distributivity plus dependency reading. However, for UNIVERSAL items this reading is unavailable, so here one would predict low acceptability for all conditions, except for *dist*. The only way we could imagine higher scores is if participants somehow allow exceptions on the distributive quantification. In this case, you'd expect a slippery scale from universal acceptance for the case of *dist*, lower acceptance for *cumul1* and then continuously lower acceptance for *cumul2*, *noreading1* and *noreading2*.[8] The differing predictions are summarized in Table 1.

3.3 Methodology

3.3.1 Participants

44 native Dutch speakers participated in the experiment. 38 of them were students from the University of Groningen who either volunteered or received a course credit for their participation. 6 participants were volunteers from Utrecht University.

3.3.2 Materials and Procedure

The experiment consisted of graph-sentence pairs, as described in Sect. 3.1. Participants had to decide whether the sentence was true or false given the situation captured in the graph. Two sentence types were tested (PLDEF vs. UNIVERSAL) in five scenarios (*dist*, *cumul1*, *cumul2*, *noreading1* and *noreading2*). Two items per scenario were created (10 items in total). Two lists were created out of the items, so

[8] One could also imagine that readers compare aggregate values, most likely, average speeds of typing and writing. In that case, we would expect that all sentences would be accepted. This is because the experiment was set up in such a way so that the average speed corresponding to the verb in matrix clause always exceeded the average speed corresponding to the verb in the *than* clause. We will say more about this type of reading, which is arguably an instance of a collective interpretaton, in Sect. 3.5.

that in each list only one sentence type was present for each item. Every participant was assigned to one of the lists.

Apart from 10 experimental items, the experiment consisted of 2 practice items and 24 fillers. The fillers were unambiguously true/false (e.g., for Fig. 1 one filler might be the true sentence *Participant 3 typed slower than he wrote*). Fillers and experimental items were randomly ordered and each stimulus appeared on a separate screen (with no backtracking possible).

The whole experiment was run in Ibex and hosted on Ibex Farm (see http://spellout.net/ibexfarm/).

3.4 Results

Just one participant made more than 3 mistakes in the 24 fillers. Except for this one individual, we kept all the participants for the analysis.

Figure 2 shows the results. The percentages indicate proportionally how many participants responded that the sentence correctly describes the graph.

For the analysis, we focus on the UNIVERSAL factor since this is the part at which theories make different predictions. We consider logistic regression with one factor—Reading. We consider two different models. In the first one, Reading consists of two levels, *distributive reading* (consisting only of *dist*) and *no reading* (consisting of all the other cases, i.e., *cumul1*, *cumul2*, *noreading1*, *noreading2*). This is the model that is appropriate for theories that assume no **. In the second model, *cumul1* and *cumul2* are treated as a separate factor from *noreading1/2*, that is, Reading consists of three levels. This is the model appropriate for Dotlačil and Nouwen (2016). Somewhat unsurprisingly (given the graphical summary in Fig. 2), we see that using the three-level factor of Reading improves the model fit compared to the two-level factor ($\chi^2(1) = 96$, $p < 0.001$).

As we noted in Sect. 3.2, the theory lacking ** for comparatives could predict higher scores in *cumul1* than, say, *noreading1* if it somehow allowed exceptions on the distributive quantification. But in that case there should be a slippery scale from universal to *noreading2*. This is not the case when we look at Fig. 2. Here's one way to quantify this claim. If the acceptability decreased with the number of exceptions, we might expect that responses in readings *cumul1*, *cumul2*, *noreading1* and *noreading2* would form some form of linear function. Therefore, fitting it using logistic regression with one independent variable (the number of exceptions to make the dist reading true) would be appropriate. On the other hand, if Dotlačil and Nouwen (2016) are right, such a linear fit simplifies the picture. Instead, we can consider a general additive model, in which the model itself is left to find the best smooth function over the number of exceptions (using the mgcv package, see Wood 2006). This logistic model has one dependent variable, the number of exceptions to make the dist reading (with 5 knots), and RESPONSE is the dependent variable. It turns out, perhaps unsurprisingly, that the logistic general regression model fits our data significantly worse than the logistic general additive model ($\chi^2(2.8) = 38$, $p < 0.001$). One potential worry is

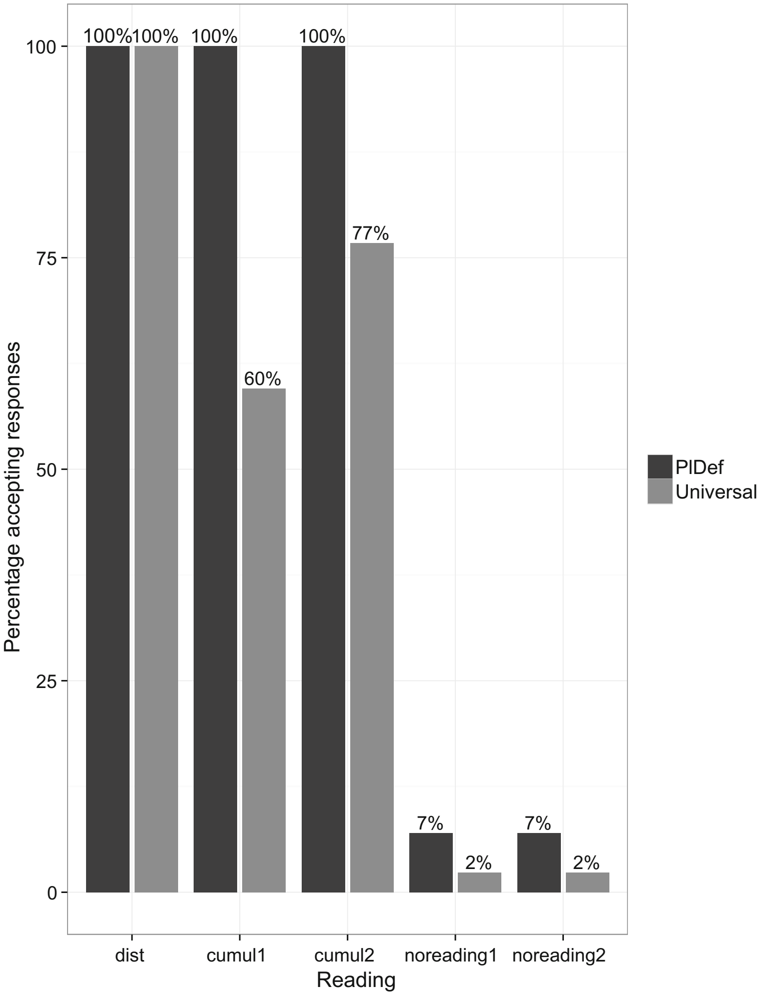


Fig. 2 Experimental results

that we might be overfitting the model in the former case. Importantly, though, the logistic general additive model is not significantly better than the simple model we considered above: logistic regression with one variable, Reading, which has three levels (*distributive reading*, *cumulative reading* and *no reading*) ($\chi^2(1.8) = 2.5$, $p > 0.1$). In conclusion we can say that in our search for the right model to fit the data with the distributive quantifier, the model that assumes that there is a distributive and cumulative reading (and nothing else) is the best. This supports Dotlačil and Nouwen (2016).

Finally, we note that it is clear from Fig. 2 that *cumul1/2* readings are more acceptable in PLDEF than in UNIVERSAL. This is compatible with our account under the assumption that * is preferred over ** and this preference is further corroborated by the higher acceptability of *dist* readings in UNIVERSAL.

3.5 Discussion: Collective Readings?

Readers familiar with Scontras et al. (2012) might recall other cases in which comparatives do not seem to be interpreted distributively. For instance, one may judge (28) to be true of a depiction of blue and red dots, even if there is one red dot that is smaller than every blue dot, as long as the average size of red dots exceeds that of blue dots.

(28) The red dots are bigger than the blue dots.

The experiments of Scontras et al. (2012) suggest that plural comparatives like (28) are indeed sometimes interpreted collectively. This means that subjects tend to interpret such sentences in terms of a comparison between an aggregate degree of size for the red dots and an aggregate degree of size for the blue ones.

It is not immediately clear whether the observations in Scontras et al. (2012) are relevant to our present study. First of all, the sentences in their experiments were always phrasal comparatives. Our theory in Dotlačil and Nouwen (2016) is a theory of clausal comparatives and so our current experiment deliberately only contains clausal comparatives as stimuli. Second, the sentences used by Scontras et al. have definite plurals in the *than* clause. It is not clear whether the collective reading is available once this definite is replaced by a distributive quantifier, as in the crucial stimuli in our experiment.

Ignoring these questions, could our results be understood as cases of collective comparison? We do not think so. First, while dot sizes might be easily imaginable as aggregate, supporting the collective interpretation for comparatives, our setup stressed individuals' writing/typing achievements. The focus on individual performances makes it unlikely that participants would consider collective comparisons in our experiment. Second, in our items, *all* conditions, including the *noreading* ones, were created in such a way that the collective reading would be true. For a sentence like *Participants typed faster than they wrote* the total typing speed was always faster than the total writing speed, in any of the test conditions. Consequently, the average typing speed exceeded the average writing speed for such a sentence.[9] Consequently, if the average-based collective reading is an option, we would expect that no condition would be rejected. But this is not the case: both *noreading* conditions were almost universally rejected.

[9] The items balanced the order of the comparison. Sometime typing was compared to writing, as in this example, but sometimes it was the other way around. In each case, however, the average speed corresponding to the verb in matrix clause exceeded the average speed corresponding to the verb in the *than* clause.

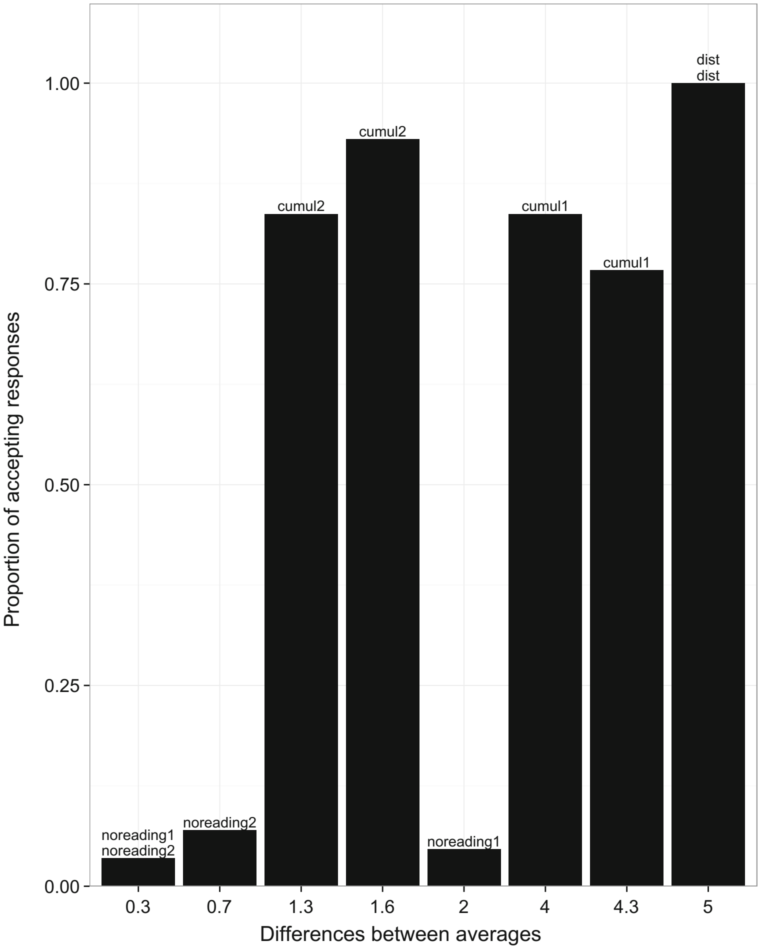


Fig. 3 Responses compared to differences between average speeds

In some cases, however, the difference between the average speeds is hard to gauge. It could be that the collective reading only results in "correct" responses when the difference between the averages is clear enough. That is, the likelihood of accepting the sentence increases when the difference between the average increases.

To test this, we looked at the average-differences in the test item and how these influenced responses, see Fig. 3. If the average difference played a role, we would expect that the proportion of accepting responses increases with the difference. This is clearly not the case. In particular, the difference of 2 is almost fully rejected even though lower differences between averages are almost fully accepted. To test this further, we considered a logistic regression model in which the difference in averages is a linear predictor. The model is significantly worse than the model we considered above as the one supporting Dotlačil and Nouwen (2016) (i.e., the one which has a factor with three levels, *distributive reading*, *cumulative reading*, *no reading*): $\chi^2(1) = 175$, $p < 0.001$. Thus, categorizing our data into three reading types clearly has much more predictive power than considering the difference in averages.

4 Conclusion

We discussed the semantics of comparatives and a new analysis that employs plural degrees (Fitzgibbons et al 2008; Beck 2014; Dotlačil and Nouwen 2016). Focusing on our own account, we argued that comparisons of plural degrees predicts a hitherto undiscussed reading, cumulative comparison. Controlling for several factors, we presented an experiment in which the relevant reading clearly surfaces. We take this as supporting evidence for the plural degree analysis of the comparative.

Acknowledgements Rick Nouwen gratefully acknowledges a grant from the European Research Council under the European Unions Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. 313502. Jakub Dotlačil gratefully acknowledges funding from the Netherlands Organisation for Scientific Research (NWO) VENI Grant 275.80.005. We also would like to thank one anonymous reviewer as well as Louise McNally for helpful comments at an earlier stage.

Appendix: Experimental Items

The following table contains all the experimental scenarios. As explained in Sect. 3.1, the scenarios are given by means of barplots. We represent these barplots here using the shorthand introduced before: each pair in the triple represents a participant, where the left number is their typing speed and the right number their handwriting speed. By orientation we mean whether the sentences that accompany the barplot compare typing to handwriting (left) or handwriting to typing (right), see below. The scenario types are those introduced in Sect. 3.1.

Scenario	Orientation	Scenario type
$\langle 6-5, 10-5, 12-3\rangle$	left	dist
$\langle 3-5, 4-10, 2-9\rangle$	right	dist
$\langle 6-5, 10-7, 12-3\rangle$	left	cumul1
$\langle 3-5, 7-10, 2-9\rangle$	right	cumul1
$\langle 8-6, 6-5, 5-4\rangle$	left	cumul2
$\langle 4-5, 6-7, 7-10\rangle$	right	cumul2
$\langle 4-8, 9-5, 7-6\rangle$	left	noreading1
$\langle 4-8, 6-9, 7-6\rangle$	right	noreading1
$\langle 7-8, 9-5, 2-3\rangle$	left	noreading2
$\langle 10-8, 4-5, 3-3\rangle$	right	noreading2

The stimuli themselves were as in (29). Here, (29-a) and (29-b) go with left-oriented bar plots and (29-c) and (29-d) go with right-oriented barplots. The stimuli in (29-a) and (29-c) are of the PluDef type and those in (29-b) and (29-d) are of the Universal type.

(29) a. De deelnemers typten sneller dan ze schreven.
The participants typed faster than they wrote.
b. De deelnemers typten sneller dan elk van hen schreef.
The participants typed faster than each of them wrote.
c. De deelnemers schreven sneller dan ze typten.
The participants wrote faster than they typed.
d. De deelnemers schreven sneller dan elk van hen typte.
The participants wrote faster than each of them wrote.

There were 36 fillers were, half of them presented in scenarios that made them false, half of them in scenarios that made them true true. The sentences that made up the fillers were of the following types:

(30) a. Alle deelnemers typten sneller dan ze schreven.
All participants typed faster than they wrote.
b. Precies 2 deelnemers typten sneller dan ze schreven.
Exactly 2 participants typed faster than they wrote.
c. Deelnemer 2 typte sneller dan ze schreef.
Participant 2 typed faster than she wrote.
d. Geen deelnemer typte sneller dan ze schreef.
No participant typed faster than she wrote.
e. Deelnemer 3 typte sneller dan deelnemer 1 schreef.
Participant 3 typed faster than participant 1 wrote.
f. Meer dan 1 deelnemer schreef sneller dan ze typte.
More than 1 participant wrote faster than she typed.

References

Beck, S. (2014). Plural predication and quantified than-clauses. In L. Crnič & U. Sauerland (Eds.), *The art and craft of semantics: A Festschrift for Irene Heim*, MIT Working Papers in Linguistics (Vol. 70, pp. 91–115).

Beck, S., & Sauerland, U. (2000). Cumulation is needed: A reply to Winter (2000). *Natural Language Semantics*, *8*(4), 349–371.

Brasoveanu, A., & Dotlačil, J. (2015). Strategies for scope taking. *Natural Language Semantics*, *23*(1), 1–19.

Dotlačil, J., & Nouwen, R. (2016). The comparative and degree pluralities. *Natural Language Semantics, 24*(1), 45–78.

Fitzgibbons, N., Sharvit, Y., & Gajewski, J. (2008). Plural superlatives and distributivity, In T. Friedman & I. Satoshi (Eds.), *Proceedings of Semantics and Linguistic Theory* (Vol. 18, pp. 302–318).

Heim, I. (2006). *Remarks on comparative clauses as generalized quantifiers*. Unpublished manuscript, MIT.

Krifka, M. (1989). Nominal reference, temporal constitution, and quantification in event semantics. In R. Bartsch, J. van Benthem, & P. van Emde Boas (Eds.), *Semantics and contextual expression* (pp. 75–115). Dordrecht: Foris.

Larson, R. (1988). Scope and comparison. *Linguistics and Philosophy*, *11*(1), 1–26.

Lasersohn, P. (1998). Generalized distributivity operators. *Linguistics and Philosophy*, *21*(1), 83–93.

Link, G. (1983). The logical analysis of plurals and mass terms: A lattice-theoretic approach. In R. Bäuerle, C. Schwarze, & A. von Stechow (Eds.), *Meaning use and interpretation of language* (pp. 302–323). Berlin: Walter de Gruyter.

Matushansky, O., & Ruys, E. G. (2006). Meilleurs voeux: quelques notes sur la comparaison plurielle. In: O. Bonami & P. Cabredo Hofherr (Eds.), *Empirical issues in syntax and semantics* (Vol. 6, pp. 309–330). http://www.cssp.cnrs.fr/eiss6/index_en.html.

Scha, R. (1981). Distributive, collective, and cumulative quantification. In J. Groenendijk, T. Janssen, & M. Stokhof (Eds.), *Formal methods in the study of language* (pp. 483–512). Amsterdam: Mathematical Centre.

Scha, R., & Stallard, D. (1988). Multi-level plurals and distributivity. In: *Proceedings of the 26th Annual Meeting of the ACL*. Buffalo, NY.

Schwarzschild, R. (1996). *Pluralities*. Dordrecht: Kluwer Academic Publishers.

Schwarzschild, R., & Wilkinson, K. (2002). Quantifiers in comparatives: A semantics of degree based on intervals. *Natural Language Semantics*, *10*(1), 1–41.

Scontras, G., Graff, P., & Goodman, N. D. (2012). Comparing pluralities. *Cognition*, *123*(1), 190–197.

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

Sternefeld, W. (1998). Reciprocity and cumulative predication. *Natural Language Semantics*, *6*(3), 303–337.

Tunstall, S. (1998). *The interpretation of quantifiers: Semantics and processing*. Dissertation, University of Massachusetts, Amherst.

Winter, Y. (2000). Distributivity and dependency. *Natural Language Semantics*, *8*(1), 27–69.

Wood, S. (2006). *Generalized additive models: An introduction with R*. Boca Raton, Florida: CRC Press.