XPrag-ADJ19: The semantics and pragmatics of gradable adjectives: Integrating perspectives from linguistic theory, psycholinguistics and modeling

Workshop organized as part of the SPP 1727 "XPrag.de: New Pragmatic Theories based on Experimental Evidence" in Cologne, Germany, May 23-24, 2019.

Organizers: Anton Benz (Leibniz-ZAS Berlin), Nicole Gotzner (Leibniz-ZAS Berlin), Petra Schumacher (Cologne) and Stephanie Solt (Leibniz-ZAS Berlin), Barbara Tomaszewicz-Özakın (Cologne)

PROGRAM and ABSTRACTS

see here for last minute changes

Thursday, May 23

09:30 Welcome (with coffee)
10:00 **Louise McNally**: Scalar alternatives and scalar inference involving adjectives
11:00 **Arnold Kochari, Ashley Lewis and Herbert Schriefers:** The time-course of adjective-noun composition in case of gradable and non-gradable adjectives
11:30 **Michael Henry Tessler, Noah Goodman and Roger Levy:** Comparison class inference for gradable adjectives
12:00 Lunch Break
13:30 **Poster Session**
15:00 Coffee break
15:30 **Jérémy Zehr and Nattanun Chanchaochai:** Ambidirectionality and Thai mid-scale terms: when 'warm' means less hot
16:00 **Steven Verheyen:** Vague gradable adjectives: Experimenting with probabilistic models
17:00 end / workshop dinner

Friday, May 24

09:30 **Kristen Syrett:** Setting the standard and making comparisons in language acquisition
10:30 **Catherine Davies, Jamie Lingwood and Sudha Arunachalam:** What kinds of adjectives do preschoolers encounter in the input, and how do they process what they hear?
11:00 Coffee Break
11:30 **Jérémy Zehr and Paul Egre:** Contradictory Descriptions with Absolute Adjectives
12:00 **Giorgos Spathas:** Proportional modification of gradable adjectives: the case of percentrages
12:30 Lunch Break
13:30 **Seungjin Hong, Jean-Pierre Koenig, Gail Mauner and Aron Marvel:** Computing Category membership for *tall*: An error minimization approach
14:00 **Michael Franke:** Rational Griceans are vague
15:00 Closing

Alternates:
Merle Weicker and Petra Schulz: Not all gradable adjectives are vague – Experimental evidence from children and adults
Carla Umbach and Umut Ozge: Scalar and non-scalar equatives in Turkish

Posters:
Alena Anishchanka and Steven Verheyen: Color term basicness in experimental and corpus-based research
Helena Aparicio, Roger Levy and Elizabeth Coppock: How to find *the rabbit in the big(ger) box*: Reasoning about contextual parameters for relative adjectives under embedding
Nicole Gotzner, Stephanie Solt and Anton Benz: Interplay of scalar and manner implicature
Myung Hye Yoo: Are all absolute predicates truly absolute?
Claudia Lehmann: Gradable adjectives in ironic constructions
Mora Maldonado, Alexander Martin and Jennifer Culbertson: An experimental approach to inferences to the standard in comparative constructions
Les Sikos, Noortje Venhuizen, Heiner Drenhaus and Matthew Crocker: Reevaluating pragmatic reasoning in web-based language games
Barbara Tomaszewicz-Özakin and Petra B. Schumacher: World knowledge and the absolute-relative distinction in adjectives
Carla Umbach and Umut Ozge: Scalar and non-scalar equatives in Turkish
Merle Weicker and Petra Schulz: Not all gradable adjectives are vague – Experimental evidence from children and adults

Color term basicness in experimental and corpus-based research

Alena Anishchanka (University of Antwerp, Belgium)
Steven Verheyen (KULeuven, Belgium)

The study explores the measurement of color term basicness – the cognitive construct which has been central to color categorization research across a number of disciplines. Introduced in ethnoscience and anthropology (Berlin and Kay 1999 [1969]), it was developed in cognitive psychology, anthropology and linguistics and has become the precursor of the prototype model of categorization. While the basic color term model provided a common reference point for cognitive theories of color categorization, the concept of basicness has been defined and operationalized by different research traditions relative to their own theoretical frameworks and methods. Thus, anthropological and psychological studies have predominantly relied on experimental evidence from color naming and elicitation tasks when measuring the basicness of color categories (e.g., Berlin and Kay 1999[1969]; Rosch Heider and Olivier 1972; Corbett and Davies 1995, 1997; Roberson et al. 2000; Lin et al. 2001); while linguistic studies tend to apply linguistic measurements of basicness derived from frequency counts in text collections (corpora) and dictionaries (e.g., Kerttula 2002, Steinvall 2002).

Current color categorization studies routinely use evidence from related disciplines to support their models, however the question of the compatibility of the different experimental and linguistic measurements of color basicness remains underexplored. One notable exception is the study by Corbett and Davies (1995, 1997), which suggested a possible distinction between measurements derived from texts (frequency of use and number of derived forms), color naming tasks (response time, frequency and consistency of naming) and elicitation (frequency and position in the lists).

Building on the study by Corbett and Davies (1995, 1997), this paper extends the analysis of basicness measurements for color terms in English in several ways. First, we include a more comprehensive set of basicness measures derived from experimental and corpus-based color studies that became available in the past 20 years. In addition to the data analyzed in Corbett and Davies (1995, 1997), we consider replications of the original experiments, and new types of experiments such as large-scale online color naming experiments (e.g., Mylonas et al. 2010) and free word association experiments (De Deyne and Storms 2008). Corpus-based data are obtained from much larger up-to-date corpora such as the British National Corpus (Kerttula 2002) and the Bank of English (Steinvall 2002) as well as specialized corpora including fiction (Moskovič 1977) and online retail and marketing (Anishchanka 2013). Second, the study includes new types of basicness measurement not studied previously. In particular, we analyze the geometrical structure of color categories in the three-dimensional color space (the size of the color category) and centrality characteristics of color words in word association networks (De Deyne and Storms 2008). Third, the more extensive data samples available from large-scale online experiments and corpora allow us to extend the number of color terms included in the analysis in comparison to Corbett and Davies (1995, 1997) who focused primarily on the 11 basic color terms (BCTs).

The correlations between the basicness measures are analyzed using the method of multidimensional scaling (MDS). The study is organized in three series of analyses, where standard and constrained MDS models are fitted to different subsets of the basicness measurements. Analysis 1 focuses on the frequency counts for the 11 BCTs in the different types of experimental studies (color naming, elicitation, free word association) and in different types of

corpora. Analysis 2 extends the list of basicness measurements to 58, including non-frequency-based measurements available for the 11 BCTs, such as response time and consistency in color naming tasks; the average position in the elicited lists; derivational productivity of the color word calculated from dictionary and corpus data; and color category size in two- and three-dimensional color spaces. Analysis 3 extends the list of color words to 47 terms, moving beyond the 11 BCTs.

The three series of analyses reveal a consistent correlational structure of the basicness measurements defined by the three main types of data sources: color naming tasks, elicitation tasks and corpus data. These three groups of measurements align with the pattern identified in Corbett and Davies (1995, 1997). The analyses show that basicness measurements derived from text corpora – even as different as Twitter and British poetry – are more highly correlated with each other than with the measurements obtained in color naming and elicitation tasks.

At the same time, the more representative datasets and the extended list of basicness measurements in this study provide further insights into the factors that might explain the differences between basicness measurements in the three major groups of the available data. We propose that basicness measurements are affected by the presence/absence of a color stimulus in the elicitation procedure and the contextualized vs decontextualized usage of color words in experimental and corpus-data.

The results of the analyses suggest that basicness is a multidimensional construct and show that there are systematic differences between basicness measurements applied in different research traditions in color categorization. It is expected that their meaningful interpretation will contribute to comprehensive modeling of color categories using evidence from experimental and corpus-based studies.

References

Anishchanka, Alena. 2013. "Seeing It in Color: A Usage-Based Perspective on Color Naming in Advertising." KU Leuven. PhD Dissertation.

Berlin, Brent, and Paul Kay. 1999 [1969]. *Basic Color Terms*. Stanford: CSLI.

Corbett, Greville, and Ian Davies. 1995. "Linguistic and Behavioural Measures for Ranking Basic Colour Terms." *Studies in Language* 19(2): 301-357.

Corbett, Greville, and Ian Davies. 1997. "Establishing Basic Color Terms: Measures and Techniques." In *Color Categories in Thought and Language*, edited by Clyde L. Hardin and Luisa Maffi, 197–223. Cambridge University Press.

De Deyne, Simon and Gert Storms. 2008. "Word Associations: Network and Semantic Properties." *Behavior research methods* 40(1):213–31.

Kerttula, Seija. 2002. *English Colour Terms: Etymology, Chronology, and Relative Basicness. Mémoires de La Société Néophilologique de Helsinki, Vol. LX*. Helsinki: Société Néophilologique. PhD Dissertation.

Lin, H., M. R. Luo, L. W. MacDonald, and A. W. S. Tarrant. 2001. "A Cross-Cultural Colour-Naming Study. Part I: Using an Unconstrained Method." *Color Research & Application* 26 (1): 40–60.

Moskovič, Vulf A. 1969. *Statistika i Semantika [Statistics and Semantics]*. Moscow: Nauka.

Mylonas, Dimitris, and Lindsay MacDonald. 2010. "Online Colour Naming Experiment Using Munsell Samples." In *Conference on Colour in Graphics, Imaging, and Vision*, 2010:27–32. Society for Imaging Science and Technology.

Roberson, Debi, Ian R.L. Davies, and Jules Davidoff. 2000. "Color Categories Are Not Universal: Replications and New Evidence from a Stone-Age Culture." *Journal of Experimental Psychology General* 129 (3): 369–98.

Rosch (Heider), Eleanor, and Donald C. Olivier. 1972. “The Structure of the Color Space in Naming and Memory for Two Languages.” *Cognitive Psychology* 3: 337–45.
Steinvall, Anders. 2002. *English Colour Terms in Context*. Umeå University: Skrifter från moderna språk 3. PhD Dissertation.

How to find *the rabbit in the big(ger) box*:
Reasoning about contextual parameters for relative adjectives under embedding

Haddock (1987) noticed that *the rabbit in the box* succeeds in referring even in the presence of multiple boxes, so long as only one contains a rabbit; uniqueness w.r.t. *box* is not required when *the box* is embedded in such a description. The present work investigates interpretive preferences for similarly embedded noun phrases containing a positive or comparative relative adjective (e.g., *the rabbit in the big/ger box*). We find that embedded positive adjectives exhibit a sensitivity to contextual manipulations that embedded comparatives lack, and we derive this sensitivity using a probabilistic model of the contextual parameters guiding the interpretation of the embedded NP.

Experiment. In our experiment ($N = 75$), participants heard definite descriptions while looking at visual contexts containing five pictures. The embedded noun was masked using static noise, so the instruction was always ambiguous between two potential referents (Target 1 and 2 in Figure 1). Participants clicked on the target they judged more likely. In each of the conditions in Figure 1, the display contains a pair of boxes and a pair of bags. In the +COMPETITOR conditions, there is a third bag, bigger than the other two. Although it does not contain a rabbit and cannot serve as a referent for the noun phrase as a whole, this competitor introduces uncertainty regarding the threshold for *big*.

In the SAME/DIFFERENT conditions, the two bags have the same animal in them (rabbits), and the boxes have different animals in them (a rabbit and a frog). In the SAME/SAME condition, the two pairs of boxes both have the same animal (rabbits). When the same animal is in both members of a pair, the descriptive content of the gradable adjective is informative in a noun phrase resolving to a member of that pair, identifying which to pick.

Results are presented in Figure 2. Unsurprisingly, participants exhibited a clear sensitivity to informativity, preferring resolutions on which the adjective helps to identify a referent. Furthermore, a significant COMPETITOR $\times$ ADJECTIVE interaction was found for SAME/DIFFERENT conditions such that the presence of a competitor object increased clicks to Target 1 for the positive form adjective but not for the comparative ($p < 0.05$). The same effect occurred in the +COMPETITOR SAME/SAME condition, compared to chance: presence of a competitor acted as a deterrent, with the positive form.

RSA model. We implement a Rational Speech Act Model (e.g. Frank & Goodman 2012) that derives the observed effects in human behavior as a result of uncertainty about contexts and threshold values for the embedded modified NP (e.g. *big box*). For any given set of five referents R contained in each of the three displays tested, a context C is defined as any element in $\mathcal{P}(R)$. We assume a flat prior over contexts. For a given description d of the form *the* N_1 *in the big* N_2, we assume that $[\![d]\!]^{C,\theta} = r$ iff (i) $r \in [\![N_1]\!]$; (ii) r is inside N_2; (iii) $[\![big]\!]^{C,\theta}([\![N_2]\!]) = 1$, where θ is the threshold value for the relative adjective; (iv) uniqueness holds. Following Bumford (2017), we assume that uniqueness of the embedded NP is checked w.r.t. e.g. rabbit-box pairs. Following Muhlstein et al. (2015), we put a uniform prior on contexts, and low prior probability on referential failure:

$$P(r) = \begin{cases} \epsilon & \text{r}= FAIL \\ \text{uniform} & \text{otherwise} \end{cases} \qquad P(r|C) = \begin{cases} \frac{P(r)}{\Sigma_{r' \in C} P(r')} & \text{r} \in C \\ 0 & \text{otherwise} \end{cases}$$

The truth-conditions for descriptions containing a comparative (i.e. *the* N_1 *in the bigger* N_2) differ from those assumed for the positive form in that the comparative is only defined

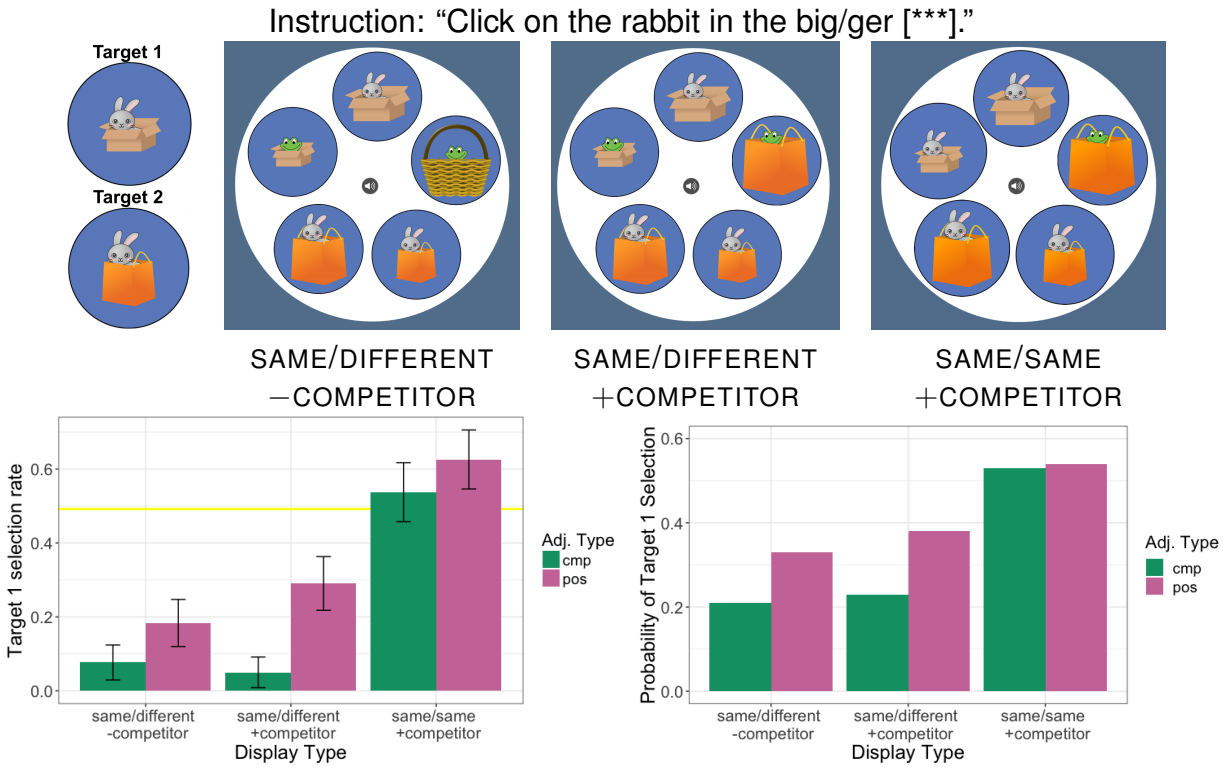


Figure 2: Left: Experimental results (yellow line indicates chance). Error bars show 95% CIs. Right: Simulation results.

in contexts that contain exactly two objects in the extension of N_2 (e.g. exactly two boxes) that differ in size. Given these assumptions, the model described in (1-3) ensures that given an ambiguous incomplete description of the form ‘*the rabbit in the big/ger* ***’, listeners reason about possible contextual partitions and the conditions that are more likely to make the message true, i.e. the probability of threshold values in a circumscribed context as well as how informative the description is in the context. For the positive form, contexts that allow for higher threshold variability (e.g., contexts with three bags, where the threshold can be set in two ways) assign a lower probability to the relevant referent compared to contexts where no such uncertainty exists (e.g. contexts that contain two bags, and for which there exists only one possible threshold resolution). In the case of the comparative, no such effect is predicted to arise, since the description is only defined in contexts that contain exactly two bags and for which there is no uncertainty about possible threshold values.

(1) $L_0(r \mid d, C, \theta) \propto [\![d]\!]^{C,\theta}(r, \theta, C)\, P(r)$ (Literal listener)

(2) $S_1(u|r, C, \theta) \propto L_0(r|d, C, \theta)$ (Speaker)

(3) $L_1(r, C, \theta|d = N_1\ in\ Adj\ **) \propto$
$\sum_{N_2} P(d = N_1\ in\ Adj\ N_2|r, C, \theta)P(r|C)P(\theta|C)P(C)$
$\sum_{C,\theta} L_1(r, C, \theta \mid d = N_1\ in\ Adj\ **)$ (Pragmatic listener)

Selected references: Bumford, D. 2017. Split-scope definites: Relative superlatives and Haddock descriptions. *Linguistics and Philosophy* 40(6). 549–593. ◇ Haddock, Nicholas J. 1987. Incremental interpretation and Combinatory Categorial Grammar. In *Proceedings of IJCAI*. ◇ Muhlstein, Larry, Christopher Potts, Michael C. Frank & Roger Levy. 2015. Pragmatic coordination on context via definite reference. Poster presented at XPRAG 2015.

What kinds of adjectives do preschoolers encounter in the input, and how do they process what they hear?

Catherine DAVIES[1], Jamie LINGWOOD[1], & Sudha ARUNACHALAM[2]
[1]University of Leeds, UK; [2]New York University, US

Adjectives are essential for describing and differentiating concepts. However, they have a protracted developmental course relative to other open word classes, which has been attributed to their semantic, syntactic, and pragmatic variability. Despite the importance and relatively late appearance of adjectives in children's repertoires, their acquisition has received little attention to date.

This study uses corpus and eye movement analyses to investigate how the formal properties of adjectives interact with children's language processing. Experiment 1 measured three- and four-year olds' naturalistic experience of adjectives in multiple child-directed speech contexts (allowing reflection on how features of the input might help / hinder adjective acquisition). Experiment 2 measured how 3-year-olds actually comprehend descriptive vs. contrastive adjectives online, presented pre- and post-nominally.

In light of experimental evidence suggesting that **predicative**, **postnominal** frames and **contrastive** functions scaffold adjective development (Ramscar et al., 2010; Ninio, 2004), experiment 1 used corpus analysis to survey the variability of adjectives across a range of interactive and socioeconomic contexts. We examined adjectives in three forms of child-directed speech (CDS): 1) the text from 16 books for three-year-olds; 2) CDS used by parents of three-year-olds during free play (n = 16); 3) CDS used by parents of four-year-olds across the socioeconomic spectrum during shared bookreading (n = 50). Syntactically, adjectives occurred more frequently in prenominal (M = 0.63, SD = 0.22) than in postnominal frames (M = 0.32, SD = 0.20, p <.001), though postnominal adjectives were more frequent for less familiar adjectives. Semantically, children heard more absolute (M = 0.45, SD = 0.26) and relative adjectives (M = 0.45, SD = 0.27) than non-gradable ones (M = 0.10, SD = 0.18; both ps < 0.001). Pragmatically, descriptive adjectives (M = 0.94, SD = 0.11) appeared much more frequently than contrastive ones (M = 0.06, SD = 0.11; p < .001). These patterns held across free play CDS, shared book reading CDS, and in children's book texts. Findings present a partial mismatch between the forms of adjectives found in real-world CDS and those forms that should be most developmentally useful.

To investigate this mismatch, experiment 2 used a 2 (frame: prenominal; postnominal) x 2 (pragmatic function: descriptive; contrastive) x 2 (age: 3;6-year-olds; adults) design to identify the online processes by which by children and adults integrate and interpret adjectives across syntactic and pragmatic contexts. Previous research has shown that preschoolers do not always integrate adjectives and nouns, and instead over-rely on noun information (Thorpe et al., 2006; Fernald et al., 2010). When arrays are complex they may not integrate until 5 or 6 years of age (Huang & Snedeker, 2013; Ninio, 2004; Sekerina & Trueswell, 2012), but what matures to enable this is unknown. So, as well as analysing

children's gaze data to track their online interpretation, we analyse i) the development of their semantic knowledge, and ii) processing speed as predictors of individuals' adjective-noun integration.

Unlike previous research that has focused on adjectives in prenominal position, we also measure children's processing of postnominal adjectives, in line with the frequency of this construction in English. Languages that frequently place adjectives before nouns present an extra challenge to adjective processing because the referent is unknown during the adjective region. For this reason, we hypothesise that modified noun phrases will be processed more quickly when adjectives appear postnominally than when they occur prenominally.

We analyse the effect of pragmatic function, where adjectives contrast an object with another of the same class (*the big cup* in the context of a smaller one), or describe an entity on its own merits (in the absence of a competitor cup). We hypothesise that adults will show earlier integration in the contrastive than the descriptive condition by using contrastive inference, and that children will show no difference between conditions due to their still-maturing inferential skills (Kronmüller et al., 2014).

The main outcome of experiment 2 will be the identification of the strategy children use to interpret adjectives online. There are two strategic possibilities: 1) filter out all prenominal material and wait for the noun before fixating the target object, regardless of the adjective's informativeness (the easier/safer but slower strategy), or 2) deduce the informativeness of the adjective online, then use it flexibly in incremental adjective interpretation, requiring adult-like pragmatic abilities and processing capacities. Currently in the data collection phase, we expect to have analysed data from n=37 for each age group by the end of May 2019.

All results are discussed in light of their implications for sentence processing, clinical practice, and for models of adjective acquisition.

References

Fernald, A., Thorpe, K., & Marchman, V. A. (2010). Blue car, red car: Developing efficiency in online interpretation of adjective-noun phrases. *Cognitive Psychology, 60,* 190–217. doi: 10.1016/j.cogpsych.2009.12.002

Huang, Y., & Snedeker, J. (2013). The use of referential context in children's on-line interpretation of scalar adjectives. *Developmental Psychology, 49,* 1090-1102. doi: 10.1037/a0029477

Kronmüller, E., Morisseau, T., & Noveck, I. A. (2014). Show me the pragmatic contribution: A developmental investigation of contrastive inference. *Journal of Child Language, 41,* 985-1014. doi: 10.1017/S0305000913000263

Ninio, A. (2004). Young children's difficulty with adjectives modifying nouns. *Journal of Child Language, 31,* 255–285. doi: 10.1017/S0305000904006191

Ramscar, M., Yarlett, D., Dye, M., Denny, K. & Thorpe, K. (2010) The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science, 34,* 909-957. doi: 10.1111/j.1551-6709.2009.01092.x

Sekerina, I. A. and Trueswell, J.C. 2012. Interactive processing of contrastive expressions by Russian children. *First Language, 32,* 63–87. doi: 10.1177/0142723711403981

Thorpe, K., & Fernald, A. (2006). Knowing what a novel word is not: Two-year-olds 'listen through' ambiguous adjectives in fluent speech. *Cognition, 100,* 389-433. doi: 10.1016/j.cognition.2005.04.009

The interplay of scalar implicature and negative strengthening in different types of gradable adjectives

Nicole Gotzner, Stephanie Solt and Anton Benz (Leibniz-ZAS)

In Gotzner, Solt and Benz (2018 a,b), we examined the extent to which the structure underlying the semantics of gradable adjectives predicts the inferences that listeners derive (following up on van Tiel et al., 2016). In particular, we tested whether a weak scale-mate triggered a **scalar implicature** negating a stronger term on the same Horn scale (e.g., *John is tall* -> John is not gigantic). Conversely, we tested whether the negated stronger terms yielded a strengthened interpretation negating the weaker term – a phenomenon called **negative strengthening** (e.g., *John is not gigantic*-> John is not tall; see Horn, 1989, Levinson 2000). Our results showed that endorsement of scalar implicature for a given scale was anti-correlated with the degree of negative strengthening. Further, several factors related to scale structure predicted inference rates such as the type of standard instantiated by the weak term, whether the stronger term was endpoint-denoting or extreme and the polarity of the scale.

Most theories of implicature assume that scalar implicature and negative strengthening are based on two different pragmatic principles. For example, in Horn's account the Q principle derives scalar implicature while negative strengthening is based on the R-principle (or the M principle in Levinson, 2000). There is surprisingly little discussion in the theoretical literature on (i) how semantic factors should play a role in pragmatic strengthening and (ii) how variability across Horn scales can be formally modeled. In this presentation, we therefore wish to stimulate an open discussion on key findings of our research and what they tell us about theories of pragmatic strengthening.

The key questions addressed in our presentation are the following:

1. Why are scalar implicature and negative strengthening anti-correlated?
2. Why do upper-bounded Horn scales yield high rates of scalar implicature?
3. Why do extreme stronger terms yield low rates of scalar implicature but high rates of negative strengthening?
4. Why do negative scales behave differently?

Key references

Gotzner, N., Solt, S. & Benz, A., (2018a). The interplay of scalar implicature and negative strengthening in different types of gradable adjectives. Frontiers Research Topic Scalar Implicatures, *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2018.01659. OSF repository: https://osf.io/eb2hk/

Gotzner, N., Solt, S. & Benz, A. (2018b). Adjectival scales and three types of implicature. *Proceedings of SALT 28,* 409-423*.* MIT, Boston, MA.

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

Computing category membership for *tall*: An error minimization approach

Seungjin Hong, Jean-Pierre Koenig, Gail Mauner, Aron Marvel (University at Buffalo, SUNY)

shong23@buffalo.edu jpkoenig@buffalo.edu

How do we decide which objects in a comparison class are tall? Some experimental work has suggested that whether or not an object is above the mean for a category or set of objects is part of the basis for judgments of tallness (Barner and Snedeker, 2008; Syrett et al., 2010); some recent computational modeling studies have suggested that the determination of the threshold of tallness is a by-product of a process of optimizing how to communicate what somebody's actual height is by stating (s)he is tall (Lassiter & Goodman, 2017; Qing and Franke 2014; Schmidt et al., 2009). In contrast, we view setting the threshold of tallness as the result of categorization rather than communication and examine, in two experiments, some factors that affect people's decisions about where the threshold of tallness for novel categories of objects is. We propose that the threshold of tallness is set in a way that minimizes categorization errors when participants do not know in advance which object is or is not tall.

The two experiments we conducted use the same procedure. Following Barner and Snedecker (2008), three distinct arrays of 20 objects, each belonging to different novel categories, were displayed on separate tables. The layout of the object arrays was random across both participants and arrays as was the order in which arrays were presented. For each array, a participant placed the objects they deemed to be tall in a new location. The goal of Experiment 1 was to determine whether a large discontinuity in height in an array of objects affected participants' judgments of which objects were tall. In each of the three arrays in Experiment 1, there was a large discontinuity among the heights of the novel objects. In the 60-40 array, this discontinuity occurred between the bottom 60% and top 40% of objects. Similarly, in the 75-25 and 85-15 arrays, the discontinuities occurred between the bottom 75% and top 25% of objects and the bottom 85% and top 15% of objects respectively. Three Fisher exact tests showed that significantly more participants chose a particular height as a threshold for tallness if it corresponded to a discontinuity than if it did not (all ps < 0.001). χ_2 goodness of fit tests also showed that more participants (68%) chose as a threshold for tallness the object where the discontinuity in height occurred in the 75-25 array than in either the 60-40 array (46%) or the 85-15 array (28%), $p < 0.002$ and $p < 0.001$. Taken together, these results show that (1) participants preferred to set the threshold of tallness where a discontinuity in height occurs, suggesting that non-distributional properties (e.g., a discontinuity in height) can influence categorization, and (2) the effect of the discontinuity is modulated by where, in a distribution of objects, it occurs, as it was larger for the 75-25 array than either of the two other arrays. This latter result is interesting as it comports with Barner and Snedecker's observation that in arrays where objects varied in height from 1" to 9" in 1" increments, the average minimum height for tallness across their participants was 7.12".

The goal of Experiment 2 was to determine whether participants prefer to set the threshold of tallness between the object that corresponds to the 75th percentile of heights in an array (henceforth, 75% Total) or to 75% of the tallest object in an array (henceforth, 75% Tallest). Like Barner and Snedecker's results, Experiment 1's results cannot distinguish between these two possibilities since the two criteria converged on the same object in the critical 75-25 array. To distinguish between these possibilities, we constructed three arrays. In the first two arrays, the large discontinuity in height was at either the 75% Total object or the 75% Tallest object. In the control array, there was no discontinuity in the heights of the objects. A comparison of the number of participants that chose as their threshold objects in intervals centered around either the 75% Total object or the 75% Tallest object showed that, in the absence of a discontinuity, participants prefer to set the category boundary at the 75% Total

object (χ^2 $p < 0.0001$). A comparison of the number of participants that chose as threshold the 70%, 75%, or 80% Total object in the control array with no large discontinuity in height versus the array where there was a large discontinuity in height at the 75% Total object showed that the spread of the distribution around the 75% Total object was smaller when a discontinuity occurred at the 75% Total object (19, 14, 8 participants vs. 2, 38, 3 participants respectively, Fischer's exact test $p < 0.0001$). A comparison of the number of participants that chose as threshold objects an interval around the 75% Tallest object in the control array vs. in the array in which the large discontinuity in height occurred at the 75% Tallest object showed that a discontinuity significantly affected participants' choices (4 vs. 12, χ^2 $p < 0.0001$). But the presence of a discontinuity at the 75% Tallest object did not override the overall preference for the 75% Total object as a threshold for tall. More participants still chose as threshold an object within an interval centered around the 75% Total object than the 75% Tallest object (N = 26 vs. N= 13, χ^2 $p<0.05$).

Overall our results show that participants prefer a threshold that ensures that no more than about 25% of objects are categorized as tall and that the existence of a large discontinuity in height can modulate, but not necessarily override, this preference. But why do participants prefer a threshold that ensures that no more than 25% of objects are tall? Drawing inspiration from Huttenlocher et al. (2000) who found that participants' errors in reproducing one-dimensional stimuli is best explained by their goal of minimizing errors, we propose that our participants' preferences are similarly best explained by their goal of minimizing the chance of miscategorizing an object as tall/not-tall when it is or is not tall. Our hypothesis rests on two assumptions. First, participants have a prior belief that an object is tall only if it is within the top 50% of heights. Second, for a category like tallness and for a new comparison class, there are no features to help determine category membership aside from this prior belief. When there is no feature to help decide which objects are tall, the optimal strategy is to set the threshold so that, on average, errors are as few as possible given "true" membership in the category. This amounts to maximizing the harmonic mean *F* between the goal of including in the category only objects that are actually tall (a.k.a. precision) and the goal of including as many actually tall objects as possible in the category (a.k.a. recall). We tested this hypothesis by computing which proportion of objects selected as tall would on average maximize the harmonic mean *F* for all possible subsets of the top 50% of objects that may constitute the denotation of *tall*. We found that selecting the top 27.5% of objects would, on average, result in the highest harmonic mean, a proportion very similar to what we and Barner and Snedecker observed empirically. We interpret the robust effect of the presence of a discontinuity in height as the consequence of the presence of additional information about "true" membership (either because of Gestalt-like principles of perceptual grouping or the assumption that the discontinuity in height indexes two distinct populations). We surmise that the continued importance of the preference to set the boundary around the 75% Total object even in the presence of a discontinuity at the 75% Tallest object is due to participants' uncertainty as to whether the discontinuity is a reliable cue of "true" category membership. Additionally, our experiments and computational model suggest that a categorization approach to the interpretation of scalar predicates might explain the fact, well-known since Aristotle, that pairs of scalar adjectives like *short* and *tall* do not cover the entire scale. This fact may be the result of people's attempt to minimize errors when determining which of a new category of objects are tall (or short).

The time-course of adjective-noun composition in case of gradable and non-gradable adjectives

Arnold Kochari[a,b], Ashley Lewis[b,c], Herbert Schriefers[b]

[a] Institute for Logic, Language and Computation, University of Amsterdam; [b] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen; [c] Haskins Laboratories

Gradable adjectives are striking in their dependence on contextual information: the meaning of an adjective like "large" would be different if combined with the noun "mouse" or the noun "horse". Specifically, the threshold of application (i.e. how large the object needs to be in order to count as 'large') will differ depending on the typical sizes for the type of object that the noun refers to. Hence, it is necessary to know which noun the gradable adjective is combined with in order to fully determine its meaning. This is different from non-gradable adjectives like "dead", the meanings of which remain relatively stable regardless of the nouns they are composed with. The present project is aimed at investigating the consequences of this difference in context dependence between gradable and non-gradable adjectives for incremental processing of adjective-noun phrases by the brain. We do so using magnetoencephalography (MEG) data.

Composition of an adjective and a noun (e.g., "dead animal") as opposed to processing a noun by itself (e.g., "zgftr animal") has previously been found to be supported by the left anterior temporal lobe (LATL) with its activity peaking at 200-250 ms after noun onset (Bemis & Pylkkänen 2011; Westerlund & Pylkkänen 2014). This is a relatively early time-window given that noun meaning retrieval is thought to occur at around 300-400 ms after its onset. Interestingly, however, the effect at this early time-window was found when the noun was combined with a non-gradable adjective ("dead animal"), but not when it was combined with a gradable adjective ("large animal"; Ziegler & Pylkkänen 2016). This is what we would expect based on the difference between gradable and non-gradable adjectives in their context-dependence: composition cannot yet occur this early for the gradable adjectives because it critically depends on retrieval of the meaning of the noun.

In addition, a modulation of the activity in the same brain area by noun specificity (less specific "animal", "fruit" as opposed to more specific "horse", "apple") was observed at 350-470 ms after noun onset when the noun was combined with a gradable adjective, but not when it was combined with a non-gradable adjective. This time-window corresponds to the time we would expect most of noun meaning to already have been retrieved. One speculative idea for why there was such a difference between low- and high-specificity nouns is that the two have a different size or type of comparison class. Given this assumption, an effect in this time-window would indeed be expected in composition with gradable adjectives only, since only they require computation of the threshold of application.

The above described set of results is highly relevant for theories of gradable adjectives as they, to our knowledge, for the first time support theoretical predictions regarding computation of the threshold of application in incremental processing of gradable adjectives. One goal of the present project is to replicate these findings in an independent study. We do so with an experiment in a different language (Dutch) and with an improved set of materials controlling for additional potential confounds such as lexical frequency of nouns and plausibility of adjective-noun phrases. We also extend the experimental design with an additional condition where participants see pseudoadjective-noun phrases (this manipulation will allow us to disentangle specifically syntactic and semantic composition, but is not be

relevant for the difference between classes of adjectives that we focus on here).

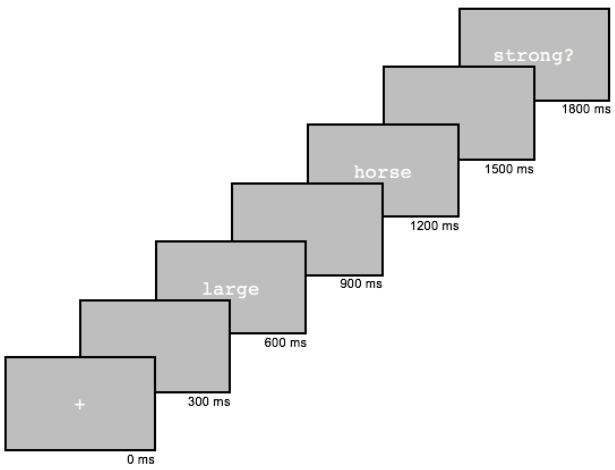
In the present study, participants (N=40) see an 'adjective' followed by a noun (see *Figure 1*) and subsequently respond to a comprehension question about this phrase. We present 80 nouns (40 high- and 40 low-specificity) in 4 conditions (see *Table 1*). We record MEG signals using a whole-head MEG system with 275 axial gradiometers and conduct the analyses in the same time-windows as in the original Ziegler & Pylkkänen (2016) study.

Successful replication of the observed difference between gradable and non-gradable adjectives will not only strengthen our confidence in this effect, but also also open avenues for potential follow-up studies which can, for example, investigate the difference in processing relative and absolute gradable adjectives or look for evidence that the modulation of the effect by nouns specificity is indeed due to comparison class differences. The MEG data allows us to look at processing in terms of both time-course of processing with a good resolution and brain areas involved. These data can not only test the processing validity of semantic theories (as we do in this project), but also inform semantic theories back.

Table 1. Example adjectives and nouns in each of the conditions.

Condition	**Adjective**	**Low spec noun**	**High spec noun**
gradable	groot(e) [large]	dier [animal]	paard [horse]
non-gradable	dood(e) [dead]		
pseudoadjective (control)	dirg(e)		
letter string (control)	hzgmr		

Figure 1. Trial structure. The comprehension question is included to ensure participants combine the meanings of the adjective and the noun. NB: English is used only for demonstration purposes.



References

Bemis, D. K., & Pylkkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *Journal of Neuroscience*, *31*(8), 2801-2814.

Westerlund, M., & Pylkkänen, L. (2014). The role of the left anterior temporal lobe in semantic composition vs. semantic memory. *Neuropsychologia*, *57*, 59-70.

Ziegler, J., & Pylkkänen, L. (2016). Scalar adjectives and the temporal unfolding of semantic composition: An MEG investigation. *Neuropsychologia*, *89*, 161-171.

Gradable Adjectives in Ironic Constructions – a corpus study

Claudia Lehmann
(University of Osnabrück)

Verbal irony is a linguistic phenomenon that is said to involve some kind of incongruity of assessment between what is said and its actual, implied meaning (e.g. Giora 1995, Attardo 2000, Wilson & Sperber 2012). A prime means to convey an ironic assessment is by using gradable adjectives. Consider, e.g. the widely cited

(1) What lovely weather this is.

where the gradable adjective *lovely* conveys the ironic meaning. It could be assumed that terms denoting the upper part of the scale bring about stronger pragmatic effects (in the sense of Colston 2015) than those denoting the middle part since the discrepancy between what is said and what is implicated is greater. Canestrari, Bianchi & Cori (2017) provide experimental evidence against this assumption.

This present paper offers a corpus study and thereby contributes to investigating the role gradable adjectives play in creating ironic meanings in natural settings. For the present purpose, *The Corpus of Contemporary American English* (COCA, Davies 2008-) has been used to find instances of the ironic construction *XP pro BE not* (e.g. *Eloquent he's not*). All instances that contained an adjective phrase in the XP slot of the construction were categorized as either non-ironic or ironic. The adjectives were further categorized into non-scalar, high, middle, and low adjectives. The results are summarized in the following table:

	non-scalar	**low**	**middle**	**high**
non-ironic	2	1	12	5
ironic	1	0	9	8

Even though there is a slight tendency to use more adjectives denoting the upper part of the scale in ironic contexts, this tendency does not reach a significant level. These results suggest that adjectives which denote the upper part of the scale are just as preferred as middle terms are for conveying ironic meanings, a finding that is in line with the findings of Canestrari, Bianchi & Cori (2017). Due to the limited number of instances of the *XP pro BE not* construction, a similar procedure will be used for other ironic constructions including the *NP BE* Adj_{sup} *N* construction (e.g. *He is not the most brilliant dog*) to see whether the present results can be replicated.

References

Attardo, S. (2000). Irony as relevant inappropriateness. *Journal of pragmatics*, *32*(6), 793-826.

Canestrari, C., Bianchi, I., & Cori, V. (2017). De-polarizing verbal irony. *Journal of Cognitive Psychology*, *30*(1), 43-62.

Colston, H. L. (2015). *Using figurative language*. Cambridge University Press.

Davies, M. (2008-) *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Available online at https://corpus.byu.edu/coca/.

Giora, R. (1995). On irony and negation. *Discourse processes*, *19*(2), 239-264.

Wilson, D., & Sperber, D. (2012). Explaining irony. *Meaning and relevance*, 123-145.

Reevaluating Pragmatic Reasoning in Web-based Language Games

Les Sikos, Noortje Venhuizen, Heiner Drenhaus, Muqing Li, and Matthew W. Crocker
(Saarland University)
sikos@coli.uni-saarland.de

Recent work testing formalizations of Gricean maxims [1] using web-based reference games has led to mixed results. Some studies indicate that Bayesian (e.g., rational speech act (RSA)) models closely predict human (pragmatic) behavior [e.g., 2], while others suggest that participants rarely go beyond the literal meanings of words in such studies [e.g., 3-4]. For instance, [2] presented participants with three objects (Fig1) in 7 different context types. Using a one-shot paradigm (each participant sees a single trial), they collected separate judgments from speakers, listeners, and for salience. Results of the RSA model, which combines a speaker model (likelihood that speakers use a particular word to refer to the target) with empirically measured salience (Eq1), were highly correlated with aggregate listener judgments (Fig1d; R=0.99). This was interpreted as indicating that participants reasoned pragmatically in this task. However, the reasoning required in [2] ranged from simple (e.g., Fig1b) to more complex (e.g., Fig1c), such that the close fit of predicted to observed results might be driven by the simpler inferences. Consistent with this possibility, [3] attempted a close replication of [2], focusing on more challenging items like Fig1c, and found that the basic RSA model was a poor predictor of their data. Furthermore, [4] found that while listeners responded pragmatically in conditions similar to Fig1b, they were only at chance for conditions similar to Fig1c.

To account for these results, [3] and [4] proposed various modifications to RSA (e.g., adding parameters for speaker/listener degree of rationality). Here, we investigate another possibility: Listeners in such web-based tasks may not reason as pragmatically as presumed. Instead, they may simply interpret the utterance based on a combination of its literal meaning and the salience of particular referents. In other words, a simpler rather than more complex model may better explain human behavior than RSA. To test this hypothesis we employed the same general methods as [2] and systematically explored a wider variety of context types (34 in total). 3387 participants recruited via Amazon Mechanical Turk were randomly assigned to Speaker (N=1143), Listener (N=1111), and Salience (N=1133) tasks (Fig2). We then compared observed responses to predictions from the basic RSA model and a Literal Listener (LL) model that does not incorporate a model of the speaker. This basic LL model predicts that listeners should be equally likely to select any referent that a given word (e.g. “green”) can refer to. In order to provide a more direct comparison to RSA, which relies heavily on salience, we also tested a LL+Salience model that weights its probabilities based on salience (Eq2). For completeness, we also considered an RSA model that assumes uniform salience, and salience values alone.

Table 1 and Fig2d show that while RSA provided a good fit to the entire dataset (replicating [2]), both LL models performed better. Furthermore, when we analyzed only the contexts for which the predictions from RSA and LL+Salience models differed (i.e, the more challenging inferences), LL+Salience performed best (Table 2 bottom; Fig2e). In fact, salience alone was a better predictor than RSA. Moreover, comparing RSA and RSA-uniform-salience models suggests that salience essentially corrects for incorrect predictions in the basic RSA model. To the extent that one-shot web-based experiments accurately elicit the depth of pragmatic reasoning seen in typical human interactions, these findings indicate that a simpler model than RSA can better explain human behavior.

a. **Speaker Task.** Imagine you are talking to someone and you want him to refer to the middle object. Which word would you say, "green" or "circle"?

Listener Task. Imaging someone is talking to you and uses the word "green" to refer to one of these objects. Which object are they talking about?

Salience Task. Imaging someone is talking to you and uses a word you don't know to refer to one of the objects. Which object are they talking about?

b. 

c. 

d.

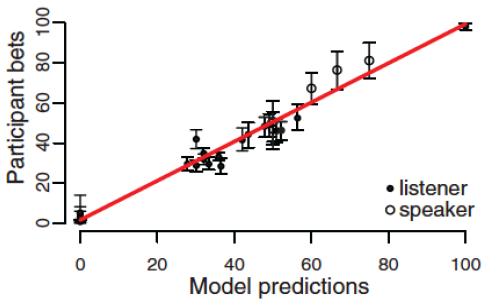


Fig 1. Overview of [2]. (a) Instructions. (b) Simple inference required. (c) Complex inference required. (d) RSA model predictions plotted against observed results.

$$\underbrace{P(r_s|w,C)}_{\text{Listener model}} = \frac{\overbrace{P(w|r_s,C)}^{\text{Speaker model}}\overbrace{P(r_s)}^{\text{Salience}}}{\sum_{r' \in C} P(w|r',C)P(r')}$$

Eq 1. RSA model for inferring the speaker's intended referent r_s in context C, given speaker's uttered word w.

$$\underbrace{P(r_s|w,C)}_{\text{Listener model}} = \begin{cases} \frac{\overbrace{P(r_s)}^{\text{Salience}}}{\sum_{r' \in R} P(r')} & \text{if } r_s \in R \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where: } R = \{r \in C \mid w \text{ can refer to } r\}$$

Eq 2. LL + salience model provides a distribution over the set of referents in context C that can be referred to with word w, weighted based on salience.

Table 1. Overall model fits (ranked in order of best fit).

	R	Adj R sq	t	p	
LL + Salience	0.89	0.79	16.36	0.00	***
LL uniform salience	0.88	0.77	15.52	0.00	***
RSA	0.87	0.75	14.44	0.00	***
RSA uniform salience	0.83	0.68	12.28	0.00	***
Salience alone	0.29	0.07	2.57	0.01	*

References

[1] Grice (1975). [2] Frank & Goodman (2012).

[3] Qing & Franke (2015). [4] Frank et al (2017).

a. **Speaker Task**. Imagine you are talking to Robert and you want him to pick out Item B. If you can only use one word, which word would you say, "green" or "fish"?

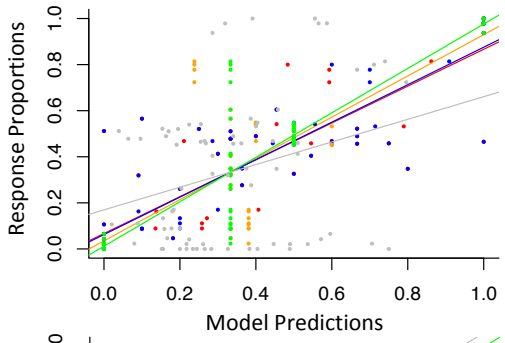
Listener Task. Robert wants you to pick one of the objects below but he can only say one word. He says, "green". Which object do you think he is talking about: A, B, or C?

Salience Task. Robert wants you to pick one of the objects below, but due to background noise you cannot understand what he said. Which object do you think he is most likely talking about: A, B, or C?

b. 

c. 

d.



e.

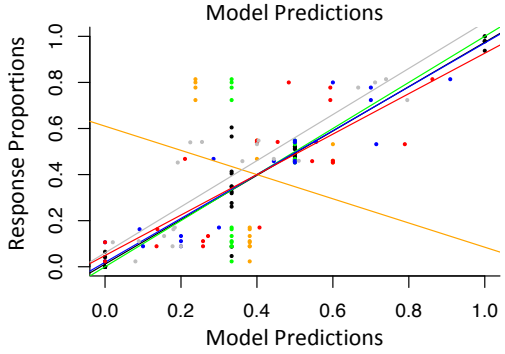


Fig 2. Overview of current study. (a) Instructions. (b) Simple inference required. (c) Complex inference required. (d) Predictions vs observed results over all visual contexts. (e) Predictions vs observed results for visual contexts in which models had identical predictions (black), and contexts in which model predictions differed. RSA, LL + salience, LL with uniform salience, RSA with uniform salience, Salience alone.

Table 2. Model fits for contexts in which models had identical predictions (top) and different predictions (bottom).

	R	Adj R sq	t	p	
Same prediction	0.97	0.94	28.56	0.00	***
Different predictions:					
LL + Salience	0.93	0.86	10.65	0.00	***
Salience alone	0.89	0.77	8.13	0.00	***
RSA	0.79	0.61	5.49	0.00	***
LL uniform salience	0.31	0.05	1.40	0.18	
RSA uniform salience	-0.23	0.00	-1.02	0.32	

An experimental approach to inferences to the standard in comparative constructions

Mora Maldonado, Alexander Martin and Jennifer Culbertson
Centre for Language Evolution — University of Edinburgh

What is the impact of adjective class and polarity on the inference pattern of comparative constructions? We provide the first experimental evidence regarding the availability of inferences to the standard for relative, absolute and *mixed* 'subjective' gradable adjectives.

Background In positive constructions, gradable adjectives are interpreted relative to a standard. For instance, (1a) is true iff John's height is above a certain standard of height.

(1) a. John is tall. b. The towel is dirty.

A central question in gradability semantics is whether this standard is context-sensitive or can be determined without reference to the context. Recent approaches to this question distinguish two main adjective classes[4,3]: *relative adjectives* (e.g., 1a), interpreted with respect to a context-sensitive standard; and *absolute adjectives* (e.g., 1b), which are claimed to have lexically encoded standards that always lie on their scale's minimum or maximum endpoint.

A critical piece of evidence for this distinction comes from the inference patterns of comparative constructions, which can establish a direct comparison of degrees. Relative comparatives are not expected to trigger an inference towards the standard, making (2a) fully compatible with John not being tall (both John and Bill may be short). If one assumes a lexicalised standard for absolute adjectives, absolute comparatives should still carry an inference to a 'standard': indeed, (2b) entails that the towel's dirtiness exceeds that standard. Crucially, the inference to the standard for absolute comparatives interacts with polarity: since the dirty/clean scale has only one absolute endpoint (where dirtiness "begins"), the antonym of 'dirty' in (2c) does not carry such inference.

(2) a. John is taller than Bill. $\not\rightarrow$ John is tall. **Neg/Pos Relative**
b. My towel is dirtier than yours. $\rightarrow$ My towel is dirty. **Neg. Absolute**
c. My shirt is cleaner than yours. $\not\rightarrow$ My shirt is clean. **Pos. Absolute**

This relative/absolute taxonomy does not exhaust gradable adjective space: 'subjective' or 'judge-dependent' predicates, for example, show a *mixed* behavior, alternatively patterning with relative or with absolute adjectives[1,5]. Judgments regarding the availability of inferences to the standard for *mixed* comparatives (e.g., 3) are even more variable and controversial than the ones obtained for classical relative and absolute examples.

(3) John's painting is uglier than Bill's. $\overset{?}{\rightarrow}$ John's painting is ugly. **Mixed**

Our goal here is twofold. First, we aim to provide experimental evidence regarding the role of the relative/absolute distinction in the semantics of comparatives[7,6]. Second, we will investigate the inference pattern of *mixed* comparatives, which remains mostly unexplored in the experimental literature[5]. We test the availability of inferences to the standard for comparatives instantiating different adjective types and polarities by testing judgements on inferences to the positive form.

Experiment We run an inferential task, where participants were presented with a statement (prompt), and asked to evaluate how justified it was to draw a given conclusion (target inference) based on a 5-point Likert scale (e.g., Fig 1)[2,6]. Critical trials instantiate one of three possible adjective classes (ADJ.CLASS: relative,

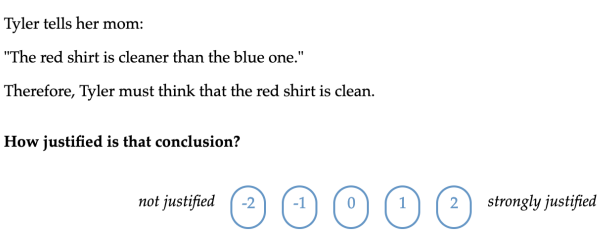


Figure 1: Example of critical trial

absolute and mixed) and one of two polarities (POL: positive, negative). We used 5 pairs of antonyms per ADJ.CLASS (30 items in total).

In each critical trial, the prompt was a comparative and the target inference was the corresponding positive construction. Participants' judgments on these positive constructions were then taken to indicate the extent to which the inference to the standard holds for the comparative.

63 control trials were included to elicit ratings across the entire scale. For instance, upward inferences in negative contexts were used to elicit low ratings (e.g., Ann doesn't own high heels $\nrightarrow$ Ann doesn't own shoes), and upward inferences in positive contexts were meant to elicit high ratings (e.g., John owns a red car $\rightarrow$ John owns a car). Control performance served as our exclusion criterion.

Results and discussion Proportion of responses in critical trials are shown in Fig. 2 (N=50, after exclusion).[1] The comparison between relative and absolute trials reveals a significant ADJ.CLASS $\times$ POL interaction ($\chi^2 = 48$, $p < .001$). As predicted by the taxonomy sketched above, negative absolute comparatives trigger significantly more inferences to the standard than positive absolute and relative comparatives. Overall, mixed comparatives give rise to significantly stronger inferences to the standard than relative comparatives (main effect of ADJ.CLASS: $\chi^2 = 6$, $p = .016$), but not than absolute comparatives ($\chi^2 < 1$, $p = .6$). The inference pattern of mixed adjectives however is not strictly analogous to the one of absolute adjectives. A significant ADJ.CLASS $\times$ POL interaction ($\chi^2 = 28$, $p < .001$) suggests that polarity has a weaker effect on mixed than on absolute comparatives.

Figure 2: Average inferences towards the standard by ADJ.CLASS and POL. Each dot represents a participant

Conclusions Our findings provide evidence suggesting that relative, absolute and mixed or 'subjective' gradable adjectives constitute different classes. To our knowledge, this is the first experimental approach that directly compares these three classes in comparative constructions.

In consonance with recent analyses and diagnostics, we have provided evidence that relative and absolute adjectives differ in the nature of their standards: absolute but not relative adjectives seem to encode a fixed standard that plays a role in the interpretation of comparative constructions. Like absolute adjectives, mixed comparatives also trigger inferences to the standard. Should 'mixed' gradable adjectives also be modeled as encoding a context-independent standard? Our experiment does not allow us to tell whether inferences arising for mixed and absolute comparatives are of exactly the same nature. Arguably, these inferences might have different strengths (entailment vs pragmatic inference), which might explain the interaction with polarity. In the presentation, we will discuss this issue. We will also present and discuss some additional item-specific behaviour that indicates that one should be cautious about generalisations based on single items.

[1]Our analysis plan was preregistered: https://osf.io/gkfjr/?view_only=94ae407832584276b58f5a3992ac95a9

[1] M. Bierwisch. (1989) *The semantics of gradation. Dimensional adjectives*. [2] C. Cummins and N. Katsos. (2010) Comparative and superlative quantifiers: Pragmatic effects of comparison type. *Journal of Semantics*. [3] C. Kennedy. (2007) Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*. [4] C. Kennedy and L. Mcnally. Scale Structure, Degree Modification, and the Semantics of Gradable Predicates.*Language* 2005. [5] S. Y. Liao and A. Meskin. (2017) Aesthetic Adjectives: Experimental Semantics and Context-Sensitivity. *Philosophy and Phenomenological Research* [6] J. Rett and A. Brasoveanu. (2018) Evaluativity across adjective and construction types: An experimental study. *Journal of Linguistics* [7] K. Syrett, C. Kennedy, and J. Lidz. (2009) Meaning and context in children's understanding of gradable adjectives. *Journal of semantics* [8] A. Toledo and G. W. Sassoon. Absolute vs . Relative Adjectives. *SALT Proceedings*

Adjectives in degree result clause constructions with negative polarity minimizers: Experimental investigation and a Lexical Resource Semantics account

Monica-Mihaela Rizea
University of Bucharest

Manfred Sailer
Goethe University Frankfurt/Main

Scales play an important role in recent research in formal semantics and experimental pragmatics. In this paper, we discuss a less investigated phenomenon, for which a scalar analysis is very natural: high degree readings of adjectives that act as primary predicates in result clause constructions (RCX), where the secondary predication in the result clause (RCl) is represented by an emphatic negative polarity item (E-NPI). We propose a twofold analysis, combining an experimental investigation, in the first part, and a formal, constraint-based analysis in the second part. The current research is part of a comparative study on Romanian and English E-NPIs. For simplicity reasons, in this proposal, we present the observed phenomena, using English examples, even if the Romanian constructions have some particularities that are not found in English.

1. Recent experimental studies such as Gotzner et al. 2018 report results related to how scalar implicatures affect the interpretation of gradable adjectives, focusing on factors such as polarity and boundedness. In our study, we analyse the role played by scalar implicatures in the interpretation of adjectives that express their high degree of intensity by means of a special type of degree result clause construction, i.e., ADJ + finite RCl, where the RCl hosts an emphatic NPI, such as a negated minimizer -- see example (1). We show that, in these particular structures, two scalar inferences concur in deriving the intensity readings: the use, in the RCl, of a minimizer, which corresponds to the speaker's subjectively defined low degree, can be, first, pragmatically strengthened under negation to express an endpoint on a contextually salient scale[1]; second, an indirect scalar implicature is responsible for the high degree reading of the adjective in the matrix: *the lower the degree of the result represented by the negated minimizer, the higher the intensity of the quality emphasized* or, in other words, if the intensity_i of a quality results in something that is emphatic and low on a contextual scale, this intensity_i is interpreted as high.

In an utterance such as (1) The fog is **so thick that <u>you can't see your hand in front of your face</u>**, the RCX *so thick that you can't see your hand in front of your face* can receive the high degree interpretation of "extremely thick" [2], as a quality of the *fog*.

We have shown above that this interpretation can be derived by a scale-reversal implicature. We will now focus on the (initial) inference that involves the proposition in the RCl: The negated minimizer i.e., NOT+ the 'minimum' action that for the speaker counts as an event of seeing -- *see <u>your hand in front of your face</u>* and nothing more (or what Eckardt 2005 called a 'minimality implicature') -- can be pragmatically strengthened to implicate that one "can see <u>nothing at all</u>" (i.e., from the *minimum something* to *absolutely nothing*, on a visibility scale); this process would correspond to endorsing, in the experimental task, the *outmost low point* on the scale -- that is, the *strongest endpoint-denoting scale-mate*, which would translate in interpreting the proposition in the RCl as "there is an extremely low degree of visibility". However, we cannot exclude the possibility that some speakers can trigger an implicature that would be *close to (i.e., not corresponding to) the scalar endpoint* (even if, naturally, still stronger than other alternatives). Therefore, even if they would endorse a

[1] By *strengthening*, we understand here *the phenomenon by which an utterance receives a stronger interpretation than its semantic meaning* (see Gotzner et al. 2018).

[2] For the current stage of the investigation, we use *very* and *extremely* interchangeably, since not all speakers can perceive a relevant difference between the two (i.e, *extremely* is commonly interpreted as an upper-most boundary on a scale of intensity, and not as exceeding the endpoint of a conceivable scale).

strong scale-mate, this would not be the (**lower**) **bound**. This would reflect in interpreting the proposition in the RCl as 'one can see almost nothing at all' or 'there is a quite/pretty low degree of visibility'. We explain this choice as largely depending on the set of alternatives that the speaker conceives on the (here, visibility) scale -- see (2), and on whether the minimizer is perceived as the *strongest i.e., most emphatic claim possible.* To generalize, a minimizer NPI makes a set of alternatives accessible – here, ranges of visibility. Consequently, when the NPI is used in an RCl, the result clause makes an emphatic statement with respect to a contextually relevant set, namely the alternatives provided by the NPI: (1) could also be rephrased as (2), with other E-NPIs or just regular emphatic negative statements: (2) *The fog is* ***so thick that you can't see your hand in front of your face/within a step/two steps ahead/ whithin half a meter***, etc. However, what is perceived as *the most emphatic* statement largely depends on a subjectively-defined threshold varying with the speaker. *Fog so thick you can't see within half a meter,* which, on its purely semantic meaning, seems to indicate a *less low visibility* in comparison to the alternatives we formulated in (2) could, for some speakers, still imply that "there is absolutely no visibility at all", that is, to make the *strongest claim.* In other words, *not seeing within half a meter* could be perceived as *emphatic enough* as to represent *the lowest degree of visibility*, thus ignoring other possible alternatives, since the utterance in the RCX is interpreted as serving a mere emphasis purpose in the communication process, then offering no relevant semantic information on the actual distance where visibility is possible; this means that all the above alternatives would be interpreted as the strongest possible, i.e., neutralized to equally strong within the given set[3].

An initial experiment tests the hypothesis that there is an interaction between (i) speakers' decision whether to apply pragmatic strengthening in the proposition in the RCl and (ii) the endorsement of the strongest scale-mate, representing the degree of the adjective in the matrix -- that is, whether in the constructions where the adjective and the negated minimizer are represented in a cause-result relation in discourse, the interpretation of the degree of the adjective -- as 'extreme/high degree' vs. 'relatively high degree '-- is correlated to whether the negated minimizer expression in the RCl is associated by the subjects to 'the strongest possible claim' (i.e., to an absolute endpoint on a contextually salient scale), or to 'strong, but not strongest' claims. In another task, we verify if, for an individual minimizer NPI, there are specific degree RCXs -- within the set of its admitted collocations with degree adjectives and nouns modified by the respective adjectives -- where speakers consistently (i.e., significantly) expect the strongest claim possible while 'clearly excluding' (i.e., not 'also accepting') less strong claims. We hypothesize that the semantic distance between the lexical scale contributed by the minimizer and the noun that the adjective modifies is also a factor in the above constellation.

In what concerns the method, as rightfully pointed out in related studies (Gotzner et al. 2018: 12), presenting the test sentences *within a conversational context* would be desirable; however, we have decided to present the tested items *isolated from context* on the grounds of better serving our research purpose. Previous experimental research (e.g., Liu et al.2013) also pointed out that, in the absence of contextual clues, subjects have the tendency of constructing their own best-case scenario (i.e., what they find as the most representative situation) in order to support their linguistic judgements. One issue that we intend to test here is, precisely, if there is a strong pattern in speakers' decision of applying pragmatic strengthening for certain RCXs (i.e., endorsing the upper-bound scale-mate), and we assume the best way of obtaining this is on what subjects will, themselves, describe as the most

[3] These assumptions are based on our discussions with informants and on the results of an initial native speaker judgement online survey.

suitable context they resort to for their decisions, which will also offer insight on the possible alternatives they construct, without influencing them by providing a context[4].

2. The second part of our study will provide a scalar extension of a standard account of degree RCXs to capture high degree readings of adjectives, and will propose a constraint-based version of a pragmatic, scalar approach to E-NPIs. In particular, we will show how minimizers interact with another scalar construction, degree result clauses, and we will provide support for the integration of an (exhaustification) operator for scalar emphasis inside embedded clauses. We will propose an account that combines the pragmatic theories of NPI licensing, such as Krifka (1994), Eckardt (2005), Chierchia (2006), with those of a degree-semantic analysis of adjectives Meier (2003), Kennedy & McNally (2005). Our analysis will be formulated in Lexical Resource Semantics (LRS) (Richter & Sailer 2004).

References • Chierchia, Gennaro. 2006. Broaden your views. implicatures of domain widening and the "logicality" of language. Linguistic Inquiry 37. 535–590. • Gotzner, Nicole, Anton Benz & Stephanie Solt. 2018. Scalar diversity, negative strengthening and adjectival semantics. Frontiers in Psychology, Art. 1659. • Kennedy, Christopher & Louise McNally. 2005a. Scale structure, degree modification, and the semantics of gradable predicates. Language 2(82). 345–381. • Krifka, Manfred. 1994. The semantics and pragmatics of weak and strong polarity items in assertions. In Proceedings of Salt IV, 195–219. • Liu Mingya, Eva Csipak & Regine Eckardt. 2013. Polarity in Context. *Beyond 'Any' and 'Ever': New Explorations in Negative Polarity Sensitivity.* vol. 262, De Gruyter Mouton, Berlin: 351- 368. • Meier, Cécile. 2003. The meaning of too, enough, and so . . . that. Natural Language Semantics 11. 69–107. • Richter, Frank & Manfred Sailer. 2004. Basic concepts of lexical resource semantics.In Arne Beckmann & Norbert Preining (eds.), Esslli 2003.

[4] We exclude from this analysis the cases that we have identified as result clause constructions with a bleached interpretation of the result clause material.

Proportional modification of gradable adjectives: the case of percentages

Giorgos Spathas, Leibniz ZAS

This paper investigates percentages (e.g. *fifty percent*) in their use as modifiers of gradable adjectives. I provide a first characterization of the readings that arise and argue for an analysis of percentages as quantifiers over degrees that specify ratios of lengths of intervals.

The readings. The **Degree Reading (RD)**, exemplified in (1), specifies a particular degree in the scale associated with the adjective. RD is possible with closed adjectives (1), and not with relative, partial, or total adjectives (2). Notice that (1a) and (1b) mutually entail each other.

(1) a. The door is 66 percent open. **Ratio**: $\frac{the\ door's\ openness}{maximum\ openness}$

b. The door is 34 percent closed.

(2) John is 60 percent # tall / # wet/ ?? dry. *relative / partial / total*

The **Comparison Reading (CR)** is a marginal reading that involves the ratio of the measurement of two arguments, as in the naturally occurring example in (3) with a positive relative adjective. CR is (marginally) available with all adjectives except negative relative and total adjectives. We exemplify the unavailable readings in (4) and (5).

(3) I have created a [background] image that is 2000 pixels wide […]. Next you need to create an area on the background image […], **so you need to create an area on the background image that is 23 percent wide**. [http://tinyurl.com/y47wva2r]

(4) #Compared to the background, the new area is 23% narrow. **Ratio**: $\frac{the\ area's\ width}{the\ image's\ width}$

(5) #Compared to your blanket, mine is only 20 percent # dry / # clean.

The **Partition Reading (PR)** shows no sensitivity to the boundedness of the scales, nor to the positive-negative distinction. We exemplify in (6) with a relative adjective. In the talk, we provide evidence that PR involves a derived measure function rather than the one associated with the adjective and requires a different treatment. We put it aside for now.

(6) The river is 70 percent wide, and 30 narrow. **Ratio**: $\frac{extent\ of\ the\ river\ that\ is\ wide}{full\ extent\ of\ the\ river}$

Finally, percentages appear in **differential comparatives**, where they are unambiguous and show no restrictions. Again, we exemplify with relative adjectives, in (7).

(7) a. John is 20 percent taller than Bill. **Ratio**: $\frac{difference\ between\ John\ and\ Bill's\ heights}{Bill's\ height}$

b. John is 20 percent shorter than Bill.

Intervals, not degrees. The distribution of DR and CR provides novel evidence against the idea that gradable adjectives relate individuals to unique degrees, (8), or that percentages specify ratios of degree points. Assume (8) and a scale of OPENNESS from zero openness to maximum openness (from 0° to 90°) and a door open at 60°. This gives the correct result for (1a), assuming that the numerator is the degree of the door's openness and the denominator is the maximum degree in the scale. However, it fails with the negative adjective *closed*; *closed* picks up the same degree as *open*, e.g. 60°, in a scale with a reverse ordering relation. If so, maximum CLOSEDNESS in the denominator is zero and the ratio should be undefined, contrary to fact. The problem persists even if a direction-sensitive *max* function returns the minimum degree of CLOSEDNESS (i.e. 90°). In this case, *The door is 66 percent open* is semantically identical to *The door is 66 percent closed*, clearly the wrong result. Moreover, any account that allows *max* functions, as they, e.g., appear in the treatment of comparatives, to pick up the relevant degrees in the ratios, will fail to explain the differences in the distribution of DR and CR between (a) closed and total adjectives, which both have maximum degrees, and (b) closed and partial adjectives, which both have minimum degrees.

(8) $[\![\text{open}]\!] = \lambda d\, \lambda x\, [\, \text{OPEN}(x) = d\,]$

Proposal. We adopt the entries in (9) (von Stechow 1984, Heim 2000, a.m.o.), and assume that the scales underlying classes of adjectives differ not only in the OPEN-CLOSE distinction (Kennedy and McNally 2005), but also in terms of boundedness (Rotstein and Winter 2004) , as in (10) (where $a > 0$). Antonymic pairs further differ in the direction of scale. In our example, the door's 'openness' and 'closedness' are identified with the intervals in (11a-b), respectively.

(9) a. $[\![\text{open}]\!] = \lambda d\, \lambda x\, [\, \text{OPEN}(x) \geq d\,]$ b. $[\![\text{closed}]\!] = \lambda d\, \lambda x\, [\, \text{CLOSE}(x) \geq d\,]$

(10) a. Closed: pos $[\, \overrightarrow{0, a}\,]$ / neg: $[\, \overleftarrow{0, a}\,]$ b. Relative: pos $(\, \overrightarrow{0, \infty}\,)$ / neg: $(\, \overleftarrow{0, \infty}\,)$

c. Partial: $[\, \overrightarrow{a, \infty}\,)$ d. Total: $[\, \overleftarrow{a, \infty}\,)$

(11) a. $\{\, d \mid \text{OPEN}(\text{door}) \geq d\,\} = [0, 60]$ b. $\{\, d \mid \text{CLOSE}(\text{door}) \leq d\,\} = [60, 90]$

The innovation lies in assuming that percentages specify ratios of **lengths of intervals**, as in (12) (which assumes that the scales are dense). Given that the denominator in (1) is the length of the relevant scale, (1a-b) now differ in the required way, as indicated by the ratios in (13).

(12) The **length** of an interval I, $l(I)$, with endpoints a, b is $b - a$ if I is bounded and ∞ if I is unbounded.

(13) a. $\frac{l[0,60]}{l[0,90]} = \frac{60}{90}$ = (appr.) 66% b. $\frac{l[60,90]}{l[0,90]} = \frac{30}{90}$ = (appr.) 34%

Restrictions on DR and CR. The account predicts that percentages will be infelicitous any time an unbounded interval appears in the relevant ratio. Assuming that the denominator in DRs is the length of the relevant scale, we correctly predict that all adjectives other than closed, will be infelicitous. Assuming that the denominator in CRs is always the length of the interval associated with the comparative element in the relevant dimension, we correctly predict that CRs will be out with partial and negative relative adjectives, since the corresponding intervals will be unbounded, as in (15), where b is measurement of the relevant object.

(14) a. For, e.g., *short*: $\{\, d \mid \text{WIDTH}(x) \geq d\,\} = [b, \infty)$ for some $b > 0$ *neg relative*
b. For, e.g., *dry*: $\{\, d \mid \text{DRY}(x) \geq d\,\} = [b, \infty)$ for some $b > 0$ *total*

The lexical entry. We propose to treat percentages as generalized quantifiers over degrees, which captures their ability to coordinate with Measure Phrases, their availability with modifiers like *exactly* (von Stechow 2005 for MPs) and their ability to take wide scope (Beck 2009 for MPs). The denominator is determined contextually, allowing for both DRs and CRs.

(15) $[\![\, n \text{ percent}\,]\!] = \lambda D_{d,t}\, [\, \frac{l(D)}{l(C)} \geq n\,]$ *where C a variable over degree intervals*

Differential comparatives. A unified treatment with percentages in comparatives necessitates an entry for *-er* that creates a differential interval, as in (17) (Breakstone et al. 2011). A problem arises with negative adjectives as in (8b). These are expected to be infelicitous, as specified above for CRs. We treat (8b) in a decompositional framework of antonymy (cf. Büring 2007, Heim 2008), slightly modified so that *shorter* can be the spell-out of LESS TALL. TALL in the main clause allows ellipsis of TALL in the *than*-clause, eliminating the problem of unbounded intervals in the denominator.

(16) $[\![\, \text{-er}_{\text{diff}}\,]\!] = \lambda P_{dt,t}\, \lambda T_{d,t}\, \lambda M_{d,t}\, [\, T \subset M \,\&\, P(M \backslash T)\,]$ *where* $(M \mid T) =_{def} \{x \mid x \in M \,\&\, x \notin T\}$

(17) $[\![\, \text{LESS}\,]\!] = \lambda P_{dt,t}\, \lambda T_{d,t}\, \lambda M_{d,t}\, [\, M \subset T \,\&\, P(T \backslash M)\,]$

Open issues. The account cannot capture examples with modals and quantifiers in the *than*-clause. As far we can see so far, no current account of (20-21) can be successfully combined with an analysis of percentages, in a way consistent with the results described above.

(18) Lucinda drove 10 percent less fast than she is allowed to.

(19) John is 10 percent taller than every girl.

Comparison class inference for gradable adjectives

Michael Henry Tessler[1,2], Roger Levy[1], and Noah Goodman[2]
[1]Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology
[2]Department of Psychology, Stanford University

A 24 °C (75 °F) day is warm, while a 16 °C (60 °F) day is not. Unless it's January: 16 °C could be warm for January. *Warm* is a relative adjective, and its felicity depends upon what the speaker uses as a basis of comparison—the *comparison class* (e.g., other days of the year vs. January). Comparison classes are necessary for understanding gradable adjectives, but as with relevant aspects of context more generally, comparison classes often go unsaid (e.g., in "It's warm").

Any particular referent of discourse can be conceptualized or categorized in multiple ways, giving rise to multiple possible comparison classes. How do listeners decide on the appropriate comparison class? In this project, we investigate the first aspect of this open-ended inference problem: deciding between a relatively specific comparison class (e.g., *warm for winter*) and a relatively general class (e.g., *warm for the year*). Theoretical work in semantics has focused on how information from the comparison class gets integrated with a compositional semantics and what representations might be preferred (Bale, 2011; Solt, 2009). To our knowledge, the question of how listeners reconstruct a comparison class from just the adjective and world knowledge has not been addressed either empirically or formally.

We develop a Rational Speech Act model wherein a listener combines hierarchical, category knowledge with pragmatic reasoning to infer the comparison class implicitly used by the speaker. The model combines a previously proposed method for inferring common ground (Degen, Tessler, & Goodman, 2015) with an uncertain threshold mechanism for deriving context-sensitive interpretations for gradable adjectives (Lassiter & Goodman, 2015). In this model, the speaker could produce an adjective with an explicit comparison class ("warm for winter"), which has the effect of changing the common ground (or, the listener's prior distribution over the degree e.g., plausible temperatures in a season). In this way, reasoning about the comparison class is cashed out in terms of reasoning about alternative utterances (in this case, alternative utterances using different comparison classes). The model predicts that hearing "it's warm" (in Winter) signals that the speaker meant *warm for winter*, while hearing "it's cold" is more likely to signal *cold for the year*. The opposite relationship is predicted to hold in summer, where "it's cold" should signal cold *for summer* more so than "it's warm" (Fig. 1).

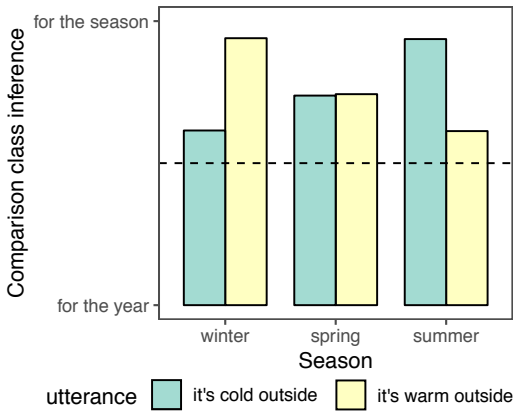


Figure 1: Listener inferences about the comparison class upon hearing either a positive-form adjective ("warm") or negative-form adjective ("cold") during different seasons.

We test these qualitative predictions in a free-production paraphrase experiment. On each trial, participants were given a sentence introducing the subordinate category (e.g., *Tanya lives in Maryland and steps outside in Winter*), followed by a sentence which described an object or situation with an adjective (e.g., *Tanya says to her friend, "It's warm."*). Participants were asked "What do you think Tanya meant?"; they were provided with a sentence frame ("It's warm relative to the other ____") and asked to fill in the blank. We used positive- and negative-form gradable

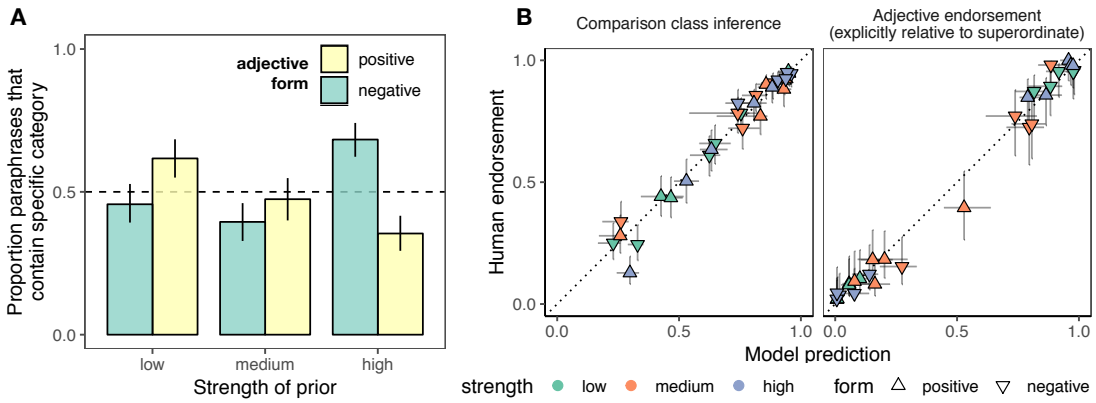


Figure 2: A: Human production of subordinate comparison classes averaged across items. X-axis orders different subordinate categories by their intuitive average value relative to a superordinate category (e.g., "low" = Winter; "high" = Summer; degree = temperature). B: Model fits for the two-alternative comparison class inference task (left) and adjective endorsement task (right). Error bars = 95% credible intervals.

adjectives describing eight scales (height, weight, size, duration, temperature, price, darkness, loudness). Each scale was paired with a superordinate category and three subordinate categories that were intuitively situated near the high-, low-, and intermediate parts of the degree scale (e.g., winter, spring, and summer for temperature). We recruited 63 US IP-addresssed participants from Amazon's MTurk with a 95% work approval rating and self-reported native language of English.

The free production data show the richness of the human capacity to reconstruct context. An expensive crystal flower vase could be expensive for a *crystal flower vase*, a *flower vase*, a *vase*, or just for an *item* or *gift*. As a simple test of our main hypothesis (Fig. 1), we coded each response as to whether or not the text contained the specific category (e.g., did the response mention the words *crystal*, *flower*, and *vase*? did the response for a warm day in Winter mention *winter*?), and we assume that responses which do not mention the specific category refer to more general categories. The qualitative predictions of the model (described above) were borne out (Fig. 2A).

To test the quantitative predictions of the model, we designed a two-alternative forced choice version of the experiment ($n = 264$). The quantitative predictions depend on both the relative prior distributions over the degree scale (e.g., temperatures of days in Winter vs. over the year) and the prior distribution over comparison classes. As an extremely rough approximation for the prior over comparison classes, we take the corpus frequency of the noun phrase. To infer the priors over the degrees, we ran an additional adjective endorsement task ($n = 100$) where participants are asked to judge if an adjective applied to a typical member of the subordinate category (e.g., a typical day in Winter) relative to the more general category (e.g., cold relative to a typical day of the year). We then used a Bayesian data analytic model together with an RSA model for adjective endorsement to triangulate the priors over the degrees that would accommodate both data sets. The model provides a strong quantitative fit to the comparison class inference data (Fig. 2B), demonstrating that these inferences can be investigated with quantitative methods.

References

Bale, A. C. (2011). Scales and comparison classes. *Natural Language Semantics*, *19*, 169–190.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the 37th CogSci.*

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*.

Solt, S. (2009). Notes on the Comparison Class. In *International workshop on vagueness in communication.*

World knowledge and the absolute-relative distinction in adjectives in offline and online processing

Barbara Tomaszewicz & Petra B. Schumacher (Universität zu Köln)

Gradable adjectives fall into two classes: **relative** adjectives, such as 'short', that require context for their interpretation, and **absolute** adjectives, 'empty', 'spotted', that can have context independent meanings (Rotstein & Winter 2004, Kennedy & McNally 2005, Kennedy 2007). We show that the absolute-relative distinction is blurred in an offline rating task, but emerges during online processing.

The interpretation of both relative and absolute adjectives is subject to the effects of world knowledge, i.e. to what extent familiar everyday objects are associated with particular thresholds of gradable properties. To judge whether the sentence 'This glass is short' is true, you need to know the comparison set of other relevant glasses to establish a **contextual threshold** for shortness. To judge a glass as 'empty' you only need to know if the **maximum threshold** of emptiness is reached; other glasses in the context are irrelevant. However, your world knowledge still affects interpretation, e.g. how empty the glass should be. Similarly, the **minimum threshold** for 'spotted' can shift (but doesn't have to) depending on the context. The shifting thresholds of minimum and maximum absolute adjectives can be treated as a **pragmatic** phenomenon (Kennedy 2007, Leffel et al. 2016, 2017, a.o.) or a **semantic** one where world knowledge determines the threshold just like with relative adjectives (the probabilistic knowledge of the threshold is derived from the prior degree distribution) (Lassiter & Goodman, 2013, 2015; Qing & Franke, 2014a, 2014b). On the latter account, the difference in context sensitivity results from the difference in the stability of the optimal threshold. Assuming that everyday objects are associated with a **typical threshold** of a property (e.g., all the scarves in Fig.1 can be described as 'short') do we find evidence for probabilistic uses of both types of adjectives or is there a clear absolute-relative distinction with both offline and online measures?

Fig. 1 Sample items in Exp 1-2

Adjective Type:	Degree of property: 1	2	3	4	5
RELATIVE *short*					
ABSOLUTE MINIMUM *spotted*					
ABSOLUTE MAXIMUM *empty*					

Exp 1-2. Ratings. We followed the design of Kim et al. (2013, 2014), Leffel et al. (2017) who found a distinction: the thresholds of relative adjectives were midway on the scale, those of absolute adjectives close to the endpoints. We created substantially more items: 187 adjective-object pairs in photographs of 5 degrees of a familiar property (Fig.1) for 14 relative and 14 absolute (8 max, 6 min) adjectives. In an offline task (in German), participants evaluated the goodness of fit of adjective and photograph by choosing between *Yes*/*No*/*Don't Know* for each of the 5 pictures in a set, counterbalanced left-right/right-left order (Exp.1, *n*=72, 84 targets, Exp.2, *n*=72, 103 targets). Averaging over the 5-point scales for each adjective-object pair, we obtain the curves in Fig.2-3. In Exp1-2, there is an effect of degree, a significant interaction, but no effect of adjective type. The averages hide a great underlying variability, therefore, we ran a clustering algorithm revealing 3 clusters (Fig.4). Each cluster contained adjectives from all classes (see table). This result supports **the probabilistic approach**.

Fig. 2 Results Exp 1

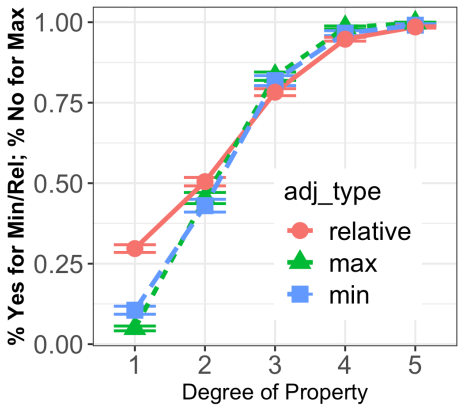


Fig. 3 Results Exp 2

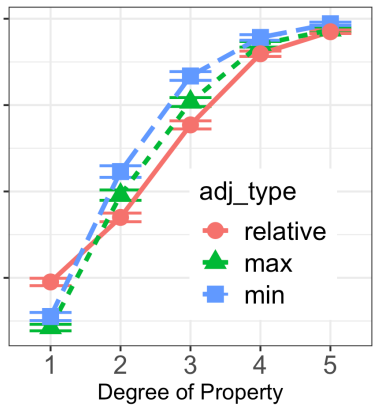
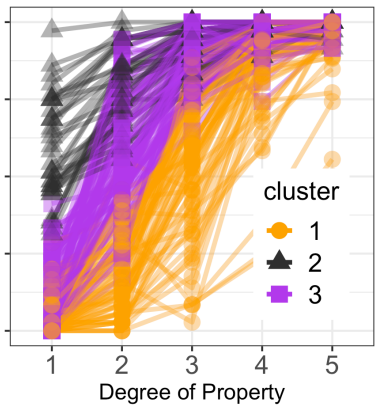


Fig. 4 Clusters from Exp 1-2

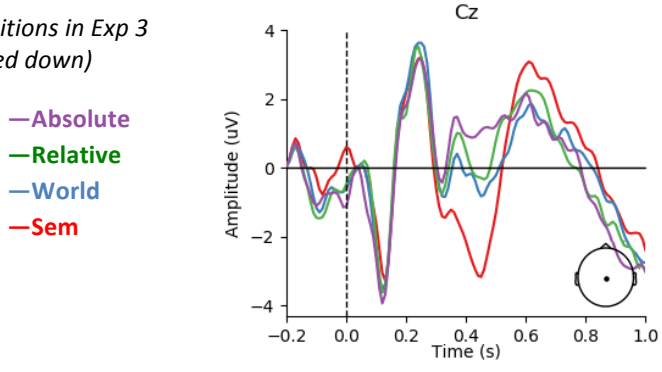


No. of adjectives in:	Relative	Absolute Min	Absolute Max	Total
Cluster 1	14	4	7	25
Cluster 2	11	3	2	16
Cluster 3	13	6	8	27

Exp 3. ERPs. Cluster 3 in Exp 1-2 contains the most uniform profiles, spanning all 5 degrees, and the largest number of adjectives (13 relative, 14 absolute). From cluster 3 we selected 10 relative and 10 absolute (5 min, 5 max) adjectives that each had 3 adjective-object pairs in the cluster. We predicted that differences in processing can be attributed to the differences in meaning, i.e. optimal thresholds. In an EEG study (in German), participants (n=30) first saw a set of 5 pictures (as in Fig.1), then a red frame appeared selecting degree 2 or 4 (counterbalanced order), followed by a serial presentation of a sentence, e.g., 'This / is / short'. The adjective either matched or mismatched the selected degree. In addition to absolute and relative adjectives, we added two conditions for which mismatches resulted in semantic and world knowledge violations (Hagoort et al. 2004), Fig.5, to compare the cognitive mechanisms underlying adjective interpretation. The relative adjectives required the assessment of the relation between the marked object and its comparison set. The absolute adjective didn't; the marked object was never the max/min of the scale. But if absolute adjectives relied on probabilistic thresholds like relative adjectives, both classes should pattern together with the world knowledge condition, which was not confirmed. Analyses revealed a reliable effect of condition between 300-500ms. The semantic mismatch elicited a clear N400 effect (Fig.6). The relative condition also elicited a negative shift, like the world knowledge condition, indicating that during online processing **relative thresholds are probabilistically resolved**, while absolute thresholds are stable. This result is compatible with both pragmatic and probabilistic accounts.

Fig. 5 Sample items from additional conditions in Exp 3

Fig. 6 Mismatching conditions in Exp 3 (negativity is plotted down)



Scalar and non-scalar equatives in Turkish

Carla Umbach (ZAS Berlin / University of Cologne)
Umut Özge (Middle East Technical University)

Equative comparison constructions occur across categories – adjectival as well as nominal and verbal. In English, adjectival equatives are (mostly) scalar, while nominal and verbal ones are (mostly) non-scalar. At the same time, in English scalar (adjectival) equatives make use of the standard marker *as*, while non-scalar (nominal/verbal) ones make use of *like* as a standard marker, cmp. (1a-c). In German as well as Polish, there is only one standard marker, which is used across categories, in scalar as well as non-scalar equatives (German *wie*, Polish *jak*), see (2a-c).

(1) a.	Anna is as tall as Berta.	(2) a. Anna ist so groß wie Berta.	(adj. / scalar)
b.	Anna's dress is like Berta's.	b. Annas Kleid ist so wie Bertas.	(nom. / non-scalar)
c.	Anna runs like Berta (does).	c. Anna rennt so wie Berta.	(verb. / non-scalar)

In Turkish, there are two standard markers, *kadar* and *gibi*, indicating scalar and non-scalar equatives,rsp. *Kadar* is an originally Arabic word roughly equivalent to English *much*. *Gibi* can be translated as *similar* or *like*. Surprisingly, in contrast to English the two standard markers can be used across categories. Thus we find scalar equatives based on adjectival as well as nominal/verbal parameters of comparison, (4a, 5a, 6a), and we find non-scalar equatives based on nominal/verbal as well as adjectival parameters, see (4b, 5b, 6b).

(4) a. Anna Berta kadar uzun.
A. B. kadar long.Cop3sg
`Anna is as tall as Berta.'
(scalar, same height)

b. Anna Berta gibi uzun.
A. B. gibi long.Cop3sg
`Anna is tall like Berta.'
(non-scalar, similar in the way of being tall)

(5) a. Anna'nın elbisesi Berta'nın-ki kadar.
A.-Gen dress.Poss3sg B.-Gen-Rel kadar.Cop.3sg
`Anna's dress is as _____ as Berta's.'
(scalar, e.g., same length or price)

b. Anna'nın elbisesi Berta'nın-ki gibi.
A.-Gen dress Poss3sg B.-Gen-Rel gibi.Cop.3sg
`Anna's dress is like Berta's.'
(non-scalar, e.g., design & color & fabric)

(6) a. Anna Berta kadar koşuyor.
A. B. kadar run.3sg.Prog
`Anna runs as _____ as Berta.'
(scalar, e.g. duration or frequency

b. Anna Berta gibi koşuyor.
A. B. gibi run.3sg.Prog
`Anna runs like Berta.'
(non-scalar, e.g. style and equipment)

The contrast between *kadar* equatives and *gibi* equatives gives rise to a number of intriguing observations concerning (i) the unexpected (from the point of view of English) occurrence of *gibi* in adjectival equatives and (ii) the unexpected occurrence of *kadar* in nominal/verbal equatives.

Ad (i) a) *gibi*, but not *kadar,* is compatible with non-gradable adjectives, e.g. *Anna Berta gibi mezun.* 'Anna is graduated like Berta' (e.g. through an intense program, etc.)

b) *gibi* allows for different comparison classes, e.g. in (4), Anna might be a girl and Berta her mother, which is strongly dispreferred with *kadar*;

c) *gibi* blocks degree modifiers like *en az* ('at least'), which are o.k. with *kadar;*

d) *gibi,* but not *kadar,* blocks measure phrases *1,90 m kadar uzun / *gibi uzun.* However, with *kadar* the sentence has only a comparative reading: *around 1.90 m taller.*

(c) and d) indicate that Turkish has degree-variables, see Beck et al.2010).

Ad (ii) e) *kadar* in nominal and verbal equatives selects exactly one dimension, which has to be metric. For example, (5b) can neither be understood as *Anna's skirt is as long and expensive as Berta's* nor as *Anna's skirt is as stylish as Berta's*;

f) licit dimensions in nominal/verbal *kadar* equatives are severely restricted by the particular noun/verb; for example, the dimension of age is licensed for kids but not for houses; similarly, scalar comparison of dresses is restricted to length and price, see (4a)

g) licit dimensions in *gibi* equatives seems to be subject to general restrictions to appearance or manner (see Umbach & Stolterfoht in prep).

There are at the moment two types of analyses available to account for the semantics of equative comparison. One is based on the standard degree-based semantics of comparatives (e.g., Bierwisch 1987, Kennedy 1999). It focuses on scalar adjectival equatives, as in (1a), and fails to handle non-scalar cases. The other type of analysis makes use of kinds or similarity classes. Kind-based accounts (Anderson & Morzycki 2015) and similarity-based accounts (Umbach & Gust 2014) are suited for non-scalar as well as scalar equatives, though when dealing with scalar cases they have to make some extra effort (postulating "degree-kinds" and referring to similarity in one-dimensional spaces, rsp.).

However, in view of the Turkish data the idea that degree-based and kind-/similarity-based accounts of equatives are competing theories can no longer be maintained. We have to acknowledge that – within the same language – two different strategies of performing equative comparison are manifest, while the choice between strategies depends on the standard marker.

The framework in Umbach & Gust (2014) can be adapted to handle the two strategies – degree-based and similarity-based – in parallel (without reducing one to the other). We propose a semantic interpretation of Turkish equatives within this framework such that

- Scalar equatives make use of a 1-dimensional measure function of type <e,d> which is either
  given by the adjective: [[*uzun*]] = $\lambda x . \mu_{height}(x)$ (Kennedy 1999),
  or underspecified: [[Meas]] = $\lambda x. \mu_S(x)$ (Solt 2015)

- Non-scalar equatives make use of an n-dimensional measure function taking individuals to points in n-dimensional attribute spaces, <e, d^n>: [[genMeas]] = $\lambda x. \mu_{Sn}(x)$
  (where S_n is a variable over n-dimensions vectors and d^n are points in n-dimensional spaces)

- *kadar* denotes a weak linear order, e.g. $\geq$, and
  gibi denotes a similarity relation in an n-dimensional attribute space:

  [[A. B. *kadar* α]] = $\mu_\alpha(a') \geq \mu_\alpha(b')$ where μ_α is 1-dim measure function

  [[A. B. *gibi* δ]] = $\mu_\delta(a') \approx_F \mu_\delta(b')$ where μ_δ is a generalized measure function and $\approx_F$ denotes similarity in the n-dimensional attribute space F

Stepping back, the semantics of Turkish comparatives is evidence that two distinct ontological categories' (degrees and similarity classes) may co-inhabit the semantic domain of an operator in a single language, which have so far been thought to be individually selected by languages and/or theories in an either-or fashion as their primary comparative semantic ontology.

Anderson, C. and Morzycki, M. (2015) Degrees as kinds. NLLT 33:79 -821.

Beck, S. & Krasikova, S. & Fleischer, D. et al. 2010. Crosslinguistic variation in comparison constructions. In J. van Craenenbroeck & J. Rooryck (eds.) *Linguistic VariationYearbook* 2009.

Bierwisch, M. (1987) Semantik der Graduierung. In M.Bierwisch & E. Lang (eds.) *Grammatische und konzeptuelle Aspekte von Dimensionsadjektiven*. Akademie Verlag Berlin, 91-286.

Kennedy, C. (1999) *Projecting the Adjective*. Garland Press, New York.

Solt S. (2015) Q-Adjectives and the Semantics of Quantity. Journal of Semantics 32 221 -273.

Umbach, C. & H. Gust (2014) Similarity demonstratives. *Lingua* 149:74-93.

Umbach, C. & B. Stolterfoht (in prep.) Ad-hoc kind formation by similarity.

Not all gradable adjectives are vague – Experimental evidence from children and adults

Merle Weicker and Petra Schulz
Goethe University Frankfurt

Summary. Semanticists widely agree that all gradable adjectives (GA) have the same semantic type, denoting relations between individuals and degrees on a corresponding scale (e.g., Cresswell 1976), and that relative gradable adjectives (RGA) show characteristics of vagueness (e.g., Kennedy 2007). However, there is disagreement on whether absolute gradable adjectives (AGA), like RGAs, are vague predicates. We investigated whether AGAs are interpreted as vague predicates and how vagueness is reflected in adults' judgements for RGAs and possibly AGAs. By studying 3- to 5-year-old children, we also examined how early in the course of acquisition the interpretation pattern found in adults emerges.

Gradable adjectives & vagueness. Sentences containing a RGA (e.g., *big/small*) show several characteristics of vagueness. First, RGAs are interpreted relative to a context sensitive standard of comparison. Context sensitivity is related to the scale structure of RGAs: they have open scales, and the standard is typically located around the midpoint of the scale (Kennedy 2007). Second, context sensitivity is reflected in antonym pairs of RGAs. They are 'non-complementary' (Cruse 1980), i.e., the negation of the adjective does not entail the assertion of its antonym (see (1) for the context in Fig. 1a). Because RGAs are context sensitive, the standard for bigness and smallness need not to be the same degree (Kennedy 2007, Solt 2011).

(1) Water balloon no. 5 is not big. ⇏ Water balloon no. 5 is small.

Third, 'borderline cases' exist for vague predicates. Take Figure 1a: in addition to the sets of objects intuitively judged as big (e.g., balloons 6-8) and as not big (e.g., balloons 1-3), there are objects (e.g., balloons 4, 5) that are more difficult to judge as big or not big. Depending on the specific account, these borderline cases are judged as (i) 'truth value gluts', i.e. big **and** not big (or big and small) or as (ii) 'truth value gaps', i.e. **neither** big **nor** not big (or neither big nor small) (Égré/Zehr 2018). In other words, borderline cases seem to allow contradictions.

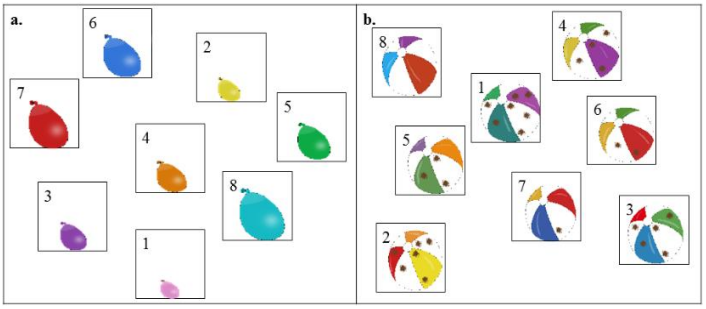


Fig. 1: Example context for *big/small* (a) and for *clean/dirty* (b).

AGAs (e.g., *clean/dirty*) have been regarded as non-vague for two reasons. First, their standard is less context-sensitive. AGAs have (partially) closed scales, and the standard is either a non-zero degree (for *dirty*) or the maximum degree (for *clean*) of the scale (Kennedy 2007). Consequently, antonyms of AGAs are often analyzed as 'complementary': in Figure 1b every ball that is not considered clean could be considered dirty (no. 1-7). Second, AGAs do not seem to give rise to borderline cases (Kennedy 2007). However, some approaches postulate a less sharp distinction between RGAs and AGAs. According to Toledo & Sassoon (2011), AGAs are context-sensitive just like RGAs. Rotstein & Winter (2004) show that antonyms of AGAs need not always to be complementaries. Moreover, vagueness may arise through pragmatic processes, which are assumed to be the same for all classes of GAs (Burnett 2012).

Previous empirical studies focused on specific aspects of vagueness in GAs. Studies with adults suggest that borderline cases for RGAs are interpreted as gluts and as gaps, but that the gap-interpretation is preferred (Solt/Gotzner 2010, Égré/Zehr 2018). Studies with children indicate a non-complementary interpretation of RGAs (*tall, short*) at age 4 (Barner/Snedeker

2008). Moreover, differences in the standard of comparison between RGAs (*big, long*) and AGAs (*spotted, full*) were found at age 3 (Syrett et al. 2006).

Study. ***Research questions.*** We investigated whether children and adults differentiate between RGAs and AGAs regarding the standard of comparison (Q1), the relation between adjectives in antonym pairs (Q2), and the existence of borderline cases (Q3). ***Participants.*** Three groups of German-speaking children (ages: 3, 4, 5 years; N = 43) and 26 adults were tested. ***Method.*** The interpretation of RGAs (*big, small*) and AGAs (*clean, dirty*) was tested with a forced picture-choice task (within-subject design, 2 trials per adjective). In each trial, 8 picture cards displaying single objects were presented in unordered fashion. Across the array of objects, the property denoted by the adjective increased (Fig. 1). Participants were asked *Please give me the Adj N* (e.g., *big water balloons*). The same visual array was presented with both adjectives of an antonym pair. ***Results.*** **Q1:** The 3 child groups and the adults did not differ regarding the position of the standard (Kruskal-Wallis, all *p*'s > .05): they located the standard of comparison around the midpoint of the scale for RGAs, and at the non-zero or maximal degree of the scale for AGAs (Fig. 2). **Q2:** All groups interpreted AGAs as complementary: every object that was not selected as clean was selected as dirty. In contrast, RGAs were interpreted as non-complementary; not every object that was not selected as big was selected as small (Fig. 2).

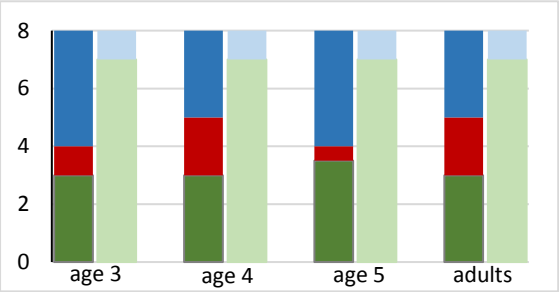


Fig. 2: Number of objects (median) classified as *big, small* or neither and *clean, dirty* or neither per age group. Blue = *big*, light blue = *clean*, green = *small*, light green = *dirty*, red = neither.

Q3: In AGA-trials for most participants (91.3%) no borderline cases existed (i.e. 'no gap/glut'). In contrast, in RGA-trials only for 7.2% of the participants no borderline cases existed, 8.7% interpreted borderline cases as big and small ('glut'), and 62.3% interpreted borderline cases as neither big nor small ('gap'). The proportion of participants with a gap-interpretation for borderline cases compared to the proportion of participants with a glut-interpretation increased with age ($X^2(3) = 8.42$, $p = .014$).

Discussion. Our study confirms previous findings that adults interpret RGAs as vague predicates and provides first evidence that adults interpret AGAs as non-vague predicates. This difference was reflected in the standard of comparison, the complementarity of antonyms, and the existence of borderline cases. The child data reveal that this interpretation pattern is already present in 3-year-olds, suggesting that–although all GAs have the same semantic type–children are sensitive to the differences of RGAs and AGAs regarding vagueness early on. Based on the parallel results for children and adults we argue that the scale structure (open vs. closed) is a core feature of GAs. Further research is necessary to provide a theoretical analysis of the gap- and glut-interpretations that takes into account our novel empirical finding that the preference for gap-interpretations increased with age. In addition, cross-linguistic research should investigate whether the pattern found for German holds universally (see Beck et al. 2009).

Selected References **Burnett, H. (2012).** The Puzzle(s) of Absolute Adjectives. On Vagueness, Comparison, and the Origin of Scale Structure. In D. Paperno (ed.), *UCLA Working Papers in Linguistics, Papers in Semantics Vol. 16* (p. 1-50). **Égré, P. & Zehr, J. (2018).** Are Gaps Preferred to Gluts? A Closer Look at Borderline Contradictions. In E. Castroviejo et al. (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure* (p. 25-58). Springer. **Kennedy, C. (2007).** Vagueness and Grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30, 1-45. **Rotstein, C. & Winter, Y. (2004).** Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. *Natural Language Semantics*, 12, 259-288. **Syrett, K.,**

Bradley, E., Kennedy, C. & Lidz, J. (2006). Shifting Standards: Children's Understanding of Gradable Adjectives. In K. Ud Deen et al. (eds.), *Proceedings of GALANA* (p. 353-364). Cambridge, MA: UConn Occasional Papers in Linguistic.

Are all absolute predicates truly absolute?

Myung Hye Yoo, *University of Delaware*

Gradable predicates have been classified into two categories in terms of scales—absolute and relative predicates (Hay et al, 1999; Kennedy & McNally, 2005; Kennedy & Levin, 2008). The absolute predicates have to achieve their own standard degrees to have their own properties. Total predicates have maximal standard of the property that must be reached out (e.g. *close*), whereas partial predicates should have at least the minimal property (e.g. *open*). Relative adjectives, on the other hand, do not have the standard property to achieve since they are context-dependent (e.g. *wide*). The scalar analysis of gradable predicates is also applied when they express the change of state, so called inchoatives.

This paper reexamines the properties of gradable predicates through the distribution of the different morphological forms of the corresponding Korean inchoatives. The traditional classification of absolute and relative predicates does not fully capture the distribution of Korean inchoatives. All relative predicates combine with *-(e)ci* morpheme after their root forms, which leads to the meaning of the change of state. Only some of the absolute predicates, however, combine with *-(e)ci,* while the rest of the absolute predicates maintain their bare forms as shown in (1)-(3).

(1) Absolute total predicates
  a. Bare: *tat-hi*, 'close', *swum-* 'hide', *pi-* 'empty', *cha-* 'fill'
  b. *-(e)ci: kkaykkushay-eci-* 'clean', *phyengphyenghay-eci-* 'flatten'

(2) Absolute partial predicates
  a. Bare: *yel-li* 'open', *tulena-* 'expose'
  b. *-(e)ci*: *telewu-eci-* 'dirty', *hulisha-eci-* 'blur'

(3) Relative predicates: *nelp-eci-* 'widen', *cop-aci-* 'narrow', *ccalp-aci-* 'shorten'

I propose that only the predicates that maintain their forms are true absolute predicates, while those that combine with *-(e)ci* to become inchoatives are in fact relative predicates even though some of them appear to be absolute. The true absolute predicates must achieve their own absolute end points to have their own predicate meanings. The predicates that combine with *-(e)ci,* however, do not necessarily require to achieve the end points to convey their own semantic properties of degrees.

(4) and (5) support this novel approach to absolute and relative predicates. Shifting a standard is only acceptable in *-(e)ci* inchoatives as shown in (4), even though *telewu-* 'dirty' seems to be an absolute predicate. Furthermore, *-ka* 'go', which presents the movement towards the absolute end point (Zubizarreta & Oh, 2007), is only acceptable in bare inchoatives as seen in (5). These findings allow us to predict that *-(e)ci* inchoatives are not compatible with absoluteness, but rather have relative meanings.

(4) a. *ku mwun-i (icen-pota) yellyessta. kulena yecenhi yellici anh-nun-ta
  The door-NOM (before-than) open-PST-DEC *But still open not –Pres-DEC
  'The door opened (than before), but it is still not open.'
 b. ku pang-i (icen-pota) telewu-eci-ess-ta. kulena yecenhi telepci-anh-ta
  The room-NOM dirty (before-than) -eci- PST –DEC. But still dirty not-DEC
  'The room got dirty (before-than), but it is still not dirty'

(5) a. ku mwun-i (ta/keuy) yellye-ka-ass-ta (bare inchoative)
  the door-NOM (completely/almost) open-go-PST-DEC

'The door is (completely/almost) getting dry'

b. *ku pang-i (ta/keuy) telewe-ka-ass-ta (-*(e)ci* inchoative)
the room-NOM (compeletely/almost) dirty-go-PST-DEC
*The room is (completely/almost) getting dirty'

We further conducted a judgement task to examine which analysis better predicts the perception of Korean gradable predicates in 'successive increase' contexts containing 'again' sentences as seen in (6), which are discussed in Petersen (2015). Under the traditional scalar analysis, only traditional relative predicates are predicted to be compatible with 'successive increase' contexts. If, however, all predicates that combine with -*(e)ci,* including those that appear to be absolute, are in fact relative predicates, then all -*(e)ci* predicates are predicted to be more acceptable than bare inchoatives in this task.

(6) a. Last week, the river widened a lot and reached the flood barrier. This week, the river widened again and overflowed onto the bank. (relative predicates)

b. This morning, I left the soaking wet shirt out in the sun for a few hours. When I took it in, it had dried somewhat but was still quite damp. When I put the shirt outside in the afternoon, it dried again. (total absolute predicates)

In this experiment, Korean native speakers were given paragraphs that contain 'successive increase' contexts in Korean as seen in (6). Then they were asked to judge the naturalness of the contexts on a 6-point scale (n=42). We manipulated the morphological forms (bare vs. -*(e)ci*) and the scale property of predicates (total vs. partial vs. relative) in a 2x3 design. The result of the experiment revealed a significant difference between bare and -*(e)ci* inchoatives ($t=4.6$, $p<.01$). It can be concluded that subjects perceived -*(e)ci* inchoatives to be more natural than bare inchoatives (-*(e)ci* inchoatives: M=2.68, s=1.53; bare inchoatives: M=2.18, s=1.14). On the other hand, the experiment did not show significant differences in terms of the scale of predicates: total vs. partial vs. relative predicates ($F=1.65$, $p>.05$). Since 'more' is more natural than 'again' in 'successive increase' contexts, the naturalness judgment was low in average. However, it is significant that participants presented a significant difference between bare and -*(e)ci* inchoatives, but not between traditional absolute and relative predicates.

This study provided the empirical evidence that Korean native speakers perceive bare and -*(e)ci* inchoatives in a different way, following the proposal of the current paper. In 'successive increase' contexts, all -*(e)ci* inchoatives, even those that appear to be absolute, were perceived to have more relative meaning than bare inchoatives. The scale properties of inchoatives, in contrast, does not significantly reflect the perception of gradable predicates by Korean native speakers. This paper proposes the reanalysis on the semantic properties of absolute and relative predicates.

References

Hay, J., Kennedy, C., and Levin, B. 1999. *Scalar Structure Underlies Telicity in "Degree Achievements"*. Proceeding of SALT 9, 127-144.

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

Kennedy, C. & Levin, B. 2008. Measure of Change: The Adjectival Core of Degree Achievements. In *Adjectives and Adverbs: Syntax, Semantics and Discourse*, (eds.). L. McNally and C. Kennedy, 1-30. Oxford: Oxford University Press.

Pedersen, W.2015. A scalar analysis of *again*-ambiguities. *Journal of Semantics*, 32, 373-424.

Zubizarreta, M.L., and Oh, E. 2007. *On the Syntactic Composition of Manner and Motion.* Cambridge: The MIT Press.

Ambidirectionality and Thai mid-scale terms: when 'warm' means less hot

Jérémy Zehr & Nattanun Chanchaochai *University of Pennsylvania*

Empirical observations It may seem an uncontroversial thing to say that *to get warmer* means to undergo an increase rather than a decrease in temperature. This, however, is a semantic point that Thai speakers may not find intuitive, for the Thai translation of 'get warmer,' *ʔùn kʰûn* (literally 'warm ascend') can describe not only increases in temperature, but also decreases from hot to moderately warm [1]. The same observation holds for *sa-lǔ:a kʰûn* ('dim ascend') and *cʰúː:n kʰûn* ('damp ascend') which can respectively describe not only increases in darkness or wetness, but also changes from highly to moderately dark or wet. Such ambidirectional interpretations are unavailable for more extreme scalemates (*hot/cold*, *dark/bright*, *wet/dry*). After considering and rejecting two other possible analyses, we propose that the Thai mid-scale predicates are indeed semantically equivalent to English *warm*, *dim* and *damp*, and give a semantics for *kʰûn* ('ascend') that accounts for cases of ambidirectionality.

To be rejected 1: *mild* rather than *warm* One might consider translating *ʔùn* as *mild*, and *ʔùn kʰûn* as *get milder*, which also exhibits ambidirectionality [2]. Such an analysis has two weaknesses: first, it would require new, parallel translations for *sa-lǔ:a* ('dim') and *cʰúː:n* ('damp'), and, second, it predicts *too mild* to be a good translation for excess-constructions built with *ʔùn*. This prediction is not borne out: while *too mild* roughly means *too moderate* [3], the Thai sentence [4] unidirectionally denotes excessively *high* temperatures in much the same way as *too warm*.

To be rejected 2: *turn A* rather than *get A-er* In another plausible type of account, *sa-lǔ:a kʰûn* ('dim ascend') would receive a non-scalar interpretation along the lines of ***turn*** *dim*. Such an analysis would be compatible with ambidirectionality [5], but further empirical observations lead us to discard it: regardless of the direction of the change, *sa-lǔ:a kʰûn* ('dim ascend') can be modified by a measure phrase referring to the *difference* in illumination [6], whereas ***turn*** *50 lumens dim* describes a *final* illumination of 50 lumens.

Our proposal: *kʰûn* as *moving away from alternative* We propose that *kʰûn* describes changes whose *initial* state satisfies a salient alternative of the scalar predicate, and whose *final* state satisfies the scalar predicate itself *instead* [7]. We make two additional assumptions: (i) {*cold, warm, hot*}, {*bright, dim, dark*} and {*dry, damp, wet*} represent salient alternative sets, and (ii) *hot*, *dark* and *wet* respectively entail *warm*, *dim* and *damp* at the literal level (i.e. Thai and English are alike). Since we have assumed two alternatives for each predicate, composition with *kʰûn* can always follow two different paths. When composing with *warm*, one path (*cold* as the alternative) results in what could be paraphrased as *warm but no longer cold*, describing an increase in temperatures; the result of the other path (*hot* as the alternative) could be paraphrased as *warm but no longer hot*, describing a decrease in temperatures. When composing with *hot*, choosing *cold* as the alternative results in the expected change, *hot and no longer cold*; choosing the *warm* alternative, however, results in a literal contradiction, paraphrasable as # *hot but no longer warm*. This result is general, given our assumptions: since it is impossible to literally satisfy an *entailing* predicate without at the same time satisfying an *entailed* one, only one path is left for *hot*, *dark* and *wet*, which therefore always yield unidirectional interpretations. As for *cold*, *bright* and *dry*, the change can only go one way, since each has two alternatives that share the same orientation (e.g. *cold and not warm/hot*).

The semantics we propose needs two refinements. For one, we need a semantic value that can combine with a measure phrase [6]. Second, native speakers' judgments suggest that the change need not *complete* a move away from the alternative nor up to satisfying the predicate itself [1]. We give our final proposal in [8] where we (i) change the type of the semantic value so that it denotes a degree corresponding to the difference between the degrees at the initial and at the final states, and (ii) quantify over consistent $standard$ functions (a method reminiscent of delineation semantics,

e.g. Klein 1980) as well as (iii) over *expansions* of the change.

Discussion Our observations on Thai show that scalar expressions give rise to semantic effects that go beyond what is attested in English. We proposed that Thai has an expression, *kʰûn*, that quantifies over its scalar complement's alternatives. We make two final remarks. First, anecdotal evidence of *English*-speaking children using "warmer" to mean *less hot* suggests a similar semantic analysis of mid-scale comparatives, which invites further investigation. Second, our semantics gives a central role to alternatives. Since the Thai counterpart of *cool* is typically not used in the same types of context as *cold*, it is not an alternative to *cold* and does not normally exhibit ambidirectionality. Remarkably, native speakers' judgments suggest that ambidirectionality becomes conceivable (if not entirely natural) for *cool* in the rare contexts that license both *cool* and *cold*. That is, if one were to manipulate the context so as to make any two unrelated scalar terms salient asymmetrically entailing alternatives (an *ad-hoc* scale) one would expect the same kind of ambidirectionality. Conversely, if a scalar predicate lacks any salient alternative, one predicts composition with *kʰûn* to be infelicitous, to the extent that the existential quantification over alternatives would yield trivial falsity. We leave this prediction open to further empirical study. Finally, while our account assumes the existence of a set of salient alternatives, it gives no indication as to how that set is determined, or what the appropriate notion of salience is. Researchers have started to tackle such issues from an experimental perspective (see Doran et al. 2012, van Tiel et al. 2012, Schwarz et al. 2016, McNally 2017) but the question remains a matter of empirical debate at the moment.

[1] tɔːn-níː man kɔ̂ː jaŋ **jen/rɔ́ːn** jùː náʔ tʰɯ̌ŋ man càʔ **ʔùn kʰûn** nít-nɯŋ kɔ̂ː-tʰɤ̀ʔ
now it EMP still **cool/hot** ASP FP although it AUX **warm ascend** a little despite
'It is still **cool/hot**, although it got slightly closer to a moderate temperature.'

[2] The weather is too **warm/cold**. I'll wait until it **gets milder**.

[3] The weather is **too mild** to have an outdoor ice rink, *or* an outdoor swimming pool.

[4] náːm kɛ̂ːw níː man **ʔùn kɤːn** náʔ
water CLS-glass this it **warm too** FP
'This glass of water is too warm' / # 'This glass of water is not hot enough.'

[5] The experiment room was very {**dark / bright**} at first, but then the light turned **dim**.

[6] mûːa-kíː man **mûːt** mâːk lɤːj tɔːn-níː **sa-lŭːa kʰûn** maː hâː-sip luːmen lɛ́ːw
just now it **dark** very EMP now **dim** **ascend** DEI 50 lumens already
'It was very dark before. Now it has become 50 lumens brighter.'

[7] $\lambda A.\ \lambda x.\ \lambda e.\ \exists B \in \text{Alt}(A)\ [B(x, e_{start}) > \text{std}(B) > B(x, e_{end})]\ \wedge A(x, e_{end}) > \text{std}(A)$.
$\approx x$ now meets A's standard but no longer meets its **alternative**'s

[8] i. $\lambda A.\ \lambda x.\ \lambda e.\ \lambda d.\quad d = \text{diff}(A(x, e_{end}), A(x, e_{start}))\ \wedge$
ii. $\exists s' \sim std, B \in \text{Alt}(A)\ [B(x, e_{start}) > s'(B) > B(x, e_{end}) \wedge A(x, e_{end}) > s'(A)] \wedge$
iii. $\exists d', h_{orizon}\ [d' = \text{diff}(h_{orizon}, A(x, e_{start}))\ \wedge\ d' \geq d\ \wedge\ h_{orizon} > \text{std}(A)]$.
$\approx$ degrees representing the amplitude of the change such that x:
- no longer meets B's standard but still meets A's for some **consistent** shift of standards
- meets A's actual standard after a change at least as big as the present one

References. **Doran, R., Ward, G., Larson, M., and McNabb, Y. (2012).** A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*, 88:124–154. **Klein, E. (1980).** A semantics for positive and comparative adjectives. *L&P*, 4:1–45. **McNally, L. (2017).** Scalar alternatives and scalar inference involving adjectives. In Ostrove et al. editors, *Asking the Right Questions: Essays in Honor of Sandra Chung*, pages 17–27. **Scwharz, F., Zehr, J., Grodner, D., and Bacovcin, H. A. (2016).** Subliminal priming of alternatives does not increase implicature responses. Poster presented the *Logic and Language in Conversation Workshop in Utrecht.* **van Tiel, B., van Miltenburg, E., Zevakhina, N., and Geurts, B. (2016).** Scalar diversity. *JoS*, 33:137–175.

Contradictory Descriptions with Absolute Adjectives

Jeremy Zehr (UPenn) and Paul Egré (ENS, PSL University, CNRS)

1. Borderline contradictions. Several experiments over the last decade indicate that borderline cases of vague predicates license contradictory descriptions such as "x is neither tall nor not tall", or "x is tall and not tall" [1,2,4]. These so-called borderline contradictions have been regimented in paraconsistent-friendly accounts of vagueness [1,3,4,5], in which "and" and "neither... nor..." descriptions are treated symmetrically. In [6], however, a marked preference was evidenced for descriptions of the form "neither P nor not P" over "P and not P" when P is a *relative* gradable adjective. [6] left open whether this pattern would extend to *absolute* gradable adjectives. In this paper, we report on an experiment that replicates the findings of [6] for relative adjectives, but shows no such asymmetry for absolute adjectives. This difference invites a revision of the account laid out in [6], by integrating data concerning the treatment of lexical antonyms.

2. Study. Drawing on [6] we presented participants in an online experiment with short vignettes describing target borderline cases for 8 adjectives, asking whether the contradictory descriptions were true, along with two unproblematic true and false control descriptions. Each participant judged either relative or absolute adjectives, either with their syntactic negation (*not tall/not flat*) or their lexical antonym (*short/bumpy*). For absolute adjectives, borderline cases were designed to be cases located very near the closed bound of the scale (see [7] and Examples below).

Figure 1 reports a bar-graph of our results. We fitted logistic regression models predicting the *Yes* answers of participants with over 50% accuracy on both controls (N=138/167). Our factors were *Negation* (syntactic vs. lexical), *Category* (relative vs. absolute) and *Description* ("and" vs. "neither" vs. "ctl-true" vs. "ctl-false"). Random effect variables were included to reflect by-adjective and by-participant variation. For *relative* adjectives, "neither" and "ctl-true" descriptions did not significantly differ (regardless of negation, no interaction) whereas only *syntactic-"and"* descriptions significantly differed from "ctl-false." For *absolute* adjectives, no significant contrast was found between "neither" and "and" descriptions (regardless of negation, no interaction). All other simple effects were significant. Acceptance of *syntactic-"and"* descriptions was significantly greater for *relative* than for *absolute* adjectives; we found no significant interaction of *Category* × *"ctl-false" vs. "and"* in the *syntactic* groups.

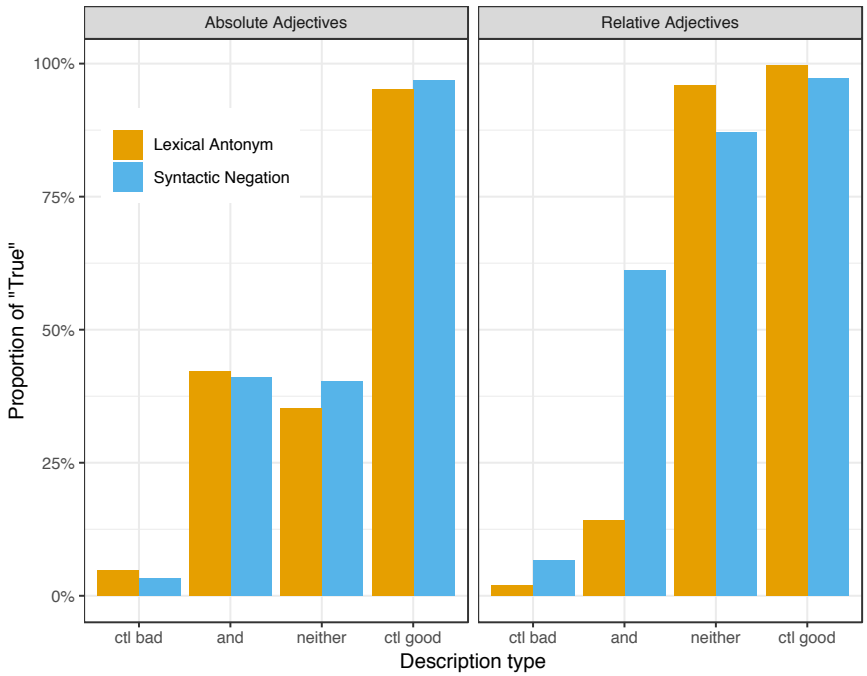


Figure 1.

3. Interpretation. Two main observations can be made about the data. Firstly, syntactic negations and lexical antonyms yield similar acceptance rates with absolute adjectives, but not so with relative adjectives (see the "and" contrast). Secondly, the preference for "neither" descriptions over "and" descriptions in borderline cases is only evidenced for relative adjectives.

To account for both effects, we adopt the strict-tolerant framework of [3], also applied in [6] and [8], in which both relative and absolute adjectives admit *strict* and *tolerant* readings, respectively narrowing and widening their extension, thereby creating gaps and gluts with their negative counterparts. In [6], the preference for "neither" descriptions in relative adjectives is explained by postulating a precedence of strict readings over tolerant readings (see [5]) and by assuming a "strictly" operator to be inserted above predicate negation (interpreting "neither tall or not tall" as "neither strictly tall, nor strictly not tall"). The account is at best incomplete, for it remains silent on lexical antonyms as well as absolute adjectives.

In light of the data, we postulate that (i) *for relative adjectives, lexical antonyms are semantic contraries, rather than contradictories*, leaving a gap between them; (ii) *for absolute adjectives, lexical antonyms are contradictories that leave no gap*, in much the same way as syntactic negations. In (i) we depart from [9]'s account, which treats every antonym $\bar{P}$ as the semantic complement of P. Under assumption (ii), the strict-tolerant account directly explains the symmetric acceptance of contradictory descriptions for absolute adjectives. To illustrate, if *dry* literally denotes a 0% amount of water, by (ii) *not dry* and *wet* denote any amount in the complement region ($> 0\%$) but an amount of 1% can still count as *dry* under a tolerant reading (creating a glut$\rightsquigarrow$"and") whereas the same 1% amount can fail to count as *not dry* or *wet* under a strict reading (creating a gap$\rightsquigarrow$"neither"). For *short* and *tall*, assumption (i) directly predicts the applicability of "neither short nor tall" in the gap region. On the other hand, tolerance may fail to fill the pre-existing gap so as to make *short* overlap with *tall*, thus explaining the massive rejection of "and" descriptions with relative antonyms.

We need one additional assumption, namely (iii) *syntactic negations of relative adjectives can be locally strengthened to their lexical antonym* (see [10]). By (iii), speakers may reinterpret "neither tall nor not tall" as "neither tall nor short." This explains the near-ceiling acceptance of *lexical-"neither"-relative* descriptions and, at the same time, does not make "and" descriptions more acceptable, thereby accounting for the lack of a significant interaction.

Overall, the present account is both more general and simpler than the one proposed in [6]: it assumes a local strengthening operation independent of the strict-tolerant machinery.

Examples.

1. 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? *Sam is neither tall nor short* []Yes []No *Sam is tall and short* []Yes []No
Sam is in the middle range []Yes []No *Sam's height is very high* []Yes []No

2. Sam is a blacksmith working in a traditional workshop where they produce swords. In the workshop, there are blades that have no bulges and there are blades that have many small bulges. Then there are blades with exactly one small bulge. Imagine that the blade that Sam is looking at has exactly one little bulge. Comparing the blade that Sam is looking at to the other blades, is it true to say the following?
The blade is neither flat nor bumpy []Yes []No *The blade is flat and bumpy* []Yes []No
The blade has exactly one bulge []Yes []No *The blade has many bulges* []Yes []No

References. **[1]** Alxatib, S. & Pelletier, F.J. 2011. The Psychology of Vagueness. *M&L* 26. **[2]** Serchuk, P. et al. 2011. Vagueness, logic and use. *M&L* 26(5). **[3]** Cobreros, P. et al. 2012. Tolerant, Classical, Strict *JPL* 41. **[4]** Ripley, D. 2011. Contradictions at the borders. *Vagueness in Communication*. **[5]** Cobreros, P. et al. 2015. Pragmatic Interpretations of Vague Expressions. *JPL* 44(4). **[6]** Egré, P. & Zehr, J. 2018. Are Gaps Preferred to Gluts? *The Semantics of Gradability, Vagueness, and Scale Structure*. **[7]** Kennedy, C. & McNally, L. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81 (2). **[8]** Burnett, H. 2014. A Delineation Solution to the Puzzles of Absolute Adjectives. *L&P* 37. **[9]** Krifka, M. 2007. Negated Antonyms. *Presupposition and Implicature in Compositional Semantics*. **[10]** Ruytenbeek, N. et al. 2017. Asymmetric inference towards the antonym. *Glossa* 2.