On Scales, Salience & Referential Language Use (A Revisit)*

Michael Franke

Institute for Logic, Language & Computation
Universiteit van Amsterdam

Abstract. Kennedy (2007) explains differences in the contextual variability of gradable adjectives in terms of *salience* of minimal or maximal degree values on the scales that these terms are associated with in formal semantics. In contrast, this paper suggests that the attested contextual variability is a consequence of a more general tendency to use gradable terms to preferentially pick out *extreme-valued properties*. This tendency, in turn, can be explained by demonstrating that it is pragmatically beneficial to use those gradable properties in referential descriptions that are *perceptually salient* in a given context.

1 Scale Types, & "Kennedy's Observation"

A prominent line of current research in formal semantics links the meaning of gradable adjectives to *degrees on scales* (cf. Rotstein and Winter, 2004; Kennedy and McNally, 2005). In simplified terms, the denotation of a gradable adjective A is taken to be a function $g_A : \mathrm{Dom(A)} \rightarrow D$ that maps any applicable argument of A to a degree $d \in D$, where $\langle D, \preceq \rangle$ is a suitably ordered *scale of degrees*. What is particularly interesting about this approach is that different adjectives may be associated with different kinds of scales. Standardly, one-dimensional scales are assumed and a distinction is made as to whether these are: (i) totally open (*tall*, *short*), (ii) totally closed (*closed*, *open*), or (iii) half-open (*bent*, *pure*). Scale types explain a number of otherwise puzzling observations, such as which adjectives can combine with which modifiers. E.g., the expression *completely A* is felicitous only if A has a totally or upper-closed scale with a maximal element: compare the felicitous *completely closed* with the awkward ?*completely tall*.

Scale types also influence the licensing conditions of utterances involving gradable adjectives in positive form. Generally speaking, a simple positive sentence like "object x has property A" is considered true whenever the contextually supplied minimal degree of A-ness, $c(A)$, is no higher than $g_A(x)$. However, the contextual standard of applicability $c(A)$ is also affected by the scale type (c.f. Kennedy, 2007): if there is a $\preceq$-maximal or -minimal degree contained in $\langle D, \preceq \rangle$,

* This paper reconsiders and complements an earlier paper on the same subject (Franke, 2011). I would like to thank the remaining *G-Team* (Gerhard Jäger, Roland Mühlenbernd, Jason Quinley), Joey Frazee and Graham Katz, as well as the anonymous reviewers for comments and discussion.

then $c(A)$ is bound to this; otherwise it is to be retrieved more flexibly from the context of utterance. In more tangible terms, "*Kennedy's observation*" (1) says that closed-scale adjectives are used rather inflexibly to denote a minimal or maximal value on the associated scale (modulo the usual pragmatic slack where lack of precision is conversationally harmless), whereas open-scale adjectives allow for more contextual variability.

(1) **"Kennedy's Observation":**

scale type		contextual standard of applicability
open	$\longleftrightarrow$	variable
closed	$\longleftrightarrow$	rigid, fixed to endpoints

For example, the contextual standard for the applicability of open-scale *tall* can vary considerably from one context (talking about jockeys) to another (talking about basketbal players), whereas that of closed-scale *closed* seems glued to the denotation of a minimal (zero) degree of openness.

2 Explaining "Kennedy's Observation"

Salience of Endpoints. Kennedy (2007) tries to explain the influence of scale topology on contextual usage conditions in terms of the *salience* of endpoints (2a) and a pragmatic principle called *Interpretive Economy* (2b) which demands that pragmatic interpretation ought to make maximal use of the available semantic resources.

(2) **"Kennedy's Explanation":**

a. **Salience Assumption:**
End-points of closed scales are salient elements provided by the conventional semantic structure.

b. **Interpretive Economy:** (Kennedy, 2007, (66), p.36)
Maximize the contribution of the conventional meanings of the elements of a sentence to the computation of its truth conditions.

The idea seems quite natural: the evaluation of expression "x is A" requires us to fix a contextual standard $c(A)$; if A is associated with a closed scale, then by (2a) the semantic structure supplies some outstanding element, which by (2b) ought to be used to set $c(A)$; if A is associated with an open scale, the semantic structure carries no such salient points and $c(A)$ can be set more variably.

We should be fully satisfied with neither (2a) nor (2b). Firstly, as for (2a), it is not clear *a priori* whether endpoints on closed scales not only appear salient to us because they are the preferred denotation of the corresponding natural language expressions. In that case, the attempted explanation would be circular. The crucial problem is that it is very hard to determine, conceptually or empirically, when an element *of an abstract semantic structure* is salient. Phrased more constructively, if salience is to play a role in an explanation of (1) it should

better be an empirically informed notion of *perceptual salience*, i.e., of salience of objects of perception that stand out relative to others and attract attention more than others do. Secondly, we should not stop at the formulation of a pragmatic principle like (2b) even if it seems plausible and yields the desired results, but continue to ask for a *functional motivation*: what is the added pragmatic value of the principle in question that enabled its evolution and sustenance?

Evolution of Pragmatic Standards. Potts (2008) addresses the latter issue. While adopting (2a), he seeks to explain (1), not via (2b), but instead by an evolutionary argument why speakers and hearers conventionally coordinate on endpoints as the contextual standard for the use of closed-scale adjectives. Towards this end, Potts consider a strategic game in which speaker and hearer simultaneously choose a standard of application for a closed-scale adjective. Payoffs are proportional to how close the players' choices are to each other, so that the maximal payoff ensues when players choose the same standard of application. Potts then shows that if a population initially has a slight bias towards choosing the endpoints (his way of implementing focality of endpoints), then the replicator dynamics (Taylor and Jonker, 1978) will eventually lead to all of the speakers and hearers of the whole population choosing endpoints as standards of application.

Potts' account has some shortcomings, unfortunately. For one, it fails to make clear what the particular pragmatic benefit of endpoint use is: it is just a consequence of the assumption that the to-be-explained outcome is already predominant in the population initially. What is more, Potts' account is either silent about or makes wrong predictions for adjectives with open scales. If we assume that the only thing that differentiates open-scale and closed-scale adjectives is the presence or absence of endpoints, then, looking at open-scale adjectives in the same way, we would simply drop the assumption that there is an initial bias in the population for a particular standard of comparison. But in that case the replicator dynamics will still eventually gravitate towards a *single fixed* standard of application, albeit not a focal endpoint, contrary to (1).

Extreme-Value Principle. In fact, adjectives with open scales, though more variable in their contextual standard of application, are not entirely unconstrained either. Take the open-scale adjective *tall*, for instance, and the question when an individual x is called *tall* when compared with a group of individuals Y. Although the precise rule of application is a question of current empirical research (e.g. Schmidt et al., 2009), it seems fair to say that x is more likely or more readily counted as tall, the more x's tallness falls within the *extreme* values of tallness within group Y. Usually it is not enough for x to be just slightly taller than the average tallness in Y. Rather, the further away x's tallness is from the average or expected value of Y, the more readily it counts as *tall*.

If this is true, or at least not too far off-the-mark, then a different explanation suggests itself for (1). If there was a tendency for gradable expressions to be used preferably to describe extreme values, then it is to be expected that closed-scale terms will mostly be used for values close to the endpoints, while open-scale terms could be used for a wider range of values simply due to the open-endedness of

the scale. In other words, I suggest that what we need not explain (1), but rather the principle in (3).

(3) **Extreme-Value Principle:**
Gradable terms are preferably/usually used to describe extreme values, i.e., values far away from the median/mean of a given distribution.

The remainder of this paper is therefore concerned with two things: (i) a proof of concept that (3) indeed leads to a general association along the lines of (1), and (ii) an attempt of explaining (3) as a concomitant of pragmatic language use. Peaking ahead, the pragmatic rationalization for (3) that I will offer eventually is indeed superficially similar to Kennedy's explanation of (1) in (2), but conceptually different. My explanation of (3) involves a notion of salience of stimuli in context (4a), paired with an account of why the use of salience is actually beneficial in conversation (4b).

(4) a. **Salience of the Extreme:**
Salience of a stimulus in a given context is proportional to its (apparent/subjectively felt) *extremeness* or *outlieriness*, i.e., to the extent that the stimulus appears unexpected or surprising against the background of the other stimuli in the context.

b. **Benefit of the Extreme:**
Describing those properties of objects that are salient is pragmatically advantageous for coordinating reference.

The main intuition that inspires (4) is this: terms are associated with extreme values because we use them, among other things, to identify referents, and for doing so the use of extreme values is a very natural and easy, yet surprisingly effective solution.[1] In order to test this intuition, I propose a simple model of referential language use, to be introduced next.

3 Referential Games

A *referential game* is a game between a sender and a receiver, both of whom observe a context c that consists of $n > 0$ objects. One of these objects is the *designated object* c_o that the sender wants to refer to. The receiver does not know which object that is. The goal is to describe the designated object by naming a property of c_o. For simplicity we assume that senders can choose only one property to describe c_o with, but can indicate whether c_o has a high or low degree of that property. If, after hearing the description, the receiver guesses the right referent, the game is a success for both players; if not, it's a failure.

[1] Elsewhere I tried to show that the use of extreme values would actually be detrimental if language was exclusively used to *describe* the actual degree of a given object as closely as possible (Franke, 2011). I focus here on the model of referential language use that was also discussed in that earlier work.

More formally, let us assume that objects are represented as points in an m-dimensional *feature space* $\mathcal{F} \subseteq \mathbb{R}^m$, $m > 0$. Each dimension of $\mathcal{F}$ corresponds to some gradable property: the value of dimension j is the degree to which the object in question has property j. A context is thus a set of n points in $\mathcal{F}$, which can easily be represented as an $n \times m$-matrix c. For example, the context in (5) contains three "objects", namely Hans, Piet and Paul, which are represented as a triple of features, namely their degrees of tallness, weight, and intelligence.

(5)

	tallness	weight	intelligence
Hans	0.2	-0.1	1.3
Piet	-0.1	0.0	0.3
Paul	0.3	-0.2	0.5

To make a distinction between open and closed scales, it is reasonable to assume that open-scale features take values in $\mathbb{R}$, while closed-scale features take values on some closed interval of reals. But open and closed scales should also plausibly differ with respect to the probability that a particular degree is observed. To keep matters simple, assume that a *random context* is obtained by sampling independently n random objects, and that a random object is obtained by sampling independently m random degree values for the relevant properties. Finally, let us assume, rather naïvely, that degrees are sampled from the distributions in (6) (see also Figure 2).

(6)

scale type	distribution
open	normal distribution (mean 0, standard deviation $1/3$)
totally closed	uniform distribution on $[0; 1]$
half-open	truncated normal distribution on $\mathbb{R}^{\geq 0}$ (mean 0.1, standard deviation $1/3$)

Together this yields a unique probability density $\Pr(c)$ for each context c (the exact nature of which will, however, not be of any relevance here). For each round of playing a referential game, a context is sampled with $\Pr(c)$ and from that context an object is selected uniformly at random as the designated one.

Finally, let the set of messages from which the sender can choose contain exactly one pair of antonymous terms for each property of the feature space. So, the set of messages is $M = \{1, \ldots, m\} \times \{\text{low}, \text{high}\}$, where, e.g., $m = \langle j, l \rangle \in M$ has a conventional meaning saying that property j is low. For example, if Hans is the designated object in context (5) above, the sender could describe him as being short or tall, skinny or fat, stupid or smart.

4 Solving Referential Games

Intuitively, I would describe Hans as *the smart guy* in the example above. (What about you?) This is because of his comparatively high value along that dimension, and his median values for the respective others. Does this intuition follow

from an assessment of what is an *optimal* way of playing a referential game? – Unfortunately it does not, which is why I suggest to instead look at a *natural* way of playing the game, namely by exploiting salience. But first: optimal play.

Optimal Solutions. Player behavior is captured in the notion of a *(pure) strategy*, as usual. A sender strategy is a function $\sigma : C \times \{1, \ldots, n\} \rightarrow M$ mapping a context and a designated object onto a message. A receiver strategy is a function: $\rho : C \times M \rightarrow \{1, \ldots, n\}$, mapping each context and each message onto an object. Given a context c with designated object o, the *utility* of playing with a sender strategy σ and receiver strategy ρ is simply:

$$\mathrm{U}(\sigma, \rho, c, o) = \begin{cases} 1 & \text{if } \rho(c, \sigma(c, o)) = o \\ 0 & \text{otherwise.} \end{cases}$$

The *expected utility* of σ and ρ is then just the averaged utility over all contexts and designated objects, weighted by the probability of occurrence:

$$\mathrm{EU}(\sigma, \rho) = \int \Pr(c) \times \mathrm{EU}(\sigma, \rho, c)\, \mathrm{d}c\,, \text{ where}$$

$$\mathrm{EU}(\sigma, \rho, c) = \sum_{i=1}^{n} \frac{1}{n} \times \mathrm{U}(\sigma, \rho, c, i)\,.$$

A usual, we say that $\langle \sigma, \rho \rangle$ is a *Nash equilibrium* iff (i) there is no σ' such that $\mathrm{EU}(\sigma, \rho) < \mathrm{EU}(\sigma', \rho)$ and (ii) there is no ρ' such that $\mathrm{EU}(\sigma, \rho) < \mathrm{EU}(\sigma, \rho')$. Call $\langle \sigma, \rho \rangle$ an *optimal solution* iff $\mathrm{EU}(\sigma, \rho, c) = \min(1, \frac{2m}{n})$ for all contexts c.

To understand this latter notion, let's take a step back and reflect on referential games. Interestingly, referential games can be considered an infinite collection of games G_c one for each context c. These games are in fact standard Lewisean signaling games (Lewis, 1969): for fixed c, G_c has a set of states (here: objects) that are drawn from a uniform distribution; it also has a set of messages; finally, the receiver tries to guess the actual state that only the sender knows. The maximum possible payoff attainable in each G_c is $\min(1, \frac{2m}{n})$. This is because there are $2m$ messages to encode n states. If $n \leq 2m$, perfect communication is possible; otherwise only $2m$ of the n states can be named successfully. Consequently, an optimal solution for a referential game is one that scores optimally in all G_c.

Since there is a pair of strategies that reaches the theoretically maximal communicative success in each G_c, it follows that an optimal solution for each referential game exists. Since there is always more than one optimal solution for any local game (we have always at least two messages, even if we have only one object), there are in fact infinitely many optimal solutions. Moreover, since an optimal solution is a pair of strategies that achieves maximal utility in *every* context, an optimal solution is in fact a Pareto-optimal Nash equilibrium. In other words, optimal solutions are the theoretically conceivable maximum, they exist and even abound.

Unfortunately, optimal solutions might be quite bizarre. Consider a simple, but non-trivial referential game with $n = 3$ objects and $m = 2$ open-scale properties. For concreteness, let us look at the infinite set of contexts in (7).

(7)

object	tallness	weight
Hans	p	0.4
Piet	$p-1$	0.3
Paul	1	4.5

where $p \in \mathbb{N}^{\geq 3}$

One of the infinitely many optimal solutions to this game has the sender do the following: if p is not a prime number, then Hans is referred to as *the tall guy*, Pieter *the skinny guy*, and Paul *the fat guy*; if p is prime, then Paul is still identified as *fat*, but now Hans is called the *the skinny guy*, and Pieter *the tall guy*. We may happily assume that the receiver, since he knows the context as well, perfectly identifies the designated object in each case. Of course, it is a ludicrous idea to assume that humans can condition their language use on whether a numerical degree representation is prime or not. But that is not the point. The point is that optimal solutions for referential games are allowed to vary arbitrarily from one context to the other. In other words, it is not enough to know that there are optimal solutions (even if there are infinitely many), we would also need to ask whether any of these is human graspable and learnable in a reasonable way from finite observations.

A Natural Solution: Salience. Much could be said here about cross-contextual learning and rule-induction in an evolutionary game, but I do not want to go there. Instead, I would like to go straight forward and show that there is a fairly natural way of playing referential games that (i) presupposes hardly any rationality on the side of the agents, that (ii) arguably requires no learning at all, as it merely exploits the agents' cognitive make-up, but that still (iii) is sufficiently successful. The strategy I have in mind is one in which sender and receiver simply choose whatever is most *salient* from their own perspective: the sender chooses the most salient property of the designated object; the receiver chooses the most salient object with that property. Neither player thus reasons strategically about what the other player does. Players merely exploit a shared cognitive bias of perception. Still, statistically this rather myopic choice rule does fairly well and also leads to the selection of extreme values. Both of these claims will be backed up below by numerical simulations.

But let me first elaborate on the notion of salience that I would like to use, which is a notion of contextualized perceptual salience inspired by recent research on visual salience in terms of *informativity* or *surprise* (e.g. Rosenholtz, 1999; Itti and Koch, 2001; Bruce and Tsotsos, 2009; Itti and Baldi, 2009). The general idea is that, when presented with a scene, those things stand out that are *unexpected*. This may be due to sophisticated world-knowledge, but may also be due to much less sophisticated expectations raised by the immediate contextual environment. In the spirit of the latter, I suggest that how salient object i's having property j to degree c_{ij} is, is a measure of how *unexpected*, c_{ij} appears against the background of the set c_j^{-1} of degrees for property j that occur in c. For example, given a context c, the set of degrees for property j are a vector of numbers c_j^{-1}, one for each object. Such a vector could be a tuple

like $\langle 2, 3, 1, 1, 2, 1, 37\rangle$ that could, for instance, represent abstractly the tallness of my 7 sons. Most values here lie around 1 or 2, so that the one value of 37 looks suspiciously like an *outlier*. I suggest that for our purposes here we may assume that the more a degree looks like an outlier in context, the more it is perceptually salient (4a).

Indeed much work in statistics has been devoted to the issue of outlier detection (c.f. Ben-Gal, 2005, for overview). For simplicity, I explore here only one very manageable approach to outlier detection in terms of the (linear) distance between points in the feature space (c.f. Knorr and Ng, 1998; Ramaswamy et al., 2000). So define the salience of object i having degree c_{ij} for property j as:[2]

$$\text{Sal}_{\text{lin}}(c_{ij}, c) = \sum_{i'} |c_{i'j} - c_{ij}| \,.$$

Let s be the *saliency matrix* for context c, given by $s_{ij} = \text{Sal}_{\text{lin}}(c_{ij}, c)$. The *salience-based choice rules* for sender and receiver are then simply this:

(8) **Salience-based choice rules:**

a. **Sender:** if o is the designated object, select property j^* uniformly at random from $\arg\max_j s_{oj}$; if $c_{oj^*} \geq \text{median}(c_{j^*}^{-1})$, send message $\langle j^*, h\rangle$, otherwise send $\langle j^*, l\rangle$;

b. **Receiver:** if $\langle j^*, h\rangle$ is the received message, select uniformly at random from $\arg\max_i \left\{ s_{ij^*} \mid c_{ij^*} \geq \text{median}(c_{j^*}^{-1}) \right\}$; if $\langle j^*, h\rangle$ is received, the same applies, except with $<$ in the set restriction.

The choice rule in (8) is rather successful (see Figure 1), despite the fact that players blindly maximize salience from their own perspective, without taking each other's strategy into account. Moreover, choices according to (8a) also corroborate (3). The values c_{oj^*} selected by (8a) are indeed extreme (Figure 2).

In conclusion, these simulation results give a proof-of-concept that it is possible to think of (3), and hence (1), as a concomitant of a pragmatically efficient language use. Crucial for this proposal is, of course, that the choice rules in (8) make use of a sense of salience that is *shared* by sender and receiver. Also it is important for an explanation of (3) that salience of a stimulus is, even if indirectly, linked to its extremity. But although both of these assumptions are, to my mind, defensible, the proposal has obvious shortcomings too. For example, it seems to predict that all pairs of antonymous gradable adjectives are *non-complementary* and that even closed-scale adjectives allow for some contextual variation. To address these further issues it would be necessary to consider a more encompassing model of language use that not only considers referential descriptions in a shared and perfectly accessible immediate perceptual context.

[2] I stick to this notion here for continuity with earlier work (Franke, 2011), but I have also tested different, more standard notions of salience, with essentially the same results as reported here, such as in terms of (linear) distance from the median of c_j^{-1}, or in terms of multiples of its interquartile range.

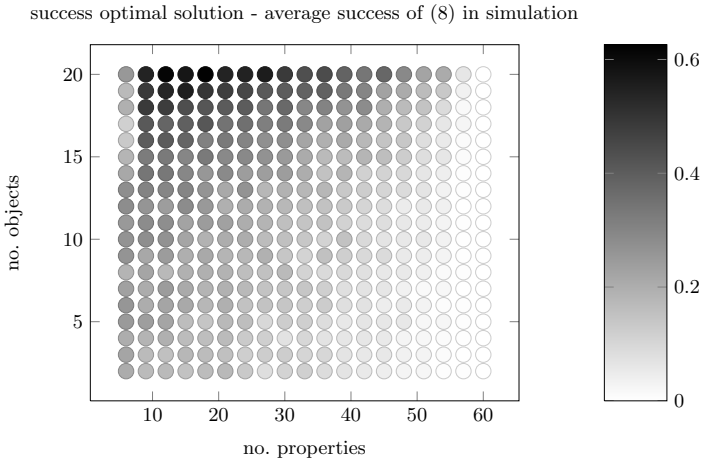

Fig. 1. Assessment of the success of choice rule (8). Each dot corresponds to a pair n, m of context size and number of properties (with $m/3$ properties for open, closed, and half-open scales each). For each pair, (8) was applied to 500 random $n \times m$-sized contexts. The proportion of successful rounds was then subtracted from the theoretical maximum for the given n and m. (8) often reached more than 80% of the theoretical optimum. With sufficient expressivity, i.e., large m, it even matches peak performance.

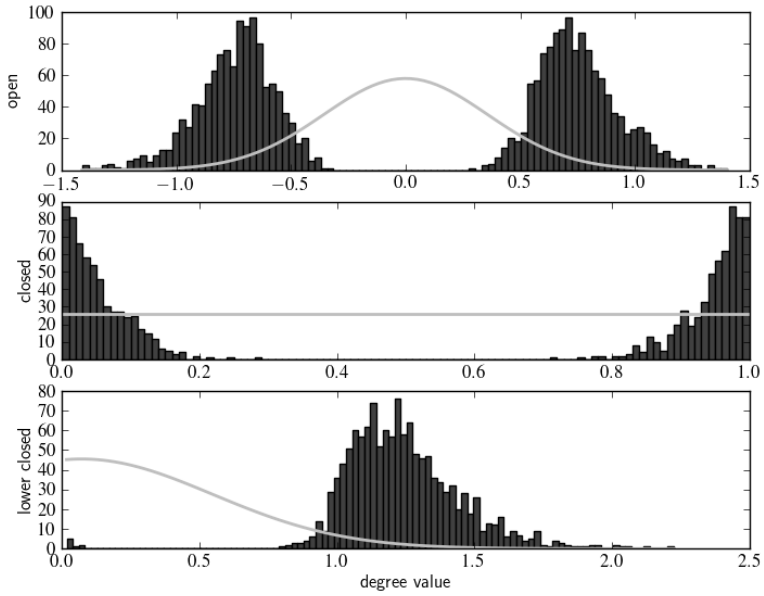

Fig. 2. Frequency with which degrees c_{oj*} were choosen by the choice rule in (8a) in 5000 randomly sampled contexts with $n = 30$ and $m = 24$ (8 properties each for open, closed, and half-open scales with prior distributions indicated by the light gray lines).

Bibliography

Ben-Gal, I. (2005). Outlier detection. In Maimon, O. and Rockach, L., editors, *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, pages 131–148. Kluwer.

Bruce, N. D. B. and Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):1–24.

Franke, M. (2011). Scales, salience and referential safety: The benefit of communicating the extreme. to appear in *Proceedings of EvoLang IX*.

Itti, L. and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 1:194–203.

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

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

Knorr, E. M. and Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th VLDB Conference*, pages 392–403.

Lewis, D. (1969). *Convention. A Philosophical Study*. Harvard University Press.

Potts, C. (2008). Interpretive Economy, Schelling Points, and evolutionary stability. Manuscript, UMass Amherst.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, 39:3157–3163.

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

Schmidt, L. A., Goodman, N. D., Barner, D., and Tenenbaum, J. B. (2009). How tall is *Tall*? compositionality, statistics, and gradable adjectives. In *roceedings of the Thirty-First Annual Conference of the Cognitive Science Society*.

Taylor, P. D. and Jonker, L. B. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Bioscience*, 40(1–2):145–156.