On Scales, Salience & Referential Language Use*

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.

Keywords: gradable adjectives, scale topology, salience, game-theory

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 $g_A : \text{Dom(A)} \rightarrow D$ 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.* Different adjectives may be associated with different kinds of scales. Usually 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 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 of mine on the same subject (Franke, 2012). I would like to thank the remaining *G-Team* (Gerhard Jäger, Roland Mühlenbernd, Jason Quinley), as well as Oliver Bott, Peter van Emde-Boas, Joey Frazee, Irene Heim, Graham Katz, Manfred Krifka, Louise McNally, Rick Nouwen and 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 compare rather inflexibly to endpoints of the associated scale (modulo the usual pragmatic slack where imprecision is conversationally harmless), while open-scale terms show 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 pragmatic interpretation 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 measuring how much an object stands out or attracts attention relative to others. 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 considers 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 some *arbitrary fixed* standard of application. But that does not seem quite right, as discussed next.

Extreme-Value Principle. Adjectives with open scales, though more contextually variable, are not entirely unconstrained. Take the open-scale adjective *tall*, say, 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 x's tallness is from the average or expected value of Y, the more it counts as *tall*.

These considerations suggest a different explanation for (1). If there is a tendency for gradable expressions to be used preferably to describe *extreme* values, then closed-scale terms will usually 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 to explain (3), not (1).

(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. The pragmatic rationalization for (3) that I will offer is 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 $c = \langle o_1, \ldots, o_n \rangle$ that consists of $n > 0$ objects. One of these objects is the *designated object* o_d 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 o_d. For simplicity we assume that senders can choose only one property to describe o_d with, but can indicate whether o_d 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.

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.

[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, 2012). I focus here on the model of referential language use that was also discussed in that earlier work.

(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 type
open	normal ($\mu = 0$, $\sigma = 1/3$)
totally closed	uniform on $[0; 1]$
half-open	truncated normal on $\mathbb{R}^{\geq 0}$ ($\mu = 0.1$, $\sigma = 1/3$)

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. As usual in game theory, the subsequent question of interest is: does this intuition follow from an assessment of what is an *optimal* way of playing a referential game? – Unfortunately, as we will see presently, although there are infinitely many optimal ways of playing referential games, optimality comes at the price of plausibility. Therefore we should rather ask whether there is a *natural* solution to referential games that corroborates our intuitions. This section looks at one arguably natural solution to referential games, namely a strategy that exploits *perceptual salience*. I show that under this strategy extreme-value use, as in (3), emerges, and that communicative success is often not much worse than the theoretical optimum.

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

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

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:

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

As 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 $\mathrm{EU}(\sigma, \rho) < \mathrm{EU}(\sigma, \rho')$.

Referential games can be considered an infinite collection of games G_c one for each context c that are 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, call $\langle \sigma, \rho \rangle$ an *optimal solution* iff it scores perfectly in all games, i.e., iff $\mathrm{EU}(\sigma, \rho, c) = \min(1, \frac{2m}{n})$ for all contexts c. It is straightforward to show that each referential game has infinitely many optimal solutions each of which is a (Pareto-optimal) Nash equilibrium.

Optimal solutions are the theoretically conceivable maximum, they exist and even abound. Unfortunately, optimal solutions might be quite implausible. To see this, consider the context in (7) with Hans as designated object.

(7)

	tallness	weight
Hans	0.1	-0.1
Piet	-1.0	-1.0
Paul	1.0	1.0

Intuitively, there is no description that plausibly refers uniquely to Hans. Consider the alternatives: *the short guy* and *the skinny guy* most plausibly refer to Piet, and *the tall guy* and *the big guy* most plausibly refer to Paul. The problem is that poor average Hans doesn't stand out at all. However, by convention, telepathy or magic, an optimal solution would, for instance, yield perfect referential success by describing Hans as *tall* and using only *big* for Paul. This is not exclusive to this specific example and it is not the only reason why optimal solutions might be quite implausible. (Notice that referential games do not require agents to use messages "reasonably" in line with their "semantic meaning". Whence that optimal solutions need not stick with "reasonable semantic meaning" either.) The lesson to learn from this is that, when it comes to referential games, it is not necessarily helpful to look at optimal communication. Even if we know that there exist optimal solutions (and even if there are infinitely many), we should rather be interested in more *natural* strategies, but use optimal strategies as a yardstick to measure communicative success.

A Natural Solution: Salience. Much could be speculated about what a natural strategy is for referential games. But instead I want to look at just one strategy that strikes me as plausible and appealing because (i) it presupposes hardly any rationality on the side of the agents, as it merely exploits the agents' cognitive make-up, but still (ii) it is remarkably successful. The strategy in focus is one in which players 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 given 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 here to explore the idea 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:

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

We are then interested in the *saliency matrix* s for context c, given by $s_{ij} = \text{Sal}_{\text{lin}}(c_{ij}, c)$. For example, the salience matrix for example context (5) is:

(8)

	tallness	weight	intelligence
Hans	0.4	0.2	1.8
Piet	0.7	0.3	1.2
Paul	0.5	0.5	1.0

The *salience-based choice rules* for sender and receiver are then defined as follows. The sender picks the most salient property j^* for the designated object o and chooses the corresponding message m^* indicating whether the property j^* of o_d is high or low (in the given context set $c_{j^*}^{-1}$). For example, if Hans is the designated object in (5), the sender looks at the first row of (8) and selects the column (i.e., property) with the highest number, leading to expression choice *the smart guy*. On hearing m^*, the receiver picks the object with the highest salience value for j^* from all those objects whose value for j^* is as indicated (high or low). In our example, when the receiver hears *the smart guy*, he consults the column for *intelligence* in (5) and (8), drops all rows whose values in (5) are below average and selects the row (i.e., referent) with the maximal salience value among the remaining, which is indeed the correct referent in this example. Whenever objects are exactly equally salient, players randomize.[2]

This procedure leads to communicative failure for the context (7) when the designated object is Hans, but is rather successful in general (see Figure 1), despite the fact that players blindly maximize salience from their own perspective, without taking each other's strategy into account. As shown in Figure 1, salience-based choices easily reach 90% of the theoretically possible amount of referential success unless the number of objects in context far exceeds the number of properties which to describe objects with.

Moreover, salience-driven choice of referential expressions also corroborate (3). The values c_{oj^*} selected by the sender's salience-based choice rule are indeed extreme (Figure 2). These results give a proof-of-concept that it is possible to think of (3) as a concomitant of pragmatically efficient language use. In other words, if we look at the actual objects that are described as *tall*, *smart* etc., we see that their respective degrees of tallness, intelligence etc. are indeed at the outer ends, so to speak, of their respective scales. The results in Figure 2 thus also make plausible that salience-driven referential communication leads to a difference between closed- and open-scale terms, in line with Kennedy's observation (1): closed-scale terms are used for a rather narrow range of values close to the end point of the scale; open scale terms are used more variably to describe a broader, in fact, open-ended range of values.

[2] To be precise, salience-based choices were implemented as follows. If o is the designated object, the sender selects property j^* uniformly at random from $\arg\max_j s_{oj}$. If $c_{oj^*} \geq \text{median}(c_{j^*}^{-1})$, she sends message $\langle j^*, h \rangle$, otherwise $\langle j^*, l \rangle$. If $\langle j^*, h \rangle$ is the observed message, the receiver selects 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 previous set restriction.

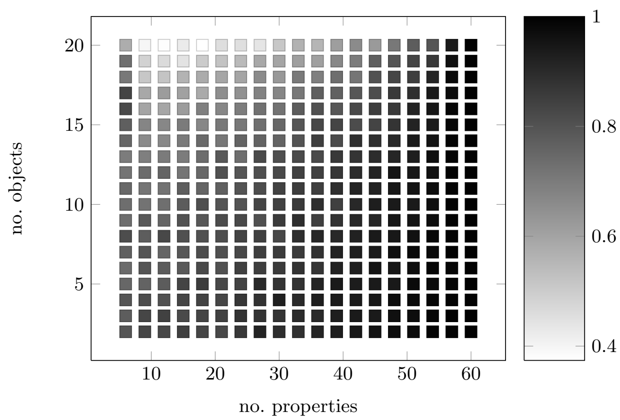

Fig. 1. Success of the salience-based choice rules relative to optimal solutions. Each square corresponds to a pair $\langle n, m \rangle$ of context size n and number of properties m (with $m/3$ properties for open, closed, and half-open scales each). For each pair, the salience-based choice rule was applied to 500 random $n \times m$-sized contexts. The color encodes the proportion of observed success rate divided by theoretical optimum for n and m.

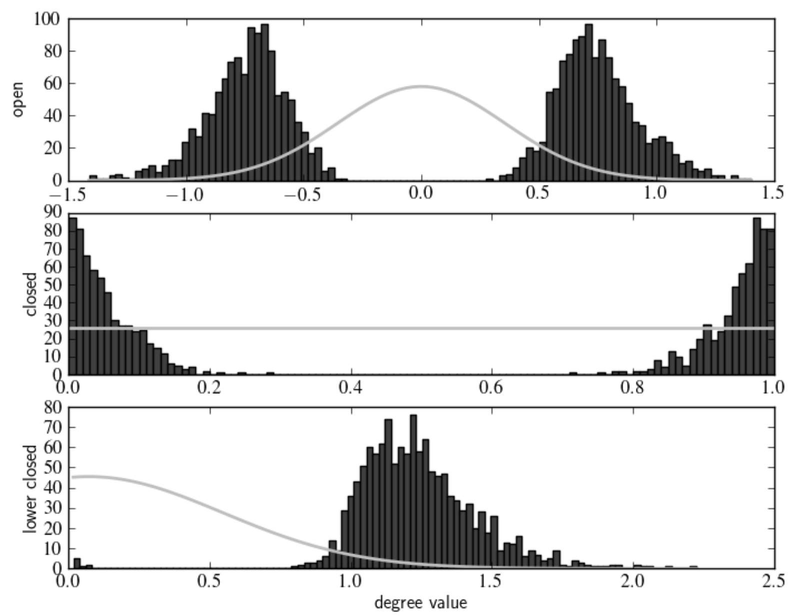

Fig. 2. Frequency with which degrees c_{oj^*} were choosen by the sender's salience-based choice rule in 5000 randomly sampled contexts with $n = 30$ and $m = 24$ (8 properties each for open, closed, and half-open scales with prior distributions indicated by the light gray lines).

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. (2012). Scales, salience and referential safety: The benefit of communicating the extreme. *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.