Extremes are typical

a game theoretical derivation

van Rooij, R.; Brochhagen, T.

DOI
10.1007/978-3-030-50200-3_16

Publication date
2021

Document Version
Final published version

Published in
Concepts, Frames and Cascades in Semantics, Cognition and Ontology

Extremes are Typical. A Game Theoretical Derivation

Robert van Rooij and Thomas Brochhagen

Abstract In this paper we argue that a typical member of a class, or category, is an extreme, rather than a central, member of this category. Making use of a formal notion of representativeness, we can say that a typical member of a category is a stereotype of this category. In the second part of the paper we show that this account of typicality can be given a rational motivation by providing a game-theoretical derivation.

Keywords Typicality · Representativeness · Extreme · Game theory

1 Typicality: Prototypes Versus Stereotypes

In *cognitive* psychology, a typical representative of class X is normally called its *prototype*. At least since the work of Rosch (1973) in psychology, a prototype of a category is standardly seen as an item that is most similar to all other members of the category: a *central member* of the category. It is standardly assumed that category membership is a graded affair, and that goodness-of-exemplar judgments depend on similarity to the prototype.

But is a typical member of a category really a central member of this category? A simple Google search seems to question this view. The man that comes up very prominently when one does a simple Google search of a typical, or real man is Rambo. Whatever one can say of Rambo, he is *not an average man*. Very similar pictures of real *tall* men, and real *scientists* give rise to similar conclusions.

Our Google search should obviously not be taken too seriously, but it is in line with many experimental findings in cognitive psychology of what we think of typical examplars. First, Hampton (1981) found that at least for abstract categories, central

R. van Rooij (✉) · T. Brochhagen
Institute for Logic, Language and Computation University of Amsterdam, Amsterdam, Netherlands
e-mail: r.a.m.vanrooij@uva.nl

T. Brochhagen
e-mail: thomasbrochhagen@gmail.com

S. Löbner et al. (eds.), *Concepts, Frames and Cascades in Semantics, Cognition and Ontology*, Language, Cognition, and Mind 7,
https://doi.org/10.1007/978-3-030-50200-3_16

tendencies are not a good predictor of goodness-of-exemplar judgments. Second, Barsalou (1985) showed that *ideals*, rather than central exemplars, are better determinants of category goodness in goal-based categories such as 'foods to eat on a diet' (food with zero calories) and 'ways to hide from the Mafia'. Lynch et al. (2000), Palmeri and Nosofsky (2001) and Burnett et al. (2005) found that ideals, or psychological extreme points, may define category goodness even in natural categories.[1] These studies also show that sometimes categorization can be based on ideals, and that people judge the ideal rather than the average members as the typical ones. Perhaps more interesting for this paper is the finding that when categories were learned in relation to alternative contrast categories, extreme members were counted as typical (cf. Ameels and Storms (2006)), and people were best able to categorize based on such ideals (cf. Goldstone et al. (2003)). This all suggests that if we want to model what it means to be a 'real', or typical, X, one should not just pick an average exemplar of type X.

If Rambo is not a prototypical man, he is certainly a *stereotypical* one. The Oxford English Dictionary defines a stereotype as a 'widely held but fixed and oversimplified image or idea of a particular type of person or thing'. The so-called 'social cognition approach' to stereotypes (e.g. Schneider et al. (1979)), rooted in social psychology, views a social stereotype as a special case of a cognitive schema. Such schemas are intuitive generalizations that individuals routinely use in their everyday life, and entail savings on cognitive resources. Hilton and von Hippel (1996) define stereotypes as 'mental representations of real differences between groups [. . .] allowing easier and more efficient processing of information. Stereotypes are selective, however, in that they are localized around group features that are the most distinctive, that provide the greatest differentiation between groups, and that show the least within-group variation.' Thus, according to Hilton and von Hippel (1996), stereotypes are rather extreme representatives of a class.

Within social psychology, McCauley et al. (1980) have defined the following measure of how stereotypical x is for class X: $\frac{P(x|X)}{P(x)}$. An easy proof shows that this measure behaves monotone increasingly with respect to $\log \frac{P(x|X)}{P(x|\neg X)}$,[2] meaning that the x with the highest value for the former notion also has the highest value for the latter notion. The latter notion goes back to Turing, and has been called the *weigh of evidence* of x for X by Good (1950). The same notion has been called the

[1] Of course, Plato already thought of universals as represented by *ideals* (the Forms).

[2] To show this, note first that $P(x|X) - P(x)$ behaves monotone increasingly with $P(x|X) - P(x|\neg X)$.

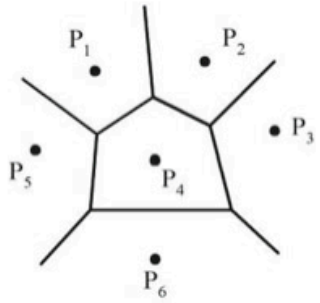
$$\begin{aligned} P(x|X) - P(x) &= P(x|X) - [(P(x|X) \times P(X)) + (P(x|\neg X) \times P(\neg X))] \\ &= P(x|X) - [\alpha P(x|X) + (1-\alpha)P(x|\neg X)], \text{ with } 0 \leq \alpha \leq 1 \\ &= (1-\alpha)P(x|X) - (1-\alpha)P(x|\neg X) \\ &= \beta[P(x|X) - P(x|\neg X)], \text{ with } 0 \leq \beta \leq 1. \end{aligned}$$

Obviously, $\frac{P(x|X)}{P(x)}$ behaves monotone increasingly with $P(x|X) - P(x)$, just as $\frac{P(x|X)}{P(x|\neg X)}$ behaves monotone increasingly with $P(x|X) - P(x|\neg X)$. Given the nature of logarithmic functions, the latter, in turn, behaves monotone with $\log \frac{P(x|X)}{P(x|\neg X)}$.

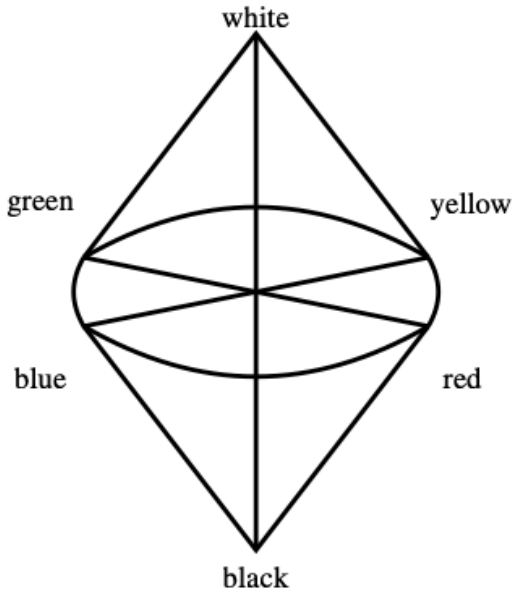
representativeness of x for X by Tenenbaum and Griffiths (2001). Adding things up, it all suggests that typical, or representative, members of their classes, are, in fact, their stereotypes, members that provide the greatest differentiation between the classes.

2 Typicality and Structured Meaning Spaces

Gärdenfors (2000) proposes that primitive categories (or natural properties) are always formed in contrast to alternative contrast categories in a priori given conceptual spaces. He suggests that—perhaps as a result—these basic categories are typically *convex sets*. A set X is convex if and only if for two arbitrary members x_1 and x_2 of X, any x_i that is somewhere between x_1 and x_2 is also a member of X. Gärdenfors claims that for primitive categories, the relevant conceptual spaces give rise to Voronoi tessellations. A Voronoi tessellation not only partitions a structured space into convex sets, it also has prototypes at the center of each convex set. Here is a typical example:

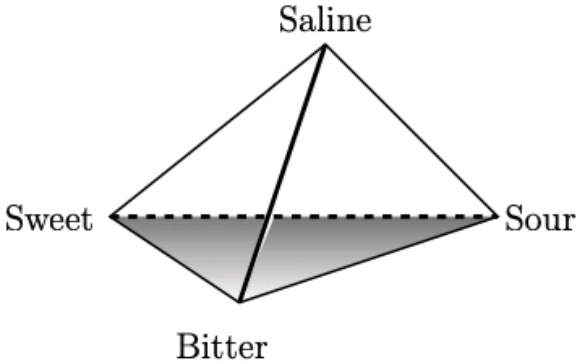


Two of the main examples discussed by Gärdenfors are colors and tastes. He claims that the color space and the phenomenological taste space give rise to Voronoi tessellations. We would like to question, however, whether the most typical colors and tastes are the central members, as proposed by Gärdenfors. First, consider one-dimensional spaces closed on at least one side. In linguistics (e.g. Kennedy and McNally (2005)), the meanings of contrastive adjective pairs such as 'open' and 'closed', 'dry' and 'wet' and 'full' and 'empty' are based on such one-dimensional spaces. The *endpoints* of such meaning spaces, however, will always be marked linguistically, by absolute adjectives, and thus be typical representatives of the classes such (absolute) adjectives denote. Second, inspection suggests that the focal points of the colors in the color space are not in the center. Below is a picture of the representations of colors as the full color spindle.



This picture strongly suggests two things: (i) that colors can be thought of as convex sets in the color space, and (ii) that the prototypes of the colors are (except for gray) always *at the edges* of the color spindle, and thus *not in the center* of the convex sets. Indeed, Regier et al. (2007) found that the best examples of English' white and black, respectively, are the lightest and darkest chips of a chart of colors.[3] Similarly, the so-called 'color emotion wheel' (from Sacharin et al. (2016), though not shown here), suggests that the color pixels which give rise to the highest emotions are on the edges of the color spindle (or circle, in this case). That picture also suggests that the pixels of the highest emotional value of the three basic colors red, blue and green, are *as far away from each other as possible*.

Finally, according to Henning (1916) the phenomenal gustatory space should be described by the following tetrahedron:



[3]A reviewer suggests that white and black are not real colors. This reviewer moreover suggests that 'true' colors only sit along the rim of the middle disc of the color spindle. All 'true' colors are maximally saturated, and only these should be considered. We are somewhat surprised about this suggestion. We agree that 'real', 'true', or stereotypical, red is red with full saturation, but we don't see any reason why we should limit the color space to full saturated colors to begin with. To us, this should be the *result* of an analysis, not the beginning.

Again, it seems that the basic tastes are convex regions of the relevant meaning space, and that their typical representatives are at the edge of the taste space, and far away from each other.

Bickerton (1981) already proposed that 'simple' expressions can only denote *connected*, or *convex*, regions of cognitive space, and hypothesized that the preference for convex properties is an *innate* property of our brains. Perhaps this is the case. Still, we would like to delve somewhat deeper and provide an analysis where convexity of meanings doesn't have to be stipulated, but can be *explained*. The goal of this paper is to provide rational motivations for why standard meanings give rise to convex sets and why typical representatives are as far away from each other as possible.

Linguists like Jakobson (1941) and Martinet (1955) long observed that naturally occurring vowels in languages over the world are always far away from each other in the acoustic space available for vowels. Liljencrants and Lindblom (1972) showed that one can explain this linguistic 'universal' by adopting a principle of maximal perceptual contrast. Likewise, Regier et al. (2007) show that a model that categorizes the color space based on maximization of similarity within category and dissimilarity across categories gives rise to surprisingly accurate predictions for the predicted colors, and gives rise to categories as convex sets. Abbott et al. (2016) show that in trying to predict the focal colors, or the best examples of named color categories across many languages, a model making use of Tenenbaum and Griffiths's (2001) notion of *representativeness* mentioned above outperforms several natural competitors such as models based on likelihood or on prototypes thought of as central members. Although very appealing, we feel that these explanations need to be based on the idea that language is used for communication between agents. This is the starting point of Lewis's (1969) analysis of meaning making use of signaling games. In this paper we seek to motivate why meanings tend to be convex and why extreme exemplars of these meanings, or categories, are considered to be representative by making use of such signaling games.

Jäger (2007) and Jäger and van Rooij (2007) introduced so-called sim-max games, signaling games using an Euclidean meaning space with a similarity-based utility function. They show that by using a simple learning dynamic the evolved equilibria of these games give rise to descriptive meanings which are convex sets.[4] For *sim-max games*, it is shown as well that with uniformly distributed points in the meaning spaces, the imperative meanings derived from the equilibria will be in the center of their descriptive meanings, and can be thought of as prototypes. As argued above, we indeed want an explanation of convex meanings, but now with typical representatives as extremes.[5] Zuidema and de Boer (2009) observed that Liljencrants and Lindblom (1972)'s explanation of naturally occurring vowels as extremes in the acoustic space in terms of maximal contrast makes game theoretical sense in a noisy environment. In this paper we would like to provide a game theoretical

[4]Elliot Wagner (p.c.) has shown, however, that this does not hold in general, if a more standard evolutionary dynamic is used.

[5]One might think that the problem can be solved by adopting a non-flat probability distribution. As observed by Franke (2012), however, this won't do.

explanation of a phenomenon involving maximal contrast as well. But there is an important difference: whereas in phonology the contrast involves the *signals*, in our case the contrast involves the *meanings* of the signals. For simple one-dimensional meaning spaces, Lipman (2009) already provided such a game theoretical derivation, not making use of similarity or confusability at all. Surprisingly enough, his analysis even explains convexity. Unfortunately, we don't see how to extend his derivation to more complex spaces. Franke (2012) *does* explain the preference for extreme points in multi-dimensional spaces.[6] However, he does so by doing it, so to say, in terms of a derived preference for extremes in one-dimensional spaces. What we would like to do is, we think, more ambitious: to explain the preference for the extremes in one go. We think that something like this is required to provide a natural explanation of the preference for extremes in complex spaces where the dimensions are not obviously made up of previously given dimensions that are independent of each other. Such a dependence of the dimensions we find, for instance, in the color space which Gärdenfors (2000) takes to be consisting of a set of *integral* dimensions.

3 Extremes and Iterated Best Response

One way to understand why languages exhibit the properties they do is by analyzing them in the context of cooperative social reasoning. That is, by taking the idea seriously that language is used for communication between interlocutors, and that these interlocutors will reason about each other's linguistic choices to reach mutual understanding (e.g., Lewis, 1969; Grice 1975; Parikh 1991; Rooy van 2004; Benz et al. 2005). To illustrate how such a process of mutual reasoning may naturally lead to convex meanings with extreme typical representatives, this section sketches out the predictions of the Iterated Best Response (IBR) model (Franke 2009; Franke and Jäger 2014) on these matters.

At its core, IBR aims to explain linguistic outcomes in a Gricean fashion: as outcomes of mutual reasoning about rational language use. Formally, patterns of language use can be represented by mappings from messages (utterances) to states (meanings) in the case of receivers, $\rho : M \to T$; and by mappings from states to messages for senders, $\sigma : T \to M$. Plainly put, these are comprehension and production strategies that tell us how two interlocutors behave. That sender and receiver are rational means that, given (their beliefs about) another interlocutor's behavior, they will try to maximize communicative success. If, e.g., the sender believes the chances of the receiver interpreting utterance m_1 correctly to be higher than those of utterance m_2, she will send the former. Letting R and S be the set of all receiver and sender strategies, the set of best responses to a sender/receiver belief is defined as follows:

[6] Explaining convexity is not aimed for in Franke (2012).

$$\mathrm{BR}(\sigma_b) = \{\rho \in R \mid \forall m\colon \rho(m) \in \mathrm{argmax}_{t \in T}\ EU_R(t, m, \sigma_b)\}; \tag{1}$$

$$\mathrm{BR}(\rho_b) = \{\sigma \in S \mid \forall t\colon \sigma(t) \in \mathrm{argmax}_{m \in M}\ EU_S(t, m, \rho_b)\}, \tag{2}$$

where σ_b and ρ_b are the receiver's, respectively the sender's, beliefs about her interlocutor's behavior and $EU(t, m, \cdot)$ codifies the expected utility of either interpreting a message m as t or sending a message m in state t (see below).

Equations (1) and (2) may look unwieldy at first glance, so let us unravel them before moving on. A belief about a sender/receiver strategy is an expectation of how this sender/receiver will act given a state/message. Beyond the fact that they are beliefs about another agent's behavior, these are just mappings from states/messages to messages/states as well. A best response to an interlocutor's (expected) behavior is the strategy that will ensure the best payoff from an interaction with such an interlocutor: the one with the highest expected utility. There might be many ways to use language that maximize utility conditional on a particular belief σ_b or ρ_b; the sets $\mathrm{BR}(\sigma_b)$ and $\mathrm{BR}(\rho_b)$ collect them all.

Having identified the set of best courses of action given a belief about an interlocutor's behavior, we still need to distill from them how an agent should act. For convenience, we write the resulting strategies as behavioral ones. In words, a sender's strategy σ is the one that sends a message m in state t if there is a best response $\sigma' \in \mathrm{BR}(\rho_b)$ that sends it. Otherwise, message m is not sent in t. Formally, $\sigma(m \mid t, \rho_b) = {}^1/_{|\{m' \mid \sigma'(m') = t;\ \wedge\, \sigma' \in \mathrm{BR}(\rho_b)\}|}$ if there is a strategy $\sigma' \in \mathrm{BR}(\rho_b)$ such that $\sigma'(t) = m$, and otherwise 0. Analogously for $\rho(t|m, \sigma_b)$, with the additional proviso that if a message is not believed to be sent at all, the receiver will pick an interpretation at random (cf. Franke and Jäger 2014).

As a final ingredient, we need to specify what sender and receiver care about. Assuming that interlocutors have no preferences over messages and that all they care about is faithful information transfer, utility can be captured by a single function that tracks how closely sender state and receiver interpretation match; e.g., $\delta(t, t') = 1$ iff $t = t'$ and otherwise 0. We then have

$$EU_R(t, m, \sigma_b) = \sum_{t'} \frac{Pr(t')\sigma_b(m|t')}{\sum_{t''} Pr(t'')\sigma_b(m \mid t'')}\delta(t', t); \tag{3}$$

$$EU_S(t, m, \rho_b) = \sum_{t'} \rho_b(t'|m)\delta(t, t'). \tag{4}$$

In words, the expected utility of sending/interpreting message m given state t is just the average of our communicative success given our beliefs about our interlocutor's linguistic behavior. That is to say, expected utility gives us the average payoff we expect when producing or comprehending, conditional on our beliefs about our communicative partner. As stated in (1) and (2), best responses are made up of those strategies that maximize expected utility; those that guarantee the best outcome based on what we care about.

All of this is just to formally capture the idea that a message is sent only in states in which it is believed to have the highest chances to be understood; and that, analogously, a receiver interprets a message as the state that she believes is most likely to be conveyed by it. If there are many optimal choices, players pick randomly from them. If a choice has to be made but none is optimal they pick at random from the entire pool of actions at their disposition. From here, we just need to consider the consequences of nesting beliefs to arrive at pragmatic reasoning: reasoning about the reasoning (and so on) of others to inform our linguistic choices. Formally, a level-$n+1$ reasoner in IBR is defined as acting upon the belief that her interlocutor is of level-n with reasoning levels starting at $n = 0$. Put differently, we have that $\sigma_{n+1}(\cdot \mid \cdot, \rho_n)$ and $\rho_{n+1}(\cdot \mid \cdot, \sigma_n)$.

Beliefs about an interlocutor's strategy at level 0 are usually constrained or biased in some way to start the reasoning chain. If just any belief were permitted, meaningful inference would seldom get off the ground (cf. Sect. 1.2 Franke 2009). Let us consider a simple case in which the sender has seen how the receiver interprets messages and the receiver is aware of this. For instance, she has seen the receiver interpret the utterance *tall woman* as an entity of a particular height and *small man* as an entity of another height. As we shall see, we need not constrain this receiver strategy beyond requiring that it associates each message with a distinct information state. Mutual awareness of this arbitrary separating strategy suffices to lead to the adoption of convex strategies with extreme typical representatives as long as extremes are salient. Saliency could be cashed out in different ways: It may be that extremes are focal points that draw the attention of reasoners due to their psychological noteworthiness relative to other states (cf. Schelling 1980; Mehta et al. 1994); or it might be that extremes confer a functional advantage and attract the reasoners by virtue of their drive to maximize expected utility. The latter might happen, e.g., if perception is noisy in that states that are near to each other are easily confused. This would make extremes attractive in virtue of their special position at the edge of a space, making them the least confusable (see, e.g., Franke et al. 2011, Gibson et al. 2013, Franke and Correia 2018 for other proposals where noise, or error, has been argued to play an explanatory role). Abstracting away from the details of particular noise signatures, their consequences can be captured by a graded utility function that is inversely proportional to a distance measure over the state space under the assumption that coordinating on extremes confers a higher utility than coordinating on less extreme points. We background the details of this function because these two general conditions are sufficient to illustrate our argument. In which way extreme points are salient is ultimately an empirical issue. At this stage proposing a particular function seems too strong a commitment in light of these unknowns.

With these notions at hand, consider the case of four heights, $T = \{1, 2, 3, 4\}$, and two messages, $M = \{m_1, m_2\}$. Figure 1 illustrates how mutual reasoning can lead to convex strategies with extreme typical representatives when reasoning over two initial receiver strategies ρ_0. Intuitively, a level-1 rational sender strategy against a belief of her interlocutor's behavior, $\sigma_1(\cdot \mid \cdot, \rho_0)$, will first ensure that messages sent in a state correspond to correctly interpreted messages; $t_1 \mapsto m_1$ and $t_3 \mapsto m_2$ in the upper example of Fig. 1; and $t_2 \mapsto m_2$ and $t_3 \mapsto m_1$ in the lower one. Second, remaining

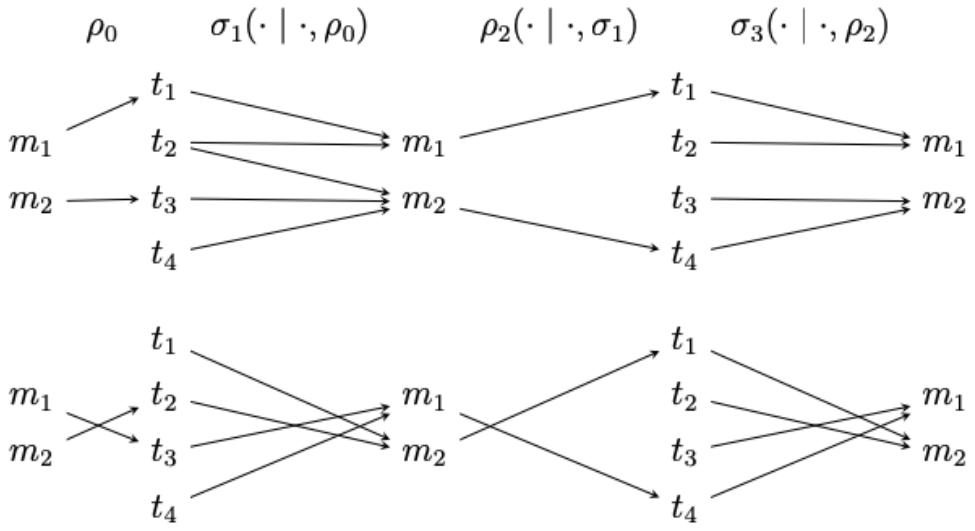


Fig. 1 Illustration of IBR-sequence for two separating initial receiver strategies ρ_0. Depicted outcomes correspond to endpoints of the reasoning process

states will be associated with messages whose interpretation is closest to them. In the upper example in Fig. 1 state t_2 lies in between ρ_0's interpretation of m_1 and m_2, so it is associated with both. A (level-2) receiver who reasons about such a message allocation will naturally associate her messages with the interpretations that are most rewarding: the extremes. Subsequent sender reasoning leads to the association of remaining states such that the state space is partitioned into convex regions. As noted above, this may, e.g., be a consequence of reasoned noisy perception or that of a particular graded utility function. More iterations will not change the sender and receiver strategies anymore. They are in equilibrium.

Just as in Lewis, (1969), we can ascribe two types of meanings to a message in these equilibrium pairs: its *descriptive meaning* is the set of states in which this message is sent and its *imperative meaning* is the response to this message by the receiver. Just as in standard sim-max games, descriptive meanings are now convex sets. But whereas imperative meanings in Jäger (2007) and Jäger and van Rooij (2007) were central points, i.e., prototypes, now they are extreme points, i.e., stereotypes.

This outcome is not limited to one-dimensional spaces such as this ordering of heights. Instead, it obtains in any discrete space with a distance measure, should there be at least as many extreme points as messages. For instance, the color spindle, the taste space, or any discrete subset of a multi-dimensional interval. In any such space, mutual reasoning will iteratively lead to a rational receiver's association of (at least some) messages with extremes. A rational sender follows suit by uniquely identifying extremes with these messages, as well as by improving the space's tessellation with respect to these associations. This process continues as long as the receiver has not yet associated each message with an extreme, being driven by the improved partition each round of back-and-forth reasoning provides. In the end, mutual reasoning bottoms out with convex sender strategies with extreme typical representatives

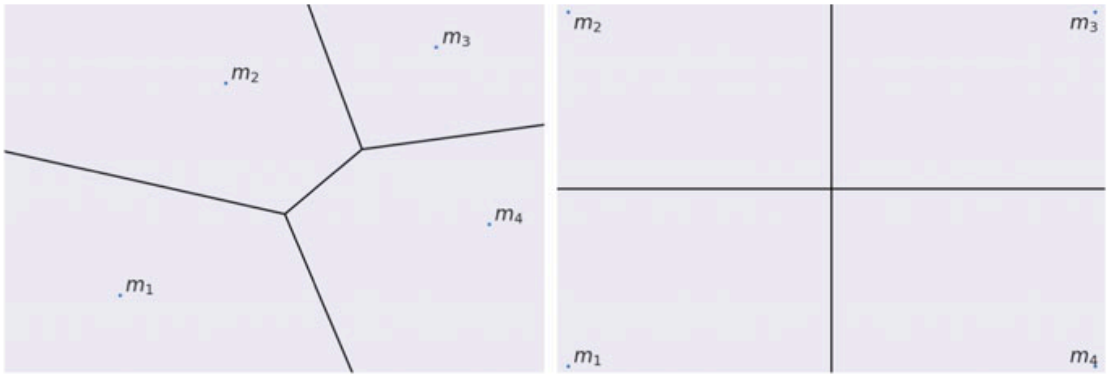


Fig. 2 Illustration of IBR-sequence in a two-dimensional space. Labeled nodes in the left-hand picture depict an initial receiver strategy ρ_0. The resulting convex sender strategy $\sigma_1(\cdot \mid \cdot, \rho_0)$ corresponds to the four regions enclosing each node. Labeled nodes in the right-hand picture correspond to ρ_2 and regions enclosing them depict σ_3

for receiver strategies. Figure 2 sketches out how convex descriptive meanings and extreme imperative ones result from mutual reasoning in such a space.

In the previous section we mentioned that best examples of named color categories are well-predicted by a model based on the following measure of representativeness, $\log \frac{P(x|X)}{P(x|\neg X)}$, which is very similar to a measure used to define stereotypicality. It is interesting to observe that our game-theoretical analysis predicts that the imperative meaning of messages in equilibrium are the most representative ones for their descriptive meanings. To show this, one has to think of $P(t|m)$ and $P(t|\neg m)$ either in terms of sender strategies or in terms of receiver strategies. In the former case, one can interpret $P(t|m)$, for instance, as the probability that t is the actual state if m is sent. However, it is easier to think of $P(t|m)$ and $P(t|\neg m)$ in terms of receiver strategies. In that case, $P(t|m)$, for instance, is just $\rho(t|m, \sigma_b)$, with ρ and σ as the equilibrium receiver and sender strategies, respectively. Once one assumes that senders and receivers use a *quantal* instead of a *maximizing* best response function,[7] in the upper example of Fig. 1, for instance, t_1 and t_4 maximize $\log \frac{\rho(t|m_1,\sigma_b)}{\rho(t|\neg m_1,\sigma_b)}$ and $\log \frac{\rho(t|m_2,\sigma_b)}{\rho(t|\neg m_2,\sigma_b)}$, respectively, and are thus predicted to be the most representative states for m_1 and m_2. In other words, they are the *stereotypes* of the (descriptive) meanings of the

[7]The need for quantal best response is due to a technical complication resulting from the use of maximizing expected utility: it often causes the measure of representativeness to be undefined. To see this, notice that the most representative, or stereotypical, state for message m would now be $argmax_{t \in T} \log \frac{\rho(t|m,\sigma_b)}{\rho(t|\neg m,\sigma_b)}$. But as illustrated in, for instance, the upper example of Fig. 1, $\rho_2(t_1|m_2, \sigma_1) = 0$, meaning that the denominator of $\frac{\rho_2(t_1|m_1,\sigma_1)}{\rho_2(t_1|\neg m,\sigma_1)}$ is 0, which makes the fraction undefined. This problem is solved if we make sure that for no t and m it ever will be the case that $\rho(t|m, \sigma) = 0$. This is what comes out if we assume that instead of being expected utility maximizers, senders and receivers choose probabilistically modeled by quantal response functions (QRFs). These functions are motivated by the idea that (perhaps due to observation errors) decision makers sometimes make mistakes in choosing their best action. These functions are popular in behavioral economics and are gaining popularity in linguistics as well, as they more readily connect rational language use models with empirical data (see, e.g., Franke et al. 2011; Frank and Goodman 2012; Franke and Jäger 2016).

messages. This result is not limited to our simple example using a one-dimensional meaning space, but generalizes to more-dimensional spaces: *stereotypes follow from (boundedly) rational language use*.

4 Conclusion and Outlook

In this paper we followed Gärdenfors and others in the assumption that (simple) properties denote convex sets in conceptual spaces, but argued that typical representatives of categories are (many times) extreme rather than central members of such categories, i.e., stereotypes. Moreover, we provided a rational motivation for convexity of meaning and of stereotypes as typical representatives making use of game theory.

We believe that these motivations are interesting for more general linguistic reasons. For instance, it is not uncommon to believe that generic sentences like 'Birds fly' and 'Sharks are dangerous' express typicalities and it is well-known that generics are excellent tools to express and generate stereotypes. In Rooij van and Schulz (2020) an analysis of generic sentences is proposed based on *contingency*, a measure of representativeness adopted from causal associative learning theory that behaves monotone increasingly with the measures of stereotypicality and representativeness discussed in this paper. This suggests that we could provide a game theoretical motivation for generic language use as well. There is at least one complication, though. Whereas we thought of stereotypes as *members* of a category, for Rooij van and Schulz (2020) it is crucial to think of stereotypes as *sets of* perhaps mutually inconsistent *features*. In the future we would like to see how crucial this distinction is.

In this paper—just as in Jäger (2007) and others—we have fixed the number of messages that play a role in the game beforehand, which determined the number of cells in the resulting partition of the meaning space in equilibrium. Intuitively, that should not be the case: in how many cells the meaning space will be partitioned should be an *outcome* of the game as well, depending on the structure of the meaning space and the utility of each partition. Corter and Gluck (1992) defined a notion of *category utility* to derive Rosch's so-called 'basic-level' categories. It is interesting to observe that this notion is closely related to the notions of 'representativeness', 'contingency' and 'stereotypicality' discussed above. In the future we would like to explain natural partitions of different types of meaning spaces, making use of this notion of *category utility*.

References

Abbott, J., Regier, T., & Griffiths, T. L. (2016). Focal colors across languages are representative members of color categories. *Proceedings of the National Academy of Sciences*, *113*, 11178–11183.

Ameels, E., & Storms, G. (2006). From prototypes to caricatures: Geometrical models for concept typicality. *Journal of Memory and Language*, *55*, 402–421.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 629–654.

Benz, A., Jäger, G., & van Rooij, R., (Eds.) (2005). *Game theory and pragmatics*. Springer.

Bickerton, D. (1981). *Roots of language*. Karoma Publishers.

Burnett, R., Medin, D., Ross, N., & Blok, S. (2005). Ideal is typical. *Canadian Journal of Experimental Psychology*, *59*(3–10).

Corter, J., & Gluck, M. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, *111*, 291–303.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Franke, M. (2009). *Signal to act: Game theoretic pragmatics*. Ph.D. thesis, University of Amsterdam.

Franke, M. (2012). On scales, salience and referential language use. In Aloni, M., Roelofsen, F., & Schulz, K., (Eds.), *Proceedings of Amsterdam Colloquium 2011* (pp. 311–320). Springer.

Franke, M., & Correia, J. (to appear). Vagueness and imprecise imitation in signalling games. *British Journal for the Philosophy of Science*.

Franke, M., & Correia, J. (2018), Vagueness and imprecise imitation in signalling games, *British Journal for the Philosophy of Science*, 69(4), 1037–1067

Franke, M., & Jäger, G. (2014). Pragmatic back-and-forth reasoning. In *Pragmatics, Semantics and the Case of Scalar Implicatures: Nature Publishing Group*.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*(1).

Franke, M., Jäger, G., & van Rooij, R. (2011). Vagueness, signaling and bounded rationality. In B. Onoda, & D. J. McReady (Eds.), *New frontiers in artificial intelligence* (pp. 45–59). Springer.

Gärdenfors, P. (2000). *Conceptual spaces*. The Geometry of Thought: MIT Press.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Goldstone, R., Steyvers, M., & Rogosky, B. (2003). Conceptual interrelatedness and caricatures. *Memory and Cognition*, *31*, 169–180.

Good, I. (1950). *Probability and the weighing of evidence*. London: Griffin.

Grice, P. (1975). Logic and conversation. In P. Cole, & J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). Nature Publishing Group.

Hampton, J. (1981). An investigation of the nature of abstract concepts. *Memory and Cognition*, *9*, 149–156.

Henning, H. (1916). Die Qualitätenreihe des Geschmacks. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, *74*, 203–219.

Hilton, J., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, *47*, 237–271.

Jäger, G. (2007). The evolution of convex categories. *Linguistics and Philosophy*, *30*, 551–564.

Jäger, G., & van Rooij, R. (2007). Language structure: Psychological and social constraints. *Synthese*, *159*, 99–130.

Jakobson, R. (1941). *Kindersprache, aphasie und allgemeine lautgesetze*. Uppsala

Kennedy, C., & McNally, L. (2005). Scale structure and the semantic typology of gradable predicates. *Language*, *81*, 1–37.

Lewis, D. (1969). *Convention: A philosophical study*. Cambridge: Harvard University Press.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language*, *48*, 839–862.

Lipman, M. (2009). Why is language vague? Unpublished manuscript.

Lynch, E., Coley, J., & Medin, D. (2000). Tall is typical: Central tendency, ideal dimensions, and graded category structure among tree experts and novices. *Memory and Cognition*, *28*, 41–50.

Martinet, A. (1955). *Économie des changements phonétiques*. Berne: Francke.

McCauley, C., Stitt, C., & Segal, M. (1980). Stereotyping: From prejudice to prediction. *Psychological Bulletin*, *87*, 195–208.
Mehta, J., Starmer, C., & Sugden, R. (1994). Focal points in pure coordination games: An experimental investigation. *Theory and Decision*, *36*(2), 163–185.
Palmeri, T., & Nosofsky, R. (2001). Central tendencies, extreme points, and prototype enhancement effect in ill-defined perceptual categorization. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *54*, 197–235.
Parikh, P. (1991). Communication and strategic inference. *Linguistics and Philosophy*, *14*(5), 473–514.
Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*, 1436–1441.
Rooij, van R. & Schulz, K. (2020), Generics and typicality: *A bounded rationality approach, Linguistics and Philosophy*, 43(1), 83–117.
van Rooy, R. (2004). Signaling games select horn strategies. *Linguistics and Philosophy*, *27*, 493–527.
Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In *Cognitive development and the acquisition of language*. Academic Press.
Sacharin, V., Schlegel, K., & Scherer, K. R. (2016). Geneva emotion wheel rating study. *Report.*
Schelling, T. C. (1980). *The strategy of conflict*. Harvard University Press.
Schneider, D., Hastorf, A., & Ellsworth, P. (1979). *Person Perception*. Adison-Wesley Pub. Co.
Tenenbaum, J., & Griffiths, T. (2001). The rational basis of representativeness. In J. Moore, & K. Stenning, (Eds.), *Proceedings of the 23th Annual Conference of the Cognitive Science Society*, pp. 1036–1041.
Zuidema, W., & de Boer, B. (2009). The evolution of combinatorial phonology. *Journal of Phonetics*, *37*, 125–144.