# The evolution of adjectival monotonicity[1]

Fausto CARCASSI — *University of Edinburgh*
Marieke SCHOUWSTRA — *University of Edinburgh*
Simon KIRBY — *University of Edinburgh*

**Abstract.** A striking property of scalar adjectives is that they allow for inferences like the following: If John is tall, and Mary is taller than John, then Mary is tall. This inference can be made because if an individual falls in the extension of the adjective, then any individual that has the property to a greater degree also will. We call this universal of adjectival semantics *monotonicity*. In this paper, we present an evolutionary account of monotonicity and support it with three computational models. In the first model, we study which of the possible meanings of scalar adjectives evolve under a pressure for simplicity alone, and we observe degenerate meanings that are unlike natural language adjectives. In a second model, we combine the pressure for simplicity with a pressure for communicative accuracy. Under these pressures, mostly non-monotonic meanings evolve. In the third model, we equip the agents with pragmatic reasoning skills. In the third condition monotonic meanings prevail. We conclude that adjectival monotonicity is caused by a combined pressure for semantic simplicity and communicative accuracy, given human pragmatic skills.

**Keywords:** Iterated Learning, cultural evolution, universals, Rational Speech Act.

## 1. Introduction

Scalar adjectives like “tall”, “fast” and “full” are used in predicative position to relate individuals to degrees of properties. The degree is specified precisely in *measure* uses, e.g. “Mary is 190cm tall”, and imprecisely in *bare* uses, e.g. “Mary is smart”. Bare uses convey that the degree to which the individual possesses the property falls in a certain portion of the scale, the *bare extension* of the adjective in the conversational context of the utterance. For instance, “Mary is tall” means that Mary’s height is greater than a certain threshold, and “Mary is short” means that Mary’s height is lower than a certain threshold. A striking property of scalar adjectives is that they allow for inferences like the following: If John is tall, and Mark is taller than John, then Mark is tall. These inferences can be made due to the property of *increasing (decreasing) monotonicity*: if an individual falls in the extension of the adjective, then any individual that has the property to a greater (lower) degree will also fall in the extension of the adjective. More formally, we call a bare extension $P$ monotonically increasing (decreasing) iff for any two degrees $d_i$ and $d_j$, if $d_i \geq d_j$ ($d_j \geq d_i$) and $d_j \in P$, then $d_i \in P$. $P$ is monotonic iff it is monotonically increasing or decreasing. We call monotonic those adjectives whose bare extension is monotonic in every conversational context.[2]

In the semantic literature, although scalar adjectives have been modelled in several different

[1]We would like to thank Kenny Smith, Jennifer Culbertson, Robert Truswell, Shane Steinert-Threlkeld, Jakub Szymanik, and Isabelle Dautriche for helpful discussion.

[2]In some of the literature (see below), *all* adjectives are increasing monotonic. This is because negative adjectives refer to different scales from their positive counterparts. For instance, “short” is increasing monotonic because its bare extensions consists of every degree of shortness higher than a threshold. There are numerous arguments for this analysis (Kennedy, 2001; Bierwisch, 1989). In the following, we will talk as if adjectives of opposite polarity extended on the same scale, e.g., the scale of height for “tall” and “short”. However, this is for ease of presentation and is easy to reformulate in a way that is consistent with the literature.


*Proceedings of Sinn und Bedeutung 23*, vol. 1, pp. 219–230. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès).

ways (Klein, 1980; Kennedy and McNally, 2005; Kennedy, 2007), the monotonicity assumption is considered uncontroversial. However, monotonicity is not semantically necessary. Prima facie, English could have included a scalar adjective "schtall" such that "Mary is schtall" means that Mary is either shorter than 150cm or taller than 190cm. This possibility raises the question of why monotonicity is a general property of scalar adjectives. The topic of generally occurring patterns in semantic structure (or semantic universals; von Fintel and Matthewson (2008)) has seen a recent increase in attention. Important questions about how we can functionally explain these patterns are being addressed, quite recently, in learning experiments as well as various types of computational model.

Properties to do with scalarity have been studied particularly in the domain of quantifiers. While not identical, quantificational and adjectival monotonicity are similar. Consider a quantifier $Q$ of type $\langle\langle e,t\rangle,\langle\langle e,t\rangle,t\rangle\rangle$. $Q$ is (right-)monotone iff for any sets $A$ and $B$ either $Q(A,B) \wedge B \subseteq B' \implies Q(A,B')$ or $Q(A,B) \wedge B' \subseteq B \implies Q(A,B')$ (we omit the model $M$ for simplicity). If we order the sets by inclusion and call $Q^A$ the set $\{B \mid Q(A)(B) = \text{True}\}$, $Q$ is monotone iff for any sets $A$, $B$ and $B'$, either $B \in Q^A \wedge B \leq B' \implies B' \in Q^A$ or $B \in Q^A \wedge B' \leq B \implies B' \in Q^A$. In this formulation, quantificational monotonicity is strikingly similar to adjectival monotonicity.

Steinert-Threlkeld and Szymanik (forthcoming) focus on monotonicity from the point of view of learning, corresponding to the level of single agents in the model below. They pose the challenge to provide a model of learning on which monotonic quantifiers are easier to learn than non-monotonic ones. They propose *long short-term memory (LSTM) recurrent neural networks (NN)* as a model of learning, and show that LSTM NNs require fewer observations to learn monotonic quantifiers than non-monotonic quantifiers. Furthermore, they argue that LSTM NNs have the advantage of being domain-general and biologically plausible. In virtue of their generality, LSTM could provide a unified learning model for quantifiers, scalar adjectives and other scalar phenomena. This would in turns provide a cognitively plausible foundation for the computational model below.

Further experimental evidence for an advantage of monotonicity in learning comes from Chemla et al. (forthcoming). The authors conducted an experiment in which participants were presented with collections of objects and were taught a rule (resembling a quantifier) associated with specific collections . They show that participants were significantly faster at learning rules corresponding to monotonic quantifiers than those corresponding to non-monotonic quantifiers. This result supports the hypothesis that monotonicity in scalar concepts simplifies learning.

Brochhagen et al. (2016) develops an account similar to the one presented in this paper. It focuses on the large class of scalar expressions, encompassing quantifiers, scalar adjectives, and numerals. The aim is to explain the lack of upper bounds in the semantics of scalar expressions. The authors also propose a model that combines Rational Speech Act agents with pressures for communicative accuracy and simplicity. The approach we follow in this paper is related to that presented in Brochhagen et al. (2016), but there are crucial differences, which we will discuss in the discussion section below.

In this paper, we present a picture of the evolution of monotonicity as an adaptation to the competing pressures for learnability and communicative accuracy. Section 2.1 studies the effects of a pressure for learnability alone on the semantic structure of scalar adjectives. The languages that evolve are monotonic but they do not convey information about the world, and are therefore unlike natural language. In section 2.2, we add a pressure on the languages to be communicatively accurate. This causes a prevalence of non-monotonic languages. Finally, in section 2.3 we implement more sophisticated, pragmatically skilful agents. Monotonic languages evolve. Presenting three different models that explore how these pressures affect the resulting languages, we conclude that monotonicity is a consequence of the combined pressures for learnability and communicative accuracy in a population of pragmatically skilful agents.

## 2. A model for the evolution of monotonicity

### 2.1. Model 1: Pressure from learning

The need to be culturally transmitted creates a pressure on language to be learnable. Learnability itself depends, among other things, on the cognitive biases of the learner. The Iterated Learning (IL) modelling paradigm is a way to study how the cognitive biases of the learners shape the evolution of cultural phenomena (See Smith (2018) for a recent overview). In this section, we use an IL model and find that on its own it does not suffice to account for the emergence of monotonicity.

A standard IL model consists of a number of *chains*. Each chain consists of a number of generations $h_0, h_1, \ldots, h_n$. The life of agents in each generation $h_{i>0}$ has two stages. In the first stage, agents in $h_{i-1}$ are selected to be *cultural parents* of the agents in $h_i$, and proceed to teach the language to their cultural children. In the second stage, the agents in $h_i$ become the cultural parents of agents in $h_{i+1}$. In the case of $h_0$, the languages of the agents are picked at random from the set of all possible languages. We consider the frequency of each language type across all runs of the simulation for all generation after a burn-in period.

#### 2.1.1. Language model

In our model, the world $W$ has two components. First, an ordered set $O = \{D, \geq_D\}$ modelling a scale, where $D$ is a set of three degrees. Second, a uniform probability distribution $P_D$ over $D$ modelling the probability of specific degrees being observed. The set $M$ of meanings is the set of sets of degrees, i.e. $\mathscr{P}(D)$ (See figure 1). Each language $l$ is a set $\{f_l, S, B_l\}$, where $S$ is a set of three signals that is identical for all languages, $B_l$ is a set of three possibly identical meanings $\in M$ and $f_l : S_l \mapsto B_l$ is a function from each signal to its corresponding meaning. In each language, each signal is associated with exactly one meaning, which means that there is no homonymy. However, two meanings in $B_l$ can be identical, which means that synonymy is possible. We added the restriction that in every language there is at least one meaning to refer to each degree. The languages are holistic in the sense of Kirby et al. (2015) because signals have no internal structure. Each language therefore models a system of three adjectives in their bare use, e.g. “x is small”, “x is big”, “x is huge”, with a fixed conversational context.

#### 2.1.2. Bayesian learning model

We model the learning process in a Bayesian framework. In the first model, each agent has an initial prior probability distribution over languages which models its expectation about the language to learn. As an agent observes more and more data produced by its cultural parent, the prior distribution gets updated and tends to give higher probability to the language that the cultural parent is speaking. A learning event in generation $h_i$ unfolds as follows. An agent $a_p$ is selected from $h_{i-1}$ to be the cultural parent, and an agent $a_c$ is selected from $h_i$ to be the cultural child. Then, $a_p$ observes degree $d$ and sends a signal $s_p$ to $a_c$ describing $d$ in $a_p$'s own language. Agent $a_c$ receives both $s_p$ and $d$ and updates its probability distribution over languages. For each language $l$, updating follows Bayes' rule:

$$p(l \mid d, s_p) \propto p(d, s_p \mid l)p(l) \propto p(d)p(s_p \mid d, l)p(l) \tag{1}$$

Since the probability of observing a degree is uniform across degrees, eq. 1 becomes:

$$p(l \mid d, s_p) \propto p(s_p \mid d, l)p(l) \tag{2}$$

After learning from the data, the child picks a language to use when teaching the following generation. Two kinds of agents can be defined based on how the pick their language given a posterior distribution. *Sampling* agents sample a language from the posterior distribution. *MAP* (maximum a posteriori) agents pick the language with the highest posterior (Kirby et al., 2014). We show the simulation results with both MAP and sampling agents.

### 2.1.3. Simplicity based prior, and likelihood

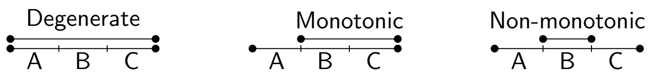
The prior distribution is based on work showing that learners prefer simpler, i.e., more compressible, languages (Culbertson and Kirby, 2016; Kirby et al., 2015). The prior probability for each language $l$ is:

$$p(l) \propto 2^{-\gamma \mathscr{L}(l)} \tag{3}$$

Where $\gamma$ models the strength of the bias for simpler languages and $\mathscr{L}(l)$ is the description length of language $l$. $\mathscr{L}(l)$ is the sum of the description lengths of the meanings expressed by the signals of $l$. We calculate the description length of each meaning as the number of bits needed to encode it.

We encode meanings as follows. Every meaning is a portion on the scale, namely that portion of the scale where the meaning applies (see figure 1). We call *transitions* the degrees where a meaning goes from applying to not applying or from not applying to applying. To encode a meaning, we first encode the position of all the meaning's transitions. Then, we specify whether the meaning applies at the start of the scale. This coding scheme allows to compress any meaning in $1 + c\log(n-1)$ bits, where $c$ is the number of transitions in the meaning and $n-1$ is the number of possible transitions in the scale (fig. 1). Due to this coding scheme, monotonic meanings are attributed a higher prior probability than non-monotonic ones. If a meaning has no transitions, it can be compressed in one bit, which says whether the meaning contains all the degrees or is empty. We call such meanings *degenerate*.

Finally, the likelihood $p(s \mid d, l)$ is the probability of signal $s$ being sent by a speaker of language $l$ after observing a degree $d$. Calculating the likelihood requires a model of production, describing how agents pick a signal given a degree. There are two relevant cases to model. First, the probability of the agent producing a signal whose meaning does not contain the observed degree is 0. Second, if one or more available meanings are compatible with the observed




1 bit $<$ $1+\log(2)$ bits $<$ $1+2\log(2)$ bits

Figure 1: “A”, “B”, and “C” are the scale’s three degrees. Each meaning is represented as a line extending over the degrees that belong to it. We estimate the complexity of a meaning as the length of a lossless encoding of the meaning.

degree, the agents need to choose which one to use for communication. For the simple agents in this model, the only semantic criterion for choosing between meanings is compatibility, so all compatible signals are equally likely to be chosen. This behaviour can be modelled as:

$$p(s \mid d, l) = \begin{cases} 0, & \text{if } d \notin f_l(s) \\ \dfrac{1}{|\{h \mid h \in B_l \wedge d \in h\}|} & \text{if } d \in f_l(s) \end{cases} \tag{4}$$

The production model in equation 4 has a number of desirable consequences. If a language only has one signal to refer to the observed degree, then the production probability is 1. A language that could not have produced a combination of signal and degree is judged impossible. Moreover, if two languages are equally probable a prior but one can only refer to refer to the observed degree with a single signal while the other has multiple signals, the former language is more probable.

The probability of each language is evaluated by the learner on a sequence of tuples $\langle \text{degree}, \text{signal} \rangle$. Given equation 4, the probability of a sequence $G = \langle \langle s_1, d_1 \rangle, \ldots, \langle s_n, d_n \rangle \rangle$ being produced by a speaker of language $l$ is:

$$p(G \mid l) = \prod_{\langle s_i, d_i \rangle \in G} \frac{1}{|D|} p(s_i \mid d_i, l) \tag{5}$$

### 2.1.4. Results

We ran a pure IL condition for 3000 generations with a population of 10 agents for both MAP and sample agents. We discarded the first 500 generations as a burn-in. In figure 2 we plot the frequency of each language type across all agents and generations. The results show that in a simple IL condition both MAP and sample agents across all generations mostly learn degenerate languages. This result is predicted by the literature on IL. IL alone creates a pressure for languages to get increasingly learnable, i.e. conform to the prior expectations of the agents, and the prior favours simple languages.

However, degenerate languages are not what we observe in real world adjectival systems. Real world adjectives allow speakers to convey information about amounts of properties. The interim conclusion is therefore that a pressure for simplicity alone does not account for the evolution of non-degenerate monotonicity. The crucial advantage of non-degenerate meanings is their

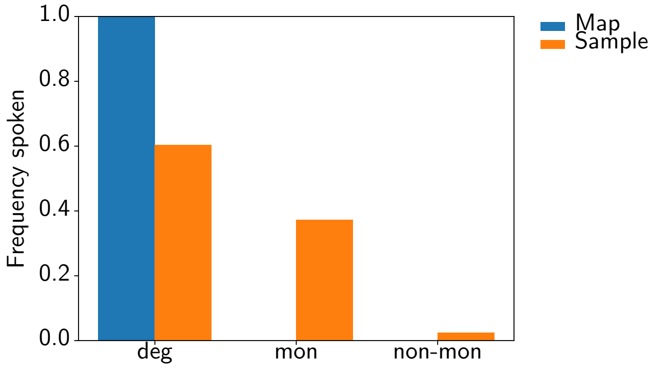



Figure 2: Frequency of monotonicity types under a learnability pressure alone averaged over 100 runs of the simulation.

greater expressiveness. We implement this idea in the next section by adding a pressure for communicative accuracy.

### 2.2. Model 2: Communicative pressure

#### 2.2.1. Communication

In the present section, we add a selective pressure for accurate communication. After agents picked a language from their posterior distributions, communication in the population proceeds as follows. First, two agents $a_S$ and $a_H$ are picked to be the speaker and the hearer respectively. Agent $a_S$ observes a degree $d_S \in D$ produced by the world and picks a meaning $m_S$ with uniform probability among the ones in $a_S$'s language that contain $d_S$ . Then, $a_S$ produces a signal $s_S$ that expresses $m_S$ in $s_S$'s own language, as described in equation 4. Then, $a_S$ sends the signal to $a_H$. Finally, $a_H$ considers the meaning $m_H$ expressed by $s_S$ and picks a degree $d_H$ with uniform probability among the degrees compatible with $m_H$ in $a_H$'s own language. The communication event is successful if $d_S = d_H$. Communication can be unsuccessful, either when speaker and hearer use different languages, or when more than one degree is compatible with the signal that was communicated. Given this picture of communication, the expected communicative success of an agent $a_H$ with language $l_H$ listening to an agent $a_S$ with language $l_S$ is:

$$c(a_H, a_S) = \sum_{\langle d_i, s_j \rangle \in D \times S} p(s_j \mid d_i, l_S) p(d_i \mid s_j, l_H) \tag{6}$$

where $p(s_j \mid d_i, l_d)$ is the probability described by equation 4 of $a_S$ producing signal $s_j$ and $p(d_i \mid s_j, l_H)$ is the probability of $a_H$ guessing degree $d_i$.

The pressure for communicative accuracy is implemented by selecting agents to be cultural parents as a function of their communicative success:

$$p(a_j \text{ is selected}) \propto \exp(\varepsilon \cdot c(a_j, a_i)) \tag{7}$$

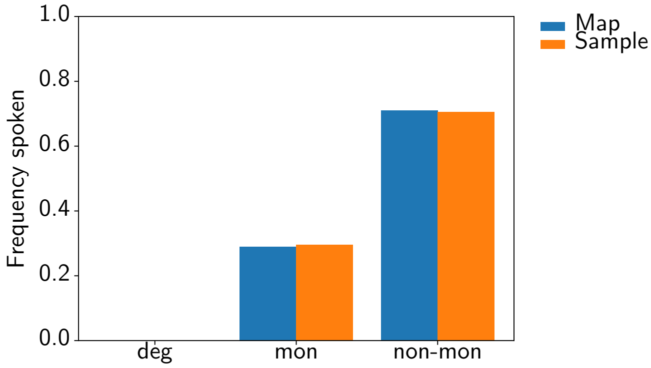



Figure 3: Frequency of monotonicity types with pressures for learnability and communicative accuracy averaged over 100 runs of the simulation.

where $\varepsilon > 0$ determines the strength of the selection and $a_i$ is the cultural parent of $a_j$. The consequence of this way of calculating fitness is that languages that often provoke a failure in communication are taught less often to the following generation.

### 2.2.2. Results

We ran the model for 3000 generations with a population of 10 agents for both MAP and sample agents. Fig. 3 shows the frequency of each language type across all runs of the model, again with a burn-in period of 500 generations. Adding a pressure for communicative accuracy decreases the frequency of monotonic languages. Monotonic languages are communicatively less accurate than non-monotonic languages, and this affects the fitness of its agents negatively.

In section 2.1, degenerate monotonic meanings prevail in virtue of their simplicity. In the present section, non-monotonic languages prevail in virtue of their greater accuracy. However, both models fail as an evolutionary account of adjectival monotonicity. The communicative suboptimality of completely monotonic languages is tied with the way in which agents produce and understand signals. Agents in this model are literal, in the sense that they produce and understand signals as a function of their compatibility with observations alone. In section 2.3, we study the effects of implementing a more realistic model of communication that takes into account human pragmatic skills.

## 2.3. Model 3: Pragmatic agents

The literal agents in the models above base their linguistic behaviour purely on the semantics of their language, without exploiting the additional information that comes from interacting with cooperative rather than merely truthful agents (Grice, 1991). In the present section, we add a more sophisticated model of production and understanding, the so-called *Rational Speech Act* (RSA) model (Goodman and Frank, 2016).

### 2.3.1. Rational Speech Act agents

RSA is a way of modelling the way in which pragmatic communication follows from agents capable of thinking about each other's minds. An RSA pragmatic listener $L_1$ reasons about a pragmatic speaker $S_1$ which in turn reasons about a literal listener $L_0$. Literal listeners $L_0$ are agents of the sort we encountered in section 2.1.

Pragmatic speakers $S_1$ observe a degree $d$ and calculate the utility that each signal $s$ has for a literal listener $L_0$:

$$\mathscr{U}_{S_1}(s;d) = \log(p_{L_0}(d \mid s)) \tag{8}$$

were $P_{L_0}(d \mid s)$ is the probability that the literal listener attributes to degree $d$ after having received signal $s$.[3] Signals that maximize the listener's posterior for the degree observed by the speaker have higher utility. Pragmatic speakers then choose the signal to utter with a probability proportional to the utility for the literal listener:

$$p_{S_1}(s \mid d) \propto \exp(\alpha \mathscr{U}_{S_1}(s;d)) \tag{9}$$

where $\alpha$ determines the strength of the increase in the probability of picking an utterance given an increase in utility.

Finally, pragmatic listeners $L_1$ perform Bayesian inference on the basis of the behaviour of $S_1$ agents. After receiving a signal $s$ with meaning $m$ from a speaker, $L_1$ calculates the probability of each degree:

$$p_{L_1}(d \mid s) \propto p_{S_1}(s \mid d) p_{L_1}(d) \tag{10}$$

where $p_{L_1}(d)$ is the prior probability that the listener attributes to the degree being observed.

### 2.3.2. Results

We ran 100 chains of 3000 generations each for both MAP and sample agents, with a population of 10 pragmatic agents per chain. We excluded the first 500 generations of each chain as a burn-in. Figure 4 shows the proportion of the languages spoken in the remaining generations by type.

The third model makes the correct prediction, namely that systems of adjectives evolve to be non-degenerate and monotonic. Implementing pragmatic skills, which gives artificial agents the ability to calculate scalar implicatures, allows agents to accommodate the prior preference for monotonic extensions without losing in terms of communicative accuracy. Monotonic, non-degenerate languages are the best trade-off between communicative and learnability pressure only if agents are pragmatically skilful.

## 3. Discussion

With our models, we have provided an evolutionary account for the monotonicity property of scalar adjectives: monotone adjectival meanings constitute the best solution for learnability and communicative accuracy, if we assume that language users are capable of pragmatic reasoning. As mentioned above, Brochhagen et al. (2016) develop an account that is similar to ours but

[3] We simplify the original model by assuming that the utterance cost is the same for all adjectives.

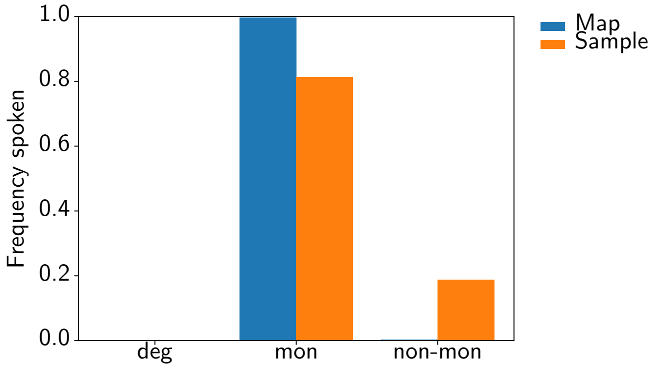



Figure 4: Frequency of monotonicity types with pragmatic agents averaged over 100 runs of the simulation.

differ in some crucial respects. The meanings are less structured than in the models above. The structure of each meaning is a function of its relation to an upper bound; each meaning can cover what is below the upper bound, what is above, both, or neither, and is therefore encoded with two bits. This modelling choice has two consequences. The first is that there are no degenerate meanings in the sense we have used above. A meaning that is true for both states in Brochhagen et al. (2016) is not degenerate, but rather simply one that lacks an upper bound, e.g. the meaning of English “some”. In the models above, we concluded that a simplicity pressure alone was insufficient because it resulted in degeneracy. On the other hand, Brochhagen et al. (2016) exclude a pressure for simplicity alone because it results in all the signals getting the same meaning, i.e. the monotonic meaning. The two models offer therefore different arguments for the insufficiency of a simplicity pressure alone: avoidance of degeneracy in our model and of synonymy in Brochhagen et al. (2016).

Additionally, since Brochhagen et al. (2016) focus on a very general sense of scalarity, there is no obvious way to add structure to the model in a way that encompasses all the relevant semantic structures. As a consequence, the higher complexity of non-monotonic meanings and the size of this difference are stipulated in the model. While this is an explicit modelling choice to avoid introducing assumptions, it makes it harder to extend the measure of complexity beyond two signals. Since the model in Brochhagen et al. (2016) only has two states, there is not much need for an explicit functional form for complexity. One can simply specify how much more complex the non-monotonic meaning is than the monotonic one, and a great variety of complexity measures could fit the two picked complexity values for some parameters specification. On the other hand, calculating the complexity level of three or more meanings requires a decision about their relative complexity, ideally as a function of the differences in their semantic structure. Using scales to model the meaning structure of scalar adjectives allowed us to model the relations between the different meaning structures and their complexity. In sum, having a simple semantic model makes the model in Brochhagen et al. (2016) suitable to discuss different cases of scalarity, but working with only two states and two signals implies that their model cannot detect differences between degeneracy and monotonicity. Conflating degeneracy and monotonicity is problematic, given that experimental work shows that under a pressure for learning only, people prefer degenerate systems (Kirby et al., 2015).

In a related and more recent paper, Brochhagen et al. (2018) narrow their focus to quantifiers, and provide an explicit measure of complexity based on the set-theoretic analysis of generalized quantification. In this paper, the semantic model has more complexity, and includes three states, i.e. a representation of meaning that is more similar to what we did. However, other differences from the model above are introduced. Crucially, in Brochhagen et al. (2018) degenerate meanings are excluded from the set of possible meanings. The spread of degenerate languages was our reason for introducing a pressure for communicative accuracy. Instead, Brochhagen et al. (2018) introduces the communicative pressure to explain how the population converges to a single shared language. A second difference between our and Brochhagen et al. (2018)'s paper is how communicative fitness is calculated. In Brochhagen et al. (2018) what matters is how well agents can speak with the other agents in the same population. Letting agents interact within their generation is useful when studying convergence to a single language. On the other hand, we calculate communicative accuracy between cultural parent and offspring, implementing communication in a more restricted manner (and allowing for more variability in the population). Spike et al. (2017) compare various implementations of communication pressure in a computational model, and their results suggest that the direction of communication (horizontal vs. vertical) does not make a huge difference for obtaining a conventional signaling system. I.e., the two models may produce very similar results. However, how and whether the differences work out in the case of scalar meanings is still an open question.

Finally, we would like to go back to the observation that scalar adjectives are monotonic. Even though this statement seems largely uncontroversial among semanticists, the claim is not unchallenged. A claim against monotonicity as a general property of scalars can be found in Verheyen and Égré (2018). The authors propose an analysis of the meaning of gradable adjectives based on prototypes. More specifically, gradable adjectives have prototypes not in isolation, but only once they are relativised to a specific comparison population. Given the prototypes, the scale is carved into categories by letting each point belong to the category of its closest prototype[4]. This process induces a so-called *Voronoi* tessellation on the space (Gardenfors, 2004). Two features of Voronoi tessellations are problematic when they are used to model adjectival semantics. Firstly, Voronoi tessellations produce extensions that are convex. For systems of two prototypes on a one-dimensional space, the convex categories will be monotone. On the other hand, a system with three prototypes or more necessarily produces at least one convex non-monotone category. Verheyen and Égré (2018) consider systems of two adjectives of opposite polarity, e.g., "cold"/"warm". However, systems of real adjectives also include more extreme adjectives such as "hot". An account of the semantics of scalar adjectives has to explain why further adjectives do not limit the other adjectives in the system, making them non-monotonic. Our model explains this under the assumption that scalar adjectives are encoded in terms of threshold. However, an account based on prototypes cannot account for this using that explanation.

The second problem with an analysis based on prototypes is that Voronoi tessellations cover the

[4]The model of Verheyen and Égré (2018) is more complex, since each category is associated with multiple prototypes. However, this aspect of their model is not relevant to the present discussion.

whole space. This prima facie contradicts the existence of what has traditionally been called the *zone of indifference* (Sapir, 1944), i.e., a part of the scale between adjectives of opposite polarity that is not covered by either. (See Franke, 2012 for a game-theoretical account). For instance, people of average height are neither tall nor short and therefore fall in the zone of indifference. A possible way to reconcile the prototype account with the existence of the zone of indifference is to claim that the latter corresponds to a vague threshold. The vagueness of the threshold should explain the intuition that the degrees in the zone of indifference do not clearly fall under either of the antonyms. This is what Verheyen and Égré (2018) propose. The discussion therefore hangs on whether one thinks that at least some degrees in the zone of indifference clearly do not belong to either antonym. More experimental data and clearer empirical predictions from the threshold account are required to establish which analysis conforms better to linguistic intuitions.

We defined a ubiquitous property of scalar adjectives we named monotonicity. We then developed an evolutionary account of monotonicity. The simulations we presented give a novel explanation for the universal of monotonicity in scalar language. The mechanism underlying the spreading of monotonicity rests on a combination three facts, namely that monotonic meanings are simpler than non-monotonic meanings, that of language is shaped by both IL and pressure for accurate communication, and that human beings are capable of pragmatic reasoning.

**References**

Bierwisch, M. (1989). The Semantics of Gradation. In W. J. M. Levelt, M. Bierwisch, and E. Lang (Eds.), *Dimensional Adjectives*, Volume 26, pp. 71–261. Berlin, Heidelberg: Springer Berlin Heidelberg.

Brochhagen, T., M. Franke, and R. van Rooij (2016). Learning biases may prevent lexicalization of pragmatic inferences: A case study combining iterated (Bayesian) learning and functional selection. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, pp. 2081–2086.

Brochhagen, T., M. Franke, and R. van Rooij (2018). Coevolution of Lexical Meaning and Pragmatic Use. *Cognitive Science 42*(8), 2757–2789.

Chemla, E., B. Buccola, and I. Dautriche (Forthcoming). Connecting content and logical words. *Journal of Semantics*.

Culbertson, J. and S. Kirby (2016). Simplicity and Specificity in Language: Domain-General Biases Have Domain-Specific Effects. *Frontiers in Psychology 6*.

Franke, M. (2012). Scales, Salience and referential safety: The benefit of communicating the extreme. In T. C. Scott-Phillips, M. Tamariz, E. A. Cartmill, and J. R. Hurford (Eds.), *The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9)*, pp. 118–125.

Gardenfors, P. (2004). Conceptual spaces as a framework for knowledge representation. *Mind and Matter 2*(2), 9–27.

Goodman, N. D. and M. C. Frank (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences 20*(11), 818–829.

Grice, H. P. (1991). *Studies in the Way of Words*. Harvard: Harvard University Press.

Kennedy, C. (2001). On the monotonicity of polar adjectives. In J. Hoeksema, H. Rullmann, V. Snchez-Valencia, and T. van der Wouden (Eds.), *Perspectives on Negation and Polarity Items*, pp. 201–221. Amsterdam: John Benjamins.

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

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

Kirby, S., T. Griffiths, and K. Smith (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology 28*, 108–114.

Kirby, S., M. Tamariz, H. Cornish, and K. Smith (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition 141*, 87–102.

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

Sapir, E. (1944). Grading, A Study in Semantics. *Philosophy of Science 11*(2), 93–116.

Smith, K. (2018). The cognitive prerequisites for language: Insights from iterated learning. *Topics in Cognitive Science 21*, 154–160.

Spike, M., K. Stadler, S. Kirby, and K. Smith (2017). Minimal Requirements for the Emergence of Learned Signaling. *Cognitive Science 41*(3), 623–658.

Steinert-Threlkeld, S. and J. Szymanik (Forthcoming). Learnability and Semantic Universals. *Semantics and Pragmatics*.

Verheyen, S. and P. Égré (2018). Typicality and Graded Membership in Dimensional Adjectives. *Cognitive Science 42*(7), 2250–2286.

von Fintel, K. and L. Matthewson (2008). Universals in semantics. *The Linguistic Review 25*(1-2).