The origins of vagueness

Peter R. Sutton

1 Introduction

At least one of the following two properties are usually assumed to be necessary for an expression to count as *vague*:

(1) *borderline cases*: there are things that are as much f as they are not-f);

(2) *blurred boundaries*: it is not clear where the boundary between fs and non-fs is.

On this basis, a great many natural language expressions are vague. For instance, there are examples of nouns/NPs, adjectives, VPs, PPs, and adverbial expressions that have blurred boundaries. E.g., there is no sharp cut-off between (*a*) *kitten* and (*an*) *(adult) cat*, between *tall* and *not tall/short*, between *speak loudly* and *shout*, between *under the table* and *not under the table*, or between *quickly* and *not quickly*.

A daunting amount has been written on vagueness in philosophy, linguistics, and cognitive science, and a large portion of this work has centred on (proposing solutions to) the *sorites paradox*, a logical puzzle generated by a form of slippery slope reasoning from clear cases to clear non-cases of a predicate by evoking increments so small that, so the intuition goes, they could not make a difference to whether or not one should apply that predicate. For example, if a room is (clearly) large, then a room smaller by only one square centimetre must also be large, but then by repeated applications of this step, a 1 metre-square room is large, contrary to fact. The argument is troubling as it appears to show either that there is an inconsistency in how we reason about the extensions of expressions that leads us into contradiction, or that at least one of the premises of the argument is false, which, in most commonly used logics, implies that there is a sharp boundary between fs and non-fs, meaning that, contrary to our intuitions, vague predicates do not have blurred boundaries or genuine borderline cases after all.

Even setting the sorites paradox aside, there is a tension between the two main diagnostics of vagueness in (1) and (2) and classical semantic theory in the Montague-Lewis tradition. Namely, if the extensions of natural language expressions such as common nouns, adjectives and VPs are modelled in terms of classical sets (or functions from classical sets to classical sets), it is not clear how such an expression could have borderline cases or why its application conditions are blurred/non-sharp. Given the prevalence of vague expressions in natural languages, a proper treatment of the phenomenon of vagueness is therefore theoretically pressing. In this chapter, we will examine a broad family of analyses of vagueness, all of which make use of classical (Bayesian) probability theory.

Although there were some early applications of probability theory to vagueness (Black, 1937; Borel, 1907/2014), in the last eleven to twelve years, the use of this and related formal tools has seen a resurgence of interest. A core idea runs through each of these approaches – the use and/or meaning of vague expressions inherently involves uncertainty.

This uncertainty has been characterised, for instance, as uncertainty about what the truth-conditions of an expression are, or, from a more agent-oriented perspective, uncertainty about whether a competent speaker would apply a vague expression in a given instance. Furthermore, some of the advocates of probabilistic approaches to vagueness have argued that probabilistic analyses of vagueness afford us more clarity on what could be termed the *origins of vagueness*, namely, why natural languages have developed vague expressions in the first place, and it is this issue that will be addressed in this chapter.

Section §2 provides an overview of many of the probabilistic proposals for treatments of vagueness. (For an overview of vagueness theories including non-probabilistic approaches to vagueness, see Burnett & Sutton, 2020). Although probabilistic solutions to the sorites will briefly be discussed in §2.5, our main area of interest here will be what, from a probabilistic standpoint, underpins why vagueness arises in the first place. Most of the work to be discussed (see §2.2) forms a loose consensus on this issue, namely, that some form of **noise** generates uncertainty about how to apply a predicate. In information theoretic terms, for an information channel where a message is encoded as a signal and sent to a receiver where it is decoded, noise increases uncertainty regarding the message, since it either modulates or confounds the signal, making the message harder to recover (unless the signal encodes enough redundancies (Shannon, 1948)). Although different theories propose different explanations of how noise enters the system, the unifying idea here is that this affects agents' uncertainty regarding predication and that this uncertainty, under certain other constraints, characterises vagueness. In §2.3, we will discuss a more minority view, namely that vagueness arises as a **byproduct of semantic learning**: an agent's learning data in some sense underspecifies how a predicate should be used in a certain situation and so the semantic representations one develops do not carve out sharply defined boundaries. Although the noise and semantic learning hypotheses are not mutually exclusive (both could contribute to vagueness in natural language), they do offer different perspectives on where the principal source of vagueness lies.

The second part of this chapter, 3, presents the results of a series of simulations that were run with the goal of providing a clearer picture of whether noise or a byproduct of semantic learning contributes to the evolution of vagueness. These simulations were run with a probabilistic iterated learning model (for non-probabilistic iterated learning models, see Kirby, 2007; Kirby & Hurford, 2002). In this model, for $n + 1$ generations of agents, a_0..., a_n, each agent $a_{i<n}$ provides a sample of their language as learning data for a_{i+1} and each agent $a_{j>0}$ learns on the basic of a sample of the language from a_{j-1}. Perhaps surprisingly, the results of these simulations indicate that both noise and learning data sparsity appear to contribute to the development of vagueness in a language. That is to say that, even if we assign a completely precise (non-vague) language for a_0 in the model, the introduction of noise and a restriction on the amount of learning data each agent receives generates a stable and vague language across generations.

2 Overview of probabilistic approaches to vagueness

2.1 Background

An influential philosophical proposal for the analysis of vagueness is *epistemicism*, the theory that vague predicates have sharp boundaries/thresholds that are unknowable since

extension facts supervene on language usage facts, and one cannot, at least practically speaking, infer the former from the latter (Williamson, 1994, 1992).[1] From a probabilistic standpoint, some aspects of the theory are appealing, since it is a fairly natural move to recast *ignorance regarding the application conditions of a predicate* with *uncertainty regarding the application conditions of a predicate*, where this uncertainty, and reasoning in conditions of metalinguistic uncertainty can be modelled with Bayesian probability theory.

However, despite the apparent affinity with a probabilistic approach to vagueness, epistemicism has also been criticised by their proponents on the basis that it is unclear how this theory can explain how meanings can be learned from instances of uses of predicates, and how agents using vague expressions can nonetheless coordinate sufficiently to communicate successfully (see, e.g., Lassiter, 2011; Sutton, 2013). Bayesian approaches assuage these concerns in one of two ways: (i) They model agents as a being uncertain regarding where the threshold for f lies, but nonetheless as approximating its location (reducing uncertainty). On this conception, probabilities are probabilities of what the truth conditions of f are (see, e.g., Fernández & Larsson, 2014; Lassiter, 2011)[2]; (ii) They interpret probabilities as reflecting an agents estimation of how likely a competent speaker, or, alternatively, one's interlocutor, would use f, given a forced choice between a set of alternatives of say f, not-f and saying nothing (see Bernardy et al., 2018; Égré, 2017; Lassiter & Goodman, 2017; Sutton, 2018; Sutton, 2013, among others). This latter perspective seems to have become more dominant in recent years. In either case, the kinds of uncertainty being assumed in these approaches will be referred to in this chapter as *metalinguistic uncertainty*.

Our focus here, however, will be on the insights Bayesian approaches can provide on the origins of vagueness (how and why vagueness arises in the first place), especially given their rootedness in theories of learning and communication. Theories fall into two main (non-exclusive) categories: those that locate the origins of vagueness in a source of noise, and those that attribute it to a byproduct of semantic learning.

2.2 Noise giving rise to vagueness

This section gives a summary overview of some of the main probabilistic models that posit some form of noise as an explanation for why natural language predicates are vague.

Probabilistic linguistic knowledge. One way to assuage some of the aforementioned concerns with epistemicism is to assume a sharp boundary for vague predicates at each context such that agents imperfectly estimate where this boundary is (Eijck & Lappin, 2012; Fernández & Larsson, 2014; Frazee & Beaver, 2010; Lassiter, 2011). Here, we will briefly outline one of these more communication-oriented views as proposed by Lassiter (2011).

Building on the dynamic logic theory in (Barker, 2002), Lassiter (2011) defines a probabilistic belief space that can be updated dynamically to model both uncertainty about the world (which possible world is actual) and metalinguistic uncertainty (what the precise meaning of an expression is in a communicative context). An agent reasons

[1] However, for a different conception of epistemicism based on necessarily unknowable boundaries, see Sorensen (1988, 2001).

[2] A problem this gives rise to is that there are, de facto, sharp truth conditions for every vague predicate, at least relative to a context.

about the most likely precise interpretation of a vague expression in the context, given their beliefs about what the world is like. The basic idea here is that agents may be able to approximately converge on what the relevant threshold is for a vague expression. Although, Lassiter does not discuss noise explicitly, one way to think about this is that the inherent uncertainty we have regarding others' metalinguistic beliefs is generated by a source of noise that is seldom if ever eliminated entirely.

Approximative measurement. Égré (2017) develops a probabilistic model of vague judgement based on an idea from psychology that magnitudes, such as height, and loudness, are mentally represented only with some degree of approximation. Égré (2017) explicitly proposes that there is noise in the information channel between how the world is (sizes/height of objects etc.) relative to a unit of measurement, and the representation of these entities by agents.

Noise is represented by the value of a random variable summed over the number of units of measurement being considered, i.e., for each measurement unit (e.g., centimetres, decibels etc.), there is some chance of over- or underestimating it. These over- and underestimations compound, so that for an entity of some size on the given scale, there is a probability distribution over the size it will be estimated to be such that the greater the variance in this distribution, the greater the amount of noise there is in the agent's estimation system. The probability of applying some predicate to an entity is derived via a context- and comparison class dependent *criterion value* which can be thought of as a kind of threshold, above which $Pr(\text{`}x \textit{ is } f\text{'}) > Pr(\text{`}x \textit{ is not } f\text{'})$.

Égré's (2017) proposal, furthermore, makes progress on the proposals considered above insofar as he pays closer attention to the complexities of the interactions between context and the comparison class for vague predicates and also provides an explanation of how the contextual standards of agents can diverge on the basis of *interest relativity* (See also Fara, 2000).[3]

Introducing Measurement Error. Bernardy et al. (2018, 2019) implement a compositional semantics for natural language that includes a treatment of vague gradable adjectives such as *tall*.[4] Conditional probabilities in (Bernardy et al., 2018) represent the likelihood that a competent speaker of the language would endorse an assertion, given certain evidence/assumptions/hypotheses. The predicates and individuals are represented as vectors, and the probability that a predicate applies to an individual is estimated via Markov chain Monte Carlo (MCMC) sampling over these vectors, conditioned by certain observations. Two advantages of this system are, firstly, that semantic judgements are inherently graded, which as argued by Sutton (2018) is a promising basis for modelling vagueness, and, secondly, an arbitrarily rich notion of context is built into the model via the inclusion of data (observations) upon which probabilistic judgements are conditioned (a feature which is shared by the richly-typed approaches in Fernández & Larsson 2014, see also Cooper et al. 2015 and Schuster et al. 2020).

To model vagueness, Bernardy et al. (2018, 2019) introduce a Gaussian error into probabilistic evaluation, i.e., an error due to the Gaussian distribution, which introduces additional uncertainty into a predication. The effect of this is to introduce an added

[3]Égré (2017) also proposes further applications of his theory. For instance, how the model can be applied in the context of Égré's other work on borderline contradictions (see, e.g., Egré et al., 2013).

[4]It is beyond the scope of this chapter to go into the full details of their model (which includes a treatment of quantifiers, for instance), and so we will restrict our focus to (vague) predication.

amount of uncertainty into the system. As such, this approach is an explicit example of a noise-based theory of vagueness.

Game theoretic and Bayesian Pragmatics approaches. A further proposal that incorporates noise as part of the explanation of vagueness is made within game theoretic models (Correia & Franke, 2019; Franke & Correia, 2018; Franke et al., 2011; Qing & Franke, 2014). One of the main questions that arise regarding vagueness from a game-theoretic perspective is why we should have vagueness at all, given that, on standard assumptions, vague expressions are sub-optimal communication tools compared to precise ones. A solution proposed by Correia & Franke (2019) and Franke & Correia (2018) is that vagueness can be seen as *boundedly rational* under the hypothesis that agents have an error rate in the calculation of their utilities (See Frazee & Beaver, 2010, for an early version of a similar idea.).

Relatedly, within a Rational Speech Act (RSA) model, Lassiter & Goodman (2017) propose that the interpretation of a vague adjective in context is the result of a balancing of two pressures: "the listener's preference for interpretations which are likely to be true, and the speaker's preference for interpretations that are informative" (Lassiter & Goodman, 2017, p. 3815). To take their example of *tall*, hearers reason about the likely height of an individual, given a description of them as *tall* (as opposed to *not tall*, or no description), and an estimation of a contextual parameter (approximately the threshold, above which, the probability that a speaker would use *tall* instead of *not tall* is greater than 0.5). Although Lassiter & Goodman (2017) do not discuss noise explicitly, they intend their model to be an adaptation of a traditional information theoretic model of communication (Shannon, 1948), and a reasonable interpretation of their model is that part of the noise in the system generates uncertainty about what the threshold value of the adjective is in the context, given a use of that adjective to describe somebody.

To the extent that Correia & Franke (2019), Franke & Correia (2018), Franke et al. (2011), and Qing & Franke (2014) have a speaker-centred approach and Lassiter & Goodman (2017) have a hearer-centred approach, the two analyses are, to some extent, different sides of the same coin. The 'error rate in the calculation of utilities' of the game-theoretic approach could, in Lassiter and Goodman's terms, be seen as attributable to the hearer's speaker model only imperfectly estimating the threshold of the relevant expression in the context, thus introducing uncertainty with respect to the point at which the use of a vague expression will successfully communicate what it is that they intend (such that this uncertainty propagates 'up' to the pragmatic hearer).[5]

2.3 Vagueness and semantic learning

We now turn to the work that proposes an alternative source of vagueness, namely that it arises as a byproduct of semantic learning.

Semantic learning as a source of vagueness. Eijck & Lappin (2012, §5.2) propose that vagueness can be seen as "the residue of probabilistic learning". Rather than noise,

[5]Correia & Franke (2019) also discuss further applications of such models such as how to simulate whether there are evolutionary advantages to vague languages (i.e. to different levels of imprecision employed in game-theoretic strategies), and furthermore which levels of imprecision emerge as dominant within simulated populations.

the idea is that the sample of a language from which a learner must establish how to properly use expressions does not necessarily fully determine their extensions because of 'gaps' in the learning data. Given a domain of entities, ordered with respect to the degree/extent that they instantiate some property (e.g., height, colour etc.), and a witnessed instance of the use of a predicate to describe one of these entities, on the plausible assumption that agents to infer a higher probability of applying the same predicate to comparatively similar entities than to comparatively dissimilar entities (see Decock & Douven, 2014, for a related idea articulated in terms of conceptual spaces), vagueness can be seen as arising from the very process of learning to navigate a semantic space with a limited number of data points (as, indeed, humans must). A similar idea is echoed in (Sutton, 2013, ch. 5) in which it is argued that, given a restricted amount of learning data, a plausible means of modelling how some data points provided stronger reasons than others for forming semantic judgements regarding previously unseen cases is probabilistic reasoning.

Semantic learning and noisy approximation. Fernández & Larsson (2014) propose that expressions such as *tall* are vague as a result of two factors: (i) agents approximate a threshold value (the point at which entities transition from more probably f to more probably not-f), based on a set of witnessed situations (i.e., the learning data) in which the predicate has been used; (ii) when it comes to making judgements, agents do not perfectly track the difference between the threshold value and the relevant properties of the entity being judged (modelled by an error function, a parameter of which is a noise rate[6]). While (ii) bears similarities to some of the proposals regarding noise discussed above, (i) constitutes a different proposal for a source of metalinguistic uncertainty, namely, similar to the idea outlined by Eijck & Lappin (2012, §5.2), an agent's learning data can underspecify what the application conditions for a predicate are. As an example of how these two components work together, if noise is very high, then even if an individual's height is quite far from the threshold value, they will not be judged as being tall with a probability significantly above 0.5. If noise is very low, then even a tiny difference between an individual's height and the threshold will push the probability to 1 or 0.

2.4 Summary: Two sources of uncertainty

A unifying theme in all of the proposals outlined in this section is that vagueness arises as a result of some form of uncertainty being introduced into our cognitive representations. On the one hand, uncertainty can arise due to a source of noise, such as an imperfect mapping between the exact properties in the world an entity may have, and the inexact means we have for representing that property. Alternatively, uncertainty can arise due to a an insufficiency of learning data. As we have seen with Fernández & Larsson's (2014) proposal, for instance, these two sources of uncertainty could both be part of the explanation of why so many natural language expressions are vague. One of the main goals of §3 is to delve deeper into these potential sources of vagueness and the ways in which they interact.

[6]Fernández & Larsson (2014) ground this error rate based upon the empirical findings in (Schmidt et al., 2009).

2.5 The sorites paradox

To end this section, we turn to the sorites paradox. Rather than detailing specific probabilisitic 'solutions' to the paradox, I will outline what I take to be the commonalities between such solutions, as well as the main challenges that they face.

The structure of the paradox. The sorites paradox can be generated for any expression, such that, for some relevant dimension, we have the intuition that a small difference between two entities along that dimension cannot mark a difference in whether to apply that expression. For instance, for *tall*, the dimension is height/vertical size, for *heap*, number of entities in the heap, and for *under the table*, the relative locations of the table and the entity said to be under it. In each case, the intuition is that there are clear cases for applying these expressions, and also clear non-cases. Paradox arises, if we follow the further intuition that any small change along the relevant dimension cannot take us from a clear case to a clear non-case (or even to a non-clear case), and we end up concluding that something that is a clear non-case (or a non-clear case) is a clear case, contrary to our initial assumption.

The argument has two structures, the 'short' and 'long' sorites arguments. In either case, we start with an ordering of entities $D = \langle d_0, d_1, ..., d_n \rangle$ such that for any d_i, d_{i+1} differs only slightly along some relevant dimension for a predicate f (e.g., differs in height by 1mm for *tall*). By assumption, d_0 is clearly f and d_n is clearly not-f. In the long sorites, the argument proceeds by repeated applications of modus ponens based upon *tolerance conditionals* of the form $f(d_0) \rightarrow f(d_1)$, ... $f(d_{n-1}) \rightarrow f(d_n)$. In the short sorites, the tolerance conditional premises are replaced with either a universally quantified conditional $(\forall x[f(x_i) \rightarrow f(x_{i+1})])$ or an inductive premise.

The probabilistic treatment of the paradox: With respect to both versions of the sorites, in a probabilistic setting, the categorical propositions in the argument are interpreted probabilistically. Both assumptions can have an arbitrarily high probability π. The appeal of the long sorites is explained because the probability of each tolerance conditional can be as high as $\pi - \epsilon$, for some small value ϵ. Therefore, each application of modus ponens feels like a good inference, since the drop in probability value between $f(d_i)$ and $f(d_{i+1})$ is so minute. However, over repeated applications of modus ponens, the probability of $f(d_i)$ approaches $1 - \pi$ as i approaches n, therefore the conclusion has a probability of at most ϵ. This explanation of the long sorites captures the slippery slope appeal of the argument, while still diagnosing where we go wrong: inferences in which the conclusion is slightly less probable than the premises are, in general, reliable, but we can go wrong if we string too many of these inferences together.

Challenges for the probabilistic treatment of the paradox: There are at least two main challenges that probabilistic analyses of the sorites face. I take them in turn.
The short sorites: Along with other theories of vagueness such as supervaluationism (Fine, 1975; Kamp, 1975), probabilistic treatments of vagueness face a challenge when it comes to explaining the reason why the premise in the short sorites is no easier to deny than any one of the premises in the long sorites (Edgington, 1997). The basis of the problem is that, unlike any single tolerance conditional, all of which have high probabilities, the universally quantified premise in the short sorites should have a very low probability on the assumption that the probability of a universally quantified statement is the same as

the probability of the conjunction of its instances. Although this problem is not trivial to explain away, one possible explanation could be attributed to a performance error or faulty heuristic in the way we evaluate multiple propositions with high probability values (see Sutton, 2013, ch. 8 for a similar point put in terms of cognitive efficiency.).
Higher-order vagueness: A version of a higher-order vagueness problem arises on a probabilistic approach to the sorites when one is forced to say something about assertion conditions. For instance, if ϕ is assertable iff $Pr(\phi) > \theta$, for some threshold θ, then there should be a sharp cut-off point in the short sorites such that $f(a_i)$ is assertable, but $f(a_{i+1})$ is not. Yet, that result seems to be antithetical to the vagueness of f. For an extensive discussion of this problem and the options probabilistic approaches have for addressing it, see (Sutton, 2018).

3 A probabilistic, iterated learning model

3.1 Overview and hypotheses

In this section, I present the results of simulations that were designed to test two of the common hypotheses that we have seen in the literature:

(H1) Vagueness arises as a result of noise – when an agent hears a vague predicate being used, due to noise, they are unsure exactly what the situation being described is like.

(H2) Vagueness arises as a result of learning based upon incomplete data – when an agent learns a predicate, they do not witness (enough) uses of that predicate. As a result, for some situations, agents are unsure whether to apply the predicate.

Both of these sources of uncertainty can, in principle, result in similar effects: the hearer's semantic representation encodes some uncertainty, and, while there may be canonical uses of the predicate in which there is little uncertainty, at the 'edges', competing predicates' extensions bleed into one another, giving rise to the borderline cases and blurred boundaries that are hallmarks of vague predicates. The goal of the simulations, which were run with a probabilistic iterated learning model for simple artificial languages, is to test whether one, both or neither of (H1) and (H2) is sufficient to derive vagueness in multiple predicates defined over a meaning space.

A further aim of this work, that distinguishes it from the approaches discussed in section 2, is to investigate the meaning spaces over which multiple predicates are defined. In most other approaches, at most two predicates are tested e.g., *tall* vs. *short* or *tall* vs. *not tall*. While this is understandable as a simplifying assumption, it may be beneficial to see what the effects of, e.g., noise, are when the extensions of several predicates are competing with each other for part of the domain.

In the most general terms, the simulations were as follows. We assume a series of generations of agents, one agent per generation, and an ordered set of entities. For example, these entities could be a meaning space of discrete shades of colour between red and yellow (ordered from the red shades through into the orange into the yellow). The first agent has a completely precise, non-vague language. To take the previous example, this would mean a unique predicate for each shade, such that that predicate only applies to entities of that shade of colour. The first agent is then given a set of messages to encode in their language. These signal-message pairs, constitute a sample of the use of

the first agent's language from which the second agent must learn, as best they can given the data, the extensions of each of the predicates to which they are exposed. This learned language of second agent, which may differ from the language of the first agent, is then sampled as learning data for the third agent and so on. The purpose of this set-up is to see under what conditions a vague language arises, i.e., one on which predicates that have both clear cases, clear non-cases, and unclear cases.

The parameter that was used to test (H1) was to introduce noise into the learning data. To take the example of shades of colour, the result of noise is that if an agent is trying to refer to some particular shade of, say orange, with a predicate, f, the learner will be uncertain exactly which shade of colour that use of the predicate is being used to denote.[7] Given, however, that we are assuming an ordered meaning space (typical of that associated with vague predicates), we will assume that the closer as shade of colour is to the intended message, the more likely it is that the learner with extend the extension of f to this shade.

The parameter used to test (H2) controls the size of the *learning bottle neck* (Kirby, 2007; Kirby & Hurford, 2002). With no bottleneck, every learner witnesses at least one use of every predicate from the previous generation's language. As the bottleneck narrows, the probability of witnessing at least one use of every predicate from the previous generation's language gets lower. The effect of the bottleneck is therefore to introduce gaps in the learning data. To take our example of shades of colour once again, suppose that a learner has witnessed a predicate f_1 applied to a shade of red, and a predicate f_2 applied to a shade of orange. Based upon no further data, the agent has no direct evidence for which predicate they should use to describe the shades of reddish-orange, and so must reason whether to use f_1 or f_2, based on what shades f_1 and f_2 have been used to apply to and how close the shades of reddish-orange are to these shades. The result will be that there will be at least on shade of colour that the agent infers is neither clearly f_1 nor clearly f_2, all else being equal.

These two parameters were modulated in order to test four conditions: 1. No noise, no bottleneck; 2. Bottleneck, no noise; 3. Noise, no bottleneck; 4. Bottleneck and noise. In all conditions, the criteria of success was whether the language of the agent in the last generation had the hallmark characteristics of vagueness, which, here, we shall assume to be predicates that have both clear cases, clear non-cases, and unclear classes. These test conditions were designed to test whether noise alone, a bottleneck alone, both noise and an bottleneck, or neither noise nor a bottleneck are sufficient to account for the emergence of a vague language, even when the starting conditions for each run were a completely precise language.

Perhaps surprisingly, these simulation runs suggest that both noise and a bottleneck are needed for vagueness to emerge. In §3.2, I briefly discuss the differences between the iterated learning model used in these tests and others in the literature. In §3.4, the details of the four main test conditions are given. The results are presented in §3.5 and these are discussed in section §3.6. The formal details of the probabilistic iterated learning models are given in the appendix.

3.2 Comparison with other learning models

Iterated learning models. Iterated learning models (ILMs) (Kirby, 2007; Kirby & Hurford, 2002) simulate ways in which a language can evolve over generations of agents.

[7] This treatment of noise differs from that standardly assumed. This issue is further discussed in §3.2

A parameter of these models controls the size of the *semantic bottleneck* (learners may not witness every string or meaning). However, learning is neither probabilistic nor undertaken in noisy conditions: learners receive a set of string-meaning pairs and instantly learn these pairs. Kirby and Hurford's aim was to evaluate whether a language can evolve that is both *expressive* and *stable*. A language is expressive if agents, after the learning phase, are able to provide a string for any meaning that is in the meaning space. A language is stable if each learner ends up with a reasonably similar set of string-meaning pairs as the agent from whose language their training sample is provided. A major finding of these models was that compositionality is required for stable and expressive languages to emerge.

The probabilistic ILM presented here differs in three key respects from those in (Kirby, 2007; Kirby & Hurford, 2002):

i. **An ordering on the meaning domain** (here, the set of situations).[8] The motivation for this ordering relation is to simulate the similarity relations there are for entities in the denotations of vague expressions, e.g., entities that can be ordered from least to greatest heights, or hues that can be ordered from, say, red to yellow.

ii. **Learning is probabilistic.** If an agent is tasked with describing a situation s_i and if they have learned no predicates to describe it, then the agent infers what predicate to use on the basis of the situations close to s_i that they have witnessed being described (with at least some probability).

iii. **Noise can be introduced.** When noise is present in the simulation, agents do not receive, as their learning data, a set of predicate-situation pairs (i.e., string-meaning pairs in Kirby and Hurford's terminology). Rather, if the speaker is tasked with describing a situation s_i and chooses a predicate f_j, the learner receives only a pair of f_j and a distribution over situations with a mean centred around s_i (the standard deviation of this distribution is the noise parameter of the model).

Runs of the simulation were conducted and noise and the width of the semantic bottleneck were modulated in order to test (H1) and (H2) on the basis of whether, after some number of generations, stable, vague languages emerge. *Stability*, like in previous models will de defined in terms of whether there are significant changes between the languages of agents from one generation to another. *Vagueness* will be cast in terms of whether the resulting language has predicates that each have clear cases, clear non-cases and blurred boundaries.[9]

Modelling of Noise. As pointed out by Shalom Lappin (p.c.), this representation of noise differs from that introduced by Shannon (1948). For example, for an ordering of situations (e.g., shades of colours), s_1, s_2, s_3, and a predicate f, on a standard modelling of noise, if the original message-signal pair was $\langle s_2, f \rangle$, then, in line with some stochastic function, the learner receives either $\langle s_1, f \rangle$, $\langle s_2, f \rangle$, or $\langle s_3, f \rangle$. However, if we take seriously

[8] For brevity, I will refer to the denotations of predicates as situations, but these are, more correctly, best thought of as situations of a certain type, namely those that witness some entity with a relevant property such as a size, weight, or shade of colour etc.

[9] *Expressiveness* is not relevant for these probabilistic iterated learning models, since, unlike in non-probabilistic ILMs, agents are able to reason about likely language usage in conditions of uncertainty. Furthermore, compositionality does not factor in these models, since each data point for a learner is assumed to be a simple predication (that for some situation s_i and predicate f_j, s_i is (of type) f_j).

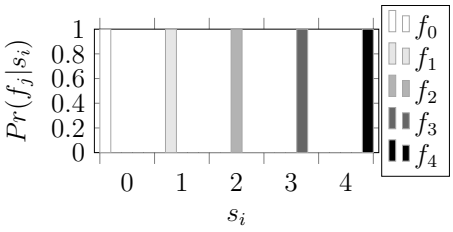


Figure 1: The starting condition (the language of the first agent) for a space of five situations s_0, s_1, s_2, s_3, s_4. There are also five predicates f_0, f_1, f_2, f_3, f_4.For each situation s_i there is exactly one predicate f_j such that $Pr(f_j|s_i) = 1$. In other words, the language is completely precise and predicates have non-graded extensions. This language has a vagueness score of 0.

the idea in Égré 2017, that the relevant notion of noise is one in which the perceptual systems of humans only represent properties with some degree of approximation, then this modelling of noise does not quite fit what we want. Taking our previous example, the idea is that if an object with a shade s_2 is described as being of colour f, then, rather than assuming that a learner represents this as either $\langle s_1, f\rangle$, $\langle s_2, f\rangle$, or $\langle s_3, f\rangle$ based on some stochastic function, a more accurate model would be to assume that the learner associates f with some probability distribution over shades of colour. In other words, for any use of, say, a colour term to refer to an object, with a noisy representation system, we have evidence that it applies to some graded convex portion of the colour space centred around the actual shade of colour of that object. In future studies it would be interesting to see whether a model with a stochastic error based representation of noise would produce different results.[10]

3.3 The model

This section outlines the model in relatively informal terms. Please see the appendix for more details.

Languages. Languages are represented as sets of conditional probability distributions: $Pr(F|S)$ (The probability, for each predicate $f \in F$, and for each situation $s \in S$ that the agent applies f, given s). Samples of languages that provide the learning data for the next agent are probabilistic. First a random sample of situations is generated. The size of this sample is governed by a parameter `data_size` (see below). For each situation s_i in this sample, a predicate is chosen based upon the probability distribution for that language for s_i (from $Pr(f \in F|s_i)$).

Starting conditions. For all simulation runs, the language of the first agent (a_0) is completely precise (without vagueness). In other words, for the first agent, for each situation s_i, there is one predicate f_j such that $Pr(f_j|s_i) = 1$ (when describing a situation, there is never any uncertainty about which predicate applies to that situation), and for each predicate f_j, there is one situation s_i such that $Pr(s_i|f_j) = 1$ (for speakers of this language, when one hears a predicate being used, there is no uncertainty about what situation it describes). In other words, a_0 has enough predicates to singularly describe every type of entity in that space (e.g., a different predicate for every hue).

[10] A further possibility would be to test a combination of both error rate and probabilistic perturbation.

An example of a starting condition language for a space of five situations is given in Figure 1. (Nb., in all of the simulation runs reported below, the starting number of situations and predicates was 15: $S = \{s_0, ..., s_{14}\}$ and $F = \{f_0, ..., f_{14}\}$).

Parameters. The parameters of the model that govern the four test conditions are `data_size`, `delta`, and `noise`:

`data_size` The number of situation-predicate pairs a learner witnesses. When `data_size` is high compared to $|S|$, there is little or no bottleneck. As `data_size` gets closer to the size of the situation space, the bottleneck narrows.

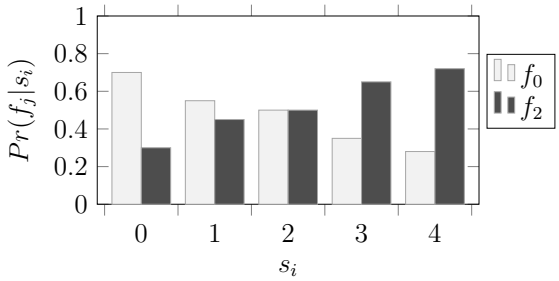


Figure 2: A possible ending condition for a situation space s_0-s_4. The depicted language has two predicates f_0, f_2. Both predicates fail conditions (3a) and (3c) and so the vagueness score is 0: neither predicate is vague.

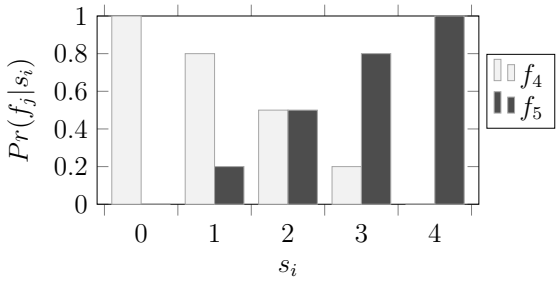


Figure 3: A possible ending condition for a situation space s_0-s_4. The depicted language has two predicates f_4, f_5. Both predicates meet all conditions in (3) and so the vagueness score is 1: both predicates are vague.

`delta` This parameter controls reasoning about gaps in the learning data. I.e., for an extension gap G between two predicates f_i, f_j, how quickly $Pr(f_i|s \in G)$ and $Pr(f_j|s \in G)$ approach 0.5. See the appendix for details.

`noise` This parameter controls the level of noise in the learning data. If the intended message is the pair of a situation s_i and a predicate f_j, then the noisy data is a normalised discrete probability distribution $Pr(s_k \in S|f_j)$ with a mean for $Pr(s_i \in S|f_j)$ and a standard deviation of `noise`.

Criteria of success. For each run of each test condition, the languages of the final 20 generations were checked for stability, namely, whether the number of predicates was stable. The main criterion of success was whether for each run of each test condition, the *ending condition* (the final language of this run) was vague. Vague predicates, as is defined here, are predicates that have both clear cases, clear non-cases, and non-clear classes. The criteria for this was as follows. For each run, for each predicate f_j:

(3) a. There is a clear case for f_j: $\exists s_i.Pr(f_j|s_1) > 0.8$

Table 1: The four different simulation condition types for runs of the model. For all conditions, the initial number of predicates and situations in the language of the first agent was set to 15 and simulations were run for 350 generations.

Simulation conditions	No Bottleneck	Bottleneck
No Noise	1.	2.
Noise	3.	4.

b. There is a non-clear case for f_j: $\exists s_i.0.2 \leqslant Pr(f_j|s_1) \leqslant 0.8$

c. There is a clear non-case for f_j: $\exists s_i.Pr(f_j|s_1) < 0.2$

The two principal ways in which predicates failed to meet these conditions were, first, they were precise and so had no non-clear cases (see, e.g., Figure 1). I.e., (3b) was false. Second, they could fail to have clear positive cases at all (i.e., (3c) was false). An example of such an outcome is given in Figure 2.

This criterion was quantified as a score in the range $[0, 1]$: the proportion of all predicates in all runs that were vague (that satisfy the conditions in (3)). This will be referred to as the *vagueness score* for a setting of the model. An example of a possible resulting language with a vagueness score of 1 is given in Figure 3).

(4) ***Vagueness score:*** For a setting of the model, the proportion of vague predicates that result on average. Where n is the sum of the number of predicates in all ending conditions, and m is the number of these predicates that satisfy the conditions in (3), the vagueness score is m/n.

3.4 Test conditions

The four types of simulation conditions are given in Table 1. For all conditions, the number of predicates and situations in the initial language was set to 15 ($S = \{s_0, ..., s_{14}\}$ and $F = \{f_0, ..., f_{14}\}$). The number of generations was set to 350. A total of 258 simulation runs were conducted with 43 different parameter settings (6 runs per condition).

Condition 1. is a control condition where the expectation is that the final agent has the same precise language as the first agent. The `noise` parameter was set to 0.01, which, in computational terms, amounts to zero noise.[11] Runs of the simulation were tested with `data_size` values of 150 and 200 (both at least 10 times greater than the number of predicates and situations in the initial language, 15).

Condition 2. tests whether the bottleneck alone can yield a stable, vague language. The `noise` parameter was set to 0.01. The variables in Condition 2 were `data_size` and `delta` with runs for `data_size` values of 45, 60 and 75, each with `delta`-values of 0, 0.5, and 1.[12]

[11]For a situation distance 1 away from the situation in the initial data, the probability of the relevant predicate applying to this situation was so small, that it was treated by the computer as zero.

[12]In pre-testing, wider bottlenecks (i.e. lower `data_size` values) tended to lead to languages collapsing into single-predicate languages.

Table 2: Tables of results for the vagueness scores in Conditions 1-4: 'No bottleneck, no noise', 'Bottleneck, no noise', 'Noise, no bottleneck', and 'Noise and bottleneck'. Vagueness scores are the proportion of vague predicates to feature in the final language across all simulation runs for each setting of the model. Aside from the the 'No bottleneck, no noise' (control) condition 1, the least successful condition is 'Bottleneck, no noise' (Cond. 2), closely followed by the 'Noise, no bottleneck' Condition 3. The highest vagueness scores were produced in the 'Noise and bottleneck' Condition 4.

(a) Condition 1: No bottleneck, no noise. The effect of `data_size` on the vagueness score.

	`delta` = 1
`data_size` = 150	0.00
`data_size` = 200	0.00

(b) Condition 2: Bottleneck, no noise. The effect of `data_size` and `delta`-value on the vagueness score.

	`delta` = 0	`delta` = 0.5	`delta` = 1
`data_size` = 45	**0.62**	0.27	0.29
`data_size` = 60	0.38	0.09	0.19
`data_size` = 75	0.00	0.19	0.10

(c) Condition 3: Noise, no bottleneck. The effect of `data_size` and `noise`-value on the vagueness score.

	`noise` =	0.25	0.275	0.3	0.4
`data_size` = 150		0.28	**0.68**	0.57	0.12
`data_size` = 200		0.35	0.33	0.20	0.03

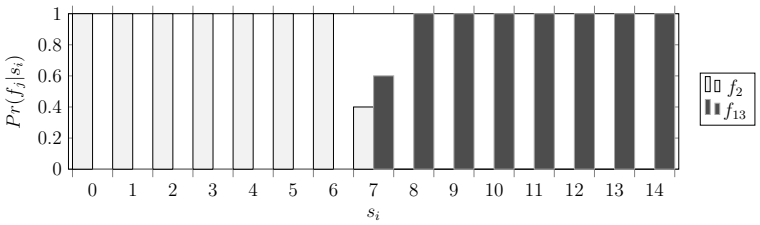
(d) Condition 4: Noise and bottleneck. The effect of `data_size` and `noise`-value on the emergence of vague predicates. Results given for `delta` = 0.5/1.0

	`noise`	0.30	0.35	0.40
	`delta`	0.5/1.0	0.5/1.0	0.5/1.0
`data_size` = 45		0.50/0.46	0.67/0.60	0.67/0.44
`data_size` = 50		0.43/0.50	0.73/0.75	0.80 /0.25
`data_size` = 60		0.31/0.40	**1.00/0.85**	0.73/0.62
`data_size` = 75		0.41/0.72	0.67/0.77	0.56/0.80

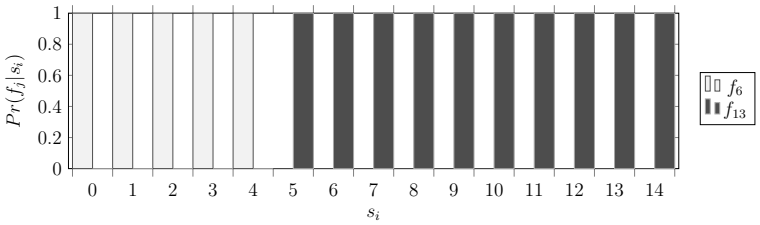
Condition 3. tests whether noise alone can yield a stable, vague language. Runs were conducted with `data_size` values of 150 and 200. `delta` was kept fixed at 1. The variable for these runs was the `noise` value. Runs were conducted for `noise` values 0.25, 0.275. 0.3 and 0.4.

Condition 4. combines noise and the bottleneck. Runs were conducted with `data_size` values of 45, 50, 60, 75, and 90. For each of these data sizes, noise values of 0.3, 0.35 and 0.4 were tested. Simulations were run with `delta`-values of 0.5 and 1.0.

Figure 4: Condition 2: Bottleneck, no noise. There were two types of typical ending conditions when `data_size` = 45, `delta` = 0 (when there was a bottleneck, but no noise).



(a) First type of ending condition: A language with two predicates, in this run it was f_2 and f_{13}. Both predicates have sharply graded boundaries. In this case, only s_7 is a borderline case.



(b) Second type of ending condition: A language with two predicates, in this run it was f_6 and f_{13}. Neither predicate is vague.

3.5 Results

Condition 1: No noise, no bottleneck. All runs in the `data_size` = 200 and all but one in the `data_size` = 150 condition produced the same results, namely that the language of generation 350 was identical to the, non-vague language: 15 predicates, each of which uniquely identifies one situation. One run resulted in 14 non-vague predicates, one with an extension covering two situations. The vagueness scores for these settings are given in Table 2a. (All runs scored 0.) As expected, in this control condition, no vagueness arose when there is no noise and no bottleneck.

Condition 2: Bottleneck, no noise. With these settings, a number of trends emerged (see Table 2b). First, in general, the lower the `data_size`, the higher the vagueness score was. Second, when data size was low (45 or 60), a low `delta` value increased vagueness score. The highest result over 6 runs was for `data_size` = 45, `delta` = 0 resulting in a vagueness score of 0.62. This is not a terrible result, however, the resulting languages had only very narrow boundaries. Two typical examples of the resulting languages are given in figure 4. A further effect of the bottleneck was that the end condition language had a smaller vocabulary (2-3 predicates) than the starting condition (15 predicates).

Condition 3: Noise, no bottleneck. Perhaps surprisingly, the results in this condition were somewhat chaotic, even with relatively low levels of noise such as 0.25 or 0.275. Vagueness scores were not significant improvements on Condition 2, and in many runs, especially with the higher `data_size` of 200, other trends emerged such as partial synonymy and predicates with non-convex extensions.

When the `data_size` was 150, the best vagueness score arose at the relatively low noise level of 0.275. (See Table 2c). However, at 0.68, this vagueness score is only marginally

Figure 5: Condition 3: Noise, no bottleneck. The typical end condition for runs with these settings (`data_size` = 150, `noise` = 0.275) was a language that displayed displayed some vagueness. For example, predicate f_1 below is vague. However, this model setting also resulted in non-convex predicates such as f_9 and f_{11}: f_9 has an extension gap in s_8 and s_9 and f_{11} has an extension gap in s_{10}.

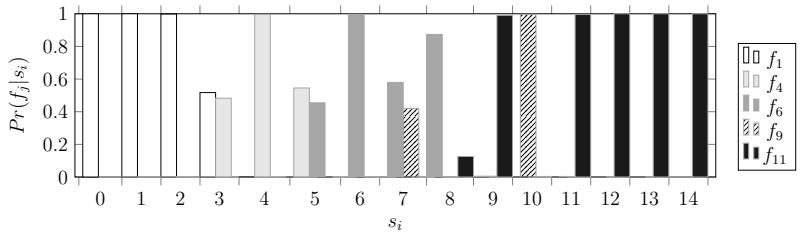
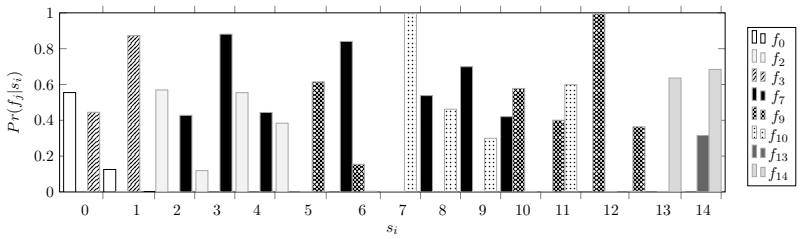


Figure 6: Condition 3: Noise, no bottleneck. The typical end condition for this setting of the model (`data_size` = 200, `noise` = 0.275) was a language with some vagueness. For example, predicates f_3, f_9 and f_{10} below count as vague. However, other predicates failed to be vague because they lack clear cases. For instance predicates f_0, f_2 and f_{14}. Languages also had non-convex predicates such as f_7 and f_9 below.



better than the best score in the bottleneck, no noise condition (0.62). In other words, like in the best settings for Condition 2., more often than not, predicates had blurred boundaries. However, in this condition, unlike in Condition 2., there were cases of non-convexity (where a predicate has an extension gap filled by another predicate), see Figure 5. In the worst runs, something evocative of partial synonymy arose, where two predicates were competing for similar parts of the situation space. Under the success criterion, these predicates were not classed as vague, since they failed to have clear cases in the meaning space.

When the data size was increased to 200, the vagueness scores decreased in all but the lowest noise level settings. The maximum score was only 0.33 (only one third of predicates met the vagueness criteria, see Table 2c). A further difference that arose with the increase of `data_size` to 200, was that there were typically more predicates in the final languages (an average of 7.4 predicates per run as opposed to an average of 5.7 when `data_size` was 150). It is possible, therefore, that the decrease in vagueness score was due, in part, to an overcrowding the situation space. In this setting, more non-convex predicates resulted as we see for f_7 and f_9 in Figure 6, for instance (f_7 has an extension gap in s_5 and s_7, and f_9 has an extension gap in s_7, s_8 and s_9). Furthermore, once again, something evocative of partial synonymy often arose, where two predicates were competing for similar parts of the situation space. Under the success criterion, these predicates were not classed as vague, since they failed to have clear cases in the meaning space. For example, f_0, f_2, f_{13}, and f_{14} in Figure 6 failed to meet this criteria, since none exceeded a probability of 0.8 in any situations. In other words, unlike the results in Condition 2 (bottleneck, no noise), lower scores resulted not from many sharp, non-vague predicates, but from many predicates that had only non-clear cases or clear non-cases.

In summary, when there was noise and no bottle neck, providing agents with too much noisy learning data seems to give rise to some vagueness, but also a lot of uncertainty where, for large swaths of the situation space no one term is inferred to apply. For `noise` values above 0.275, the resulting languages were chaotic, so much so that they were often hard to interpret. Interestingly, the introduction of noise decreased the vocabulary size of end condition languages (typically 5-8 predicates) compared to the starting condition (15 predicates). The decrease in vocabulary size was smaller than in condition 2, however.

Condition 4: Bottleneck and noise. This condition yielded the most successful results. Not only was there a combination of settings that yielded resulting languages where every predicate had graded boundaries (see Table 2d), but on all settings, the average values generated were higher than in any of the other conditions. This sometimes resulted in what could be described as a prototypical image of a system of vague predicates as we see in the three-predicate language in figure 7. As in Condition 2., the vocabulary size of the ending condition of languages was smaller with a lower `data_size`. For instance, when `data_size` was 45, typically, final languages had just two predicates, see Figure 8.

The effect of `delta` on the results in Condition 4 was unclear, although, in general, for higher `noise` values, the lower `delta` setting of 0.5 yielded slightly higher vagueness scores than when it was set to 1.0.

Interestingly, if we compare the results for this condition with those for noise and no bottleneck (Condition 3), we see that, not only were vagueness scores higher, but, in general, there is a greater tolerance for noise in the system when a bottle neck was also present. In Condition 3, the best results were obtained with `noise` set to 0.275, with results

Figure 7: Condition 4: Noise and Bottleneck. The typical end condition for this setting of the model (`data_size` = 60, `delta` = 1.0, `noise` = 0.35) was a language with relatively few predicates, all of which were vague. For example, the final language depicted below has three predicates f_4, f_8, and f_{10} all of which have clear cases, non clear cases and clear non-cases.

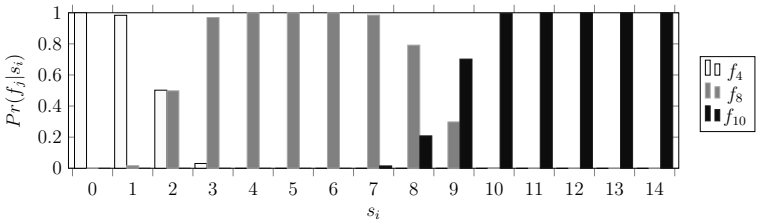
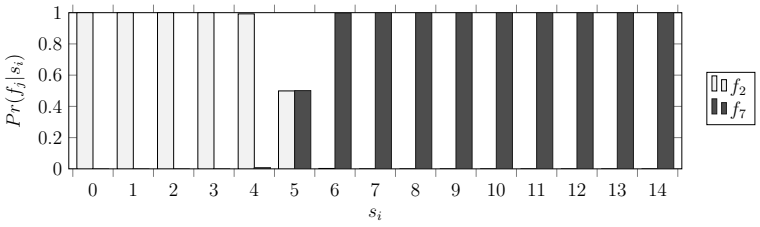
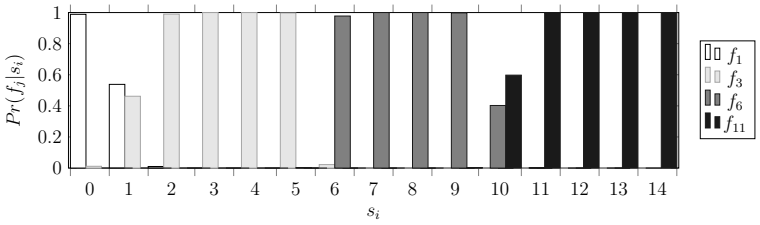


Figure 8: Condition 4: Noise and bottleneck. The typical end condition for this setting of the model (`data_size` $\in \{45, 75\}$, `noise` = 0.3) was a language with 2-5 predicates, all of which were vague. The vocabulary size of the languages increases with `data_size`.

(a) When `data_size` = 45, `delta` = 1.0, the typical end condition was a language with two predicates (f_2 and f_7 below), both of which were vague.



(b) When `data_size` = 75, `delta` = 1. , the typical end condition was a language with three or four predicates (the language below has four predicates f_1, f_3, f_6 and f_{11} below), all of which were vague. In the end point of the run depicted below, there is a sharp boundary between predicates f_3 and f_6. However, there is a graded boundary between predicates f_1 and f_3 and between predicates f_6 and f_{11}.



getting worse for higher `noise` settings such as 0.3 and 0.4. However, in Condition 4, results peaked with with a `noise` setting of 0.35. With with a `noise` setting of 0.40, there was an increased tendency for languages to collapse into single predicate languages, an effect that arose to a lesser extent when `noise` was set to 0.35. In fact, the lower figures in Table 2d for `noise` = 0.35 can be attributed almost entirely to this occurrence. For the lower noise setting of 0.3, the low vagueness scores in Condition 4 arise because, although predicates have graded boundaries, the gradience of these boundaries was too sharp to fall into the definition used here.

3.6 Discussion

The most interesting upshot of these simulations is that, although a semantic bottleneck alone or noise alone can generate some vagueness, a combination of noise and a semantic bottleneck produced the best results. In other words, a combination of (H1) and (H2) best explains the emergence of vagueness. This can be seen in the contrast between Tables 2b and 2c, on the one hand and Table 2d on the other. With both noise and a bottleneck, 50% (12/24) of the different parameter settings resulted in vagueness scores > 0.66 and the mean vagueness score across all parameter settings was 0.61. For the noise only condition only 12.5% (1/8) of the settings exceeded a vagueness score of 0.66. The mean vagueness score in this condition was 0.32. For the bottleneck only condition, no settings resulted in a vagueness score above 0.66 and the mean score was 0.24. This result does, however, raise some questions.

Why is a semantic bottleneck alone insufficient? In order to generate predicates with graded boundaries as a result of a semantic bottleneck, there need to be gaps in the learning data (of at least one situation). However, for this to occur frequently enough, the size of the learning data needs to be small. As we see in Table 2b, for the bottleneck-only condition, the highest vagueness scores occurred with a `data_size` parameter value of 45. Vagueness scores dropped to nearly zero when this parameter was set to 75. However, there is a lower limit on how low the data size can be. With a value lower than 45, the simulation runs collapsed into trivial one-predicate languages. Therefore, a semantic bottleneck alone can only generate some amount of vagueness, since narrowing the bottleneck further leads to learning a trivial language.

Why is a noise alone insufficient? Why do noise and a semantic bottleneck together do better? The reason why noise alone is insufficient is less clear. When the `noise` was set too low (at 0.25), predicates were not vague because they had sharp boundaries. When noise was increased, predicates were not vague because they did not have clear cases. One observation that may explain this, and also why a combination of noise and a bottleneck does better, is that the noise-only condition resulted in languages with significantly higher vocabulary sizes than in either the bottleneck only or noise and bottleneck conditions (2-4 predicates vs 5-9 predicates), i.e., that one effect of the semantic bottleneck is to constrain the vocabulary size of languages. This observation suggests the possibility that lower vocabulary sizes allow for more noise in the system. In other words, there seems to be an inverse relationship between how many vague predicates can be accommodated within a semantic space, on the one hand, and how much noise can give rise to this vagueness on the other: For a fixed situation space, the more predicates one has, the less noise there can be without these predicates failing to have any clear cases. Therefore, one way in which we can increase noise, and thereby the gradedness of predicate boundaries, is to constrain the size of the vocabulary independently of the noise level. As we saw from the bottleneck-alone condition, narrowing the bottleneck is one way to achieve a reduction on vocabulary size, and so it makes sense that a combination of a bottleneck and noise can allow for more vagueness.

If these conclusions are correct, then we have evidence that the introduction of the semantic bottleneck plays two roles in the emergence of stable vague languages: First, as we saw from the bottleneck-only Condition 2, in line with Hypothesis (H2), a semantic bottleneck paired with probabilistic reasoning about data gaps can contribute to the mergence of vagueness. Second, an unexpected effect was that the bottleneck, in virtue of constraining the vocabulary size of a language, facilitates more tolerance of noise within

the system, noise which, in turn, contributes to the emergence of vagueness. An obvious question this raises is whether one can separate out these two contributions of the semantic bottleneck. I address this in §3.7.

3.7 Possible criticisms of and improvements to the model

Finally, let us consider some possible objections to the model as well as ways potential ways in which the model could be refined.

An over-simplified view of communicative need. The model presented here assumes that, for each generation, the list of situations that a speaker is tasked with describing to a learner is drawn (pseudo)-randomly from a list of the situations in the situation space. However, plausibly, this assumption is an oversimplification. For example, it is not obvious that the colours and hues we encounter in the world are distributed evenly, or that, even if they were, we would be equally likely to encounter descriptions of each of them (either evolutionary or social factors might make some hues more prominent and/or likely to be described by others). Therefore, the model could be adjusted to allow for the list of situations to be described to be drawn from a non-flat distribution. It is possible, with this adjustment, that some of the more chaotic tendencies we saw in the noise-only condition and in the higher noise settings of the noise and bottleneck condition may be reduced.

An over-simplified model of communication. One way in which the probabilistic iterated learning model presented above is a refinement of non-probabilistic models is that it does not treat learning as trivial (e.g., the probabilistic model does not assume that a learner can learn an expression from one instance of a form-meaning pair). However, the model of communication in this probabilistic iterated learning model was trivial insofar as a learner received, simultaneously, all of their learning data. Arguably, one way in which the meanings of expressions can become more self-regulated is via interaction and corrective feedback loops between agents (Kempson et al., 2019). But the model presented above has no such mechanisms for coordination and interaction between interlocutors.

An over-simplified speaker (and hearer) model. The speaker and hearer models we assumed were effectively what amount to 'literal hearers' and 'literal speakers' in the Bayesian pragmatics literature (see, e.g., Lassiter & Goodman, 2017, and references therein). For example for an agent a_j with a vague language, given some situation to describe s_i, the predicate chosen by the agent was simply a sample from the distribution $Pr_{a_j}(f_j \in F|s_i)$. Arguably, however, this is too simple insofar as if the speaker is taking into account what the learner may infer from the use of one predicate over another, given the alternatives available, then, given a situation s_i, this may make the probability that a speaker will use a predicate f_j much lower than $Pr_{a_j}(f_j|s_i)$, especially if for some other predicate f_k, $Pr_{a_j}(f_k|s_i)$ is much higher than $Pr_{a_j}(f_j|s_i)$. This is because, for example, a hearer would be more likely to infer that the speaker does not intend to convey s_i, but rather a more paradigm case of f_j.

Relatedly, as pointed out by Shalom Lappin (p.c.), when learning multiple predicates, it can be advantageous to filter out inferences that have low probabilities on the basis of the learning data. A threshold-based filter plus renormalisation of likelihoods inferred from the learning data could be incorporated into future models. Plausibly, this would reduce

occurrences of non-convexity in the noise, no bottleneck condition (Condition 3.), since in this condition, relatively large data sizes over many generations of agents makes the occurrence of at least some predications in the learning data based on very low conditional probabilities, quite likely.

Hard-wiring a limit on the size of the language. Since the model presented above was designed such that the language of a_0 was maximally precise, given the size of the situation space, the model was not flexible enough test a starting condition in which a large situation space is covered by relatively few precise predicates (a hard-coding of the downwards pressure on vocabulary size). It is possible that, were this to be tested, the noise-only and bottleneck-only testing conditions may have produced more concordant results.

A vague starting condition. Finally, since human languages, we may assume, never had a non-vague starting point, the model does not resolve what factors, given a vague language as a starting condition, contribute, as stable attractors, to such vagueness persisting across generations.[13] The results of these simulations do suggest that noise and a semantic bottleneck do not prevent vagueness in a language from persisting across generations in the iterated learning model, however, it would certainly be interesting, in future work, to compare the outcomes of a vague and non-vague starting condition.

4 Conclusion

In relation to the literature discussed in section 2, this simulation study provides some evidence in favour both of accounts that posit some form of noise as an underlying factor in why natural language predicates are vague, and those that attribute at least part of the explanation to the residue of semantic learning (see §3.6). In other words, the hypotheses that best account for the vagueness are a combination of (H1) and (H2). More broadly, given that both the bottleneck and noise generate conditions of uncertainty in language learning, we have evidence that vagueness arises as a result of reasoning in these conditions of uncertainty, a claim that virtually all probabilistic accounts of vagueness make in one form or another.

However, this study also uncovers something that has received less notice in the literature. As remarked in the introduction to section 3, a common simplifying assumption in previous approaches is to consider minimal pairs such as *tall* and *short* (possibly with the addition of a 'say nothing' alternative as in (Lassiter & Goodman, 2017)) . However, when more predicates are involved, an extra crucial ingredient for stable vague languages to emerge is some kind of downwards pressure on vocabulary size (see §3.6). This chimes well with Zipf-inspired information theoretic analyses of the origins of ambiguity. For instance, Piantadosi et al. (2011) propose that the balancing of two pressures: *clarity* and *ease*, which can be paraphrased as pressures toward a larger vocabulary, which maps straightforwardly onto only one or a few meanings on the one hand, and a smaller vocabulary, the tokens of which are, for example, more frequent and easier to access.[14] The

[13] Thank you to Daniel Lassiter (p.c.) for raising this issue.

[14] An important part of Piantadosi et al.'s (2011) proposal is that a prima facie decrease in clarity resulting from a smaller vocabulary can be assuaged if context reduces uncertainty about what is meant. In the context of vagueness, this idea is likely also to be highly relevant when considering the role of context in the interpretation of vague predicates.

simulations presented here suggest that the semantic bottleneck may be one source of a pressure towards *ease* (i.e., a smaller vocabulary size), and that this pressure contributes to the emergence of vague, stable languages.

5 Further reading

- Early foundations of probabilistic approaches to vagueness and a discussion thereof: (Black, 1937; Borel, 1907/2014; Égré & Barberousse, 2014).
- A more philosophically oriented proposal from Edgington based on degrees of closeness to clear cases of truth that has influenced several of the papers cited above (Edgington, 1997, 1992). See also a review of probabilistic approaches to vagueness from a more philosophical perspective (Sutton, 2018).
- Extension of Rational Speech Act approaches to morphological and lexical negation for expressions such as *happy* (Tessler & Franke, 2018), to probability expressions such as *likely* (Herbstritt & Franke, 2019), and to generics (Schuster & Degen, 2020; Tessler & Goodman, 2019).
- Probabilistic pragmatics applied to quantifiers (Tiel et al., 2021) (but see also (Emerson, 2020) for a distributional approach); estimating thresholds for vague quantifiers (Schöller & Franke, 2017).
- Collections of papers on vagueness: (Dietz & Moruzzi, 2009; Keefe & Smith, 1997; Nouwen et al., 2011)
- A handbook article that places probabilistic approaches to vagueness in the context of other theories: (Burnett & Sutton, 2020).

Appendix: The probabilistic iterated learning model

This probabilistic iterative learning was coded in Python.[15] For ease of presentation, the model is represented below in mathematical and set-theoretic notation.

Preliminaries: The model assumes the following sets. F, S, and A are taken as primitives. L, I, and D, are defined further below:

$F = \{f_0, ..., f_n\}$	A set of predicates
$S = \langle s_0, ..., s_n \rangle$	An ordered set of situations
$A = \langle a_0, ..., a_n \rangle$	An ordered set of agents
$L = \{l_{a_0}, ..., l_{a_n}\}$	A set of languages, one for each agent
$I = \{i_{a_0,a_1}, ..., i_{a_{n-1},a_n}\}$	A set of intended data, one for pair of agents where $i_{a_n,a_{n+1}}$ is data a_{n+1} intends to convey to a_n
$D = \{d_{a_0,a_1}, ..., d_{a_{n-1},a_n}\}$	A set of learning data, one for pair of agents where $d_{a_n,a_{n+1}}$ is the learning data a_{n+1} receives from a_n.

[15] A link will be added here to a GitHub repository containing the code and the results from simulation runs etc.

Parameters: The following are parameters within the model. The parameters relevant for testing the hypotheses (H1) and (H2) are `data_size`, `noise`, and `delta`:

`NumPredSit`	the cardinality of F and S (always identical for a_0)
`NumGen`	the cardinality of A
`data_size`	the number of samples given to a_i from the language of a_{i-1}
`noise`	the degree to which an agent a_i can discern which situation a_{i-1} is describing with a predicate. This governs how precisely intended datasets are mapped to learning datasets.
`delta`	When there is a gap in the learning data set (when a learner has no direct information regarding what predicate to apply to a situation), the learner must reason about what predicate to use. This parameter affects how strongly witnessed uses of 'near-by' predicates influence this reasoning.

A language for an agent, a_k is a set of (posterior) conditional probability distributions $\{Pr_{a_k}(f_i|s_j) : f_i \in F, s_j \in S\}$. For agents $a_{i>0}$ these posterior probabilities are calculated from priors and the learning data (which provide the likelihoods) via Bayes' Rule. In cases where the learning data is 'gappy' and so provides no information about at least one situation in S, this gap is filled on the basis of reasoning about how best to extend the extensions of predicates to this/these situations.

Language for $\mathbf{a_0}$ is a set of conditional probabilities $Pr_{a_0}(f_i \in F|s_j \in S)$ (for each predicate and situation, the probability that the agent will use that predicate to describe that situation). The number of situations and predicates for a_0 is always the same. $Pr_{a_0}(f_i|s_j) = 1$ if $i = j$ and $Pr_{a_0}(f_i|s_j) = 0$ if $i \neq j$. See Figure 1.

Intended data. For each a_i, a_{i+1}, a set of tuples $\{\langle s_i, f_j\rangle : s_j \in S, f_i \in F\}$. The situations $s \in S$ are randomly generated. The size of this list is governed by the parameter `data_size`. The predicate in the tuple is derived from the language of a_i. The intended data is akin to what learners receive in non-probabilistic iterated learning models in which learning is trivial (i.e., the intended message is always paired with the signal). In this probabilistic model, learners do not receive the intended data, but rather the learning data which may be modulated by noise.

Learning data (likelihoods) are, for an agent a_k probability distributions $Pr_{a_k}(s_i \in S|f_j \in F, i_{a_{k-1},a_k})$, calculated from intended data with the addition of noise. Noise is governed by the `noise` parameter. For each member of the intended data set $t_k = \langle s_i, f_j\rangle$, we derive a distribution $Pr_{a_k}(s_i \in S|f_j, t_k)$ based on a normal distribution with a mean of s_i and a standard deviation of `noise`.

Example: suppose that for a space of three situations s_0, s_1, s_2, a member of the intended data set is $t_1 = \langle s_1, f_1\rangle$. If `noise`$= 0.5$, then for a Gaussian function $\mathfrak{f}$, the corresponding

member of the learning data set is a probability distribution:

$$\begin{array}{lcl} Pr(s_i|f_1,t_1) & = & \frac{\mathfrak{f}(s_i|\mu=s_1,\sigma^2=0.5^2)}{\sum_{s_j\in S}\mathfrak{f}(s_j|\mu=s_1,\sigma^2=0.5^2)} \\ Pr(s_0|f_1,t_1) & \approx & 0.11 \\ Pr(s_1|f_1,t_1) & \approx & 0.78 \\ Pr(s_2|f_1,t_1) & \approx & 0.11 \end{array}$$

If a predicate, f_j is witnessed more than once in the intended dataset i then the learning data relating to that predicate is calculated as the normed average of $Pr(s_i \in S|f_j, t \in i)$, such that $t \in i$ is the set of tuples that witness f_j.

Example: Suppose following the first intended data point above, the second intended data point is $t_2 = \langle s_2, f_1\rangle$. This gives us:

$$\begin{array}{lcl} Pr(s_0|f_1,t_2) & \approx & 0.00 \\ Pr(s_1|f_1,t_2) & \approx & 0.12 \\ Pr(s_2|f_1,t_2) & \approx & 0.88 \end{array}$$

The normed average over both of these data points is:[16]

$$\begin{array}{lcl} Pr(s_i|f_1,t_1,t_2) & = & \frac{Pr(s_i|f_1,t_1)+Pr(s_i|f_1,t_2)}{|\{t_1,t_2\}|} \\ Pr(s_0|f_1,t_1,t_2) & \approx & 0.055 \\ Pr(s_1|f_1,t_1,t_2) & \approx & 0.45 \\ Pr(s_2|f_1,t_1,t_2) & \approx & 0.495 \end{array}$$

Priors for situations are calculated from the learning data:

$$Pr_a(s_i) = \frac{\sum_{f_j\in F} Pr_a(s_i|f_j,i)}{\sum_{s_k\in S, f_j\in F} Pr_a(s_k|f_j,i)}$$

Priors for predicates are also calculated from the learning data. Each item in the learning data consists of a pair of a predicate and a probability distribution over situations:

$$Pr_a(f_i) = \frac{\text{Number of instances of } f_i \text{ in the learning data for } a}{\texttt{data\_size}}$$

Languages for $\mathbf{a_{k>0}}$ are sets of conditional probabilities $Pr_{a_k}(f_i \in F|s_j \in S)$ (for each predicate and situation, the probability that the agent will use that predicate to describe that situation). Initial values for $Pr_{a_k}(f_i \in F|s_j \in S)$ are calculated from the learning data and priors using Bayes' theorem. If there are no situations with a prior probability of 0, these initial values are used as the language of the agent.

If there are situations with a prior probability of 0 (i.e., when some situation s_i, does not occur in the intended data set due to the semantic bottleneck), the agent reasons about which predicate to apply with what probability based on their learning datasets. This reasoning is governed by the value of the `delta` parameter and a distance measure between situations in the ordered tuple.

[16]The reason why s_2 gets a slight boost over s_1 is because s_2 in this toy example is at an end point in the ordering of S.

Example: Suppose we have a case where the priors for s_1 and s_2 are zero, and so we only have a partial language:

$Pr(F\|S)$	s_0	s_1	s_2	s_4
f_0	1	–	–	0
f_1	0	–	–	1

For $s_i \in \{s_1, s_2\}$, the values for $Pr(f_1|s_i)$ and $Pr(f_2|s_i)$ are calculated from the values for $Pr(f_0|s_0)$, $Pr(f_0|s_4)$, $Pr(f_1|s_0)$ and $Pr(f_1|s_4)$ along with a distance measure between situations and the `delta` value. Where $Dist(s_i, s_j) = abs(i - j)$,

$$Pr(f_0|s_1) \propto exp(ln(Pr(f_0|s_0) - (\texttt{delta} \times Dist(s_1, s_0))$$

If similar calculations are made for $Pr(f_0|s_2)$, $Pr(f_1|s_1)$ and $Pr(f_1|s_2)$, the above case yields:

`delta` = 1

$Pr(F\|S)$	s_0	s_1	s_2	s_4
f_0	1	0.73	0.27	0
f_1	0	0.27	0.73	1

`delta` = 0.5

$Pr(F\|S)$	s_0	s_1	s_2	s_4
f_0	1	0.62	0.38	0
f_1	0	0.38	0.62	1

`delta` = 0

$Pr(F\|S)$	s_0	s_1	s_2	s_4
f_0	1	0.50	0.50	0
f_1	0	0.50	0.50	1

References

Barker, Chris. "The Dynamics of Vagueness". *Linguistics and Philosophy* 25.1 (2002), pp. 1–36.

Bernardy, Jean-Philippe, Rasmus Blanck, Stergios Chatzikyriakidis & Shalom Lappin. "A Compositional Bayesian Semantics for Natural Language". *Proceedings of the First International Workshop on Language Cognition and Computational Models*. Ed. by Manjira Sinha & Tirthankar Dasgupta. 2018, pp. 1–10.

Bernardy, Jean-Philippe, Rasmus Blanck, Stergios Chatzikyriakidis, Shalom Lappin & Aleksandre Maskharashvili. "Bayesian Inference Semantics: A Modelling System and A Test Suite". *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*. 2019, pp. 263–272.

Black, Max. "Vagueness. An Exercise in Logical Analysis". *Philosophy of Science* 4.4 (1937), pp. 427–455.

Borel, Émile. "An Economic Paradox: The Sophism of the Heap of Wheat and Statistical Truths". *Erkenntnis* 79(Suppl. 5) (1907/2014), pp. 1081–1088. ISSN: 1572-8420.

Burnett, Heather & Peter R. Sutton. "Vagueness and Natural Language Semantics". *The Wiley Blackwell Companion to Semantics*. American Cancer Society, 2020, pp. 1–34. ISBN: 9781118788516. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118788516.sem053`.

Cooper, Robin, Simon Dobnik, Staffan Larsson & Shalom Lappin. "Probabilistic Type Theory and Natural Language Semantics". *LILT* 10.4 (2015).

Correia, Jose Pedro & Michael Franke. "Towards an ecology of vagueness". *Vagueness and Rationality*. Ed. by Richard Dietz. Berlin: Springer, 2019, pp. 87–113.

Decock, Lieven & Igor Douven. "What Is Graded Membership?" *Noûs* 48.4 (2014), pp. 653–682.

Dietz, Richard & Sebastiano Moruzzi, eds. *Cuts and Clouds. Vaguenesss, its Nature and its Logic*. Oxford University Press, 2009.

Edgington, Dorothy. "Vagueness by Degrees". *Vagueness: A Reader*. Ed. by R. Keefe & P. Smith. Cambridge, MA: MIT Press, 1997, pp. 294–316.

— "Validity, Uncertainty and Vagueness". *Analysis* 52.4 (1992), pp. 193–204.

Égré, Paul. "Vague Judgment: A Probabilistic Account". *Synthese* 194.10 (2017), pp. 3837–3865.

Égré, Paul & Anouk Barberousse. "Borel on the Heap". *Erkenntnis* 79.5 (2014), pp. 1043–1079.

Egré, Paul, Vincent De Gardelle & David Ripley. "Vagueness and order effects in color categorization". *Journal of Logic, Language and Information* 22.4 (2013), pp. 391–420.

Eijck, J. van & S. Lappin. "Probabilistic Semantics for Natural Language". *Logic and Interactive Rationality (LIRA) 2012, Volume 2*. Ed. by Z. Christoff, P. Galeazzi, N. Gierasimszuk, A. Marcoci & S. Smets. ILLC, University of Amsterdams, 2012, pp. 17–35.

Emerson, Guy. "Linguists Who Use Probabilistic Models Love Them: Quantification in Functional Distributional Semantics". *Proceedings of Probability and Meaning (PaM2020)*. Ed. by Christine Howes, Stergios Chatzikyriakidis, Adam Ek & Vidya Somashekarappa. 2020, pp. 41–52.

Fara, Delia Graff. "Shifting sands: An interest-relative theory of vagueness". *Philosophical Topics* 28 (2000).

Fernández, Raquel & Staffan Larsson. "Vagueness and Learning: A Type-Theoretic Approach." *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014), Dublin, Ireland, August 23-24 2014*. 2014, pp. 151–159.

Fine, Kit. "Vagueness, truth, and logic". *Synthese* 30 (1975), pp. 265–300.

Franke, Michael & Jose Pedro Correia. "Vagueness and imprecise imitation in signalling games". *British Journal for the Philosophy of Science* 69.4 (2018), pp. 1037–1067.

Franke, Michael, Gerhard Jäger & Robert van Rooij. "Vagueness, Signaling & Bounded Rationality". *JSAI-isAI 2010*. Ed. by T. Onoda, D. Bekki & E. McCready. 2011, pp. 45–59.

Frazee, Joey & David Beaver. "Vagueness is rational under uncertainty". *Logic, Language and Meaning: 17th Amsterdam Colloquium, Amsterdam, The Netherlands, December 16-18, 2009, Revised Selected Papers, Lecture Notes in Artificial Intelligence*. Ed. by Maria Aloni, Harald Bastiaanse, Tikitu de Jager & Katrin Schulz. Springer, 2010, pp. 153–162.

Herbstritt, Michele & Michael Franke. "Complex probability expressions & higher-order uncertainty: Compositional semantics, probabilistic pragmatics & experimental data". *Cognition* 186 (2019), pp. 50–71. ISSN: 0010-0277.

Kamp, Hans. "Two Theories About Adjectives". *Formal Semantics of Natural Languages*. Ed. by Ed Keenan. Cambridge: Cambridge University Press, 1975, pp. 123–155.

Keefe, R. & P. Smith, eds. *Vagueness: A Reader*. Cambridge, MA: MIT Press, 1997.

Kempson, R., E. Gregoromichelaki & C. Howes. "Language as Mechanisms for Interaction: Towards an Evolutionary Tale". *Language, Logic, and Computation. TbiLLC 2018. Lecture Notes in Computer Science, vol 11456*. Ed. by Silva A., Staton S., Sutton P. & Umbach C. Springer, 2019.

Kirby, Simon. "The Evolution of Meaning-Space Structure through Iterated Learning". *Emergence of Communication and Language*. Ed. by C. Lyon, C. Nehaniv & A. Cangelosi. London: Springer, Verlag, 2007, pp. 253–268.

Kirby, Simon & James Hurford. "The Emergence of Linguistic Structure: An overview of the Iterated Learning Model". *Simulating the Evolution of Language*. Ed. by A. Cangelosi & D. Parisi. London: Springer, Verlag, 2002, pp. 121–148.

Lassiter, Daniel. "Vagueness as Probabilistic Linguistic Knowledge". *Vagueness in Communication*. Ed. by R. Nouwen, U. Sauerland, H.C. Schmitz & R. van Rooij. Springer, 2011.

Lassiter, Daniel & Noah D. Goodman. "Adjectival vagueness in a Bayesian model of interpretation". *Synthese* (2017), pp. 1–36. ISSN: 1573-0964.

Nouwen, R., U. Sauerland, H.C. Schmitz & R. van Rooij, eds. *Vagueness in Communication*. Springer, 2011.

Piantadosi, S., H. Tily & E. Gibson. "The communicative function of ambiguity in language". *PNAS* 108.9 (2011), pp. 3526–3529.

Qing, Ciyang & Michael Franke. "Vagueness, and Optimal Language Use: A Speaker-Oriented Model". *Proceedings of SALT 2014*. Ed. by T. Snider, S. D'Antonio & Mia Weigand. 2014, pp. 23–41.

Schmidt, L.A., N.D. Goodman, D. Barner & J.B. Tenenbaum. "How tall is tall? compositionality, statistics, and gradable adjectives." *Proceedings of the 31st annual conference of the cognitive science society*. 2009, pp. 3151–3156.

Schöller, Anthea & M. Franke. "Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of few & many". *Linguistics Vanguard* 3 (2017).

Schuster, Annika, Corina Strößner, Peter Sutton & Henk Zeevat. "Stochastic Frames". *Proceedings of the Probability and Meaning Conference (PaM 2020)*. Gothenburg: Association for Computational Linguistics, 2020, pp. 78–85.

Schuster, Sebastian & Judith Degen. "I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions". *Cognition* 203 (2020), p. 104285. ISSN: 0010-0277.

Shannon, Claude. "A Mathematical Theory of Communication". *Bell System Technical Journal* 27 (1948), pp. 379–423.

Sorensen, Roy. *Blindspots*. Oxford: Clarendon Press, 1988.

— *Vagueness and Contradiction*. Oxford: Clarendon Press, 2001.

Sutton, Peter R. "Probabilistic Approaches to Vagueness and Semantic Competency". *Erkenntnis* 83.4 (2018), pp. 711–740.

Sutton, Peter. R. "Vagueness, Communication, and Semantic Information". PhD thesis. King's College London, 2013.

Tessler, M. H. & M. Franke. "Not unreasonable: Carving vague dimensions with contraries and contradictions." *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*. 2018, pp. 1108–1113.

Tessler, Michael H & Noah D Goodman. *Learning from Generic Language*. 2019.

Tiel, Bob van, Michael Franke & Uli Sauerland. "Probabilistic pragmatics explains gradience and focality in natural language quantification". *Proceedings of the National Academy of Sciences* 118.9 (2021). ISSN: 0027-8424. eprint: `https://www.pnas.org/content/118/9/e2005453118.full.pdf`.

Williamson, Timothy. *Vagueness*. London: Routledge, 1994.

— "Vagueness and Ignorance". *Proceedings of the Aristotelian Society* 66 (1992), pp. 145–162.