Vagueness as a product of learning over a noisy channel

Peter R. Sutton
Heinrich-Heine-Universität Düsseldorf

1 Introduction: This paper investigates the hypothesis that vagueness in predicates arises naturally from a combination of information theoretic pressures based on learning and communication. The semantic bottleneck problem (that learners are exposed to insufficient data to fully resolve the interpretations of expressions) introduces a pressure on languages to encode information with fewer, more equivocal expressions. At the same time, transmission of information between speakers is not perfect (information is transmitted over a noisy channel) and so there is uncertainty about what any particular message is communicating.

We propose a model based on the Iterated Learning paradigm [i.a. 1; 5; 6] (which can model the semantic bottleneck problem over generations of agents), a noise source, and a simple information theoretic learning strategy (which encodes a form of constraint over the level of information learnt to be encoded by a particular predicate, given a set of learning data). Given a set of situations to be described and a set of predicates to encode this information, language evolves between generations of agents with the result that it either remains highly unstable (not a phenomenon we observe in natural languages), or becomes stable. If stable, the language can either be uninformative (e.g. where it contains only one predicate to communicate all situations), or informative (with multiple predicates relative to the size of the situation set). A successful balance between the learning and communication pressures results in a stable, informative language. Our findings indicate two results. When the parameters of the model are set in such a way as to produce stable, non-trivial languages: (1) there is a correlation between the 'narrowness' of the learning bottleneck and the number of predicates that languages have; (2) noise in the information channel results in vagueness (i.e. graded boundaries between predicates modelled as the probability of an agent applies a predicate, given a situation to be described).

2 Background: Since at least the work of Zipf [11, 12], numerous linguistic phenomena have been analysed in terms of the interaction between information theoretic constraints and the idea that language approximates an optimal tool for communication. Recent examples include arguments that: ambiguity is an optimal feature of human language [8]; languages optimize information density [4; 7]; general efficiency principles can explain a wide range of crosslinguistic syntactic patterns [i.a. 3]; conflict between information-theoretic pressures can explain crosslinguistic count/mass variation [10]; and vagueness evolves in language when boundedly rational agents repeatedly engage in cooperative signalling [2]. This work falls broadly within this Zipfian paradigm.

The model presented in this paper is inspired, methodologically, by Iterated Learning Models (ILMs) [1; 5; 6] and how semantic learning connects to vagueness [9]. In ILMs a competent adult agent provides a sample of her vernacular to a learner. The learner becomes a second generation competent agent and provides a sample of her language to a third generation learner etc.. The cycles progress and the impact of certain parameter values within models can be witnessed on the long-term behaviour of the language. In particular, ILMs model the semantic bottleneck in that learners are not necessarily exposed to all of the competent agent's language.

3 Hypotheses: It was hypothesised, but not further investigated in [9] that vagueness is an effect of the combination of employing a learning strategy to help to overcome the semantic bottleneck whilst learning in conditions of uncertainty. This paper seeks support for this hypothesis by implementing a Iterated Learning Model with a probabilistic learning strategy for inferring applications of predicates based on a nosy information channel.

Hypotheses: (i) For at least some settings of model parameters (e.g. the ratio of the size of the language to the size of the data set presented to the learner), probabilistic learning results in a stable, informative language. (ii) Items in the language will display the hallmarks of vague predicates, i.e., having graded boundaries.

4 Model: A computational model was written with Matlab. Elements in the model are: a collection of agents $\mathbb{A} = \{A_1, ..., A_n\}$ where, e.g. A_1 is the first generation agent and A_n is the n^{th} generation agent; an ordered set of situations to be described $\mathbb{S} = \langle s_1, ..., s_n \rangle$. Situations are assumed to form a total order with respect to similarity such that s_1 is most similar to s_2 and least similar to s_n. (Likewise s_n is most similar to s_{n-1} and least similar to s_1.) The intuitive idea behind this was to model something like a colour spectrum of shades. A distance function $D : \langle \mathbb{S} \times \mathbb{S} \rangle \rightarrow \mathbb{N}$ such that $D(s_n, s_{n \pm k}) = \delta \times k$ (where δ is a parameter of the model). A set of predicates $\mathbb{M} = \{m_1, ..., m_n\}$. A set of languages $\mathbb{L} = \{L_{A_1}, ..., L_{A_n}\}$ where L_{A_1} is the language of A_1. (All other languages are derived from L_{A_1} via a string of intergeneration learning events in a way to be made clear below.) Languages are characterised as sets of probability distributions $P(m_i|s_j)$ such that for each $s_j \in \mathbb{S}, \sum_{m_i \in \mathbb{M}} P(m_i|s_j) = 1$.

A random sample of situations, S, is generated (the sample size is a parameter of the model). A_k then provides A_{k+1} with a set of situation-predicate pairs d_{A_k} (with a predicate for every situation in the randomly generated set). This set of pairs provides the learning data for A_{k+1}. The size of S relative to $\mathbb{S}$ determines the probability that A_{k+1} will not witness every situation and so will not be exposed to the full extent of L_{A_k}. This models the semantic bottleneck. When noise is present in the model, there are two parameters. One parameter, $\mathcal{N} \in [0, 1]$ sets the probability that for a situation-predicate pair $\langle s_i, m_j \rangle \in d_{A_k}$, the learner A_{k+1} also witnesses $\langle s_{i+1}, m_j \rangle$ and $\langle s_{i-1}, m_j \rangle$. For example, when $\mathcal{N} = 0$, there is no noise in the channel and the learner receives three identical sets d_{A_k}. When $\mathcal{N} = 1$, the learner receives d_{A_k}, but also $d_{A_k,\uparrow}$ and $d_{A_k,\downarrow}$ such that:[1]

$$d_{A_k,\uparrow} = \{\langle s_{i+1}, m_j \rangle : \langle s_i, m_j \rangle \in d_{A_k}\} \tag{1}$$

$$d_{A_k,\downarrow} = \{\langle s_{i-1}, m_j \rangle : \langle s_i, m_j \rangle \in d_{A_k}\} \tag{2}$$

This reflects noise in the channel (the addition of information before reception by the learner). A second parameter, $\mathcal{U} \in [0, 1]$, reflects the extent a learner is able to be certain about the 'true' situation being described (the extent that they are able to be sure that s_i is being described by m_j as opposed to s_{i+1} or s_{i-1}). For example, when $\mathcal{U} = 0$, the learner can be certain of the situation being described (there is no noise). When $\mathcal{U} = 1$, there is maximum perplexity with respect to the situation being described.

From the data set, there is then a learning event in which A_{k+1} develops her own language $L_{A_{k+1}}$ based on d_{A_k} (and $d_{A_k,\uparrow}$ and $d_{A_k,\downarrow}$ in the case of noise). This amounts to inferring probability distributions, $P(m_j|s_i)$, for each s_i that is witnessed in d_{A_k}, $d_{A_k,\uparrow}$, or $d_{A_k,\downarrow}$:

$$P_{A_{k+1}}(m_j|s_i) = \frac{P_{A_k,d_{A_k}}(m_j, s_i) + (\mathcal{U} \times P_{A_k,d_{A_k,\uparrow}}(m_j, s_i)) + (\mathcal{U} \times P_{A_k,d_{A_k,\downarrow}}(m_j, s_i))}{P_{A_k,d_{A_k}}(s_i) + (\mathcal{U} \times P_{A_k,d_{A_k,\uparrow}}(s_i)) + (\mathcal{U} \times P_{A_k,d_{A_k,\downarrow}}(s_i))} \tag{3}$$

If a situation $s_\varnothing$ is not in the second projections of d_{A_k}, $d_{A_k,\uparrow}$, or $d_{A_k,\downarrow}$, then, the learner, A_{k+1}, searches in both 'directions' along the set of situations for the nearest s that is instantiated in the data set. Where $s_\uparrow$ is the nearest instantiated situations in one direction and $s_\downarrow$ is the nearest instantiated situation in another. For each $m_j \in \mathbb{M}$ (where all probabilities in (4) are for A_{k+1}):

$$P(m_j|s_\varnothing) = \frac{\log_2^{-1}(\log_2(P(m_j|s_\uparrow) - D(s_\varnothing, s_\uparrow)) + \log_2^{-1}(\log_2(P(m_j|s_\downarrow) - D(s_\varnothing, s_\downarrow))}{\sum_{m_i \in \mathbb{M}} \log_2^{-1}(\log_2(P(m_i|s_\uparrow) - D(s_\varnothing, s_\uparrow)) + \sum_{m_i \in \mathbb{M}} \log_2^{-1}(\log_2(P(m_i|s_\downarrow) - D(s_\varnothing, s_\downarrow))} \tag{4}$$

An example of this is shown in Figure 1. On the left hand side, s_5, s_6 and s_7 are not instantiated in the learners data set, but $P(m_3|s_4) = P(m_8|s_8) = 1$. The right hand side shows the result of the inference (where distance function parameter $\delta = 1$).

The result of this procedure is that each learner has a probability distribution for applying predicates in each situation. These distributions can then be sampled by a fresh randomly selected set of situations (possibly plus noise) and used as input data for a fresh learner.

[1] For s_1, noise only impacts $d_{A_k,\uparrow}$. For s_{10}, noise only impacts $d_{A_k,\downarrow}$.

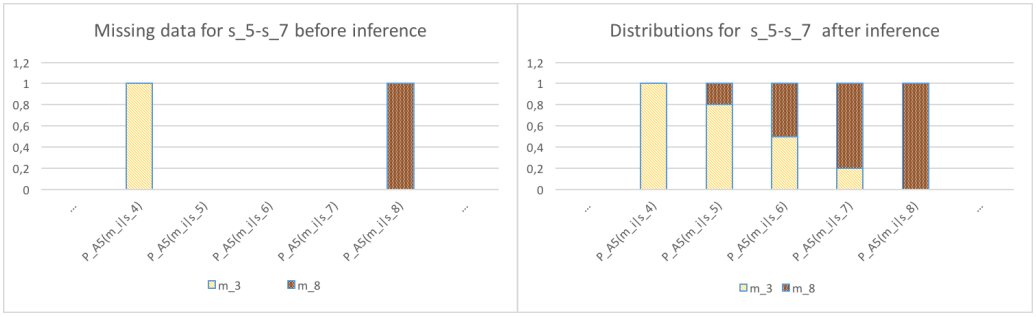

Figure 1: Example of inferring how to apply predicates when situations haven't been witnessed in the data set.

By assumption, A_1 has a completely categorical language with not vagueness and a unique predicate for every situation. This is graphically represented in Figure 2. The reason for this assumption is to be sure that any vagueness that emerges is not the result of the input from A_1.

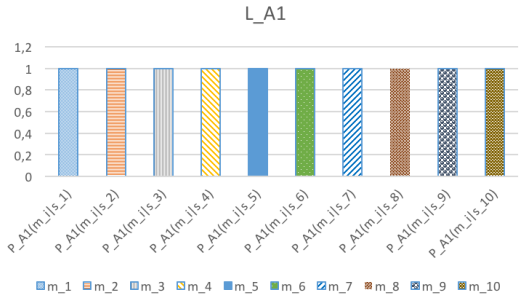

Figure 2: Starting language for A_1

5 Results: Simulations were run for a space of 10 situations (s_1-s_{10}): (A) without a bottleneck or noise; (B) with a bottleneck, without noise; (C) without a bottleneck, with noise; (D) with a bottleneck, with noise. Results are shown for A_{100}. The distance function parameter was kept at $\delta = 1$. Where relevant, $\mathcal{U}$ was kept at 0.5. Other parameters are described below. In all cases the languages that emerged were relatively stable, where boundary shifts between predicates were small between generations.

(A) Without a bottleneck or noise: The language of A_{100} was exactly as it was for A_1 in Figure 2, namely, with ten categorical predicates. Any bottleneck was in effect removed by setting the sample size to 500 (so that the probability of an agent not witnessing a situation was very low).

(B) With a bottleneck, without noise: The result for A_{100} in Figure 3 was typical of the result for a situation sample size of 30, namely, a reduction to between two and three categorical predicates.

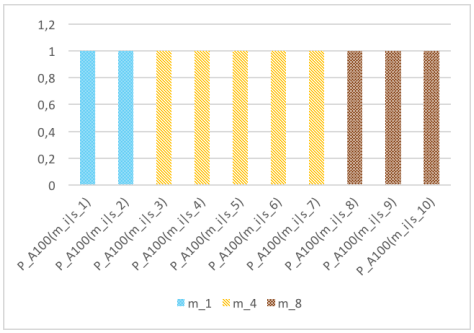

Figure 3: Bottleneck, sample of 30.

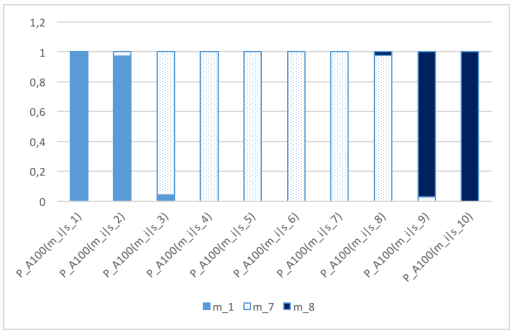

Figure 4: Noise, $\mathcal{N} = 0.1$. No bottleneck.

(C) Without a bottleneck, with noise: In these simulations, as in condition (A), the situation sample size was set to 500 to effectively remove any bottleneck. Figure 4 shows a typical result for A_{100} when $\mathcal{N} = 0.1$ (approximately 10% of signals were noisy), namely a reduction to 2-3 predicates which have graded boundaries, but only marginally so (around 0.98 and 0.02 at the boundary situations). Figure 5 shows a typical result for A_{100} when $\mathcal{N} = 0.3$ (approximately 30% of signals were noisy).

On these settings, it was more typical to end up with 2 predicates, and the boundaries between them were more graded (around 0.9 and 0.1 at the boundary situations).

(D) With a bottleneck, with noise: In these simulations, the situation sample size was set to 30, as in condition (B). The noise level was $\mathcal{N} = 0.3$, as in the second example in condition (C). Here, we typically witnessed a reduction to two predicates with graded boundaries (around 0.8-0.85 and 0.2-0.15 at the boundary situations). A typical example outcome for A_{100} is given in Figure 6.

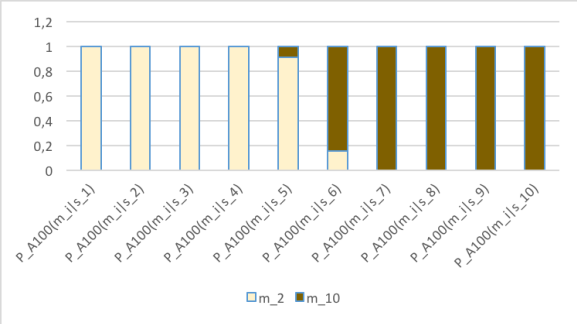

Figure 5: Noise, $\mathcal{N} = 0.3$. No bottleneck.

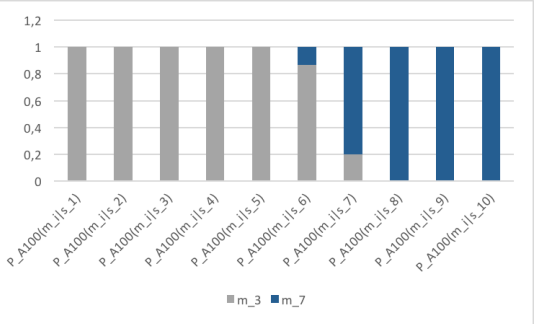

Figure 6: Noise, $\mathcal{N} = 0.3$. Bottleneck, sample of 30.

6 Discussion: The simulations showed some confirmation for hypothesis (i): for at least some settings of model parameters, probabilistic learning results in a stable, informative language. However, if noise, $\mathcal{N}$ was set too high, or the situation sample size was set too low, the outcome was an uninformative, trivial language (one with only one predicate which applies in every situation). We observed a surprising result regarding hypothesis (ii) (that items in the language will display the hallmarks of vague predicates). The introduction of a bottleneck alone did not result in the common occurrence of vagueness. This was the case despite the fact that agents reasoned probabilistically in cases where they had not witnessed description for a situation in the learning phase. However, when noise was introduced into the model, vague, graded boundaries between predicates invariably emerged. Furthermore, there appears to be a correlation between the noise level $\mathcal{N}$ and the extent to which graded boundaries emerge. A tentative conclusion we might also draw is that when the bottleneck and noise are combined, we see a small increase in the level to which boundaries are graded.

7 Further work and conclusions: Further work needs to be done to establish with more clarity how the settings of the parameters within the model interact. In particular, we have not yet assessed the impact of changing the value of the distance parameter δ. Some preliminary testing seems to indicate that low δ values tend to make the boundaries of predicates more graded.

Given the simplicity of these simulations, conclusions about natural language are hard to draw. Our results do, however, suggest an enticing possibility. It may well be the case that vagueness in natural language also arises as a byproduct of semantic learning over a noisy communication channel.

References [1] Henry Brighton and Simon Kirby. Meaning space structure determines the stability of culturally evolved compositional language". *Technical report, Language Evolution and Computation Research Unit, Department of Theoretical and Applied Linguistics, The University of Edinburgh.*, 2001. [2] Michael Franke, Gerhard Jäger, and Robert van Rooij. Vagueness, signaling & bounded rationality. proceedings of LENLS2010, 2010. [3] John A. Hawkins. *Cross-linguistic Variation and Efficiency*. OUP, Oxford, 2014. [4] Florian T. Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62, 2010. [5] Simon Kirby. The evolution of meaning-space structure through iterated learning. In C. Lyon, C. Nehaniv, and A. Cangelosi, editors, *Emergence of Communication and Language*, pages 253–268. Springer, Verlag, London, 2007. [6] Simon Kirby and James Hurford. The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi and D. Parisi, editors, *Simulating the Evolution of Language*, pages 121–148. Springer, Verlag, London, 2002. [7] Roger Levy and Florian T. Jaeger. Speakers optimize information density through syntactic reduction. *Proceedings of the twentieth annual conference on neural information processing systems*, 19:849–856, 2007. [8] S. Piantadosi, H. Tily, and E. Gibson. The communicative function of ambiguity in language. *PNAS*, 108(9):3526–3529, 2011. [9] Peter. R. Sutton. *Vagueness, Communication, and Semantic Information*. PhD thesis, King's College London, 2013. [10] Peter R. Sutton and Hana Filip. Probabilistic mereological type theory and the mass/count distinction. Forthcoming in *JLM*, 2017. [11] G. Zipf. *The psychobiology of language*. Houghton Mifflin, 1935. [12] G. Zipf. *Human behavior and the principle of least effort*. Addison-Wesley, 1949.