

# Co-evolution of lexical meaning & pragmatic use

( – draft March 23, 2017— )

## Abstract

According to standard linguistic theory, the meaning of an utterance is the product of conventional semantic meaning and general pragmatic rules on language use. To investigate how cultural evolution of language plays out under this picture of the semantics-pragmatics interface, we present a game theoretic model of the competition between types of language users, each endowed with a selection of lexical concepts and a particular pragmatic disposition to act on them. Our model traces two evolutionary forces and their interaction: (i) fitness-based pressure towards communicative efficiency and (ii) potential learning biases during the transfer of linguistic knowledge. We illustrate the model based on a case study on scalar implicatures. In this case study learning biases that favor simple semantic representations can boost the evolution of more sophisticated pragmatic reasoning types and so prevent the lexicalization of scalar implicatures.

## 1 Introduction

What is conveyed usually goes beyond what is said. A request for a blanket can be politely veiled by uttering “I’m cold;” temporal succession of events can be communicated by the order in which conjuncts appear as in “I traveled to Paris and got married;” an invitation can be declined by saying “I have to work.” An influential explanation of the relation between the literal meaning of expressions and what they may convey in context is due to Grice (1975), who characterizes pragmatic use and interpretation as a process of mutual reasoning about rational language use. For instance, under the assumption that the speaker is cooperative and relevant, “I have to work” may be interpreted as providing a reason why the speaker will not be able to accept an invitation, going beyond its literal meaning. Some of these enrichments are rather *ad hoc*. Others show striking regularities, such as the use of ability questions for polite requests (“Could you please ...?”), or certain enrichments of lexical meanings such as *and* to convey *and then*.

A particularly productive and well studied class of systematic pragmatic enrichments are scalar implicatures (Horn 1984, Hirschberg 1985, Levinson 1983, Geurts 2010). Usually, the utterance of a sentence like “I own some of Johnny Cash’s albums” will be taken to mean that the speaker does not own all of them. This is because, if the speaker owned them all, she could have

used the word *all* instead of *some* in his utterance, thereby making a more informative statement. Scalar implicatures, especially the inference from *some* to *some but not all*, have been studied extensively, both theoretically (e.g., Sauerland 2004, Chierchia et al. 2012, van Rooij and de Jager 2012) as well as experimentally (e.g., Bott and Noveck 2004, Huang and Snedeker 2009, Grodner et al. 2010, Goodman and Stuhlmüller 2013, Degen and Tanenhaus 2015). While there is much dispute in this domain about many details, a position endorsed by a clear majority is that a scalar item like *some* is underspecified to mean *some and maybe all* and that the enrichment to *some but not all* is part of some regular process with roots in pragmatics.

If this majority view is correct, the question arises how such a division of labor between semantics and pragmatics could have evolved and why it would be so pervasive across natural languages. Models of language evolution abound. There are simulation-based models studying the evolution of language in populations of communicating agents (Hurford 1989, Steels 1995, Lenaerts et al. 2005, Steels and Belpaeme 2005, Baronchelli et al. 2008, Steels 2011, Spike et al. 2016) and there are mathematical models of language evolution, mostly coming from game theory (Lewis 1969, Wärneryd 1993, Blume et al. 1993, Nowak and Krakauer 1999, Huttegger 2007, Skyrms 2010). Much of this work has focused on explaining basic properties such as compositionality and combinatoriality (e.g., Batali 1998, Nowak and Krakauer 1999, Nowak et al. 2000, Kirby and Hurford 2002, Kirby 2002, Smith et al. 2003, Gong 2007, Kirby et al. 2015, Verhoef et al. 2014, Franke 2016), but little attention has been paid to the interaction between conventional meaning and pragmatic use. What is more, many mathematical models explain evolved meaning as a regularity in the overt behavior of agents. In contrast, we will here look at language users with a richer cognitive make-up.

We spell out a model of the co-evolution of conventional meaning and pragmatic reasoning types. The objects of replication and selection are pairs of lexical meanings and general types of pragmatic behavior, which we represent using probabilistic models of pragmatic language use (Frank and Goodman 2012, Franke and Jäger 2016, Goodman and Frank 2016). Replication and selection are described by the *replicator mutator dynamic*, a general and established model of evolutionary change in large and homogeneous populations (Hofbauer 1985, Nowak et al. 2000; 2001, Hofbauer and Sigmund 2003, Nowak 2006). The approach allows us to study the interaction between (i) evolutionary pressure towards communicative efficiency and (ii) possible infidelity in the transmission of linguistic knowledge, caused by factors such as inductive learning biases and sparse learning data. Considering transmission of linguistic knowledge is important because neither semantic meanings nor pragmatic usage patterns are directly observable. Instead, language learners have to infer these unobservables from the observable behavior in which they result. We formalize this process as a form of Bayesian inference. Our approach thereby contains a well-understood model of iterated Bayesian learning (Griffiths and Kalish 2007) as a special

case, but combines it with functional selection, here formalized as the most versatile dynamic from evolutionary game theory; the replicator dynamic (Taylor and Jonker 1978). Section 2 introduces this model.

Section 3 applies this model to a case study on scalar implicatures. We discuss a setting in which the majority view of underspecified lexical meanings and pragmatic enrichments emerges if selection and transmission infidelity are combined. In particular, we show that inductive learning biases of Bayesian learners that favor simpler lexical meanings can prevent the lexicalization of scalar inferences and lead to the emergence of Gricean-like pragmatic reasoning types. Results of this case study are critically assessed in the light of the assumptions that feed our model in Section 4.

## **2 A model of co-evolving lexical concepts and pragmatic behavior**

### **2.1 Expressivity and learnability**

The emergence and change of linguistic structure is influenced by many intertwined factors. These range from biological and socio-ecological to cultural ones (Benz et al. 2005, Steels 2011, Tamariz and Kirby 2016). Social and ecological pressures determine communicative needs, while biology determines the architecture that enables and constrains the means by which they can be fulfilled. In the following, our focus lies on cultural aspects, wherein processes of linguistic change are viewed as shaped by language use and its transmission, i.e., as a result of a process of cultural evolution (Pagel 2009, Thompson et al. 2016).

The idea that language is an adaptation to serve a communicative function is fundamental to many synchronic and diachronic analyses at least since Zipf's (1949) explanation of word frequency rankings as a result of competing hearer and speaker preferences (e.g., in Martinet 1962, Horn 1984, Jäger and van Rooij 2007, Jäger 2007, Piantadosi 2014, Kirby et al. 2015). If processes of selection, such as conditional imitation or reinforcement, favor more communicatively efficient types of behavior, languages are driven towards semantic expressivity (e.g., Nowak and Krakauer 1999, Skyrms 2010). But pressure towards communicative efficiency is not the only force that shapes language. Learnability is another. Natural languages need to be learnable to survive their faithful transmission across generations. Even small learning biases implicit in acquisition can build up and have quite striking effects on an evolving language in a process of iterated learning (Kirby and Hurford 2002, Smith et al. 2003, Kirby et al. 2014).

While natural languages are pressured for both expressivity and learnability these forces may pull in opposite directions. Their opposition becomes particularly clear when considering the extreme (cf. Kemp and Regier 2012, Kirby et al. 2015). A language consisting of a single form-meaning association is easy to learn but lacking in expressivity. Conversely, a language

that lexicalizes a distinct form for a large number of different meanings is highly expressive but challenging to acquire.

## 2.2 The replicator mutator dynamic

An elegant formal approach to capture the interaction between expressivity and learnability is the *replicator mutator dynamic* (Hofbauer 1985, Nowak et al. 2000; 2001, Hofbauer and Sigmund 2003, Nowak 2006). In its simplest, discrete-time formulation, the RMD defines the frequency  $x'_i$  of each type  $i$  in a population at the next time step as a function of: (i) the frequency  $x_i$  of each type  $i$  before the update step, (ii) the fitness  $f_i$  of each type  $i$  before the update, and (iii) the probability  $Q_{ji}$  that an agent who wants to imitate, adopt, or learn the type of an agent with type  $j$  ends up acquiring type  $i$ :

$$x'_i = \sum_j Q_{ji} \frac{x_j f_j}{\sum_k x_k f_k}. \quad (1)$$

The RMD consists of two components: fitness-based selection and transmission biases. This becomes most transparent when we consider an equivalent formulation in terms of a step-wise application of the discrete-time replicator dynamic (Taylor and Jonker 1978) on the initial population vector  $\vec{x}$  and its subsequent multiplication with a mutation matrix  $Q$ :

$$x'_i = (\mathbf{M}(\mathbf{RD}(\vec{x})))_i, \quad (2)$$

where

$$(\mathbf{RD}(\vec{x}))_i = \frac{x_i f_i}{\sum_k x_k f_k} \quad \text{and} \quad (\mathbf{M}(\vec{x}))_i = (\vec{x} \cdot Q)_i = \left( \sum_j x_j Q_{ji} \right)_i.$$

If the transmission matrix  $Q$  is trivial in the sense that  $Q_{ji} = 1$  whenever  $j = i$ , the dynamic reduces to the replicator dynamic. The replicator dynamic is a model of fitness-based selection in which the relative frequency of type  $i$  will increase with a gradient proportional to its average fitness in the population. This dynamic is popular and versatile because it can be derived from many abstract processes of biological and cultural transmission and selection (for overview and several derivations see Sandholm 2010), including conditional imitation (e.g., Helbing 1996, Schlag 1998) or reinforcement learning (e.g., Börgers and Sarin 1997, Beggs 2005). If fitness  $f_i$  is the same for all types  $i$ , the replicator step is the identity map  $(\mathbf{RD}(\vec{x}))_i = x_i$  and the dynamic reduces to a process of iteration of the transmission bias encoded in  $Q$ . In this way, the process in (1), equivalently (2), contains a model of iterated learning (Griffiths and Kalish 2007).

**Example.** Consider a simple coordination game. Agents are of two types: positive and negative. If agents of different type play with each other, they obtain a payoff of 0. If negative meets

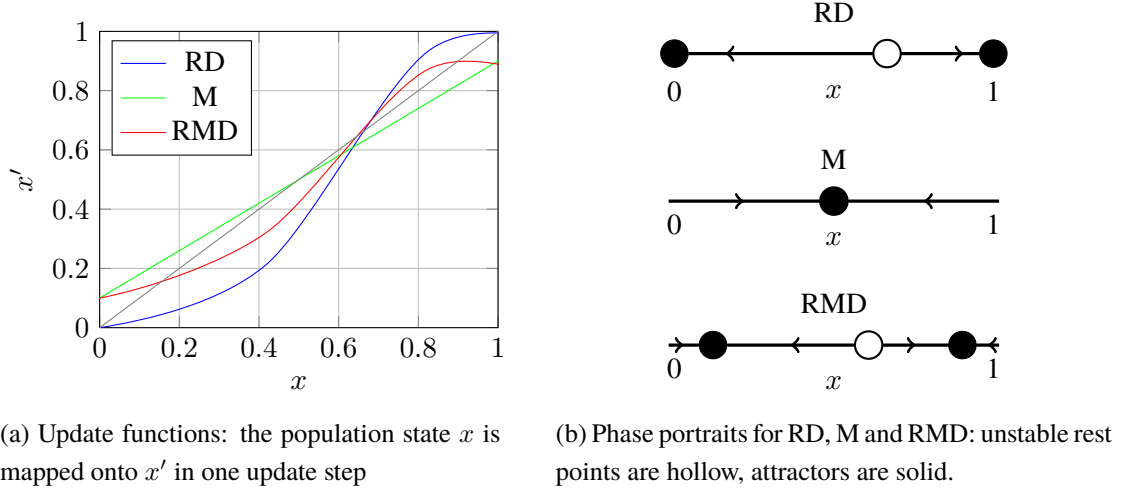


Figure 1: Example

negative, each receives a payoff of 1. If positive meets positive, they get a payoff of 2. A population state is completely characterized by the proportion  $x$  of negatives. The fitness of negatives in population state  $x$  is  $f_n(x) = x$ , that of positives is  $f_p(x) = 2 - 2x$ . The average fitness is  $\Phi(x) = xf_n(x) + (1 - x)f_p(x) = 3x^2 - 4x + 2$ . Without mutation, the replicator dynamic will then update  $x$  to  $RD(x) = x^2/\Phi(x)$ . The update function  $RD(x)$  of the replicator step is plotted in Figure 1a as the blue line. Rest points, for which  $RD(x) = x$ , are at  $x = 0$ ,  $x = 1$  and  $x = 2/3$ . The former are attractors as nearby points converge to them. Points near  $x = 2/3$  move towards 0 or 1. This is schematically pictured in the topmost phase portrait in Figure 1b.

Adding mutation changes the dynamic and its rest points. Let's assume that  $Q_{ji} = .9$  when  $j = i$ . The update effect of mutation alone is  $M(x) = .9x + .1(1 - x) = .8x + .1$  and is plotted as the linear green line in Figure 1a. It has only one stable rest point at  $x = 0.5$  (see Figure 1b). If we first take the replicator step and then the mutation step in sequence, we obtain the replicator mutator dynamic  $RMD(x) = M(RD(x)) = .9x^2 - .2x + .2/3x^2 - 4x + 2$ , which is plotted in red in Figure 1a. The rest points are at  $x = .121$ ,  $x = .903$  and  $x = .609$ . The former two are attractors (see Figure 1b).

### 2.3 Fitness & learnability of lexical meanings & pragmatic strategies

Moving beyond an abstract example, our goal is to apply the RMD to investigate the co-evolution of lexical concepts and pragmatic behavior. To do so, we need to fix three things: (i) what the relevant types are, (ii) how fitness derives from communicative success and (iii) how the mutation matrix is computed. These issues are addressed, one by one, in the following.

### 2.3.1 Types: Lexica and pragmatic strategies

Types are what evolution operates on. They define an agent's fitness, usually through a payoff accrued in single interactions with other agents. Often types can be identified as the possible acts in a game, e.g., either cooperating or defecting in a prisoner's dilemma. In other cases, they may be thought of as general properties of an agent that influences her fitness, such as being positive or negative in our previous example (whatever that means). For our present purposes, types are identified more concretely by specific assumptions about their cognitive make-up. Since we are interested in the evolutionary competition between different lexical concepts and ways of using them in communication, a type is here defined as a pair consisting of a lexicon and a pragmatic strategy.

Agents play signaling games, in which the speaker wants to communicate a world state  $s$  with a message  $m$  to a hearer who receives  $m$  but does not know  $s$  (e.g. Lewis 1969, Skyrms 2010). A lexicon associates each message with a (possibly fuzzy) set of states. A pragmatic type maps a lexicon onto a probabilistic speaker rule (a probabilistic choice of message for each state) and a probabilistic listener rule (a probabilistic choice of state for each message). There are many ways of making these general notions more concrete. Here is what we will assume in the remainder of this paper.

Lexica codify the truth-conditions of expressions. A convenient way to represent lexica is by  $(|S|, |M|)$ -Boolean matrices, where  $S$  is a set of states (meanings) and  $M$  a set of messages (forms available in the language). For example, suppose that there are two relevant world states  $S = \{s_{\exists-\forall}, s_{\forall}\}$ . In state  $s_{\exists-\forall}$  Chris owns some but not all of Johnny Cash's albums while in  $s_{\forall}$  Chris owns them all. Suppose that there are two messages  $M = \{m_{\text{some}}, m_{\text{all}}\}$  where  $m_{\text{some}}$  is short for a sentence like *Chris owns some of Johnny Cash's albums* and  $m_{\text{all}}$  for the same sentence with *some* replaced by *all*. Lexica for this case would assign a Boolean truth value, either 0 for false or 1 for true, to each state-message pair. The following two lexica are minimal examples for the distinction between a lexicalized upper-bound for *some* in  $L_{\text{bound}}$  and the widely assumed logical semantics with only a lower-bound in  $L_{\text{lack}}$ .

$$L_{\text{bound}} = \begin{array}{cc} & \begin{array}{cc} m_{\text{some}} & m_{\text{all}} \end{array} \\ \begin{array}{c} s_{\exists-\forall} \\ s_{\forall} \end{array} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array} \qquad L_{\text{lack}} = \begin{array}{cc} & \begin{array}{cc} m_{\text{some}} & m_{\text{all}} \end{array} \\ \begin{array}{c} s_{\exists-\forall} \\ s_{\forall} \end{array} & \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \end{array}$$

Pragmatic types define dispositions of produce and interpret messages given a lexicon. We distinguish between two kinds of pragmatic types. *Literal interlocutors* produce and interpret messages literally, being guided only by their lexica. *Pragmatic interlocutors* instead engage in mutual reasoning to inform their choices. Recent probabilistic models of rational language use

(Frank and Goodman 2012, Franke and Jäger 2016, Goodman and Frank 2016) capture different types of pragmatic behavior in a reasoning hierarchy. The hierarchy’s bottom, level 0, corresponds to literal language use, as in Equations (3) and (4). Pragmatic language users of level  $n + 1$  act (approximately) rational with respect to level- $n$  behavior of their interlocutors, as in Equations (5) and (6).

$$H_0(s \mid m; L) \propto pr(s) L_{sm} \quad (3)$$

$$S_0(m \mid s; L) \propto \exp(\lambda L_{sm}) \quad (4)$$

$$H_{n+1}(s \mid m; L) \propto pr(s) S_n(m \mid s; L) \quad (5)$$

$$S_{n+1}(m \mid s; L) \propto \exp(\lambda H_n(s \mid m; L)) \quad (6)$$

According to (3), a literal hearer’s interpretation of a message depends on her lexicon and her prior over states,  $pr \in \Delta(S)$ , which is here assumed flat, for simplicity. Literal interpreters thereby choose an arbitrary true interpretation for each message. Pragmatic hearers, defined in (5), use Bayes rule to weigh interpretations based on a conjecture about speaker behavior. Speaker behavior is regulated by a soft-max parameter  $\lambda$ ,  $\lambda \geq 0$  (Luce 1959, Sutton and Barto 1998). As  $\lambda$  increases, choices approximate strict maximization of expected utilities. Expected utility of a message  $m$  in state  $s$  for a level  $n + 1$  speaker is here defined as  $H_n(s \mid m; L)$ , the probability that the hearer will assign to or choose the correct meaning. For literal speakers, utility only tracks truthfulness. Literal speakers choose any true message with equal probability but may send false messages as well with a probability dependent on  $\lambda$ .

Here are some concrete examples. **[MF: please double-check the numbers!]** A literal interpreter with lexicon  $L_{\text{bound}}$  assigns  $s_{\exists-\forall}$  a probability of  $H_0(s_{\exists-\forall} \mid m_{\text{some}}; L_{\text{bound}}) = 1$  after hearing  $m_{\text{some}}$ , while a literal interpreter with  $L_{\text{lack}}$  has  $H_0(s_{\exists-\forall} \mid m_{\text{some}}; L_{\text{lack}}) = 0.5$ :

$$H_0(\cdot \mid \cdot, L_{\text{bound}}) = \begin{matrix} & s_{\exists-\forall} & s_{\forall} \\ \begin{matrix} m_{\text{some}} \\ m_{\text{all}} \end{matrix} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{matrix} \quad H_0(\cdot \mid \cdot, L_{\text{lack}}) = \begin{matrix} & s_{\exists-\forall} & s_{\forall} \\ \begin{matrix} m_{\text{some}} \\ m_{\text{all}} \end{matrix} & \begin{bmatrix} .5 & .5 \\ 0 & 1 \end{bmatrix} \end{matrix}$$

If  $\lambda = 1$  literal speaker behavior looks like this:

$$S_0(\cdot \mid \cdot, L_{\text{bound}}) \approx \begin{matrix} & m_{\text{some}} & m_{\text{all}} \\ \begin{matrix} s_{\exists-\forall} \\ s_{\forall} \end{matrix} & \begin{bmatrix} .73 & .27 \\ .27 & .73 \end{bmatrix} \end{matrix} \quad S_0(\cdot \mid \cdot, L_{\text{lack}}) \approx \begin{matrix} & m_{\text{some}} & m_{\text{all}} \\ \begin{matrix} s_{\exists-\forall} \\ s_{\forall} \end{matrix} & \begin{bmatrix} .73 & .27 \\ .5 & .5 \end{bmatrix} \end{matrix}$$

If  $\lambda = 20$ , literal speakers will choose with less slack:

$$S_0(\cdot \mid \cdot, L_{\text{bound}}) \approx \begin{matrix} & m_{\text{some}} & m_{\text{all}} \\ s_{\exists \neg \forall} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{matrix} \quad S_0(\cdot \mid \cdot, L_{\text{lack}}) \approx \begin{matrix} & m_{\text{some}} & m_{\text{all}} \\ s_{\exists \neg \forall} & \begin{bmatrix} 1 & 0 \\ .5 & .5 \end{bmatrix} \end{matrix}$$

Pragmatic hearers of level 1 have the following probabilistic interpretation behavior for  $\lambda = 1$ :

$$H_1(\cdot \mid \cdot, L_{\text{bound}}) \approx \begin{matrix} & s_{\exists \neg \forall} & s_{\forall} \\ m_{\text{some}} & \begin{bmatrix} .73 & .27 \\ .27 & .73 \end{bmatrix} \\ m_{\text{all}} & \end{matrix} \quad H_1(\cdot \mid \cdot, L_{\text{lack}}) \approx \begin{matrix} & s_{\exists \neg \forall} & s_{\forall} \\ m_{\text{some}} & \begin{bmatrix} .59 & .41 \\ 0.35 & 0.65 \end{bmatrix} \\ m_{\text{all}} & \end{matrix}$$

For  $\lambda = 20$ , there will be less slack again: Pragmatic listeners of level 1 have the following probabilistic interpretation behavior for  $\lambda = 1$ :

$$H_1(\cdot \mid \cdot, L_{\text{bound}}) \approx \begin{matrix} & s_{\exists \neg \forall} & s_{\forall} \\ m_{\text{some}} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ m_{\text{all}} & \end{matrix} \quad H_1(\cdot \mid \cdot, L_{\text{lack}}) \approx \begin{matrix} & s_{\exists \neg \forall} & s_{\forall} \\ m_{\text{some}} & \begin{bmatrix} 0.67 & 0.33 \\ 0 & 1 \end{bmatrix} \\ m_{\text{all}} & \end{matrix}$$

With  $\lambda = 1$ , pragmatic speakers of level 1 will have:

$$S_1(\cdot \mid \cdot, L_{\text{bound}}) \approx \begin{matrix} & m_{\text{some}} & m_{\text{all}} \\ s_{\exists \neg \forall} & \begin{bmatrix} .73 & .37 \\ .37 & .73 \end{bmatrix} \\ s_{\forall} & \end{matrix} \quad S_1(\cdot \mid \cdot, L_{\text{lack}}) \approx \begin{matrix} & m_{\text{some}} & m_{\text{all}} \\ s_{\exists \neg \forall} & \begin{bmatrix} .62 & .38 \\ .38 & .62 \end{bmatrix} \\ s_{\forall} & \end{matrix}$$

For high  $\lambda = 20$ , pragmatic speakers of level 1 give us:

$$S_1(\cdot \mid \cdot, L_{\text{bound}}) \approx \begin{matrix} & m_{\text{some}} & m_{\text{all}} \\ s_{\exists \neg \forall} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ s_{\forall} & \end{matrix} \quad S_1(\cdot \mid \cdot, L_{\text{lack}}) \approx \begin{matrix} & m_{\text{some}} & m_{\text{all}} \\ s_{\exists \neg \forall} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ s_{\forall} & \end{matrix}$$

It is important to note that pragmatic agents of level 1 using  $L_{\text{lack}}$  associate  $m_{\text{some}}$  preferentially with  $s_{\exists \neg \forall}$ , in contrast to their literal counterparts of level 0. This association is not perfect, and usually less strong than what agents with a lexicalized upper bound in  $L_{\text{bound}}$  can achieve, even without pragmatic reasoning. Higher order iteration of reasoning types leads to stronger associations of  $m_{\text{some}}$  and  $s_{\exists \neg \forall}$  also for the hearer. Still, the case study presented in Section 3 will consider speaker and hearer behavior at levels 0 and 1, as the latter are the simplest pragmatic reasoning types which show a tendency to communicatively attuned pragmatic enrichment. Using only level 1 reasoning types is therefore a conservative choice that works against the fitness-based selection of pragmatic language use for a notion of fitness based on communicative efficiency, which is introduced next.



### 2.3.2 Fitness & fitness-based selection based on expressivity

Under the replicator dynamic the proportion of type  $i$  in a population will increase or decrease as a function of its relative fitness  $f_i$ . In the context of language evolution, fitness is usually associated with expressivity, i.e., the ability to successfully communicate with other language users from the same population (e.g., Nowak and Krakauer 1999, Nowak et al. 2000; 2002). Under a biological interpretation the assumption is that organisms have a higher chance of survival and reproduction if they are able to share and receive useful information via communication with peers. Under a cultural interpretation the picture is that agents themselves strive towards communicative success and therefore occasionally adapt or revise their behavior to achieve higher communicative success (see Benz et al. 2005:§3.3 for discussion).

The replicator equation gives us the means to make the ensuing dynamic precise, without necessarily committing to a biological or cultural interpretation. As above, the proportion of types in a given population is codified in a vector  $\vec{x}$ , where  $x_i$  is the proportion of type  $i$ . The fitness of type  $i$  is its average expected communicative success, or *expected utility* (EU), given the frequencies of types in the current population:

$$f_i = \sum_j x_j \text{EU}(t_i, t_j) .$$

The expected utility  $\text{EU}(t_i, t_j)$  for type  $i$  when communicating with type  $j$  is the average success of  $i$  when talking or listening to  $j$ . If agents are speakers half of the time, this yields:

$$\text{EU}(t_i, t_j) = 1/2 \text{EU}_S(t_i, t_j) + 1/2 \text{EU}_H(t_i, t_j) ,$$

where  $\text{EU}_S(t_i, t_j)$  and  $\text{EU}_H(t_i, t_j)$  are the expected utilities for  $i$  as a speaker and as a hearer when communicating with  $j$ , defined as follows, where  $n_i$  and  $n_j$  are type  $i$ 's and type  $j$ 's pragmatic reasoning types and  $L_i$  and  $L_j$  are their lexica:

$$\begin{aligned} \text{EU}_S(t_i, t_j) &= \sum_s P(s) \sum_m S_{n_i}(m \mid s; L_i) \sum_{s'} H_{n_j}(s' \mid m; L_j) \delta(s, s') , \\ \text{EU}_H(t_i, t_j) &= \text{EU}_S(t_j, t_i) . \end{aligned}$$

As usual,  $\delta(s, s') = 1$  iff  $s = s'$  and 0 otherwise.

### 2.3.3 Learnability

Languages are shaped not only by functionalist forces towards greater expressivity. Another important factor is the fidelity with which linguistic knowledge is transmitted. Among others, linguistic production can be prone to errors, states or messages may be perceived incorrectly, and

multiple languages may be compatible with the data learners are exposed to. These sources of uncertainty introduce variation in the transmission of linguistic knowledge from one generation to the next. In particular, learning biases in the iterated transmission process can influence language evolution substantially.

In biological evolution, where types are expressed genetically, transmission infidelity comes into the picture through infrequent and mostly random genetic mutations. However, an agent’s lexicon and pragmatic reasoning behavior is likely not inherited genetically. They need to be learned from observation. Concretely, when agents of type  $j$  want to adopt or imitate the linguistic behavior of type  $i$ , they observe the overt linguistic behavior of type  $i$  and need to infer the covert type that most likely produced the visible behavior. Iterated learning is a process in which languages are learned repeatedly from the observation of linguistic behavior of agents who have themselves acquired the language from observation and inference. In the simplest case there is a single teacher and a single learner. After sufficient training the learner becomes a teacher and produces behavior that serves as input for a new learner. Due to the pressure towards learnability it exerts, iterated learning alone generally leads to simpler and more regular languages (see Kirby et al. 2014 and Tamariz and Kirby 2016 for recent surveys).

Following Griffiths and Kalish (2007) we model language acquisition as a process of Bayesian inference in which learners combine the likelihood of a type producing the witnessed learning input with prior inductive biases. Experimental and mathematical results on iterated learning suggest that the outcome of this process reflects learners’ inductive biases (e.g., Kirby et al. 2014). In a Bayesian setting these biases can be codified in a prior  $P \in \Delta(T)$ , which reflects the amount of data a learner requires to faithfully acquire the language of the teacher (cf. Griffiths and Kalish 2007:450). The extent of the prior’s influence has been shown to heavily depend on the learning strategy assumed to underlie the inference process. On the one hand, early simulation results suggested that weak biases could be magnified by exposing learners to only small data samples (e.g. in Brighton 2002). On the other, Griffiths and Kalish’s (2007) mathematical characterization showed that iterated learning converges to the prior in the limit, i.e., that the resulting distribution over languages corresponds to the learners’ prior distribution and is not influenced by the amount of input given to them. This difference in predictions can be traced back to differences in the selection of hypotheses from the posterior. Griffith & Kalish’s convergence to the prior holds for learners that sample from the posterior. More deterministic strategies such as the adoption of the type with the highest posterior probability, so-called *maximum a posterior estimation* (MAP), increase the influence of both the prior and the data (Griffiths and Kalish 2007, Kirby et al. 2007). In the following, we use a parameter  $l \geq 1$  to modulate between posterior sampling and the MAP strategy. When  $l = 1$  learners sample from the posterior. The learners’ propensity to maximize the posterior grows as  $l$  increases.

Let  $D$  be the set of possible data that learners may be exposed to. This set  $D$  contains all sequences of state-message pairs of length  $k$ , e.g.,  $\langle \langle s_1, m_1 \rangle, \dots, \langle s_k, m_k \rangle \rangle$ . As  $k$  increases, learners have more data to base their inference on and so tend to recover the true types that generated a given sequence with higher probability. The mutation matrix  $Q$  of the replicator mutator dynamic in (1) can then be defined as follows:  $Q_{ji}$  is the probability that a learner acquires type  $i$  when learning from an agent of type  $j$ . The learner observes a length- $k$  sequence  $d$  of state-message pairs, but the probability  $P(d | t_j)$  with which sequence  $d = \langle \langle s_1, m_1 \rangle, \dots, \langle s_k, m_k \rangle \rangle$  is observed depends on type  $j$ 's behavior:

$$P(d = \langle \langle s_1, m_1 \rangle, \dots, \langle s_k, m_k \rangle \rangle | t_j) = \prod_{i=1}^k S_{n_j}(m_i | s_i; L_j),$$

where, as before,  $n_j$  is  $j$ 's pragmatic reasoning type and  $L_j$  is  $j$ 's lexicon. For a given observation  $d$ , the probability of acquiring type  $i$  is  $F(t_i | d)$ , so that:

$$Q_{ji} \propto \sum_{d \in D} P(d | t_j) F(t_i | d).$$

The acquisition probability  $F(t_i | d)$  given datum  $d$  is obtained by probability matching  $l = 1$  or a tendency towards choosing the most likely type  $l > 1$  from the posterior distribution  $P(\cdot | d)$  over types given the data, which is calculated by Bayes' rule:

$$\begin{aligned} F(t_i | d) &\propto P(t_i | d)^l \text{ and} \\ P(t_i | d) &\propto P(t_i) P(d | t_i). \end{aligned}$$

## 2.4 Model summary

Expressivity and learnability are central to the cultural evolution of language. These components can be modelled, respectively, as replication based on a measure of fitness in terms of communicative efficiency and iterated Bayesian learning. Their interaction is described by the discrete time replicator mutator dynamic in (1), repeated here:

$$x'_i = \sum_j Q_{ji} \frac{x_j f_j}{\sum_k x_k f_k}.$$

This equation defines the frequency  $x'_i$  of type  $i$  at the next time step, based on its frequency  $x_i$  before the step, its fitness  $f_i$ , and the probability that a learner infers  $i$  when observing the behavior of a type- $j$  agent. Fitness-based selection is here thought of not as biological (fitness as expected relative number of offspring) but cultural (fitness as likelihood of being imitated or repeated) evolution, since the types that the dynamic operates on are pairs consisting of a lexicon

and a pragmatic use pattern. A type’s expressivity depends on its communicative efficiency within a population while its learnability depends on the fidelity by which it is inferred by new generations of learners. The learners’ task is consequently to perform a joint inference over types of linguistic behavior and lexical meaning.

The model has three parameters:  $\lambda$  regulates the degree to which pragmatic speakers choose messages that maximize the chance of communicative success;  $k$  is the number of observations for each language learner;  $l$  regulates where the learners’ inference behavior lies on a spectrum from probability matching to choice of the most likely parent type.

### 3 Case study: scalar implicatures

The model of the previous section formalizes the evolutionary competition between different pairs of lexical concepts and ways of using them. This section looks at a concrete case study on scalar implicatures. Concretely, we will engage in the following formal thought experiment: if a population of language users could freely combine different lexica with different pragmatic strategies, what are the conditions under which the majority view of scalar implicatures could have evolved?

Recall that the majority view is that scalar implicatures are non-lexicalized pragmatic enrichments. Scalar implicature triggers like *some*, *warm* or *may* are semantically weak expressions for which logically stronger expressions are salient, e.g., *all*, *hot* or *must*. For instance, *some* is entailed by *all*. If the sentence “Chris owns all of Johnny Cash’s albums” is true, then “Chris owns some of Johnny Cash’s albums” is also true. However, while weaker expressions such as *some* are truth-conditionally compatible with stronger alternatives such as *all*, this is not what their use is normally taken to convey. Instead, the use of a less informative expression when a more informative one could have been used can license a defeasible inference that stronger alternatives do not hold (cf. Horn 1972, Gazdar 1979). That is, a hearer who assumes the speaker to be able and willing to provide all relevant information can infer that stronger alternatives do not hold because the speaker used a weaker alternative instead. In this way, “Chris owns some of Johnny Cash’s albums” is strengthened to convey that she owns *some but not all* albums. The upper bound, so the majority view, is a pragmatic inference but not part of the conventional meaning.

[MF: I wonder if this paragraph is really still necessary after the introduction?]

In the following we consider a specific application of the model from Section 2 which allows to address the question if or when scalar inferences might (not) lexicalize. We consider what is perhaps one of the simplest non-trivial setups that speak to this matter and reflect on its limitations in Section 4. The setup is introduced in Section 3.1. Section 3.2 describes the simulations and their results.

### 3.1 Setup

To fill the model from Section 2 with life, we need to specify the sets of states, messages and lexica. Additionally, we want to explore the effects of a learning bias in favor of simple lexical concepts. One way of motivating and formalizing such a bias is introduced below.

#### 3.1.1 States, messages, concepts and Lexica

Consider a state space with three states  $S = \{s_\emptyset, s_{\exists-\forall}, s_\forall\}$  and think of it as a partition of possible worlds into cells where none, some or all of the  $A$ s are  $B$ s, for some arbitrary fixed predicates  $A$  and  $B$ . Eight concepts can be distinguished based on their truth or falsity in three world states, six of which are not contradictory or tautological.

A lexicon  $L$  is a mapping  $M \rightarrow C$  from messages to concepts. With three messages there are  $6^3 = 216$  possible lexica. Some assign the same concept to more than one message and others lexicalize the same concepts but associate them with different messages. Out of these possible lexica, three kinds are of particular relevance. First, lexica that assign the same concept to more than one message. Such lexica lack in expressivity but may be favored by particular learning biases nonetheless. Second, lexica that conventionalize upper-bounds to realize a (quasi-)partition of the state space. Finally, lexically that do not lexicalize an upper bound but allow it to be conveyed pragmatically due to the presence of a stronger lexical item. There are six lexica of the second kind and six of the third. The following three lexica exemplify each kind:

	<u><math>L_{\text{all}}</math></u>			<u><math>L_{\text{bound}}</math></u>			<u><math>L_{\text{lack}}</math></u>		
	$m_{\text{none}}$	$m_{\text{some}}$	$m_{\text{all}}$	$m_{\text{none}}$	$m_{\text{some}}$	$m_{\text{all}}$	$m_{\text{none}}$	$m_{\text{some}}$	$m_{\text{all}}$
$s_\emptyset$	0	0	0	1	0	0	1	0	0
$s_{\exists-\forall}$	0	0	0	0	1	0	0	1	0
$s_\forall$	1	1	1	0	0	1	0	1	1

Lexicon  $L_{\text{all}}$  is clearly bad for communication: all message and interpretation choices will be equally likely for all types; no information will be conveyed with it. Users of  $L_{\text{bound}}$  need not resort to pragmatic reasoning to convey an upper-bound with weak scalar expressions. Instead, this information is already provided by the semantics of their lexicon, giving them an advantage in terms of expressivity. Types with such lexica may be expected to generally achieve higher communicative success than pragmatic users with a lexicon of type  $L_{\text{lack}}$  (see Section 2.3.1).

Recall that types are a combination of a lexicon and a manner of language use. We analyze the model's predictions in populations of types with one of the two behaviors introduced earlier; literal or pragmatic. The former correspond to level 0 reasoners and the latter to ones of level 1. Accordingly, we consider a total of 432 types. Six are variants of pragmatic language users with

$C \rightarrow_2 C \wedge C$	$X \rightarrow_1 \{A, B\}$
$C \rightarrow_2 \neg C$	$X \rightarrow_1 X \cap X$
$C \rightarrow_1 X \subseteq X$	$X \rightarrow_1 X \cup X$
$C \rightarrow_1 X \neq \emptyset$	
$C \rightarrow_1 X = \emptyset$	

Table 1: Toy grammar in a set-theoretic LOT with weighted rules.

$L_{\text{lack}}$ -like lexica. These will be our target types, because they represent lexica and language use that conform to the majority view of scalar implicatures. Twelve types are either literal or pragmatic speakers with lexical of the  $L_{\text{bound}}$  kind. These will be the target types main competitors in evolutionary competition.

### 3.1.2 An inductive learning bias for semantic simplicity

There is a growing effort to develop empirically testable representational languages that allow for the measure of semantic complexity. For instance, so-called *languages of thought* (LOT) have been put to test in various rational probabilistic models that show encouraging results (see, e.g., Katz et al. 2008, Piantadosi et al. view; 2012 and Piantadosi and Jacobs 2016 for recent discussion). At its core, a LOT defines a set of operations and composition rules from which concepts can be derived. As a first approximation and for the sake of concreteness, we follow this approach to make the preference of learners for simpler semantic representations precise and define the complexity of a representation as a function of its derivation cost in a weighed generative LOT.

Our toy grammar of concepts is given in Table 1. This grammar uses basic set-theoretic operations to form expressions which can be evaluated as true or false in states  $s_\emptyset$ ,  $s_{\exists-\forall}$  or  $s_\forall$  from above. Applications of generative rules have a cost attached to them. Here we simply assume that Boolean combinations of concepts are more complex than atomic concepts and that otherwise each rule application adds the same cost unit. Table 2 lists all six concepts relevant here, their truth conditions and the simplest formula that expresses this concept from the grammar in Table 1.

The complexity measure for lexical concepts from Table 2 is used to define a learning bias that favor simpler concepts over more complex ones. The prior probability of a type is just the prior probability of its lexicon. The prior of a lexicon is a function of the complexity of the lexical

intuitive name	$s_{\emptyset}$	$s_{\exists-\forall}$	$s_{\forall}$	least complex formula	complexity
“all”	0	0	1	$A \subseteq B$	3
“some but not all”	0	1	0	$A \cap B \neq \emptyset \wedge A \neq \emptyset$	8
“some”	0	1	1	$A \cap B \neq \emptyset$	4
“none”	1	0	0	$A \cap B = \emptyset$	4
“none or all”	1	0	1	$\neg(A \cap B \neq \emptyset \wedge A \neq \emptyset)$	10
“not all”	1	1	0	$\neg(A \subseteq B)$	5

Table 2: Available concepts and their minimal derivation length

representations in its image set. Lexica with simpler concepts have a higher prior. One simple way of defining such priors over lexica (and thereby types) is:

$$P(L) \propto \prod_{c \in \text{Im}(L)} P(c), \text{ with } P(c) \propto \max_{c'} \text{Compl}(c') - \text{Compl}(c) + 1,$$

where  $\text{Compl}(c)$  is the complexity of the minimal derivation length of  $c$  according to the LOT-grammar (see Table 2). There are many ways to define priors over lexica (see, e.g., Goodman et al. 2008, Piantadosi et al. 2012) but the key assumption here, common to all of them, is that simple representational expressions should be favored over more complex ones. We should stress that these details – from the generative grammar to its complexity measure – are to be regarded as one convenient operationalization of one general approach to explicating learning biases; this is not a commitment that this general approach is necessarily superior or that, within it, this particular instrumentalization is the single most plausible.

### 3.2 Simulation results

Recall that there are three parameters: soft-max parameter  $\lambda$  affects how strongly pragmatic speakers favor more informative messages; the bottleneck size  $k$  influences how faithfully learners can identify their teacher type;  $l$  defines the learners disposition towards choosing the most likely teacher type from the posterior distribution. We expect that competitor types (types with lexica of the kind  $L_{\text{bound}}$ ) have a fitness advantage over target types (pragmatic agents with lexica of the kind  $L_{\text{lack}}$ ), especially for very low levels of  $\lambda$ . Selection based on fitness alone may therefore not lead to prevalence of target types in the population. On the other hand, lexica of type  $L_{\text{lack}}$  are simpler than those of type  $L_{\text{bound}}$  by the postulated measure from above. This may make them more likely to be adopted by learners, especially when  $k$  is low so that teacher types are relatively indistinguishable based on their behavior in general and when  $l$  is high. Still,

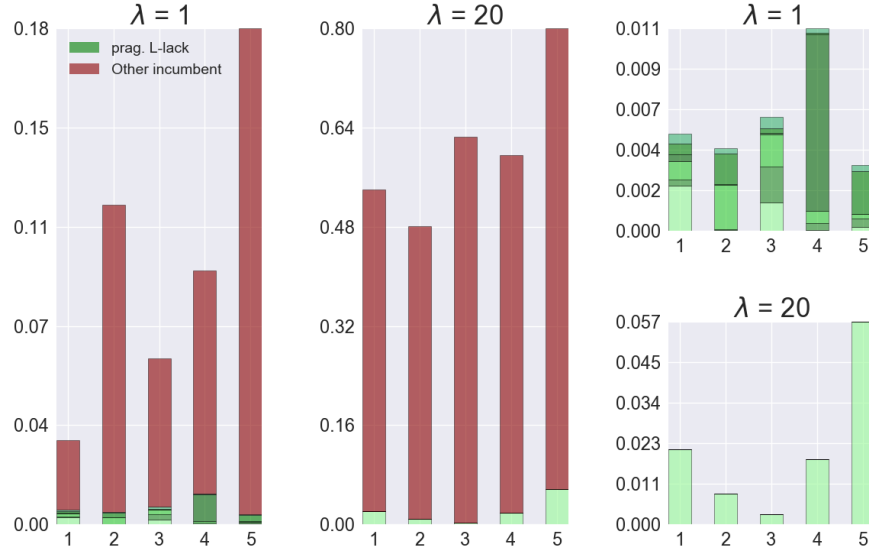


Figure 2: Stacked proportion of pragmatic  $L_{\text{lack}}$ -style types and incumbent types, when other than pragmatic  $L_{\text{lack}}$ , in 5 independent populations after 50 generations under only a pressure for expressivity. The right-most plots zoom in on only the proportion of pragmatic  $L_{\text{lack}}$  types.

lexica of the kind  $L_{\text{all}}$  are in turn more likely *a priori* than lexica of the kind  $L_{\text{lack}}$ . Simulation results will shed light on the question whether target types can emerge and for which parameter constellations.

To better understand the workings a pressure towards communicative efficiency and a pressure towards learnability, we look at the behavior of the replicator and mutator step first in isolation, and then in combination. All simulation runs are initialized with an arbitrary distribution over types, constituting a population’s first generation. All reported results are the outcome of 50 update steps for each different constellation of parameter values. This ensures that these outcomes correspond to developmental plateaus in which change is, if not absent, then at least very slow. In other words, even if the resulting states do not correspond to an eventual attracting state, they characterize long almost stationary states in which the system remains for a very long time. As specified in §2.3.3, the mutation matrix  $Q$  can be obtained by considering all possible state-message sequences of length  $k$ . Given that this is intractable for large  $k$ , the sets of data learners are exposed to are approximated by sampling 250  $k$ -length sequences from each type’s production probabilities.



### 3.2.1 Replication only: selection based on expressivity and communicative success

As expected, selection based on communicative success is sensitive to  $\lambda$  as it influences signaling behavior. This is showcased in Figure 2, which depicts the proportion of target types in 5 independent populations after 50 generations, as well as that of the type with the highest proportion in a given population, the incumbent. With low  $\lambda$  many types have very similar behavior, so that evolutionary selection lacks grip and becomes very slow. The result is a very long transition with near stagnancy in a rather homogeneous population with many types. Conversely, higher  $\lambda$  promotes more rational linguistic behavior, widening the gap in expressivity between types and promoting more homogeneous populations. As suggested by Figure 2, the incumbent in most populations is not one of the six pragmatic  $L_{\text{lack}}$ -style types. That is, a pressure only for expressivity does not lend a justification to the systematic prevalence of a lack of upper-bounds under any  $\lambda$ -value. For instance, in 1000 independent populations with  $\lambda = 20$  pragmatic  $L_{\text{lack}}$  types only made up 11 cases of incumbency, corresponding to a mean proportion of 0.003 across populations. By contrast, 913 incumbents were  $L_{\text{bound}}$ -style users with close to an even share between literal (454) and pragmatic types (459), corresponding to a mean proportion of about .48 taken together.

### 3.2.2 Iterated learning only

The effect of iterated learning without a pressure for expressivity using either posterior sampling ( $l = 1$ ) or a stronger tendency towards posterior maximization ( $l = 15$ ) is shown in Figure 3 together with the prior over types. The prior shows that while users of  $L_{\text{lack}}$  are not the most favored by the inductive bias (compared, e.g., to  $L_{\text{all}}$ ) they are nevertheless more advantaged than others, such as  $L_{\text{bound}}$ , in virtue of the relatively simple semantics they conventionalize. Crucially,  $L_{\text{lack}}$  enables its users to convey each state with a single message when combined with pragmatic reasoning provided sufficiently high  $\lambda$ . This makes it less likely to be confused with other types if the learning data is not too sparse ( $k \geq 5$ ). Put differently, learners have a higher propensity to infer pragmatic  $L_{\text{lack}}$  over types that might be hard or impossible to distinguish it based on overt linguistic behavior, such as  $L_{\text{bound}}$ . Moreover,  $L_{\text{lack}}$  is also less likely to be confused with types with different observable behavior because its pragmatic use approximates a one-to-one form-meaning mapping. As a consequence, a stronger propensity to maximize the posterior increases their proportion in the population.

However, in contrast to a pressure only for expressivity with high  $\lambda$ , learnability alone does not succeed in selecting for a single prevalent type. Indeed, all six target types tend to coexist at roughly equal proportion. Each is passed on to next generation with the same faithfulness and, differently from a pressure for communicative success, they do not stand in competition with each

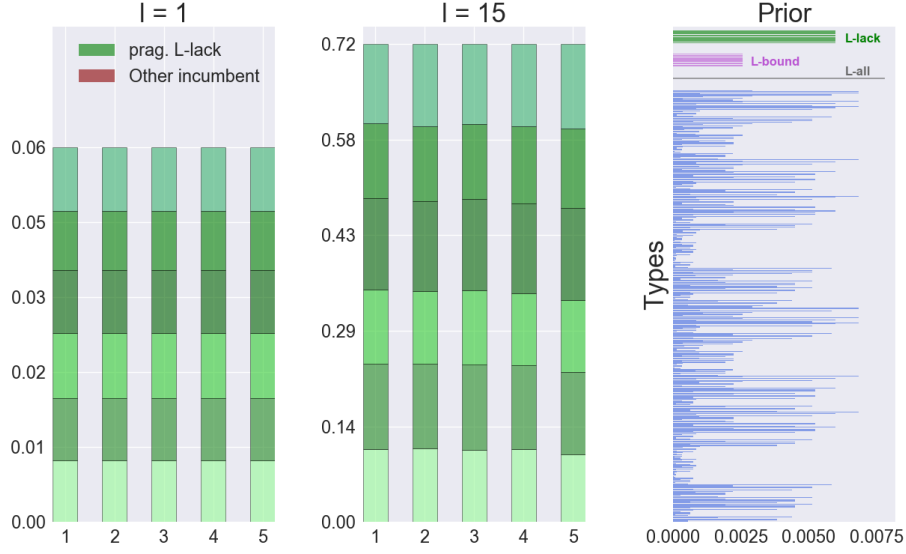


Figure 3: Stacked proportion of pragmatic  $L_{\text{lack}}$ -style types and incumbent types, when other than pragmatic  $L_{\text{lack}}$ , in 5 independent populations after 50 generations under only a pressure for learnability ( $\lambda = 20, k = 5$ ). The types' prior probability is shown in the right-most plot.

other. As implicit in Figure 3, in 1000 independent populations all incumbents were pragmatic  $L_{\text{lack}}$  users provided sufficiently high  $\lambda$ , with each reaching approximately the same proportion of users in the population. As with expressivity only, low values of  $\lambda$  make the differences in observable behavior across types less pronounced and therefore reflect the learners' inductive bias more faithfully, favoring functionally deficient but a priori preferred types such as those that use  $L_{\text{all}}$ . A pressure for learnability alone may consequently lead to a spread of communicatively suboptimal types that are easier to learn. In the extreme, when  $l = 1$  and  $\lambda = 1$  all of 1000 independent populations had users of  $L_{\text{all}}$  as incumbents.

### 3.2.3 Combining pressures of expressivity and learnability

Pressures for communicative success or learnability are not sufficient on their own to have a single target type dominate the population. When pressured toward expressivity, the slight communicative advantage of  $L_{\text{bound}}$  users leads to its prevalence. When pressured toward learnability, pragmatic  $L_{\text{lack}}$  is promoted over functionally similar but semantically more complex alternatives such as  $L_{\text{bound}}$ . However, on its own learnability does not foment the propagation of these types across the population. What is more, it leads to heterogeneous populations that foster an equal share of types of a kind, such as different pragmatic  $L_{\text{lack}}$  types.

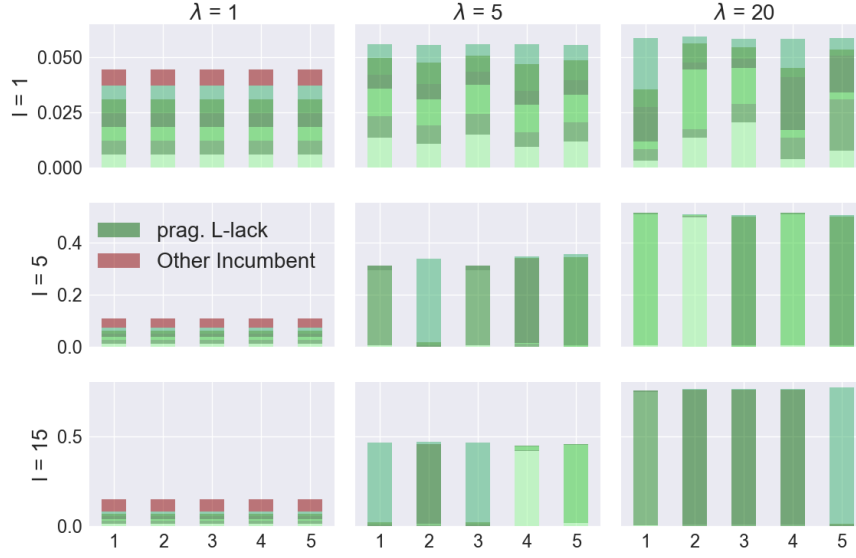


Figure 4: Stacked proportion of pragmatic  $L_{\text{lack}}$ -style types and incumbent types, when other than pragmatic  $L_{\text{lack}}$ , in 5 independent populations after 50 generations under both pressures ( $k = 5$ ).

Figure 4 illustrates the combined effects of both pressures for a sample of  $\lambda$  and  $l$  values. These results show that an inductive learning bias for simpler semantics in tandem with functional pressure can lead to the selection of  $L_{\text{lack}}$ -like semantics paired with pragmatic language use. The proportion of these types increases with  $\lambda$  and  $l$ . Pressure towards expressivity magnifies the effects of iterated learning and dampens the proliferation of types of a kind that are equal in expressivity *and* learnability. A pressure towards learnability favors the transmission of simpler semantics and thereby makes types that exploit pragmatic reasoning to strengthen lexical meaning more likely to be adopted over alternatives that do not.

As before, low  $\lambda$  and  $l$  lead to the incumbency of communicatively suboptimal types that are a priori favored, such as  $L_{\text{all}}$ . An increase in  $\lambda$  leads to the selection and incumbency of pragmatic  $L_{\text{lack}}$  but does not lead to monomorphic populations if learners sample from the posterior. Finally, a combination of sufficiently rational signaling behavior with some tendency for posterior maximization leads to increasingly monomorphic populations of types that convey upper-bounds pragmatically. The difference between the mean of the highest pragmatic  $L_{\text{lack}}$ -like type in 1000 independent populations and the highest proportion of other types across  $\lambda$  and  $l$  values is shown in Figure 5.

As for the effect of the sequence length  $k$  not mentioned so far, it influences populations in a predictable way: small values lead to more heterogeneous populations that reflect the learner’s

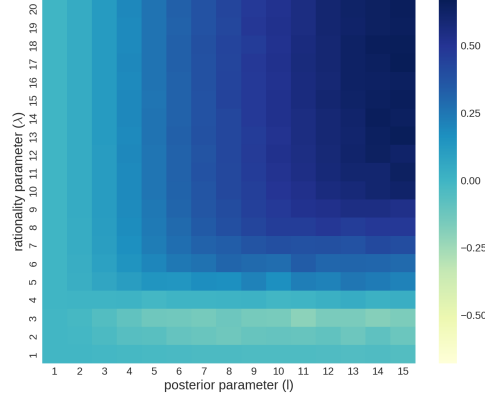


Figure 5: Difference between mean proportion of highest pragmatic  $L_{\text{lack}}$  type and highest other type in 1000 independent populations after 50 generations under both pressures ( $k = 5$ ). [TB: I will run more simulations with higher  $\lambda$  and  $l$ . I'm pretty sure the difference still increases.]

prior more faithfully. This is due to the fact that the likelihood that a small sequence was produced by any type is relatively uniform (modulo prior). By contrast, larger values increasingly allow learners to differentiate types with different signaling behaviors.

To recapitulate, other than the involvement of pressure on both expressivity and learnability, the resulting proportion of pragmatic  $L_{\text{lack}}$  speakers primarily hinges on three factors. First, the degree to which agents try to maximize communicative success. This plays a role both in expressivity as well as in producing data that allows learners to discriminate this type from others. Second, the inductive bias, which leads learners to prefer simpler over more complex semantic representations in acquisition. Lastly, the posterior parameter, as it magnifies the effects of the learning bias in tandem with replication.

### 3.3 Summary

Under the assumption of a learning bias for simpler lexical representations, our results suggest that a lack of semantic upper-bounds coupled with pragmatic reasoning can overcome communicative pressures and come to dominate a population. This prediction hinges on three assumptions. First, that language is pressured toward both expressivity and learnability. Second, that language users favor expressions/interpretations expected to have a higher likelihood of being understood/meant over ones with a lower likelihood. Lastly, that learners prefer simpler over more complex lexical representations and exhibit a tendency towards the acquisition of the type that best explains the learning data.

## 4 General discussion

The approach introduced here combines game theoretic models of functional pressure towards efficient communication (Nowak and Krakauer 1999), effects of transmission perturbations on (iterated) language learning (Griffiths and Kalish 2007), probabilistic speaker and listener types of varied degrees of pragmatic sophistication (Frank and Goodman 2012, Franke and Jäger 2014) as well as reasoning about unobservable lexical representations (Bergen et al. 2012; 2016). This model allows a structured conceptual investigation of the co-evolution of conventional meaning and pragmatic language use. Main innovations of the model are (i) its modular separation of expressivity and learnability, allowing for their isolated and combined analysis, (ii) the characterization of language learning as a joint inference over types of pragmatic behavior and lexical meaning, and (iii) the possibility to trace the co-evolution of conventional semantics and pragmatic use.

The goal to decouple but model both expressivity and learnability has also recently been addressed by Kirby et al. (2015). In contrast to our proposal, Kirby et al. model expressivity as exerting its force only in the production of learning data. The degree of mutual understanding of interlocutors central to replication and to our notion of expressivity is absent. That is, while our proposal combines bidirectional horizontal transmission with its vertical and unidirectional counterpart, Kirby et al. only consider the latter’s influence. Our reasoning behind the inclusion of the former lies in the empirical and theoretical observation that learnability alone can lead the selection of functionally defective languages, as showcased by  $L_{all}$ . This outcome has been reported in a number of laboratory experiments where subjects learned and subsequently reproduced the language produced by a previous participant over multiple iterations, leading to a proliferation of ambiguous languages that associate many meanings with a single form (see, e.g., Silvey et al. 2014 and experiment 1 in Kirby et al. 2008). In contrast, experiments involving an interactive component have been found to foster languages that enable interlocutors to communicatively distinguish meanings more accurately (e.g., Fay and Ellison 2013; for a review of laboratory results under the iterated learning paradigm and further discussion see Kirby et al. 2015, Tamariz and Kirby 2016). It is not evident how to compare these empirical findings given that they consider distinct meaning spaces, modes of transmission, iterations and feedback given to participants. However, on a general level they may suggest that there is an important difference between a language generating learnable linguistic data and its actual performance as a means of information transfer. That is, we contend that successful information transfer in a linguistic community is central to the adoption of a communication system and that this measure may not be adequately reflected by production alone.

The main result of our case study is that types that correspond to the majority view of scalar

implicatures (scalar readings are non-lexicalized pragmatic enrichments) can come to dominate a population. This can happen under the assumption that simpler semantic representations are more likely to be learned (cf. Chater and Vitányi 2003). Pragmatic language use can then be recruited indirectly by a preference for simpler lexical concepts. Under this view, semantics and pragmatics play a synergic role: pragmatic use allows maintenance of simpler concepts; pressure towards representational simplicity indirectly promotes pragmatic over literal language use. As a consequence, iterated transmission and use of language lead to a regularization that may explain the lack of lexicalization of systematic pragmatic enrichments.

While the results of this case study are interesting, they also raise a number of critical issues. First of all, it is not clear to what extent other types should be present in the final population. It seems reasonable to expect functionally suboptimal lexica such as  $L_{\text{all}}$  to be ruled out but matters are different for  $L_{\text{bound}}$ . The prediction that natural language communities are homogeneous or that speakers may entertain  $L_{\text{lack}}$ -like semantics for one scalar expression and  $L_{\text{bound}}$ -like semantics for another is not implausible (cf. ?). Alternatively, a stronger tendency for posterior maximization in learning has to be assumed. This empirical issue relates to other two aspects left undiscussed: disadvantages of pragmatic reasoning and the effect of state frequencies on the fossilization of pragmatic inferences. We tacitly assumed pragmatic reasoning to come at no cost. However, there is experimental evidence that suggests that the pragmatic derivation of upper-bounds costs effort and takes additional processing time (cf. Neys and Schaeken 2007, Huang and Snedeker 2009). This raises the question at which point such usage-based cost undercuts the learnability advantage of simpler semantic representations. Should cost play a role, then its effect is bound to depend on the frequency with which a given scalar expression is used. It may be that frequently drawn scalar implicatures fossilize to avoid cost whereas infrequent ones are derived on-line to avoid complexity in acquisition. This also opens a possible venue to address the preceding question about the expected presence of  $L_{\text{bound}}$ -like semantics in a population, but further empirical evidence is needed to assess these matters.

## 5 Conclusion

The cultural evolution of meaning is influenced by intertwined pressures. We set out to investigate this process by putting forward a model that combines a pressure toward successful information transfer with perturbations that may arise in the transmission of linguistic knowledge in acquisition. Its objects of selection and replication are pairs of lexical meanings and patterns of language use. This allows the model to trace the interaction between conventional meaning and pragmatic use. Additionally, it takes the challenge serious of neither semantics nor pragmatics being directly observable. Instead, learners need to infer these unobservables from overt data that

results from their combination. These components and their mutual influence were highlighted in a case study on the lack of lexical upper-bounds in weak scalar expressions that showed that, when pressured for learnability and expressivity, the former force drives for simpler semantic representations inasmuch as pragmatics can compensate for lack of expressivity in use. That is, the relative learning advantage of simpler semantics in combination with functional pressure in use may offer an answer to why natural languages fail to lexicalize systematic pragmatic inferences. And, more broadly, to a division of labors between semantics and pragmatics.

## References

- van der Auwera, J. (2010). On the diachrony of negation. In *The Expression of Negation*, pp. 73–110. Walter de Gruyter GmbH.
- Baronchelli, A., A. Puglisi, and V. Loreto (2008). Cultural route to the emergence of linguistic categories. *PNAS* 105(23), 7936–7940.
- Batali, J. (1998). Computational simulations of the emergence of grammar. In J. R. Hurford, M. Studdert-Kennedy, and C. Knight (Eds.), *Evolution of Language: Social and Cognitive Bases*. Cambridge, UK: Cambridge University Press.
- Beggs, A. (2005). On the convergence of reinforcement learning. *Journal of Economic Theory* 122(1), 1–36.
- Benz, A., G. Jäger, and R. van Rooij (2005). *An Introduction to Game Theory for Linguists*, pp. 1–82. Palgrave.
- Bergen, L., N. D. Goodman, and R. Levy (2012). That’s what she (could have) said: How alternative utterances affect language use. In *Proceedings of Thirty-Fourth Annual Meeting of the Cognitive Science Society*.
- Bergen, L., R. Levy, and N. D. Goodman (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*.
- Blume, A., Y.-G. Kim, and J. Sobel (1993). Evolutionary stability in games of communication. *Games and Economic Behavior* 5(5), 547–575.
- Börger, T. and R. Sarin (1997). Learning through reinforcement and replicator dynamics. *Journal of Economic Theory* 77(1), 1–14.
- Bott, L. and I. A. Noveck (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language* 51(3), 437–457.

- Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life* 8(1), 25–54.
- Chater, N. and P. Vitányi (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences* 7(1), 19–22.
- Chierchia, G., D. Fox, and B. Spector (2012). Scalar implicature as a grammatical phenomenon. In C. Maienborn, K. von Steubner, and P. Portner (Eds.), *Semantics. An International Handbook of Natural Language Meaning*, pp. 2297–2332. Berlin: de Gruyter.
- Degen, J. and M. K. Tenenbaum (2015). Processing scalar implicatures: A constraint-based approach. *Cognitive Science* 39, 667–710.
- Fay, N. and T. M. Ellison (2013). The cultural evolution of human communication systems in different sized populations: Usability trumps learnability. *PLoS ONE* 8(8).
- Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature* 407(6804), 630–633.
- Frank, M. C. and N. D. Goodman (2012). Predicting pragmatic reasoning in language games. *Science* 336(6084), 998–998.
- Franke, M. (2016). The evolution of compositionality in signaling games. *Journal of Logic, Language and Information* 25(3–4), 355–377.
- Franke, M. and G. Jäger (2014). Pragmatic back-and-forth reasoning. *Semantics, Pragmatics and the Case of Scalar Implicatures.*, 170–200.
- Franke, M. and G. Jäger (2016). Probabilistic pragmatics, or why bayes’ rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft* 35(1), 3–44.
- Gazdar, G. (1979). *Pragmatics, Implicature, Presupposition and Logical Form*. New York: Academic Press.
- Geurts, B. (2010). *Quantity Implicatures*. Cambridge, UK: Cambridge University Press.
- Gong, T. (2007). *Language Evolution from a Simulation Perspective: On the Coevolution of Compositionality and Regularity*. Chinese University of Hong Kong.
- Goodman, N., J. Tenenbaum, J. Feldman, and T. Griffiths (2008). A rational analysis of rule-based concept learning. *Cognitive Science: A Multidisciplinary Journal* 32(1), 108–154.



- Goodman, N. D. and M. C. Frank (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences* 20(11), 818–829.
- Goodman, N. D. and A. Stuhlmüller (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5, 173–184.
- Grice, P. (1975). Logic and conversation. In *Studies in the Ways of Words*, Chapter 2, pp. 22–40. Cambridge, MA: Harvard University Press.
- Griffiths, T. L. and M. L. Kalish (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science* 31(3), 441–480.
- Grodner, D. J., N. M. Klein, K. M. Carbary, and M. K. Tanenhaus (2010). “some,” and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition* 166, 42–55.
- Helbing, D. (1996). A stochastic behavioral model and a ‘microscopic’ foundation of evolutionary game theory. *Theory and Decision* 40(2), 149–179.
- Hirschberg, J. (1985). *A Theory of Scalar Implicature*. Ph. D. thesis, University of Pennsylvania.
- Hofbauer, J. (1985). The selection mutation equation. *Journal of Mathematical Biology* 23, 41–53.
- Hofbauer, J. and K. Sigmund (2003). Evolutionary game dynamics. *Bulletin of the American Mathematical Society* 40(04), 479–520.
- Horn, L. R. (1972). *On the Semantic Properties of Logical Operators in English*. Bloomington, IN: Indiana University Linguistics Club.
- Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, Form and Use in Context*, pp. 11 – 42. Washington: Georgetown University Press.
- Huang, Y. T. and J. Snedeker (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology* 58(3), 376–415.
- Hurford, J. R. (1989). Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua* 77(2), 187–222.
- Huttegger, S. M. (2007). Evolution and the explanation of meaning. *Philosophy of Science* 74, 1–27.

- Jäger, G. (2007). Evolutionary game theory and typology: A case study. *Language* 83(1), 74–109.
- Jäger, G. and R. van Rooij (2007). Language structure: psychological and social constraints. *Synthese* 159(1), 99–130.
- Katz, Y., N. D. Goodman, K. Kersting, C. Kemp, and J. B. Tenenbaum (2008). Modeling semantic cognition as logical dimensionality reduction. In *Proceedings of Thirtieth Annual Meeting of the Cognitive Science Society*.
- Kemp, C. and T. Regier (2012). Kinship categories across languages reflect general communicative principles. *Science* 336(6084), 1049–1054.
- Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic Evolution Through Language Acquisition*, pp. 173–204. Cambridge University Press (CUP).
- Kirby, S., H. Cornish, and K. Smith (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences* 105(31), 10681–10686.
- Kirby, S., M. Dowman, and T. L. Griffiths (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences* 104(12), 5241–5245.
- Kirby, S., T. Griffiths, and K. Smith (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology* 28, 108–114.
- Kirby, S. and J. R. Hurford (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi and D. Parisi (Eds.), *Simulating the Evolution of Language*, pp. 121–148. Springer.
- Kirby, S., M. Tamariz, H. Cornish, and K. Smith (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition* 141, 87–102.
- Lenaerts, T., B. Jansen, K. Tuyls, and B. D. Vylder (2005). The evolutionary language game: An orthogonal approach. *Journal of Theoretical Biology* 235, 566–582.
- Levinson, S. C. (1983). *Pragmatics*. Cambridge, UK: Cambridge University Press.
- Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge: Harvard University Press.
- Luce, D. R. (1959). *Individual choice behavior: a theoretical analysis*. Wiley.

- Martinet, A. (1962). *Functional View of Language*. Oxford: Clarendon Press.
- Neys, W. D. and W. Schaeken (2007). When people are more logical under cognitive load. *Experimental Psychology* 54(2), 128–133.
- Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press.
- Nowak, M. A., N. L. Komarova, and P. Niyogi (2001). Evolution of universal grammar. *Science* 291(5501), 114–118.
- Nowak, M. A., N. L. Komarova, and P. Niyogi (2002). Computational and evolutionary aspects of language. *Nature* 417(6889), 611–617.
- Nowak, M. A. and D. C. Krakauer (1999). The evolution of language. *Proceedings of the National Academy of Sciences* 96(14), 8028–8033.
- Nowak, M. A., J. B. Plotkin, and V. A. A. Jansen (2000). The evolution of syntactic communication. *Nature* 404(6777), 495–498.
- Pagel, M. (2009). Human language as a culturally transmitted replicator. *Nature Reviews Genetics* 10, 405–415.
- Piantadosi, S. T. (2014). Zipf’s word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review* 21(5), 1112–1130.
- Piantadosi, S. T. and R. A. Jacobs (2016). Four problems solved by the probabilistic language of thought. *Current Directions in Psychological Science* 25(1), 54–59.
- Piantadosi, S. T., J. B. Tenenbaum, and N. D. Goodman (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition* 123(2), 199–217.
- Piantadosi, S. T., J. B. Tenenbaum, and N. D. Goodman (under review). Modeling the acquisition of quantifier semantics: a case study in function word learnability.
- Piantadosi, S. T., H. Tily, and E. Gibson (2012). The communicative function of ambiguity in language. *Cognition* 122(3), 280–291.
- van Rooij, R. and T. de Jager (2012). Explaining quantity implicatures. *Journal of Logic, Language and Information* 21(4), 461–477.
- Sandholm, W. H. (2010). *Population Games and Evolutionary Dynamics*. Cambridge, MA: MIT Press.

- Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy* 27, 367–391.
- Schlag, K. H. (1998). Why imitate, and if so, how? *Journal of Economic Theory* 78(1), 130–156.
- Silvey, C., S. Kirby, and K. Smith (2014). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science* 39(1), 212–226.
- Skyrms, B. (2010). *Signals: Evolution, Learning, and Information*. Oxford: Oxford University Press.
- Smith, K., S. Kirby, and H. Brighton (2003). Iterated learning: A framework for the emergence of language. *Artificial Life* 9, 371–386.
- Spike, M., K. Stadler, S. Kirby, and K. Smith (2016). Minimal requirements for the emergence of learned signaling. *Cognitive Science*.
- Steels, L. (1995). A self-organizing spatial vocabulary. *Artificial Life* 2(3), 319–332.
- Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews* 8(4), 339–356.
- Steels, L. and T. Belpaeme (2005). Coordinating perceptually grounded categories through language: A case study for color. *Behavioral and Brain Sciences* 28(4), 469–529.
- Sutton, R. S. and A. G. Barto (1998). *Introduction to Reinforcement Learning*. Cambridge, MA, USA: MIT Press.
- Tamariz, M. and S. Kirby (2016). The cultural evolution of language. *Current Opinion in Psychology* 8, 37–43.
- Taylor, P. D. and L. B. Jonker (1978). Evolutionary stable strategies and game dynamics. *Mathematical Bioscience* 40(1–2), 145–156.
- Thompson, B., S. Kirby, and K. Smith (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences of the United States of America* 113(16), 4530–4535.
- Traugott, E. C. (2004). Historical pragmatics. In L. R. Horn and G. Wand (Eds.), *The Handbook of Pragmatics*, pp. 538–561. Blackwell Publishing.
- Verhoef, T., S. Kirby, and B. de Boer (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics* 43, 57–68.

Wärneryd, K. (1993). Cheap talk, coordination, and evolutionary stability. *Games and Economic Behavior* 5(4), 532–546.

Zipf, G. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.

## A Snippets

### A.1 On alternative explanations

We see two main explanations for the lack of upper-bounds in the lexical meaning of weak scalar expressions. The first is that their truth-conditional compatibility with stronger expressions endows them with a broader range of applicability by allowing them to occur in contexts in which their upper-bounded reading is absent. This can happen when embedded in downward-entailing contexts, when the speaker is likely uncertain about whether the upper bounded reading is true, or when the distinction between an upper-bounded reading and the simple, only lower-bounded reading, is not relevant. For instance, if for all the speaker knows Chris owns some of Johnny Cash's albums but she does not know whether he owns all, then the use of *some* lacking an upper-bound succinctly conveys her uncertainty. This may suggest a functionalist argument for why upper-bounded meanings do not conventionalize: Should contextual cues provide enough information to the hearer to identify whether a bound is intended to be conveyed pragmatically, then this is preferred over expressing it overtly through longer expressions, e.g., by saying *some but not all* explicitly. Importantly, although morphosyntactic disambiguation may be dispreferred due to its relative length and complexity (Piantadosi et al. 2012), it allows speakers to enforce an upper-bound and override contextual cues that might otherwise mislead the hearer. In a nutshell, this explanation posits that scalar implicatures fail to lexicalize because, all else being equal, speakers prefer to communicate as economically as possible and pragmatic reasoning enables them to do so. Compare this with a hypothetical language that lexicalizes two expressions for each weak scalar expression – one with and one lacking an upper-bound. We see four conditions along this functionalist explanation that may pressure languages for English-like semantics over this alternative. First, contextual cues are very reliable. Second, morphosyntactic disambiguation is seldom necessary. Third, morphosyntactic disambiguation is only marginally dispreferred. Fourth, larger lexica are costly. Overall, neither condition seems convincing as a central explanatory device for such a widespread phenomenon. The first two conditions put a heavy burden on the ability to retrieve contextual cues to a degree that seems unlikely to undercut the benefit of unambiguous communication. It is likely that human language users are very good at retrieving cues from context, but to stipulate that they are so good as to undercut the benefit of safe communication provided by this hypothetical alternative strikes us as too strong of an assumption. As for the third and fourth condition, these seem mostly like technical solutions without a proper empirical basis.

Instead, the systematicity and typological spread of scalar implicatures together with the observation that monomorphemic expressions that lexically rule out stronger alternatives are unattested across languages (Horn 1984:252-267, Horn 1972, Traugott 2004, van der Auwera 2010)

suggests that other forces may be at play. In what follows we investigate the predictions of our model under the assumption that the lack of lexicalization of scalar inferences may be accounted for by the relative representational simplicity of lexical meanings lacking an upper-bound over those that explicitly codify it. This difference is reflected in a learning bias towards more compressed lexical representations. That is, in a preference of learners for simpler over more complex explanations of the data they witness (Feldman 2000, Chater and Vitányi 2003, Piantadosi et al. 2012, Kirby et al. 2015, Piantadosi et al. view). While we do not want to argue that functional aspects as the ones discussed above do not play a role, we do see a clear benefit in exploring whether matters of transmission biases would not give us additional explanatory leverage.