Communicative pressures at the semantics-pragmatics interface

(– draft November 1, 2016—)

Abstract

According to standard linguistic theory, the conventional meaning of expressions is systematically modified by general pragmatic rules governing their use. This implies that models of the cultural evolution of meaning should likewise take into consideration that observable language use is a complex interaction of semantic representations and pragmatic rules. To this end, we present a game theoretic model of the cultural evolution of language where communicative pressures work on abstract semantic representations and pragmatic usage rules. Our model traces two evolutionary forces and their interaction: (i) fitness-based pressure towards communicative efficiency and (ii) systematic transmission perturbations when linguistic knowledge is transferred from one agent to another. The latter can contain general cognitive or learning biases, but can also contain other sources of systematic noise, e.g., resulting from errors in perception. We illustrate the model based on a case study showing that cognitive biases that favor simple semantic representations can prevent the lexicalization of pragmatic inferences. We also show that, technically speaking, it is possible that environmental factors, such as perceptual errors during acquisition, can produce evolutionary outcomes that look as if such cognitive biases are present even if they are not.

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”, a temporal succession of events can be conveyed by the order in which conjuncts appear as in “I traveled to Paris and got married”, or an invitation can be declined by saying “I have to work”. An influential explanation of the relation between the literal meaning of expressions (semantics) and what they are intended and interpreted to convey (pragmatics) due to Grice (1975) views the latter as a product of 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 the invitation – why would she have said so otherwise? Crucially, many classes of lexical meanings allow for pragmatic enrichments in a notably systematic fashion.

Productive pragmatic inferences such as scalar and manner implicatures have been at the focus of many studies on the semantics-pragmatics interface. However, an issue that has received little attention is the justification of semantic structure in light of the systematic enrichments offered by pragmatics. Similarly, multiple models of the cultural evolution of language have been put forward to analyze properties such as compositionality and combinatoriality, but the selection of particular semantics in light of pragmatics remains uncharted territory. The present investigation seeks to fill these gaps by analyzing the effects linguistic pressures have on the selection and pervasiveness of particular lexical meanings under consideration of their systematic pragmatic modification. Accordingly, our main goal is to uncover how transmission biases and functional pressure interact in the cultural evolution of meaning when pragmatics is taken into account. Crucially, we argue that transmission biases need not be cognitive in nature but can also ensue from perception and the environment.

1.1 Linguistic pressures at the semantics-pragmatics interface

The emergence and change of linguistic structure is influenced by many intertwined factors. These range from biological and socio-ecological to cultural ones (Steels 2011). 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. That is, as a result of a process of cultural evolution.

The idea that language is influenced by communicative pressures has played a pivotal role in synchronic and diachronic analyses at latest since Zipf's (1949) rationalization of the approximation of word frequency rankings by a power law distribution as 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). In recent years this line of research has led to a surge of approaches that seek to analyze the effects of such pressures from multiple angles ranging from simulations to experiments with human and robotic agents (see Steels 2015 and Tamariz and Kirby 2016 for recent overviews). Our starting point is given by the overarching argument that has crystalized from this literature: Natural languages need to be well-adapted to communicative needs within a linguistic community, but also need to be learnable to survive their faithful transmission across generations. More succinctly; natural languages are pressured for expressivity and learnability.

The opposition of expressivity and learnability becomes particularly clear when considering their consequences in the extreme (cf. Kemp and Regier 2012, Kirby et al. 2015). A language with a single form-meaning association is easy to learn but lacking in expressivity. Conversely, a language that associates a distinct form with all possible meanings one may want to convey is maximally expressive but challenging to acquire. The most prominent problem that arises from this tension is that of acquiring a language to express a potentially infinite set of meanings through finite means (Kirby 2002). However, this is not the only challenge learners confront. More central to our explanandum is the issue of selecting particular hypotheses out of a potentially infinite space of alternatives compatible with the data learners are exposed to. At the semantics-pragmatics interface this concerns the selection between functionally (near-)equivalent lexical meanings, noting in particular that what is systematically conveyed through pragmatics could alternatively be codified lexically. Notwithstanding, lexical meanings that allow for pragmatic modifications often seem to resist the lexicalization of their pragmatic component. The question is why.

We assume an integral part of the answer to lie in the effects of transmission perturbations on language when passed from one generation to the next. Such perturbations may take the form of learning biases, but may also stem from extraneous factors such as a noisy perception of the environment. In particular, we show how the presence of systematic noise can introduce regularizations that resemble those obtained under the assumption of a learning bias.

In the following, we model these components using the replicator-mutator dynamics, combining functional pressure on successful communication with effects of transmission perturbations on (iterated) Bayesian learning (Griffiths and Kalish 2007). The semantics-pragmatics distinction and its bearing on production and comprehension are captured by a probabilistic model of rational language use with different degrees of pragmatic sophisitication and languages (Frank and Goodman 2012, Franke and Jäger 2014, Bergen et al. 2016). The remainder of this section introduces these components individually together with the assumptions underlying them. These are: the representation of languages and their use (§1.2), pressures towards expressivity (§1.3) and learnability (§1.4) regulated by the replicator and mutator dynamics, respectively. After laying out the model, we analyze its predictions in a case-study on the lack of lexicalization of scalar implicatures in §2 and compare the outcome predicted by different transmission biases in §2.2 and §2.3.

1.2 Lexica and linguistic behavior

Lexica codify the truth-conditions of a language's expressions. Following Franke and Jäger (2014), a convenient way to represent lexica is by $(|S|, |M|)$-Boolean matrices, where S is a set of states of affairs (meanings) to convey and M a set of messages (forms available in the language).

We distinguish between two kinds of linguistic behavior. *Literal interlocutors* produce and interpret messages literally, being guided only by their lexica. In contrast, *pragmatic interlocutors* engage in mutual reasoning to inform their choices. Following models of rational language use such as Rational Speech Act models (Frank and Goodman 2012) and their game-theoretic predecessors, the Iterated Best/Quantal Response models (Franke 2009, Franke and Jäger 2014), these behaviors are captured by a reasoning hierarchy. The hierarchy's bottom, level 0, corresponds to literal language use. Pragmatic language users of level $n + 1$ behave rationally according to expected level n behavior of their interlocutors. Below, (1) and (3) specify the behavior of literal and pragmatic hearers of a language L. Their speaker counterparts are given in (2) and (4).

$$H_0(s|m; L) \propto pr(s)L_{sm} \tag{1}$$

$$S_0(m|s; L) \propto \exp(L_{sm}^{\alpha}) \tag{2}$$

$$H_{n+1}(s|m; L) \propto pr(s)S_n(m|s; L) \tag{3}$$

$$S_{n+1}(m|s; L) \propto \exp(\lambda\ H_n(s|m; L)^{\alpha}) \tag{4}$$

According to (1), a literal hearer's interpretation of a message m as a state s depends on her lexicon and her prior over states, $pr \in \Delta(S)$. In the following this prior is assumed to be uniform across hearers. The behavior of literal speakers, given in (2), is regulated by a parameter α which controls the sensitivity to which speakers prefer one signal over another based on its expected communicative success.

Pragmatic behavior is similar to its literal counterparts. Their difference lies in that level $n+1$ speakers/hearers reason about level n hearer/speaker behavior instead of solely relaying on their lexicon. That is, they reason about the way a rational level n interlocutor would use or interpret a message, and behave according to these expectations. Speaker behavior is further influenced by a soft-max parameter λ, $\lambda \geq 1$ (Luce 1959, Sutton and Barto 1998). As λ increases, choices made in production are more rational in that higher values lead to behavior that is increasingly in line with expected utility maximization.

We call the combination of a lexicon with its use, i.e., a level in the reasoning hierarchy, a *type*. These are the basic units on which our model's dynamics operate.

1.3 Replication & expressivity

Expressivity has received particular attention from investigations using evolutionary game theory (e.g. Nowak and Krakauer 1999, Nowak et al. 2000; 2002). Under this view, a type's ability to convey and interpret information successfully confers it a higher fitness, a measure that is relative to the success of other types in the population. In the simplest models, fitness directly translates into the proportion of types present in the population after a generational turnover. This association of communicative success within a population with changes in the proportion of types present in it creates a feedback loop that pressures the population towards greater expressivity.

The replicator equation gives us the means to make these dynamics precise. The proportion of types in a given population is codified in a vector x, where x_i is type i's proportion. As noted above, the fitness of type i is equal to its relative communicative success within this population, $f_i = \sum_j x_j \text{EU}(t_i, t_j)$. The expected communicative success of i and j is obtained by considering the average success of i conveying information to j and vice versa: $\text{EU}(t_i, t_j) = [U_S(t_i, t_j) + U_R(t_i, t_j)]/2$. $U_S(x, y)$ and $U_R(x, y)$ are, respectively, $\sum_s P(s) \sum_m S_n(m|s; L) \sum_{s'} R_o(s'|m; L)\delta(s, s')$ and $U_S(y, x)$, for n and o being the reasoning level of x and y, and $\delta(s, s') = 1$ iff $s = s'$ and 0 otherwise. This quantity is symmetric, reflecting the probability of two types' mutual understanding. The average fitness of the population is given by Φ, $\Phi = \sum_i x_i f_i$. This term serves as a normalizing constant for the discrete replicator equation: $\dot{x}_i = \frac{x_i f_i}{\Phi}$

Under its biological interpretation, this equation captures the idea of fitness-relative selection whereby fitter types produce more offspring, leading to their propagation in subsequent generations. As noted above, many aspects of language are subject to processees of transmission and change that can be likened to these biological dynamics. Amongst others, replication can be construed as modelling language acquisition, as e.g. in Nowak et al. 2002, but also as a process of horizontal adaptation in a single generation (see Benz et al. 2005:§3.3 for discussion).

In their series of papers on language evolution, Nowak and colleagues did not only consider expressivity but also recognized the central role of the fidelity by which languages are transmitted. Crucially, linguistic production can be prone to errors, states or messages may be percieved incorrectly, and multiple languages may be compatible with the data learners are exposed to. These sources of uncertainty introduce variation in their transmission from one generation to the next. In keeping the analogy to evolutionary processes this variation can be likened to mutation to the effect that a type's offspring may adopt a different type than that of its parent. Importantly, the resulting generational turnovers should depend on the relative learnability of a type instead of being a constant that is equal to all types as in Nowak et al. (2002). For this purpose, we turn to a different strand of research in cultural evolution: *iterated learning.*

1.4 Mutation & learning

Iterated learning is a process in which the behavior of one individual serves as learning input for another. This learner, upon acquisition of this behavior, then goes on to produce behavior that serves as input for a new learner. This process can be thought of as a progression through chains of parents and children; the parent produces linguistic data from which the child infers a language. The latter, now a parent, goes on to produce linguistic data for a new generation of learners. Following Griffiths and Kalish (2007) we model iterated learning as a repeated process of Bayesian inference in which learners combine the likelihood of a type producing the received learning input with prior inductive biases.

Due to the pressure towards learnability it exherts, iterated learning generally leads to simpler and more regular languages (see Kirby et al. 2014 and Tamariz and Kirby 2016 for recent surveys). Importantly, experimental and mathematical results suggest that the outcome of this process reflects learners' a priori biases. In a Bayesian setting these biases can be codified in a prior $P \in \Delta(T)$ and a way to think about it is as the amount of data a learner would require in order to adopt a language (cf. Griffiths and Kalish 2007:450). Or, in our case, a combination of a lexicon and a signaling behavior. The extent of the prior's influence has been shown to heavily depend on the learning strategy assumed to underly the inference process. On the one hand, 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 converged to the prior, i.e., 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 parametrize the posterior with a parameter $l \geq 1$. In doing so, we obtain a range of learning strategies between posterior sampling and MAP. When $l = 1$ learners sample from the posterior. As l increases towards infinity their propensity to maximize the posterior grows as well.

The data learners are exposed to is described by a set D containing sequences of state-message pairs, e.g., $\langle \langle s_i, m_v \rangle, ... \langle s_j, m_w \rangle \rangle$. These are sequences of language use witnessed by learners, the length of which we denote by k.

We integrated iterated learning in the replicator dynamics by codifying the former in a transition matrix Q, where Q_{ij} indicates the probability that a child of a parent of type i adopts type j. This quantity is proportional to the probability of i producing the learning data and that of inferring j given the data:

$$Q_{ij} \propto \sum_{d \in D} P(d|t_i) F(t_j, d), \text{ where } F(t_j, d) \propto P(t_j|d)^l \text{ and } P(t_j|d) \propto P(t_j)P(d|t_j).$$

1.5 Summary

Drawing from past research, we argued that expressivity and learnability are central to the cultural evolution of language. We propose these components to be modelled respectively as communicative efficiency-relative replication and (iterated) Bayesian learning. Taken together their interaction is described by the replicator-mutator dynamics (Hofbauer and Sigmund 2003):

$$\hat{x}_i = \sum_j Q_{ji} \frac{x_j f_j}{\Phi}$$

The units that the dynamics operate on are a combination of a lexicon and a degree of pragmatic sophistication determining its use. We call this combination a type. 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. With this model at hand we turn to the analysis of the lack of lexicalization of productive pragmatic inferences in a case study on scalar implicatures.

2 Scalar implicatures

We assume an integral part of the answer to be that learners are a priori biased towards simpler, more compressed, lexical representations. That is, rational learners should prefer simpler over more complex explanations of the data they witness (Feldman 2000, Chater and Vitányi 2003, Piantadosi et al. 2012a, Kirby et al. 2015, Piantadosi et al. under review). More narrowly, codifying more semantically is a priori dispreferred and in cases where pragmatics offers conventional means of enrichment, the communicative disadvantage that speakers would otherwise incurr by the use of simpler semantics is leveled. In this way, these domains play a synergic role in overcoming linguistic pressures which may prevent the lexicalization of pragmatic inferences. Crucially, we also show how a different kind of transmission bias in the form of a noisy perception of the environment can lead to the same outcome as that predicted by a learning bias.

The mere presence of systematic noise in the transmission of strategies can introduce regularization that looks as if the agents have a learning bias. But, most importantly, this disturbance of transmission fidelity is due to perceptual noise and properties of the environment, not learning biases. In other words, learning biases are clearly not the only transmission biases that can shape evolution alongside functional pressure. Environmental and perceptual noise can play a role too.

As touched upon in §1 we assume learners to be biased towards simpler semantics. That is, learners have a preference for types that can explain the data by codifying less in their lexical representations. More generally, a drive for simplicity has been argued to underpin speaker preferences for brevity and ease of articulation, as well as to pressure languages towards lexical ambiguity and grammatical compression (Zipf 1949, Grice 1975, Piantadosi et al. 2012b, Kirby et al. 2015). As shown by Chater and Vitányi (2003) simplicity not only provides a solution to induction problems by providing a distinction between equally explanatory hypotheses, but has also proven its empirical worth as a longstanding predictive principle in cognitive modelling.

Scalar implicatures are a particularly well-studied type of conventional pragmatic inference. They are licensed for groups of expressions ordered in terms of informativity, here understood as an entailment induced order. For instance, *some* is entailed by *all*. If it were true that 'All students came to class', it would also be true that 'Some students came to class'. However, as already noted in §1, 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 since the speaker did not choose them. In this way, 'Some students came to class' is strengthened to convey that some but not all students came to class. Conversely, speakers can rely on their interlocutors to draw this inference. The bound that rules out stronger alternatives is thusly not codified in the lexical meaning of weak alternatives but instead pragmatically supplied.

This kind of strengthening corresponds to our previous description of the pragmatic use of lexicon L_a. A pragmatic hearer who reasons about a speaker's use of message m_1 will associate it more strongly with s_1 than with s_2 since the latter is already unambiguously associated with s_2. Conversely, a pragmatic speaker will reason about her interlocutor's expected interpretation and use the messages at her disposition accordingly.

$$L_a = \begin{array}{cc} & \begin{array}{cc} m_1 & m_2 \end{array} \\ \begin{array}{c} s_1 \\ s_2 \end{array} & \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \end{array} \qquad L_b = \begin{array}{cc} & \begin{array}{cc} m_1 & m_2 \end{array} \\ \begin{array}{c} s_1 \\ s_2 \end{array} & \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}$$

Our initial question can now be narrowed to the case of scalar implicatures by asking for a justification for the lack of lexical upper-bounds in weak scalar alternatives. That is, why semantics such as those of message m_1 in L_a are regularly selected for over the alternative of codifying the bound semantically as in L_b. More poignantly, would it not serve language users better if weak(er) expressions such as *warm*, *or*, *some* and *big* were truth-conditionally incompatible with stronger alternatives such as *hot*, *and*, *all* and *huge*? This question is particularly striking considering the number of expressions that license such inferences across languages.

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 'Some students came' but she doesn't know whether all came, then the use of not upper-bounded *some* 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. For example, by saying *some but not all* explicitly. Importantly, although morphosyntactic disambiguation is dispreferred due to its relative length and complexity (Piantadosi et al. 2012b), 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 is convincing as a pivotal explanatory device for such a wide-spread 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, considering the regularity and 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 hypothesis that the lack of lexicalization of scalar inferences can be accounted for by the relative representational simplicity of lexical meanings lacking an upper-bound over those that explicitly codify it. 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 learnability would not give us additional explanatory leverage.

Note however that we do not represent the contrast between lexical representations explicitly. Instead, the learning bias towards a lack of upper-bounds in weak scalar alternatives is directly encoded in the learners' prior over types. In principle this difference could be made precise with an adequate representational language, e.g., through measures over representational complexity such as minimal description length. There is a growing effort to develop such empirically testable representational languages. For instance, the so-called *language of thought* has been put to test in various rational probabilistic models that show encouraging results (see e.g. Katz et al. 2008, Piantadosi et al. under review; 2012a and references therein). We think that our assumption is well-warranted as a working hypothesis and decide against such an enrichment at present in order to focus on the effects of linguistic pressures predicted by our model instead.

$$L_1 = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \quad L_2 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \quad L_3 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

$$L_4 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad L_5 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \quad L_6 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

Table 1: Space of possible lexica.

2.1 Analysis

We analyze the model's predictions in populations of types with two signaling behaviors; literal or pragmatic. The former correspond to level 0 reasoners who only take their lexica into consideration and the latter to level 1 reasoners. Higher level reasoning is not required to derive scalar implicatures from the lexica we consider here, nor do they leave room for substantial pragmatic refinement.

The space of lexica we consider is given in Table 1. As in the preceding examples, lexica are $(2,2)$-Boolean matrices. These are the simplest matrices that allow us to make the contrast between the presence or abscence of an upper-bound and the use of scalar implicatures precise. As implicit in the discussion so far, one may think of s_1 as a "some but not all"-state and of s_2 as an "all"-state. The literal meaning of weak scalar expressions such as English *some* then corresponds to a message true of both s_1 and s_2 in these fragments. Following our assumption for a preference for simple lexical representations, the prior biases learners against lexica in which a message holds true only of state s_1. That is, those messages that lexicalize an upper-bound that rules out the "all"-state s_2. All other semantics are assumed to be a priori equally probable. Accordingly, this prior is captured by $P(t_i) \propto n - c \cdot r$, where n is the total number of states and r is the number of messages only true of s_1 in t_i's lexicon, $c \in [0,1]$. Increments in the prior parameter c therefore bring about a stronger preference against languages that lexicalize upper-bounds, i.e., L_2, L_4 and L_6.

While there is a total of 16 possible $(2,2)$-lexica, a number of them are identical both in terms of expressivity and the learning bias. The competition between such types is determined by the initial configuration of a population, i.e., the proportion with which each type starts out. However, this fact can be obscured when averaging across simulations. We therefore focus on this smaller representative subset.[1] Lexica L_1 to L_3 are not optimal for communication because they assign the same state to all their messages. This failure to associate a single form to a meaning both semantically and pragmatically inevitably leads to their disadvantage in language use. L_4 and L_5 are our target lexica, previously respectively labeled as lexica L_b and L_a. They codify upper-bounded semantics for message m_1 and a lack thereof. Lastly, L_6 is similar to L_5 in that two messages are true of the same state but differs from it in assigning upper-bounded semantics to m_1.

Combining a signaling behavior with each of these 6 lexica yields a total of 12 distinct types. Note in particular that a type that has conventionalized upper bounds to realize a (quasi-)partition of the relevant semantic space, such as L_4, will produce speaker behavior that is *almost* indistinguishable from that of a language that lacks upper bounds, but with pragmatic speakers, such as L_5. Almost, because there may be slight differences between the probability with which speakers would (erroneously) use a semantically false description and the probability with which speakers would (erroneously) use a pragmatically suboptimal description. Due to this possibly marginal difference between pragmatic L_4 and L_5, the selection of one type over the other is

[1]Simulations conducted with the full space of possible lexica confirm that the results reported here do not hinge on their exclusion.

expected to mainly depend on the learning bias. Things are less clear for literal L_5 contrasted with literal/pragmatic L_4. The former has a learning advantage but is expected to fare worse in communicative terms in virtue of ambiguous m_1.

The dynamics are initialized with an arbitrary distribution over types, constituting the population's first generation. The results for each parameter setting were obtained from 1000 independent runs, each consisting of 20 generations. This corresponds to a developmental plateau after which no noteworthy change was registered. As specified in §1.4, the learning matrix Q can be obtained by considering all possible state-message sequences of length k. Given that this is intractable for large k, matrices were approximated by sampling 10 sequences from each type's production probabilities and a type's children being exposed only to this subset.

2.2 Learning bias

According to our hypothesis, functional pressure on successful communication combined with learning pressures in the form of a bias against upper-bounds may lead to the selection of L_5-like semantics. It is instructive to first inspect the effect of these pressures in isolation. For this purpose we focus our attention on three pragmatic types.[2] Pragmatic L_3, a type that is lacking in expressivity but is a priori preferred for its lack of upper-bounds. Pragmatic L_4, a type that is functionally advantageous but biased against. And pragmatic L_5, combining virtues of the latter two.

Expressivity only. Recall that the replicator dynamics are sensitive to λ and α as both have a bearing on a type's fitness. In particular, low α disadvantages types that rely on pragmatic reasoning to the gain of those that codify more semantically. The rationality parameter λ has a similar effect for different reasons: Low rationality leads to a less pronounced preference for the choice(s) expected to succeed best in communication. That is, λ regulates the strength by which users of L_5 associate non-upper-bounded m_1 exclusively with the "some"-state s_1 over the "all"-state s_2. Speakers of L_4 need not rely on λ for this as the association of m_1 with s_1 is already part of their language's semantics.

The influence of the rationality parameter for $\alpha = 1$ is depicted in Figure 1.A. As expected, the less expressive L_3 speakers fare the worse and are influenced the least by variations in λ. In contrast, low values of λ result in a higher proportion of L_4 speakers relative to L_5. This is expected given role of rationality in producing more deterministic behavior in users of L_5-like languages. As the rationality parameter increases, the functional difference between L_4 and L_5 is leveled. Overall, the final populations that result only from a pressure towards expressively come close to an even share of pragmatic L_4, L_5 and L_6 types (the latter follows the same trajectory as L_5 in Figure 1.A). That is, expressivity alone can not differentiate between these lexica in populations of rational pragmatic language users.

Learnability only. The effect of iterated learning without a pressure for expressivity is shown in Figure 1.B for posterior sampling. In line with our expectations, the share of L_4 speakers decreases as the bias against upper-bounds increases. In turn, this benefits L_3 and, in particular, L_4. However, note that even a strong bias against lexical upper-bounds leads only to a moderate advantage for L_5 over L_4. Furthermore, a pressure only towards learnability promotes functionally defective languages L_3.

Inspecting these pressures separately not only gives some intuitions about the parameters' influence, but also highlights some of their broader implications. First and foremost, neither

[2] Pragmatic reasoning allows language users to refine their (possibly erroneous) choices. Therefore, it is advantageous even for those types that codify an upper-bound lexically.

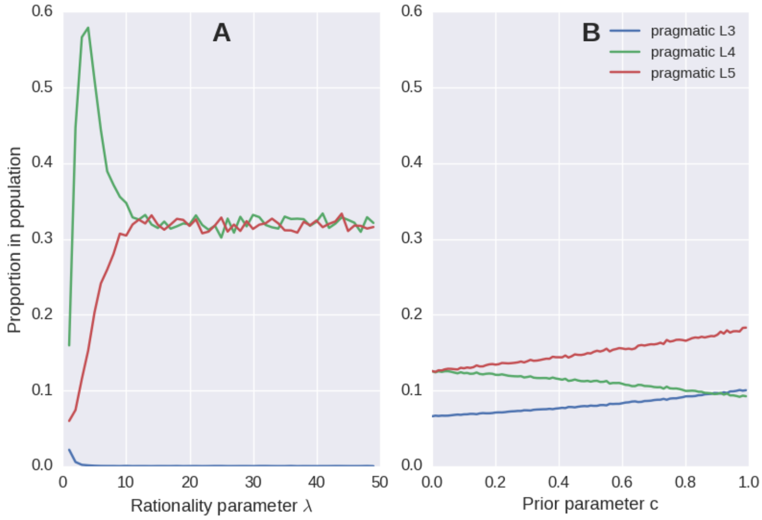

Figure 1: Mean proportions of target types after 20 generations in 1000 populations with only a pressure for expressivity in A ($\alpha = 1$) and only for learnability in B ($\alpha = 1, \lambda = 30, k = 5, l = 1$).

dynamic comes close to converging to a monomorphic population under most parameter configurations. For instance, while L_4 speakers can come to take over a substantial proportion of the population, this only happens in a restricted range of low degrees of rationality. Apart from polymorphy, both pressures make undesirable predictions. A pressure only towards expressivity leads to the to the ejection of L_1, L_2 and L_3 from the population. However, it can not explain the regular selection of L_5-like semantics over either of its functionally similar alternatives L_4 and L_6. In contrast, a pressure only towards learnability has a modest but clear effect in differentiating L_5 from these alternatives but fails to rule out functionally suboptimal types such as tautological L_3.

Expressivity and learnability. Figure 2 illustrates the effect of the learning bias for posterior sampling (2.A) and slightly more MAP-like learning (2.B). More detailed results for all types across a sample of c-values for $l = 1$ and $l = 3$ are presented in Table 2. Overall, these results suggest that in the present setup a weak bias is sufficient to lead to a selection of L_5 over L_4. As in the simulations that only considered learnability, this effect increases with the bias' strength provided L_5 users are pragmatic. Importantly, the addition of a pressure towards expressivity magnifies this effect and dampens the proliferation of functionally suboptimal types advantaged by the learning bias. As stressed above, this suggests that neither the learning bias nor functional pressure alone but their combination may lead to the lack of upper-bounds in the lexical meaning of scalar expressions.

The resulting proportion of pragmatic L_5 speakers primarily hinges on four parameters. First, the degree to which linguistic behavior is deterministic, controlled by λ and α, as it plays a role both for expressivity as well as in producing data that allows learners to discriminate this type from others. Second, the learning bias c which controls the learners preference for simpler lexical representations – leading to the selection of L_5 over L_4. Lastly, the posterior parameter l, which magnifies the effects of the learning bias in tandem with replication.

As discussed in relation to Figure 2.A, posterior sampling can lead to the incumbency of pragmatic L_5. However, not even a strong favorable learning bias combined with a pressure for

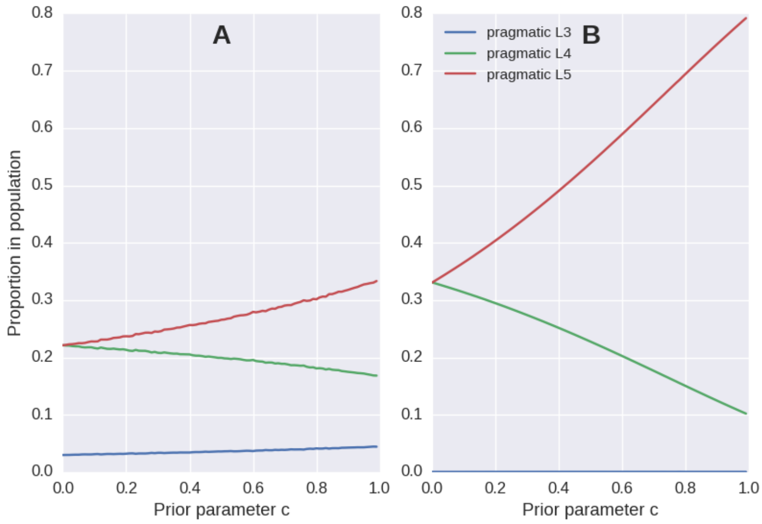

Figure 2: Mean proportions of target types after 20 generations in 1000 populations across bias values $c \in [0, 1]$ with $l = 1$ in A and $l = 3$ in B ($\alpha = 1, \lambda = 20, k = 5$).

	$l = 1$					$l = 3$				
c	0	.1	.5	.8	.9	0	.1	.5	.8	.9
lit. L_1	.03	.03	.04	.04	.04	ϵ	ϵ	ϵ	ϵ	ϵ
lit. L_2	.03	.03	.02	.01	.04	ϵ	ϵ	ϵ	ϵ	ϵ
lit. L_3	.03	.03	.04	.04	.04	ϵ	ϵ	ϵ	ϵ	ϵ
lit. L_4	.07	.07	.06	.06	.05	ϵ	ϵ	ϵ	ϵ	ϵ
lit. L_5	.04	.05	.05	.06	.06	ϵ	ϵ	ϵ	ϵ	ϵ
lit. L_6	.04	.04	.04	.04	.03	ϵ	ϵ	ϵ	ϵ	ϵ
prg. L_1	.03	.03	.04	.04	.04	ϵ	ϵ	ϵ	ϵ	ϵ
prg. L_2	.03	.03	.02	.01	.04	ϵ	ϵ	ϵ	ϵ	ϵ
prg. L_3	.03	.03	.04	.04	.04	ϵ	ϵ	ϵ	ϵ	ϵ
prg. L_4	.22	.22	.2	.18	.17	.33	.31	.23	.15	.12
prg. L_5	.22	.23	.27	.3	.32	.33	.37	.54	.7	.75
prg. L_6	.22	.22	.2	.18	.17	.33	.31	.23	.15	.12

Table 2: Mean proportions of types in 1000 populations after 20 generations across bias values $c \in [0, 1]$ with $l = 1$ and $l = 3$ ($\alpha = 1, \lambda = 30, k = 5$), $\epsilon < 0.005$.

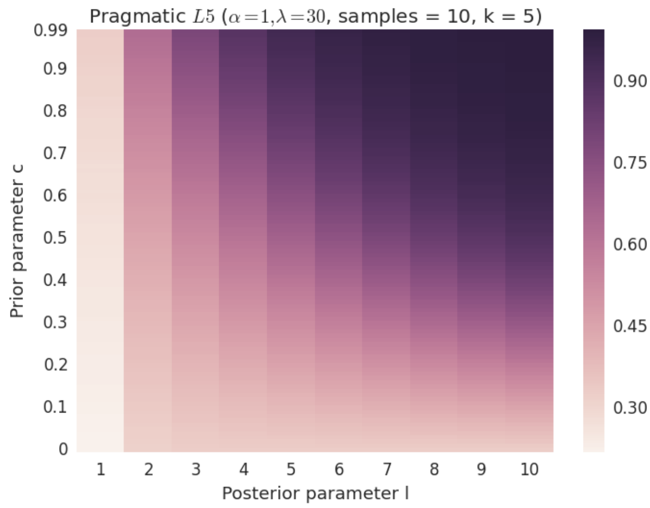

Figure 3: Mean proportion of pragmatic L_5 in 1000 populations after 20 generations ($\alpha = 1, \lambda = 30, k = 5$)

expressivity completely drives out competing types. This is not so for more posterior maximizing behavior. Crucially, as shown in Figure 3 the range of bias values within which L_5 takes over the population increases with MAP-like learning. In other words, the strength of the learning bias required for a given final proportion of L_5 speakers strongly depends on learners' inference mechanism. As for the effect of the other parameters not mentioned so far, changes in sequence length influence the population in a predictable way: smaller values lead to more heterogeneous populations that reflect the learner's prior more faithfully whereas larger ones lead to more pronounced differences amongst equally preferred types. This is expected insofar as the likelihood that a sequence of length 1 was produced by any type is relatively uniform (modulo prior) whereas the likelihood of types with lexica L_1 - L_3 to produce, for instance, a sequence of 10 observations consistently with the same state-message combination is less likely than for pragmatic types using L_4 - L_6, or literal L_4. Thus, while noteworthy, sequence length has no direct bearing on the contrast of interest.

2.3 Noise

We have two states: $s_\exists$ (some but not all) and $s_\forall$ (all). Let's assume that the probabilities for observing the state in the column when the actual state is the state in the row are given by this table:

	$s_\exists$	$s_\forall$
$s_\exists$	$1-\epsilon$	ϵ
$s_\forall$	δ	$1-\delta$

So, ϵ is the error probability when the true state is $s_\exists$ and δ is the error probability when the true state is $s_\forall$. How does this affect our model of Bayesian learning?

Let's denote the probability that the teacher (learner) observes state s_t (s_l) when the actual state is s_a as $P_N(s_t \mid s_a)$ ($P_N(s_l \mid s_a)$). This is given by the table above. We can then derive the following probabilities. Firstly, the probability that s_a is the actual state when the learner observes s_l is just Bayes rule, combining prior of s_a with the noise from above:

$$P_N(s_a \mid s_l) \propto P(s_a)\ P_N(s_l \mid s_a)\,.$$

Secondly, the probability that the teacher observes s_t when the learner observes s_l is:

$$P_N(s_t \mid s_l) = \sum_{s_a} P(s_a \mid s_l)\ P_N(s_t \mid s_a)\,.$$

Finally, this gives us the probability that a teacher of type t produces a datum that is perceived by the listener as $d = \langle s_l, m\rangle$:

$$P_N(\langle s_l, m\rangle \mid t) = \sum_{s_t} P_N(s_t \mid s_l)\ P(m \mid s_t, t)\,.$$

Generalize this to a sequence of perceived data d_l and write $P_N(d_l \mid t)$. We then define the noise-perturbed mutation matrix as:

$$Q_{ij} \propto \sum_{d_l \in D} P(d_l \mid t_i) F(t_j, d_l)\,, \quad \text{where } F(t_j, d) \text{ as before.}$$

In words, it may be the case that learner and/or teacher do not perceive the actual state as what it is. They are not aware of this, and produce/learn as if what they observed was the actual state. In particular, the learner does not reason about noise when he tries to infer the speaker's type. He takes what he observes a state to be as the actual state that the teacher has seen as well and computes which type would have most likely generated the message to this state. This can lead to biases of inferring the "wrong" teacher type, if the noise makes some types err in a way that resembles the noiseless behavior of other types. I.e., this could, in principle, induce transmission biases that look as is there was a cognitive bias in favor of a particular type, simply because that type better explains the noise.

On top of changing the mutation matrix in this way, we also need to adapt the calculation of expected utilities, taking into consideration that states are perceived noisily. So, where before we had:

$$U_S(t_i, t_j) = \sum_s P(s) \sum_m S_n(m|s; L) \sum_{s'} R_o(s'|m; L)\delta(s, s')\,,$$

we now have:

$$U_S(t_i, t_j) = \sum_{s_a} P(s_a) \sum_{s_t} P_N(s_t \mid s_a) \sum_m S_n(m|s_t; L) \sum_{s'} R_o(s'|m; L)\delta(s, s')\,.$$

Results

In a nutshel: this works for some parameter settings but not for others. I have not fully explored the whole space. I would be content with a paper that says: there are parameter values that produce behavior as if there are cognitive biases without cognitive biases. I believe that this would be an additional and very interesting contribution, even if there are noise structures that have other effects.

Importantly, we don't need costs ($c = 0$)! Unfortunately, we do need to tinker with α (for defensible, techical reasons) and we need to assume that $\epsilon > \delta$, otherwise either t_{10} or even t_{12} may dominate.

That means that if we wanted to also argue that this is a plausible explanation for lower-bounds semantics, we'd need to argue why it is natural that $\epsilon > \delta$. Maybe such a story can be given, but I presently do not see why/how any such story is obviously better than any other story I could come up with for any other relation of ϵ and δ.

One reasoning for $\epsilon > \delta$ could be this. When the true state is s_a, we would actually need negative evidence that some item/person/whatever does not have a property. Negative evidence of this kind might be relatively infrequent, especially if untrue. On the other hand, perceiving s_a when the true state is s_e could be more frequent, because it would, for example, require wrong information about the domain size. It could also be a cause of overemphasizing: speakers tend to exaggerate and want to claim *all*, when in fact that is not true (e.g. **?**:for a model that has the speakers' tendency to overemphasize as a motor of language change).

2.4 Discussion

Broadly speaking these results suggest that a lack of semantic upper-bounds coupled with pragmatic reasoning can overcome communicative pressures and stabilize in a population provided there is a bias for simpler representations. This prediction hinges on three assumptions. First, that language is pressured toward both expressivity and learnability. Second, that language use is not too unpredictable – low α or λ render languages that rely on pragmatics too prone to communicative failure and more difficult to learn. Third, that learners prefer simpler over more complex lexical representations. An important addendum to this third condition being that a combination of rationality in choice and maximization in learning requires a weaker bias towards simplicity. Under these conditions the selection of lexical meanings lacking upper-bounds in populations of pragmatic speakers is robust against parameter perturbations. This outcome is particularly encouraging in light of other advantages a lack of semantic upper-bounds may confer.

While a major share of non upper-bounded lexical meanings is predicted by the literature, it is less clear to what extent other types should be present in the final population, if at all. It seems reasonable to expect functionally suboptimal types L_1, L_2 and L_3 to be ruled out because they fail to enable their users to communicate effectively. However, this is not true of L_4.[3] Notwithstanding, the prediction that natural language communities are homogeneous or that a single speaker may entertain L_4-like semantics for one scalar expression and L_5-like semantics for another is is not implausible (references on individual level differences in the derivation of scalar implicatures). Alternatively, a stronger tendency for posterior maximization has to be assumed (cf. Figure 3). This empirical issue relates to other two aspects left undiscussed in our present analysis: 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. We conject that frequently drawn scalar implicatures might fossilize to avoid cost, while infrequent ones could still be computed on-line. This opens a possible venue to address our previous question about the expected presence of L_4-like semantics.

[3] L_6 presents a special case. In our current setup, it mirrors L_5 in enabling for the pragmatic narrowing of message m_2 rather than m_1. However, this association of s_2 with m_2 under favorable parameter conditions is achieved by ruling out the "some but not all"-state s_1 and not, as with scalar implicatures, the "all"-state s_2. L_6 speakers therefore strengthen a "some"-message to convey something paraphrasable as 'some but not [some but not all]'. Our representation of lexica as Boolean matrices is blind to this anomaly.

3 General discussion

We laid out a model that combines game-theoretical models of functional pressure towards efficient communication (Nowak and Krakauer 1999), effects of learning biases 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 different lexica (Bergen et al. 2012; 2016). This model generates predictions about lexicalization patterns and, more generally, effects of communicative pressures on language. We argued that the puzzle raised by semantics in light of pragmatics is hard to explain on purely functional grounds and that part of the answer may instead lie in learnability: Simpler semantic representations are more likely to be learned while pragmatic reasoning can counteract functional disadvantages otherwise incurred. This result is of particular relevance for the longstanding assumption of a divide and interaction between semantics and pragmatics by offering an account of why (certain) pragmatic inferences fail to lexicalize.

The main innovations of the model are its modular separation of expressivity and learnability, allowing for their isolated and combined analysis, as well as the learning process involving a joint inference over types of pragmatic behavior and lexical meaning. 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, however, Kirby et al. model expressivity as exerting its force only in the production of learning data. This model's expressivity parameter thereby fulfills a similar role to high values of λ and α in making speaker behavior more deterministic. In this way, it "favors" unambiguous languages. Crucially, the degree of mutual understanding of interlocutors central to replication and to our notion of expressivity is not taken into consideration. That is, while our proposal combines bidirectional horizontal transmission with its vertical and unidirectional counterpart, Kirby et al.'s model only considers 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 the tautological language L_3 in our analysis. This outcome has been reported in a number laboratory experiments where the participants' task was to learn and subsequently reproduce the language produced by a previous participant, leading to a proliferation of languages that associated a large number of 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 distinguish meanings 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 predictions given their variation in meaning spaces, modes of transmission, iterations and feedback given to participants. However, we take these results to suggest that there is an important difference between a language generating learnable linguistic data and its actual performance as a means of information transfer. The former solely depends on the mechanism by which speakers associate form and meaning. The latter additionally hinges on the hearers linguistic experience and her capability to interpret linguistic input based on this experience. In sum, we contend that successful information transfer in a linguistic community is central the adoption of a communication system and that this measure is not faithfully reflected by taking only production into consideration.

One of our central assumptions is the learning bias for simpler lexical representations. Albeit implicit in our formalization, the conceptual foundation for a compression preference at the lexical level is grounded in work in the language of thought tradition put forward by Piantadosi et al. (2012a; under review). Following Piantadosi and colleagues we assumed that lexical meanings are expressed in a representational language and that the learners' task is to infer lexica that best explain the data witnessed. Furthermore, we assumed learners to not only infer lexical meaning but types of linguistic use. Whether language use is culturally transmitted is an empirical matter that can not be settled here. However, this idealized model suggests that semantic structure evinced cross-linguistically that lends itself for pragmatic exploitation reflects lexical information that is both easy to learn and, perhaps more interestingly, also easy to learn how to exploit via mutual reasoning.

4 Conclusion

Language change is affected by intertwined pressures. We investigated the selection of lexical meanings lacking upper-bounds in a case study on role of semantics and pragmatics in satisfying such pressures. We showed that, when pressured for learnability and expressivity, the former drives for simpler semantic representations inasmuch as pragmatics can compensate for their lack of expressivity in use. That is, the relative learnability advantage of simpler semantics may offer an answer to why natural languages fail to lexicalize certain pragmatic inferences. More broadly, we argued that semantic patterns can be explained by taking into consideration the way in which they are used in interaction as well as their representational complexity and learnability.

References

Anton Benz, Gerhard Jäger, Robert Van Rooij, and Robert Van Rooij, editors. *An Introduction to Game Theory for Linguists*. Springer, 2005.

Leon Bergen, Noah D Goodman, and Roger Levy. 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*, 2012.

Leon Bergen, Roger Levy, and Noah D Goodman. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 2016.

Henry Brighton. Compositional syntax from cultural transmission. *Artificial Life*, 8(1):25–54, 2002. doi: 10.1162/106454602753694756.

Nick Chater and Paul Vitányi. Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1):19–22, 2003. doi: 10.1016/s1364-6613(02)00005-0.

Nicolas Fay and T. Mark Ellison. The cultural evolution of human communication systems in different sized populations: Usability trumps learnability. *PLoS ONE*, 8(8), 2013. doi: 10.1371/journal.pone.0071781.

Jacob Feldman. Minimization of boolean complexity in human concept learning. *Nature*, 407 (6804):630–633, 2000.

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

Michael Franke. *Signal to Act: Game Theoretic Pragmatics*. PhD thesis, University of Amsterdam, 2009.

Michael Franke and Gerhard Jäger. Pragmatic back-and-forth reasoning. *Semantics, Pragmatics and the Case of Scalar Implicatures.*, pages 170–200, 2014.

Gerald Gazdar. *Pragmatics, Implicature, Presuposition and Logical Form.* Academic Press, New York, 1979.

Paul Grice. Logic and conversation. In *Studies in the Ways of Words*, chapter 2, pages 22–40. Harvard University Press, Cambridge, MA, 1975.

Thomas L. Griffiths and Michael L. Kalish. Language evolution by iterated learning with bayesian agents. *Cognitive Science*, 31(3):441–480, 2007.

Josef Hofbauer and Karl Sigmund. Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, 40(04):479–520, 2003.

Laurence R. Horn. *On the Semantic Properties of Logical Operators in English.* Indiana University Linguistics Club, Bloomington, IN, 1972.

Laurence R. Horn. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin, editor, *Meaning, Form and Use in Context*, pages 11 – 42. Georgetown University Press, Washington, 1984.

Yi Ting Huang and Jesse Snedeker. Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology*, 58(3):376–415, 2009.

Gerhard Jäger. Evolutionary game theory and typology: A case study. *Language*, 83(1):74–109, 2007. doi: 10.2307/4490338.

Gerhard Jäger and Robert van Rooij. Language structure: psychological and social constraints. *Synthese*, 159(1):99–130, 2007. doi: 10.1007/s11229-006-9073-5.

Yarden Katz, Noah D Goodman, Kristian Kersting, Charles Kemp, and Joshua B Tenenbaum. Modeling semantic cognition as logical dimensionality reduction. In *Proceedings of Thirtieth Annual Meeting of the Cognitive Science Society*, 2008.

C. Kemp and T. Regier. Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054, 2012. doi: 10.1126/science.1218811.

S. Kirby, M. Dowman, and T. L. Griffiths. innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12):5241–5245, 2007. doi: 10.1073/pnas.0608222104.

S. Kirby, H. Cornish, and K. Smith. 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, 2008. doi: 10.1073/pnas.0707835105.

Simon Kirby. Learning, bottlenecks and the evolution of recursive syntax. In Ted Briscoe, editor, *Linguistic Evolution Through Language Acquisition*, pages 173–204. Cambridge University Press (CUP), 2002. doi: 10.1017/cbo9780511486524.006.

Simon Kirby, Tom Griffiths, and Kenny Smith. Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28:108–114, 2014. doi: 10.1016/j.conb.2014.07.014.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102, 2015.

Duncan R. Luce. *Individual choice behavior: a theoretical analysis.* Wiley, 1959.

André Martinet. *Functionalist View of Language.* Clarendon Press, Oxford, 1962.

Wim De Neys and Walter Schaeken. When people are more logical under cognitive load. *Experimental Psychology*, 54(2):128–133, 2007.

M. A. Nowak and D. C. Krakauer. The evolution of language. *Proceedings of the National Academy of Sciences*, 96(14):8028–8033, 1999.

Martin A. Nowak, Joshua B. Plotkin, and Vincent A. A. Jansen. The evolution of syntactic communication. *Nature*, 404(6777):495–498, 2000. doi: 10.1038/35006635.

Martin A. Nowak, Natalia L. Komarova, and Partha Niyogi. Computational and evolutionary aspects of language. *Nature*, 417(6889):611–617, 2002. doi: 10.1038/nature00771.

Steven T Piantadosi. Zipfs word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, 2014. doi: 10.3758/s13423-014-0585-6.

Steven T. Piantadosi, Joshua B. Tenenbaum, and Noah D. Goodman. Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, 123(2):199–217, 2012a.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291, 2012b. doi: 10.1016/j.cognition.2011.10.004.

Steven T. Piantadosi, Joshua B. Tenenbaum, and Noah D. Goodman. Modeling the acquisition of quantifier semantics: a case study in function word learnability, under review.

Catriona Silvey, Simon Kirby, and Kenny Smith. Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, 39(1):212–226, 2014. doi: 10.1111/cogs.12150.

Luc Steels. Modeling the cultural evolution of language. *Physics of Life Reviews*, 8(4):339–356, 2011.

Luc Steels. *The Talking Heads experiment: Origins of words and meanings.* Language Science Press, 2015.

Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning.* MIT Press, Cambridge, MA, USA, 1998.

Monica Tamariz and Simon Kirby. The cultural evolution of language. *Current Opinion in Psychology*, 8:37–43, 2016.

Elizabeth Closs Traugott. Historical pragmatics. In Laurence R. Horn and Gregory Wand, editors, *The Handbook of Pragmatics*, pages 538–561. Blackwell Publishing, 2004.

Johan van der Auwera. On the diachrony of negation. In *The Expression of Negation*, pages 73–110. Walter de Gruyter GmbH, 2010. doi: 10.1515/9783110219302.73.

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

5 Some thoughts + what may be missing

- I still need to review, read and find some references I'm withholding in the current draft (marked as "references" in color).
- The discussion (sub) sections should be expanded and polished. In particular the discussion concerning lexical representations and the language of thought. I'm aware that this part is not as readable as the rest;
- The analysis may benefit from further data and less of a qualitative overview of the parameters' effect (e.g. the regression analysis Michael conducted);
- As you will have noticed when reading footnote 3, I think the performance of L_6 being on par with L_4 is somewhat problematic;
- There's a fair share of repetition, this is partly due to the fact that scalar implicatures and some matrices are discussed in non-contiguous parts of the paper. Whether the repetition is necessary depends on whether we roughly keep the current structure and length or change things;
- In principle, we could make a direct comparison between the case-study of Griffiths and Kalish (2007) on compositionality using iterated learning and iterated learning + functional pressure. This could be relegated to the appendix. However, I'm inclined to think this is somewhat tangential and wouldn't add much. The same goes for an explicit comparison of *some* vs. *some but not all* using a language of thought-style representation. Let me know if you see worth in either option.