intuitive name	$s_\emptyset$	$s_{\exists\neg\forall}$	$s_\forall$	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	$\neg(A \cap B \neq \emptyset \wedge A \neq \emptyset)$	[TB: 10]
"not all"	1	1	0	$\neg(A \subseteq B)$	5

Table 1: Available concepts and their minimal derivation length

$C \to_2 C \wedge C$	$X \to_1 \{A, B\}$
$C \to_2 \neg C$	$X \to_1 X \cap X$
$C \to_1 X \subseteq X$	$X \to_1 X \cup X$
$C \to_1 X \neq \emptyset$	
$C \to_1 X = \emptyset$	

Table 2: Toy grammar in a set-theoretic "language of thought"

We consider three states $S = \{s_\emptyset, s_{\exists\neg\forall}, s_\forall\}$. This state space can be thought of as a partition of possible worlds into cells where none, some or all of the As are Bs, 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 (always false) or tautologous (always true). These are listed with mnemonic names in Table 1.

A lexicon L is a mapping $M \to C$ from messages to concepts. With three messages there are $6^3 = 343$ possible lexica, but many of these assign the same concept to more than one message. For simplicity, we restrict attention to only those lexica which avoid expressive redundancy. Hardcoding such an *mutual exclusivity bias* (e.g. Clark 2009), there are 20 non-redundant lexica. Functional pressure towards efficient communication will evidently favor non-redundant lexica, so this restriction is fairly innocuous but practical.

As before, the type of a player is a combination of a lexicon and a manner of pragmatic language use: literal or pragmatic. With this, there are 40 types in this model.

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 concepts in its image set. Lexica that use simpler concepts are *a priori* more likely. This can be motivated by assuming that learners beam search for suitable concepts to map onto overt signals by (probabilistically) considering simpler concepts first. Many ways for defining complexity of a concept are conceivable. If strong empirical claims were at stake here, empirically motivated measures of complexity should be used. For the sake of a non-trivial example, we follow Piantadosi et al. (under review) and related work to define complexity of a concept as a function of its derivation cost in a (weighed or probabilistic) generative "language of thought". For concreteness of example, consider the toy grammar of concepts in Table 2. This grammar uses basic set-theoretic operations to form expressions which can be evaluated as true or false in our three world states. Applications of generative rules have a cost attached to them. (Alternatively, a probability.) 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 1 lists, for each concept, the least complex formula derivable in this grammar that has the appropriate truth conditions. A simple way of defining priors over a lexicon is:

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

where $Compl(c)$ is the complexity of each concepts. [TB: I changed '$P(L) =$' to '$P(L) \propto$', as well as '$P(c) \propto \max_{c'} Compl(c') - Compl(c) + 1$'. I hope both were intended like this].

Some results

To get an idea of how this changes our previous results, I inspected some of the new setup's properties. Overall, the results are quite similar. A couple remarks:

1. So far there is no parameter to regulate the strength of the inductive bias (before controlled by c). This means that we could get stronger/weaker results than the ones presented here
2. Our target type is t_{24}, made explicit below, which currently has (i) the most a priori probable lexicon (see Figure 1), but is (ii) not the most learnable (see Figure 3), nor (iii) the fittest type (see Figure 2).
3. In terms of the dynamics, t_{24} fares worse than in our target type in the previous setup under only replication, OK but far from perfectly under mutation, but fares surprisingly well under when both dynamics are involved (see Figure 4)
4. Functional deficiency of t_{24} is due to its (Gricean) receiver behavior. In principle, this would not be so if either (i) receivers were one level higher in the reasoning type hierarchy or (ii) receivers soft-maximized.
5. As noted earlier by Michael, having our target be the most a priori probable is not so nice. On the other hand, we could stress the fact that its not the most communicatively efficient nor the most learnable language (when learning from a parent of the same type).
6. Overall, this seems to work. We have a lot more types and more complex lexica but the results are roughly the same even without involving parameter c. The most important issue is to settle on a good way to transform a LOT-like representation into a prior over lexica. Personally, I'm torn between a probabilistic conception (skipping complexity and considering the probability of each rule application directly) and something akin to what Michael proposed here. I'll read up more on how other people have done it in the past in the meantime. Let me know if you have suggestions about this. Afterwards, I can run a more proper analysis.

Details

Our target, Gricean t_{24} lexicalizes messages corresponding to the $s_\forall$, $s_\emptyset$ and $\{s_\forall, s_{\exists\neg\forall}\}$ (cf. Table 1). Its lexicon is accordingly (same as literal player t_4):

	m_{all}	m_{some}	m_{none}
$s_\emptyset$	0	0	1
$s_{\exists\neg\forall}$	0	1	0
$s_\forall$	1	1	0

Following the setup above and the complexity values in Table 1, the priors of our 40 types are shown here:

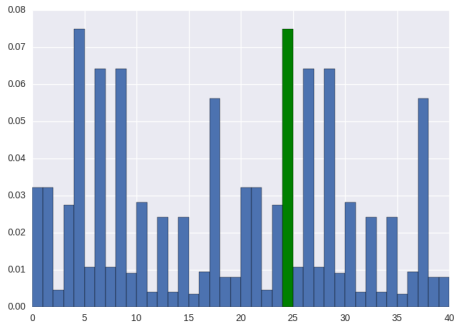

Figure 1: Prior over types (lexica). The green prior corresponds to t_{24}.

I thought it would be instructive to inspect different fitness and Q-matrices. As before, the matrix for expected utility only depends on parameters α and λ. There do not seem to be any particular surprises; t_{24} is not the best but also far from the worse (see remark 4 above). With $\alpha = 1$ and $\lambda = 30$ we get $\mathrm{EU}(t_{24}, t_{24}) \approx .88$. Here's a visualization of the entire u-matrix:

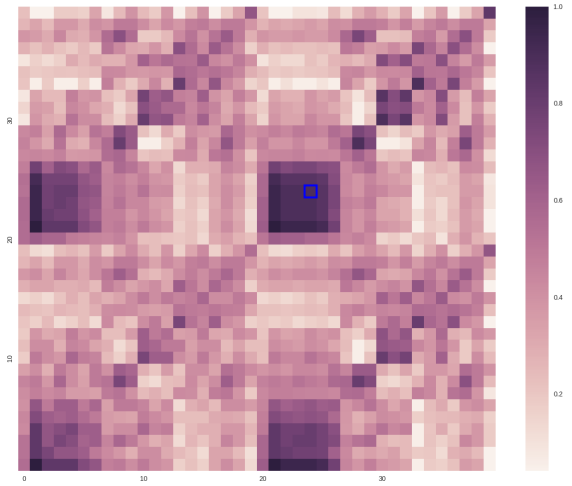

Figure 2: Utility-Matrix for $\alpha = 1$ and $\lambda = 30$. The boxed value corresponds to $\mathrm{EU}(t_{24}, t_{24})$.

For Q there are more values to test. For sequence length k there's not that much change as long as it is not too small and possible noise due to the sampling of learning data is accounted for (it's still important, but I'll background this for now). As before, the posterior parameter l is very important. If $l \geq 10$ we get very faithful transmission of our target type. Lower values lead to more mutation. Here's a visual impression of two Q matrices for different l values:

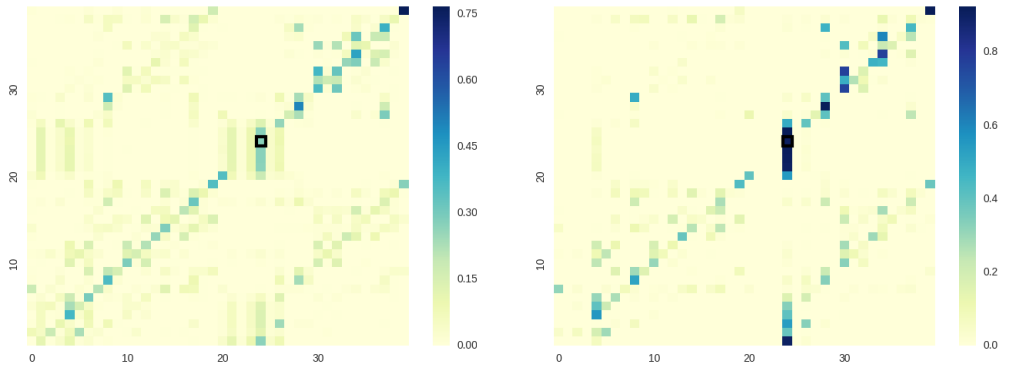

Figure 3: Q-matrices for $\alpha = 1, \lambda = 30, k = 5$, 200 samples per parent type with either $l = 1$ (left) and $l = 10$ (right). The boxed value corresponds to $P(t_{24}|t_{24})$.

Finally, we can see how the above affects actual applications of only replication, only mutation, and replication *and* mutation using the matrices shown above. To construct these plots I took the mean top 3 types from 1000 simulations using either of the three dynamics. If t_{24} was not in the top 3, I took out the third most prolific type and added t_{24} instead.

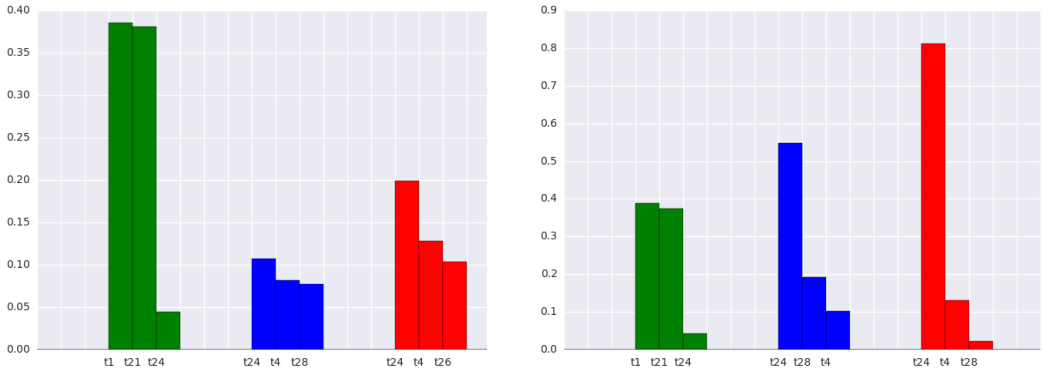

Figure 4: Mean proportion of top 2/3 types and target t_{24} under replication only (green plot), mutation only (blue plot), and replication *and* mutation (red plot) after 50 generations with $\alpha = 1, \lambda = 30, k = 5$, 200 samples per parent type and $l = 1$ (left figure) and $l = 10$ (right figure).

References

Eve V. Clark. Lexical meaning. In Edith L. Bavin, editor, *Child Language*, pages 283–300. Cambridge University Press, New York, 2009. doi: 10.1017/CBO9780511576164.016.

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.