Co-evolution of lexical meaning & pragmatic use

Revision Cover letter

We would like to thank the reviewers for their helpful comments and suggestions. They are very welcome and have certainly helped improve the article.

In general terms, we have taken care to make the exposition of the model clearer, to clarify the motivations underlying the modeling choices we make in the case study, and to clarify how the components of the model lead to the results they do. In what follows we address the reviewers' content-related comments and questions one by one, with pointers to respective changes effected on the manuscript.

Reviewer 1 comment 1 (*co-evolution*)

the co-evolutionary dynamic mentioned in the title could be better justified [...] In co-evolutionary dynamics, the evolution of one factor (here, the lexicon) should have an impact on the evolution of another (here, pragmatic strategy), and vice versa; the paper does not show evolutionary dynamics (change over time) for either of these factors, not does it show how evolution of a single factor favours the evolution of the other. The model looks more like adaptation of a kind of linguistic type (a lexicon + a pragmatic strategy) to selection pressures for communication and for learnability than co-evolution. Please justify your characterization of the process modeled as co-evolution or change to adaptation.

DL. TO DO AFTER FIGURE RMD This is an important point that we now stress and highlight more throughout the paper. Particularly in here, here, and here. The discussion surrounding Figure xxx further illustrates how one factor has an impact on the evolution of another.

In a nutshell: a pragmatic strategy (level-1 reasoning) favors the evolution of underspecified semantics of the target kind (L_{lack}) as it allows for the maintenance of simpler lexical representations that are easier to learn without functional disadvantages otherwise incurred in communication (e.g., by level-0 reasoners). In the other direction, underspecified semantics of the target kind favor the evolution of a pragmatic disposition to act on them. When paired with other semantics, pragmatic refinement can instead be (close to a) neutral trait (e.g., L_{bound} paired with high λ) or it can even lead to functional disadvantages when compared with level-0 reasoning (e.g., L_{bound} with low λ). Both cases are depicted in Figure xx.xx.

Reviewer 1 comment 2 (*modeling choices*)

the mapping between the model and the example of choice (scalar implicatures) could be better motivated

Our case study focuses on scalar implicatures because they are well-studied systematic pragmatic inferences that have received much attention, both theoretically and experimentally. This makes them particularly suitable candidates for the study of the evolution of regular pragmatic inferences because there are many findings that can guide the application and evaluation of our model to the study of these inferences. We stress these motivations in Section 1.

As for the particular choices we make to model pragmatic language use of scalar expressions, these draw from previous game-theoretic and Bayesian approaches (e.g., Franke and Jäger 2016, Goodman and Frank 2016). As above, the advantage of focusing on scalar implicatures is that we do not need to start from scratch but can scaffold on a rich tradition empirical stakes. We made the legacy of our choices more explicit in §2.3.1.

Reviewer 1 asks for a clearer description of the model. As noted earlier, we have taken care to make the exposition more accessible and to guide the reader more through the technical details of Section 2. In particular, Section 2.xx is new and should clarify much. We took the reviewer's questions, briefly answered below, as guidance for the changes made in Section 2.

How is generation transmission implemented?

As it is unlikely that lexical representations and/or pragmatic reasoning are inherited genetically, we model the transmission of linguistic knowledge across generations as iterated Bayesian learning. Technical details and motivations are given in §2.3.3.

Do agents in a population die little by little and are replaced by new agents? Or all at once?

As now made more explicit in §2.2, steps of the discrete-time replicator mutator dynamic correspond to generational turn-overs of the entire population (definitions (1) and (2)). That is to say, all agents die at once, being replaced by the following generation in a single swoop.

Do the models consider just a single teacher and a single learner at a time (I don't think this is the case, since the results refer speak of "majority type, i.e., the type with the highest proportion in the final population")?

The parenthetical remark is correct. The amount of teachers of a given type is proportional to the frequency of this type in the population after the replicator step. More generally, the population is infinite so there are infinite learners/teachers. This was implicit in the definitions of the replicator mutator dynamic in (1) and (2) before, and is now explicitly stated in Section 2.2 in the prose that motivates and explicates these definitions.

How many agents were there per population? And per generation?

The prose surrounding definitions (1) and (2) in Section 2.2 now makes explicit that the population is infinite. This allows us to track change that does not depend on varying population sizes nor their growth rate. This is why we speak of proportions of agents of a type, rather than of there being a particular number of agents of a type.

Within one run of the model, do all agents in a population have the same lambda, l and k? If so, please this homogeneous population, and discuss the possible consequences of variation in the population in these

parameters.

While our model is well-compatible with the assumption that every agent comes with her own parameter values, the reviewer is correct in pointing out that we assume they all to have the same values. Variation across agents would mean many more types: one for each combination of a possible lexicon, pragmatic strategy, λ-, l- and k-value. This would allow us to trace the evolutionary trajectory of not only lexica and pragmatic strategies, which is our goal here, but of all these factors combined. As our focus is not to show that, say, a particular λ value leads to a particular evolutionary outcome we restrict our attention to a homogeneous population where all agents have the same parameter values. We discuss this issue in Section 3.2.

Does the model assume that learners observe not only the behaviour (the message, e.g. agent A hearing agent B say: "I own some of JC's albums") but also the true state of the world (i.e. agent A also knows whether agent B owns all or some-but-not-all albums)? Or do they get feedback on their communicative success? In other words, how do they learn the correct state-message mappings that form the lexicon?

Learners witness sequences of data d. Such a sequence is composed of message-state pairings. In other words, learners do witness the true state of the world (Section 2.3.3, *learnability*).

The probability of the learner receiving learning input d depends only on the likelihood of the teacher producing it. This is determined by the teacher's production probabilities; the speaker behavior of her type. Put differently, communicative success does not come into play, only speaker behavior. It would certainly be possible to let learners witness proficient users communicate and to have them infer types from such interactions instead. However, we wanted to keep our assumptions close to the well-studied assumptions standardly made in the iterated (Bayesian) learning tradition. Iterated learning only considers production, rather than production and comprehension (see discussion that should make this explicit in the discussion section - maybe draw from thesis). Details on how learners infer types are given in Section 2.3.3.

How exactly are pragmatic strategies inherited/inferred by learners? I think all of this may be in the formulas, but it would be useful if it was expressed in prose too.

As with the other questions, we agree and have made these matters more explicit. This matter is particularly important. We now stress it throughout the article.

In short, a type is a combination of a lexicon and a pragmatic disposition to act on it. Together, both ingredients define an agent's linguistic behavior. A learner faced with input d (see above) calculates $P(\tau \mid d)$, the probability that a type τ produced input d. Crucially, neither pragmatic strategies nor lexical meanings are directly observable, so they can only be faithfully recovered insofar as the behavior evidenced by a type is (in tendency) attributable to only one pairing of a lexicon and a pragmatic strategy.

> Could you express in prose how literal and pragmatic use work in practice? I am not sure I understand why you call the strategies 'pragmatic', given that pragmatics depends on the context, and there is no context here. I would like to see the use of the term 'pragmatic' justified further in the paper.

We follow the Gricean tradition of viewing pragmatic inference as effected by mutual reasoning about rational language use (Section 1). This may involve the recruitment of contextual information, as the reviewer notes, but can also involve pragmatic enrichments that result purely from reasoning about linguistic alternatives ("why did the speaker say *some* instead of *all*"). It is in the latter sense that reasoning beyond level 0 is pragmatic. Classically, theoretical analyses of scalar implicatures put their emphasis on reasoning about linguistic choice rather than on contextual information (e.g., Horn 1972, Gazdar 1979, Franke 2009, Goodman and Stuhlmüller 2013) and we chose to do the same.

> I thought λ was the parameter that modeled communication, so it should not be present in the iterated-learning-without-communication models; why is λ set to 20 here, and why discuss lower levels of lambda?

This parameter regulates the linguistic production behavior of types, i.e., speaker behavior. Speaker behavior is relevant both for communication with hearers as well as for the learning input that a teacher produces. When considering only iterated Bayesian learning λ accordingly also plays a role because different λ values yield different likelihoods of producing data (see above and Section 2.3.3).

Also [in the Subsection *learnability only*], why isn't the proportion of Lall highest, given that it has the lowest complexity (highest prior)?

Due to their stochastic speaker behavior, the data produced by L_{all}-teachers tends to be compatible with the behavior of many other types. Intuitively, the data such teachers produce are all over the place and do not do a good job in setting them apart from other types. Consequently, even if the prior favors L_{all}, this type is not transmitted very faithfully; learners reason that it could also come from other types and may adopt those instead. If there are types that are transmitted more faithfully, the population will, over time, transition to these types rather than to L_{all}. As we put it in Section 4, iterated learning does not necessarily promote the *a priori* more likely type, but tends to promote a type t based on a gradient of how many other types might likely mutate into t, so to speak. We stress this in Section 3.2.2 and Section 4.

Reviewer 1 comment 4 (*LOT & model relation*)

This model rests heavily on the assumption that the representation or meaning of an expression is a logical formula like those in Table 2, and that these formulas yield a complexity hierarchy. This may be valid for scalar implicature, as you argue, but how realistic is it cognitively for other linguistic structures or for other socially learned items? Should we assume the non-parsimonious view that there are different complexity measures for different linguistic structures? It would be nice to see another linguistic or cultural example that could be fitted with this model.

Learnability [...] is operationalized as the complexity of the minimal logical formula for each state. Another possible measure of the difficulty/ease of learning would be comparing the regularity/systematicity of the lexicon types given in 3.1.1 (measured e.g. as mutual information (Cornish, Tamariz, Kirby 2009; using the mantel test, Kirby, Tamariz, Cornish, Smith 2015). Yet another approach to complexity would be to assume that the meanings are mutually exclusive categories without internal structure, or to assume a preference or mutual exclusivity. In these cases, I expect L-bound (competitor), to be easier to learn than L-lack (target). Please justify your choice of learnability measure in the

face of these alternatives.

We do not believe that, ultimately, we should assume the non-parsimonious view that there are different complexity measures for different structures. However, with previous work in iterated learning tradition, we do assume that (i) inductive learning biases can play a role in shaping culturally transmitted knowledge and that, as the reviewer notes, (ii) biases can be of manifold nature; another well-studied example being the mutual exclusivity bias, which could be expected to work against targets and in favor of competitors.

At present, it is impossible to disentangle the complex interaction of multiple biases. We therefore focus on the effects of a single plausible contributing factor to a preference of target lexica over competitors: a well-motivated bias favoring simplicity (Feldman 2000, Chater and Vitányi 2003, Piantadosi et al. 2012, Kirby et al. 2015, Piantadosi et al. view). By assumption, this translates to a preference not to lexicalize linguistic material over lexicalizing it; simpler formulae over complex ones. As mentioned in Section 3.1.2, we acknowledge that the complexity measure that the prior is based on is just a convenient operationalization of this particular hypothesis.

Lastly, we should stress that, just as models of iterated learning make different predictions under the assumption of different inductive biases, so does our model. The assumption and operationalization of a particular bias should, of course, always be seen critically (Section 4).

Reviewer 1 comment 5 (*transmission fidelity*)

What is your measure of fidelity/infidelity? Is it just 0 if the teacher and learner are of a different type and 1 if they are of the same? Could you have a more graded measure of fidelity, e.g. taking into account if two types are of the same 'kind' or not?

We understand transmission fidelity as the transition probability of acquiring type i when learning from type j. This is the value of a cell in the learning matrix Q. In this case, Q_{ji}. If $Q_{ji} = 1$ then i is always acquired when learning from type j. On the other extreme, $Q_{ji} = 0$ means that i is never acquired from j. Q is a stochastic matrix and we we usually see neither extreme but rather, as suggested by the reviewer, a graded notion of transmission fidelity. The details on how Q_{ji} is computed are given in Section 2.3.3, which we have expanded to make these details clearer in the prose.

Reviewer 2 comment 1 (*Interpretation of results*)

The results are interesting, though somewhat difficult to interpret. First, pragmatic language use does not evolve in the absence of learnability pressures, due to its slightly lower fitness. Second, pragmatic language use can arise from learnability pressures alone, though only for sufficiently "rational" learners. Third, pragmatic language use can evolve given both fitness and learnability pressures, though only if both learners and language speakers are both sufficiently optimal. The first two results are fairly straightforward, but the third (which is the most important in the paper) is more puzzling. Why do learnability pressures not have the same effect in the joint learnability/fitness model, as they do in the lesioned model? Why is high speaker optimality necessary for scalar implicatures to evolve, when it does not have this effect in the fitness-only model?

We have thoroughly revised the article to make the contrast between these three predictions clearer, as well as the predictions themselves. Concretely, say where this is more explicit now. This should be the new figure as well as a clearer contrast when discussing the simulation outcomes.

These questions tie in with Reviewer 1's request for a clearer exposition and the explanation of why we speak of co-evolution in this case (reviewer 1, comment 1). In a nutshell, pressure for learnability alone does not put the types found in a population in competition. Instead, the population comes to be inhabited by types that are faithfully transmitted – of which there might be many (e.g., all variants of a kind are equally learnable; if there are many variants of a type that is, in tendency, faithfully transmitted the population will stay polymorphic). What fitness-relative selection adds to this process is the competition that learnability alone lacks, independent of the type space we look at. Now, in the case of our type space we have a slight (modulo λ) communicative disadvantage for targets when in direct competition with competitors, for instance. However, the fact that targets are inferred more faithfully by naïve learners leads (i) to the gradual existence of less competitors, counteracting this, if slight, functional disadvantage, as well as (ii) to the existence of less variants of a kind (contra the *learning alone* dynamic).

Reviewer 2 comment 2 (*Model predictions*)

I agree with its claim that the results are non-trivial; it is not clear a priori that there should exist any regimes where learnability and fitness pressures balance in order to produce scalar implicatures. The paper does not, however, address why the pressures do balance in this case. Are there general theorems (from previous work) about the replicator mutator dynamic that could be brought to bear here? How do learnability and fitness pressures trade-off in general? Is there anything more general that can be said about conditions under which pragmatic language use will evolve? Does the current example generalize to other cases of pragmatic language use? The paper does not need to answer these questions in an exhaustive manner, but greater scientific understanding of the model is desired.

As noted above, we have expanded our analysis of the model and put much more emphasis on how the components that feed it play a role in driving particular evolutionary outcomes. See, in particular, sections this and that.

Reviewer 2 comment 3 (*co-evolution and pragmatic maintenance*)

It makes sense for agents to acquire their grammar/lexicon by learning, as these are social conventions. I believe, however, that the paper also assumes that an agent's pragmatic type (i.e. whether they are literal or pragmatic speakers) is also determined by learning. This is more difficult to interpret. It is reasonable for an agent to learn whether the other agents in a population are behaving pragmatically. Even if the other agents are not behaving pragmatically, however, it will be rational for the learner to behave pragmatically. This is a property of the recursive definition of the pragmatics model. The pragmatic speaker is defined so as to select utterances that will be correctly interpreted by the literal listener. If an agent learns that their interlocutors are all literal agents, then it would be rational for them to pragmatically reason about these literal agents. It is therefore not clear why the agents should, as currently proposed, copy the inferred pragmatic type of their interlocutor.

It is not necessarily true that it will always be rational for agents to behave pragmatically. We now make explicit in Section 2, but the main thing to note

is that a pragmatic level-1 hearer using lexicon L_{bound} will reason over the behavior of a *soft-maximizing* level-0 speaker of L_{bound}. The lower λ, the more the stochasticity percolates to level-1 pragmatic interpretation. By contrast, literal level-0 hearers of L_{bound} have a one-to-one form-meaning mapping from the get go. Put differently, pragmatic reasoning can actually encumber some hearers, depending on their lexicon. Particularly if λ is low.

Reviewer 2 comment 4 (*agent simplicity*)

The paper assumes that the speaker (and listener) know the type of the player that they're playing against. This does not seem like a natural assumption. Unless the speaker has repeated interactions with the listener (which is not assumed in the paper), the pragmatic type of the other agent is a latent property. One could imagine the speaker calculating expected utility with respect to the distribution on players that they expect (and speakers of different types choosing different strategies against this distribution).

As we now clarify in Section n and stress a couple of times throughout the article, we assume very little sophistication from our agents. In particular, they do *not* know the type of the player that the are playing against. They simply behave according to their subjective point of view in a boundedly rational fashion. For example, the linguistic choices of a level-1 speaker of lexicon L_{some} do not change depending on whom she interacts with; they are always defined as (boundedly) rational choice relative to the interpretative behavior of a level-0 user of L_{some} (point to definitions of behavior in section 2).

Reviewer 2 comment 5 (*posterior type sampling*)

The learner in this model always selects a particular grammar/pragmatic type. An alternative would be for them to maintain uncertainty about the type (which is optimal behavior from a Bayesian perspective). I suspect that this model is intractable (or at least much less manage-

able) than the proposed one, but it would be worth noting that this is a substantive choice point in the model.

We now briefly discuss our choice and alternatives in Section n.

Reviewer 2 comment 6 (*definitions in rational language use*)

The current pragmatic model is somewhat non-standard. In particular, neither the literal nor pragmatic speakers in Equations 3-6 use an information-theoretic utility function, as is standard in the literature. One consequence of this is that there is now an asymmetry between the literal speaker and listener: the literal listener never interprets utterances in a manner inconsistent with their literal meanings, while the literal speaker will sometimes use utterances in a non-literal way. I do not think this needs to be changed – the current utility function is perfectly sensible – but this is worth noting in the paper. It would also be desirable to know (possibly in future work) whether the modeling results are robust to switching to information-theoretic utility functions.

A related issue is in the fitness definition in Section 2.3.2. If one adopts the alternate utility function above for the agents' behavior, then it would be natural to have fitness scale in a logarithmic manner as well. In this case, the information-theoretic utility function would have the following interpretation: the fitness of a language is determined by how well it allows speakers to communicate their beliefs, rather than how often it leads the listener to make the correct guess about the world. Similar robustness questions apply here as above.

There are many alternatives when it comes to definitions of rational language use, many of which do not use an information-theoretic utility function (see Qing and Franke 2015 for an overview and discussion up to 2015). As noted in our answer to Reviewer 1's comment 2, we have taken care to highlight that our choices draw from this literature and to mention that there are other alternatives in the same spirit.

Reviewer 2 comment 7 (*Learnability*)

Throughout, the paper mostly equates the learnability of a grammar with its complexity/prior probability. In general, however, the relationship may be more complex. Certain grammars may require less data than others to be statistically identified. This is probably not an issue in the current case study, given the simplicity of the grammars and the large amount of data relative to grammar size, but this is likely to be an issue in other cases.

This is true. We are grateful that this was pointed out, as our previous rendering could lead to confusion, e.g., as to why L_{all}, being the *a priori* more likely type to be inferred, did not come out strongly even when considering learnability alone. Throughout the article, we now stress more that what matters is the posterior – and discuss the role of the prior mainly when contrasting targets and competitor, who crucially differ mostly in this respect.

Reviewer 3 comment 1 (*expected utility*)

what would be the impact of assuming that fitness is not contributed equally by being successful as speaker and being successful as a perceiver?

This is a standard assumption in the literature lacking reasons to believe that there is an asymmetry in how often, on average, agents use and hear a type of expression. In the case of scalar implicatures, we see no reason to assume such an asymmetric contribution to fitness.

As for concrete impact: The main contrast that we are interested in, i.e., competition between targets and competitors, does not hinge on agents being speakers half of the time. However, the contrast between pragmatic and literal competitor types does, because their receiver behaviors differ. Assuming that successful receiver behavior contributes differently than sender behavior would therefore modulate differences between these types.

Reviewer 3 comment 2 (*message amount*)

what motivates modelling just three messages and not just two ("some" and "all") or any $n > 3$? One obvious prediction: $n = 2$ will lead to larger prevalence of lack types, whereas increasing n beyond 3 will do the opposite. More concretely: is there any non-trivial relative frequency of

some/all that could falsify the results obtained in the paper?

As now made explicit in Section xx.xx, we chose three messages for illustrative purposes mainly. In particular, we wanted to inspect a type space where targets are not the most likely *a priori* as this may mislead readers into thinking that learnability alone, in particular the prior alone, drives the outcome. This is related to our answer to reviewer 2, comment 7. The reason why we did not go beyond $n > 2$ is for computational tractability, as calculating Q, even if approximated, is expensive. The question whether increasing the type space will decrease the frequency to which we expect to see pragmatic *some* users is interesting. It now figures in Section X.X. At present, we have no answer.

Reviewer 3 comment 3 (*LOT & complexity*)

The discussion on derivation costs is not very satisfying on my opinion. For instance, the condition of "simple representations being favored over more complex ones" is compatible with many more cost specifications.

We have expanded this section to further motivate our choices. Relatedly, see our answer to Reviewer 1's comment 4 above.

References

Chater, N. and Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1):19–22.

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

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

Franke, M. and Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1):3–44.

Gazdar, G. (1979). *Pragmatics, Implicature, Presuposition and Logical Form*. Academic Press, New York.

Goodman, N. D. and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Goodman, N. D. and Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5:173–184.

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

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

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

Piantadosi, S. T., Tenenbaum, J. B., and Goodman, N. D. (under review). Modeling the acquisition of quantifier semantics: a case study in function word learnability.

Qing, C. and Franke, M. (2015). Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning. In *Bayesian Natural Language Semantics and Pragmatics*, pages 201–220. Springer International Publishing.