Communicative pressures at the semantics-pragmatics interface: Learning biases may prevent the lexicalization of pragmatic inferences

(– draft September 2, 2016—)

1 The semantics-pragmatics divide

Abstract

Natural languages allow for conventional pragmatic enrichments in a notably productive fashion. This raises the challenge to justify the regular selection of particular lexical meanings over other alternatives. In particular, over those that codify semantically what is conveyed pragmatically. To address this challenge, we propose a general model for the analysis of linguistic pressures that integrates iterated Bayesian learning in the replicator-mutator dynamics commonly used in evolutionary game theory. This model allows for population-level analyses of rational probabilistic language users with varied degrees of pragmatic sophistication and distinct languages. In a case-study on the (lack of) lexicalization of scalar implicatures, we show how simpler semantic representations are selected for when languages are pressured towards learnability provided that pragmatic reasoning can compensate for the disadvantage in expressivity that users of such languages would otherwise incur.

In linguistic theorizing, it is common to draw a distinction between semantics and pragmatics. Broadly speaking, the former concerns the truth-conditional content of expressions, whereas the latter concerns information beyond literal meanings and their composition. Thus, under this view, the information conveyed by an utterance is seldom, if ever, solely determined by semantics, but rather in tandem with pragmatics.

Much research at the semantics-pragmatics interface has been aimed at characterizing expressions in terms of either domain, or their interplay. As a consequence, their distinction has played an important role in the field's theoretical and experimental development. Notwithstanding, an issue that has received little attention is the justification of semantic structure in light of pragmatics. That is, explanatory and predictive accounts of the selection and pervasiveness of particular semantic structures under consideration of the informational enrichment provided by pragmatics.

While their efforts have largely concentrated on compositional and combinatorial systems, a number of recent investigations have began to address the development and selection of linguistic properties (see Steels 2015 and Tamariz and Kirby 2016 for recent overviews). Our starting point is given by the overarching account of competing pressures that has crystallized across these approaches: Natural languages need to be well-adapted to the communicative needs within their linguistic community, but also need to be learnable to survive their faithful transmission across generations. More succinctly; natural languages are pressured for expressiveness as well as learnability.

We proceed by modeling these pressures using the replicator-mutator dynamics (see Hofbauer and Sigmund 2003 for an overview). This allows us to model the dynamics of linguistic pressures by combining functional pressure on successful communication, effects of learning biases on (iterated) Bayesian learning (Griffiths and Kalish 2007), and probabilistic models of language use in populations with distinct lexica (Frank and Goodman 2012, Franke and Jäger 2014, Bergen et al. 2016). In this way, the model connects the recent surge of models of synchronic probabilistic rational language use with diachronic models of cultural evolution.

2 Simplicity, expressivity, and learnability

At least since Zipf's (1949) rationalization of the observation that word frequency rankings can be approximated by a power law distribution as competing hearer and speaker preferences, the idea that languages are shaped by trade-offs of competing pressures has played a pivotal role in synchronic and diachronic analyses (e.g. Martinet 1962, Horn 1984, Jäger and van Rooij 2007, Jäger 2007, Piantadosi 2014, Kirby et al. 2015).

A cornerstone principle, both in linguistics and cognitive science more generally, is that of simplicity. In linguistics, 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 grammar simplicity (Zipf 1949, Grice 1975, Piantadosi et al. 2012a, Kirby et al. 2015)). As a broader cognitive principle, the use of simplicity as means to select betweeen hypotheses that fit the data has a long standing tradition. Crucially, Chater and Vitányi (2003) provide strong arguments for a prior towards simpler explanations.

As noted above, a feature more particular to natural communication and its population dynamics are pressures towards expressivity and learnability. Their opposition becomes particularly clear when considering their consequences in the extreme (cf. Kemp and Regier 2012, Kirby et al. 2015). On the one side, a language with a single form is easy to learn but lacking in expressivity. On the other, a language that associates a distinct form with all possible meanings its users may want to convey is maximally expressive but challenging to acquire. However, acquiring a system to express a potentially infinite set of meanings through finite means, the so-called transmission bottleneck (Kirby 2002), is not the only problem learners confront. Instead, the challenge we focus on here concerns the selection between functionally similar, if not identical, pragmatically enriched lexical meanings while striking a balance between these pressures.

A particularly well-studied type of conventional pragmatic enrichment are so-called *scalar implicatures* (cf. Horn 1972, Gazdar 1979). These inferences 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, while weaker expressions such as *some* are truth-conditionally compatible with stronger alternatives such as *all*, this is not necessarily what their use is taken to convey. Instead, the use of a less informative expression when a more informative one could have been used can license the inference that the stronger alternative does not hold. That is, a hearer who assumes the speaker to be able and willing to provide all relevant information can infer that, since the speaker did not use a stronger alternative (*all*), this alternative must not hold. In this way, 'Some students came to class' is strengthened as conveying 'Some but not all students came to class'. Analogously, a speaker can rely on her interlocutor to draw this inference without having to express the bound overtly, e.g. by stating *some but not all*. In summary, mutual reasoning about rational language use supplies an upper-bound that rules out stronger alternatives pragmatically.

We can now rephrase our initial question in terms of scalar implicatures: Why is the lack of lexical upper-bounds in weak scalar alternatives regularly selected for over the alternative of lexicalizing it? This question is particularly striking considering the number of classes of expressions that license such inferences across languages (Horn 1972, Horn 1984:252-267, Traugott 2004, van der Auwera 2010). In the following, we assume a pivotal part of the answer to be that learners are a priori biased towards simpler (more compressed) representations, corresponding to the argument that rational learners should prefer simpler over more complex explanations of data (Feldman 2000, Chater and Vitányi 2003, Piantadosi et al. 2012b, Kirby et al. 2015, Piantadosi et al. under review).

The remainder of this section introduces the individual components of the model in more technical detail. That is: (i) languages and their use, (ii) pressures towards expressivity and learnability, regulated by the replicator and mutator dynamics, respectively, as well as (iii) a bias towards simpler semantic representations, codified in the learner language learners' prior. After laying out the model, we return to its application to the lack of lexicalization of scalar implicatures.

2.1 Lexica and linguistic behavior

Lexica are taken to codify the truth-conditions of a language's expressions, i.e., its semantics. Put differently, given a state of affairs and a lexicon, one may check whether a language's expression is true or false. A convenient way to represent such lexica is by $(|S|, |M|)$-Boolean matrices, where S is a set of states of affairs and M the set of messages in the lexicon (Franke and Jäger 2014). For instance, two possible lexica fragments that codify the truth-conditions of two messages, m_1 and m_2, for two states, s_1 and s_2 are:

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

In words, m_1 is true of s_1 and s_2 according to lexicon L_a but only true of s_1 in L_b. Otherwise, the two languages are truth-conditionally equivalent.

To make the distinction betweeen semantics and pragmatics precise, we distinguish between two kinds of linguistic behavior. *Semantic interlocutors* produce and interpret messages literally. That is, their choices are solely guided by the their lexica. In contrast, *pragmatic interlocutors* engage in mutual reasoning to inform their choices. For instance, rational speakers of L_a should use m_1 in state s_2, given that m_2 is already unambiguously associated with s_1. Analogously, rational hearers can expect their interlocutors to make this choice and interpret the messages accordingly. Note that this corresponds to the description of scalar implicatures given above: Reasoning about the alternatives at the speaker's disposition can pragmatically strengthen an expression's literal meaning by ruling out stronger alternatives not mentioned.

Following models of rational language use such as Rational Speech Act models (Frank and Goodman 2012) and their game-theoretic counterparts (Benz et al. 2005, Franke 2009, Franke and Jäger 2014) this kind of signaling behavior is captured by a hierarchy over reasoning types.

2.2 Replication & expressivity

2.3 Mutation & learning

2.4 Competing pressures in cultural evolution

- discuss past research on interplay of pressures from a population-level perspective (see discussion in Kirby et al. 2015 for IL

- Learnability in iterated learning
- Expressivity in GT and some IL
- Sketch of replicator-mutator dynamics

More detailed discussion of models of cultural evolution. Short overview of past research with a focus on the difference between pure IL and functional pressure together with IL. **Possibly add a direct comparison of IL and RMD in the appendix using the setup of Griffiths & Kalish 2007. This may not be necessary.**

The emergence and change of linguistic structure is driven by many factors, from biological and socio-ecological to cultural (Steels 2011, Tamariz and Kirby 2016). Broadly put, social and ecological pressures determine communicative needs, while biology determines the architecture available for its use. Our focus is on the latter, cultural, factor, wherein linguistic structure is analyzed in terms of its use, as well as its transmission across generations.

As already mentioned in §1, research on selectional forces that apply in the cultural evolution of language has focused on two main pressures: expressivity and learnability. However, while it is generally acknowledged that both play a pivotal role, past approaches have focused exclusively, or at least emphasized, the role of one over the other (a recent exception is Kirby et al. 2015).

Expressiveness, or communicative efficiency, has been at the center of applications of evolutionary game theory to linguistics (Nowak and Krakauer 1999, Huttegger and Zollman 2013), explain RMD

In contrast, the iterated learning paradigm has focused on the effects of language transmission from generations of speakers to the next. explain IL

3 Model

3.1 Replicator-mutator dynamics

(II) Sequences and atomic observations. Before, the set of all observations was $O = \{\langle\langle s_1, m_i\rangle, \langle s_2, m_j\rangle\rangle | m_i, m_j \in M\}$. A member of O encodes that a teacher produced m_i in state s_1 and m_j in s_2, i.e., it encodes one witnessed message for each state. A datum d was a sequence of length k of members of O. Learners witnessed such data sequences. Now, more in line with Griffiths and Kalish (2007), $O = \{\langle s_i, m_j\rangle | s_i \in S, m_j \in M\}$ and d is a sequence of length k of members of O. The main difference is that now some d do not provide any production information for some states.

(III) Observations as production. Instead of taking the space of all possible sequences of length k into consideration, we take sample from O k-times according to the production probabilities of each type; $P(o = \langle s, m\rangle | t_i) = P(s)P(m|s, t_i)$. n such k-length sequences are sampled for each type. As a consequence, the data used for computing Q_i is not the same as that used for j $(i \neq j)$.

(IV) Parametrized learning $Q_{ij} \propto \sum_d P(d|t_i)F(t_j, d)$, where $F(t_j, d) \propto P(t_j|d)^l$ and $l = 1$ corresponds to probability matching and, as l increases towards infinity, to MAP.

The proportion of players of type i, x_i, is initialized as an arbitrary distribution over T. $p^\star \in \Delta(T)$ is learning a prior over (player) types dependent only on the lexicon of the type.

- $f_i = \sum_j x_j U(x_i, x_j)$

- $\Phi = \sum_i x_i f_i$
- $Q_{ij} \propto \sum_d P(d|t_i)\ P(t_j|d)$, where $P(t_j|d) \propto [P(t_j)P(d|t_j)]^l$, d is a sequence of observations of length k of the form $\langle\langle s_i, m_j\rangle, \ldots \langle s_k, m_l\rangle\rangle$, and $l \geq 1$ is a learning parameter.
- For parental learning (standard RMD): $\dot{x}_i = \sum_j Q_{ji} \frac{x_j f_j}{\Phi}$

3.2 Expressiveness as fitness-relative replication

Symmetrized expected utility. With $P \in \Delta(S)$ (uniform so far; $P = pr$):

- $U(t_i, t_j) = [U_S(t_i, t_j) + U_R(t_i, t_j)]/2$
- $U_S(t_i, t_j) = \sum_s P(s)\ \sum_m P_S(m|s; t_i)\ \sum_{s'} P_R(s'|m, t_j)\ \delta(s, s')$, where $\delta(s, s')$ returns 1 iff $s = s'$ and otherwise 0
- $U_R(t_i, t_j) = U_S(t_j, t_i)$

3.3 Iterated learning as acquisition-based mutation

3.4 Signaling behavior

Leave languages undefined until application

Signaling behavior. Exposition of signaling behavior as reasoning hierarchy, which we use to make a distinction between semantic and pragmatic language users

With $\lambda \geq 1$ (rationality parameter), $\alpha \in [0, 1]$ (pragmatic violations) and $pr \in \Delta(S)$ a common prior over S (uniform so far):

$$R_0(s|m; L) \propto pr(s)L_{sm} \tag{1}$$
$$S_0(m|s; L) \propto \exp(\lambda\ L_{sm}) \tag{2}$$
$$R_1(s|m; L) \propto pr(s)S_0(m|s; L) \tag{3}$$
$$S_1(m|s; L) \propto \exp(\lambda\ R_0(s|m; L)^\alpha) \tag{4}$$

4 Lack of semantic upper-bounds in lexical meaning

Procedural description. The game is initialized with some arbitrary distribution over player types. At the game's onset we compute Q once based on the sets of sequences D (one for each parent type). Replicator dynamics are computed based on the fitness of each type in the current population as usual. Q is computed anew for each independent run (of g generations) given that it depends on D, which is sampled from production probabilities.

Languages. We consider a population of players with two signaling behaviors, literal and Gricean (level 0 and 1 below), each equipped with one of 6 lexicons. This yields a total of 12 distinct player types $t \in T$. $|M| = |S| = 2$, i.e., a lexicon is a $(2, 2)$-matrix. These are listed in Table 1.

As in the CogSci paper, L_4 (semantic upper-bound for m_2) and L_5 (no semantic upper-bound for m_2) are the target lexica. Gricean L_5 users can convey/infer the bound pragmatically, while literal/Gricean L_4 users do so semantically.

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

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

Table 1: Set of considered lexica.

4.1 Model parameters & procedure

1. Sequence length k
2. Pragmatic production parameter α
3. Rationality parameter λ
4. Learning prior over types (lexica); cost parameter c. $p^\star(t_i) \propto n - c \cdot r$ where n is the total number of states and r that of upper-bounded messages only true of s_1 in t_i's lexicon (if only s_1 is true of a message, then this message encodes an upper-bound). Then the score for L_1, L_3, L_5 is 2, that of L_4 and L_6 is $2 - c$, and that of L_2 is $2 - 2c$; Normalization over lexica scores yields the prior over lexica (which is equal to the prior over types).
5. Prior over meanings (pr). We assume that $pr(s) = \frac{1}{|S|}$ for all s.
6. True state distribution (P). We currently assume that $P = \frac{1}{|S|}$ but it may be interesting to vary this
7. Learning parameter $l \geq 1$ with 1 corresponding to probability matching, and MAP as l approaches infinity
8. n is the sample of sequences of observations of length k sampled from the production probabilities of each type
9. Number of generations g

5 Discussion

6 Extensions

(I) Cost for pragmatic reasoning. At least in the CogSci setup the effect of adding cost to pragmatic reasoning is unsurprising: High cost for pragmatic signaling lowers the prevalence of pragmatic types. Lexica that semantically encode an upper-bound benefit the most from this. However, the cost needed to be substantial to make the pragmatic English-like lexicon stop being the incumbent type (particularly when learning is communal).

(II) Negative learning bias. Instead of penalizing complex semantics (semantic upper-bounds) one may consider penalizing simple semantics (no upper-bounds). This is useful as a sanity check but also yields unsurprising results in the CogSci setup: The more learners are biased against simple semantics, the more prevalent are lexica that semantically encode upper-bounds.

(III) Inductive bias. A second learning bias that codifies the idea that lexica should be uniform, i.e. be biased towards either lexicalizing an upper-bound for all weaker alternatives in a scalar pair or for none.

(IV) Uncertainty. The other advantage of non-upper bounded semantics lies in being non-committal to the negation of stronger alternatives when the speaker is uncertain. Adding this to the model requires the most changes to our present setup and some additional assumptions about the cues available to players to discern the speaker's knowledge about the state she is in.

(V) More scalar pairs. Taking into consideration more than one scalar pair. Preliminary results suggest that this does not influence the results in any meaningful way without further additions, e.g. by (III).

(VI) More lexica. Not necessary. Preliminary results suggest that considering more lexica has no noteworthy effect on the dynamics (tested with all possible 2x2 lexica).

(VII) State frequencies. Variations on state frequencies. This may have an interesting interaction with (III).

(VIII) Reintroduction of communal learning. One possibility: The probably N_{ij} with which a child of t_i adopts t_j could be the weighted sum of Q_{ij} (as before) and a vector we get from learning from all of the population: $L_j = \sum_d P(d|\vec{p})P(t_j|d)$, where $P(d|\vec{p}) = \sum_i P(d|t_i)\vec{p}_i$ is the probability of observing d when learning from a random member of the present population distribution.

7 Conclusion

References

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

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

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

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

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

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

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

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

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

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.

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.

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

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.

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

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

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

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.

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

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.

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

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

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.

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

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, 2012b.

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.

Anton Benz, Gerhard Jäger, Robert Van Rooij, and Robert Van Rooij, editors. *Game theory and pragmatics*. Springer, 2005.

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

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

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

S. M. Huttegger and K. J. S. Zollman. Methodology in biological game theory. *The British Journal for the Philosophy of Science*, 64(3):637–658, 2013.