Systematic transmission perturbations in the cultural evolution of language

Name Surname (mail@mail.com)
Department, street & number
city, ZIP country

Name Surname (mail@mail.com)
Department, street & number
city, ZIP country

Abstract

Over time, languages favors linguistic features that can be passed on with high fidelity from one language user to the next. The outcomes of this process are often argued to involve cognitive biases that influence a learner's inductive task. Such learning biases thereby serve as central devices to understand and predict linguistic structure. We complement this view by showing how such effects can also arise without assuming any biases but simply as an epiphenomenon of systematic disturbances stemming from environmental factors. To this end, we investigate the effects of iterated learning under noisy perception in three case studies on (i) vagueness, (ii) meaning deflation, and (iii) a lack of upper-bounds in weak scalar alternatives. We argue these technical results to underpin the importance of, either cognitive or extraneous, transmission perturbations in the cultural evolution of language and bring attention to the often overlooked possibility that channel noise can mimic effects of inductive biases.

Keywords: noise; cognitive biases; iterated learning; cultural evolution;

Introduction

Language is shaped by its use and transmission across generations. Linguistic properties therefore need not necessarily arise and stabilize solely due to functional pressure but may also be influenced and selected for by a pressure for learnability. The effects that such iterated learning has on language can be viewed as arising from a combination of general learning mechanisms and inductive cognitive biases (e.g. Griffiths & Kalish 2007, Kirby et al. 2014, Tamariz & Kirby 2016). Proposals of biases that shape language acquisition abound. Some prominent examples are mutual exclusivity (Merriman & Bowman 1989, Clark 2009), simplicity (Kirby et al. 2015), regularization (Hudson Kam & Newport 2005), and generalization (Smith 2011, O'Connor 2015).[1] In the following we show how environmental factors can produce evolutionary outcomes that look as if such cognitive learning biases are present even if they are not.

We present three case studies that show how transmission perturbations can lead to well-known linguistic phenomena: vagueness, meaning deflation, and a lack of upper-bounds in weak scalar expressions. These results are not meant to suggest noisy perception to be the sole or main determinant of these phenomena. Instead, this investigation's main contribution is conceptual and seeks to underline the pivotal role of systematic transmission perturbations of linguistic knowledge in language change while showing that such perturbations may stem from other sources, e.g., from learners' noisy perception.

[1] Depending on their formulation and the domain(s) they are proposed to apply to, biases may also interact. For instance, a domain-independent bias for simplicity may entail regularization but stand in conflict with mutual exclusivity.

Iterated learning under noisy perception

We model the transmission of linguistic knowledge as process of iterated learning (for recent overviews see Kirby et al. 2014, Tamariz & Kirby 2016). That is, as a repeated transfer of knowledge and behavior, such a language and its use, from one agent to another.

In the simplest case one may think of this process as involving a chain of parents and children. At the chain's top a parent produces linguistic data. This data is witnessed by the next agent in the chain, a child, who in turn acquires a language and behavior based on it. The child, now turned a proficient adult language user, then goes on to produce data for another child to learn from and so on. The learner's task is therefore to infer covert linguistic information, such as a language's grammar or when *red* holds true of an object, from observable language use. Importantly, there are multiple ways in which this process can induce change in a language. For instance, learning data may be sparse, speaker may make mistakes in production, or it could be that data incorrectly perceived by the learner. The fidelity by which a language or feature is transmitted against such perturbations therefore plays an important role in its emergence and stability across generations.

If linguistic behavior is kept constant across agents then the learner's task reduces to inferring a language $L \in \mathcal{L}$. A more general approach, used in our third case study, is to allow for variation in an agents production algorithm as well (Brochhagen et al. 2016). In these cases the learner's task is to perform a joint inference over types of linguistic behavior and lexical meaning. We call such a combination a type, $t \in T$.

More precisely, we follow Griffiths & Kalish (2007) in modeling language acquisition as a form of (iterated) Bayesian learning (Griffiths & Kalish 2007). Learning is hereby captured as a combination of the likelihood of a type generating the data witnessed by the learner with prior inductive biases, $P \in \Delta(T)$. This prior can be understood as a condensed codification of learning preferences. For example, learners may have a preference for simpler languages over ones with a more complex grammar, larger or more marked inventories, or cognitively taxing components (c.f. Feldman 2000, Chater & Vitányi 2003, Kirby et al. 2015). Crucially,

even weak biases can magnify and have striking effects on an evolving linguistic system. Experimental and mathematical investigations in iterated learning have therefore been taken to suggest that the linguistic structure evinced by the outcome of this process reflect learners inductive biases (Kirby et al. 2007; 2014).

The set of possible data a learner may be exposed to is represented by a set D. This set is made up of k-length sequences of the form $\langle\langle s_i, m_j\rangle, ..., \langle s_k, m_l\rangle\rangle$, where $s \in S$ is the observation of meaning s accompanied with an utterance $m \in M$. The parameter k therefore controls how much information learners have at their disposition to discern languages from one another. Generally, low k means that more languages will be compatible with the data, lowering the likelihood of a language being passed on faithfully. Conversely, inferring the language that generated the data, i.e., adopting the language used by a teacher, has a higher likelihood for larger sequences.

The likelihood $P(d|t_j)$ of datum d being produced by type t_j is given by a parent j's linguistic behavior in combination with the language it uses. Letting $P_{PA_j}(\cdot)$ and L_j stand type j's production algorithm and language:

$$P(d = \langle\langle s_1, m_1\rangle, \ldots, \langle s_k, m_k\rangle\rangle \mid t_j) = \prod_{i=1}^{k} P_{PA_j}(m_i \mid s_i; L_j)\,.$$

These components can be put together in a transmission matrix Q, where Q_{ij} is the probability that a learner acquires language i when learning from a user of language j

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

where $F(t_i \mid d)$ is the parametrized acquisition probability of t_i given datum d, obtained from the likelihood and prior:

$$F(t_i \mid d) \propto [P(t_i) P(d \mid t_i)]^l\,.$$

where l is a posterior parameter [TB: explain and then go over to noise].

Noisy transmission We denote the probability that the teacher (learner) observes state s_t (s_l) when the actual state is s_a as $P_N(s_t \mid s_a)$ ($P_N(s_l \mid s_a)$). The probability that s_a is the actual state when the learner observes s_l is therefore:

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

Accordingly, 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 $P_N(d_l \mid t)$. Then, the noise-perturbed mutation matrix is defined as:

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

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 she tries to infer the speaker's type. She takes what she observes a state to be as the actual state that the teacher has seen as well and infers 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. That is, such environmental factors can, in principle, induce transmission biases that look as if there was a cognitive bias in favor of a particular type, simply because that type better explains the noise.

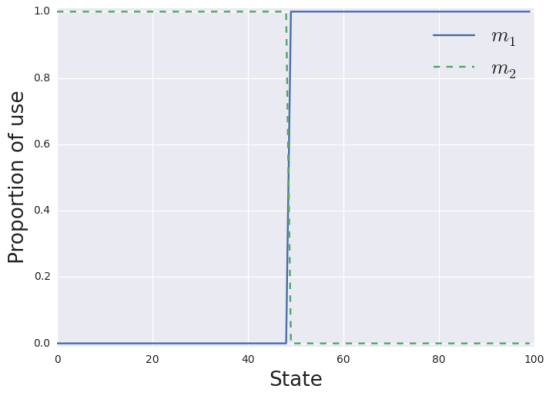
Applications

Vagueness

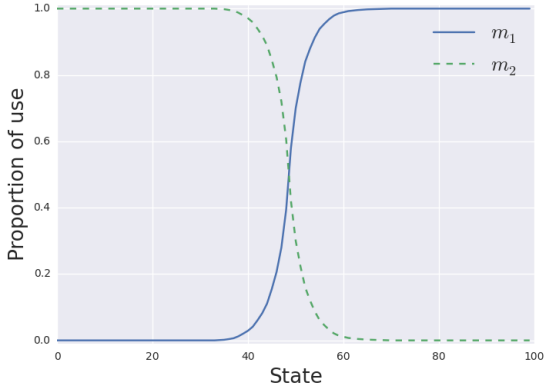
Main result. Noisy transmission perturbs initially crisp/clear linguistic distinctions, giving rise to vagueness. See Figure 1. Stabilization of the linguistic system around a particular threshold depends on functional considerations which are not modelled here but see Franke & Correia to appear.

Setup.

- $S = [0, 99]$
- $|M| = 2$
- There is one signaling behavior per threshold θ and one threshold per state, i.e., 100 types.
- $P(m_1|s,t) = 1$ iff $s \geq \theta_t$, otherwise $P(m_2|s,t) = 1$.
- $P(s_{\text{perceived}}|s_{\text{actual}})$ is the probability density of getting $s_{\text{perceived}}$ from Normal$(s_{\text{actual}}, \sigma)$
- Data generated by teachers is sampled without noise to get a representative sample. But actual likelihoods of producing the data used to compute Q are subjected to noise as above (as specified above)
- Learners are not aware of noise (as specified above)
- No replication.



(a) Initial "crisp" population



(b) Second "vague" generation

Figure 1: Noisy iterated learning with $\sigma = 0.4$, $k = 20$ and 100 sampled production sequences per parent (posterior sampling)

Deflation

Main result. Asymmetric and noisy perception can capture meaning deflation. See Figure 2.

- $S = [0, 99]$
- $|M| = 1$
- There is one type of signaling behavior per threshold θ and one threshold per state, i.e. 100 types.
- $P(m|s,t) = 1$ iff $s \geq \theta_t$, otherwise no message is sent. [TB: I'm pretty confident that adding some error-rate to this behavior wouldn't change the predictions. I left it deterministic for the time being]
- $P(s_{\text{perceived}}|s_{\text{actual}})$ is the probability density of getting $s_{\text{perceived}}$ from Normal$(s_{\text{actual}}, \sigma)$
- $P(\theta|d) \propto (\prod_{s \in d} P(m|s,\theta)) \times \text{Binom}(\text{successes} = k - |d|, \text{trials} = k, \text{succ.prob} = \sum_{s'=0}^{\theta-1} P(s'))$, where the latter is the probability of a type not reporting $k - |d|$ events for a total of k events.

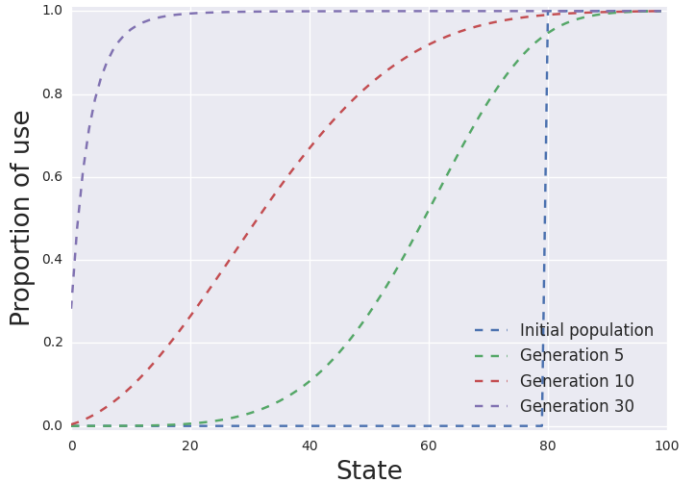


Figure 2: Noisy iterated learning with $\sigma = 0.4$, $k = 30$ and 300 sampled production sequences per parent (posterior sampling)

- Data generated by teachers is sampled without noise to get a representative sample. But actual likelihoods of producing the data used to compute Q are subjected to noise as above (as specified above)
- Learners are not aware of noise (as specified above)
- No replication. [TB: Not sure how this would work anyway. The higher θ, the less a type communicates. If that's a communicative failure, then these types are even more dispreferred than with only learning. If it's not, then we have the same fitness for each type]

Quantifiers

Main result. Noisy perception of states can mimic cognitive biases. In this case, a bias towards simplicity (no upper-bounds) as analyzed in Brochhagen et al. (2016). Pragmatic inferences stabilize in population as byproduct of noise, as shown in Figure 3. Little changes for small increases of k (10 or 15) or l (e.g. 10 instead of posterior sampling).

- $S = \{s_{\exists\neg\forall}, s_{\forall}\}$
- $|M| = 2$
- There are two lexica, one upper-bounded and one lacking upper-bound, and two signaling behaviors, literal and gricean, for a total of four lexica
- $P(m|s,t)$ is soft-maximizing literal or gricean behavior with α as exponent, using a type's lexicon – as in our other setup [TB: Alternatively, we could go for simple Boolean behavior to keep everything uniform]
- $P(s_{\exists\neg\forall}|s_{\forall}) = \delta$, $P(s_{\forall}|s_{\exists\neg\forall}) = \varepsilon$)
- Learners are not aware of noise (as specified above)

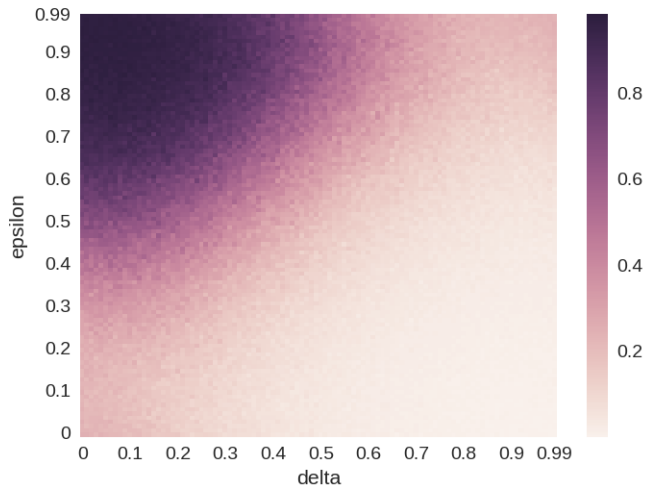


Figure 3: Influence of noise-parameters on the mean proportion of Gricean players with a lexicon lacking an upper-bound after 20 generations with $k = 5, \lambda = 20, \alpha = 1$ and 10 sampled production sequences per parent (posterior sampling).

- No replication

Discussion

[TB: To be specified. Parts can be taken from our previous draft]

Conclusion

Acknowledgments

References

Brochhagen, T., Franke, M., & van Rooij, R. (2016). Learning biases may prevent lexicalization of pragmatic inferences: a case study combining iterated (bayesian) learning and functional selection. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2081–2086). Austin, TX: Cognitive Science Society.

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

Clark, E. V. (2009). Lexical meaning. In E. L. Bavin (Ed.), *The cambridge handbook of child language* (pp. 283–300). Cambridge University Press. doi: 10.1017/cbo9780511576164.016

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

Franke, M., & Correia, J. P. (to appear). Vagueness and imprecise imitation in signalling games. *British Journal for the Philosophy of Science*.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Hudson Kam, C. L., & Newport, E. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195. doi: 10.1207/s15473341lld0102_3

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

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114. doi: 10.1016/j.conb.2014.07.014

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

Merriman, W. E., & Bowman, L. L. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the Society for Research in Child Development*, *54*(3/4), i-129. doi: 10.2307/1166130

O'Connor, C. (2015). Evolving to generalize: Trading precision for speed. *The British Journal for the Philosophy of Science*. doi: 10.1093/bjps/axv038

Smith, K. (2011). Learning bias, cultural evolution of language, and the biological evolution of the language faculty. *Human Biology*, *83*(2), 261–278. doi: 10.3378/027.083.0207

Tamariz, M., & Kirby, S. (2016). The cultural evolution of language. *Current Opinion in Psychology*, *8*, 37–43.