SOCIAL INTERACTION INFLUENCES THE EVOLUTION OF COGNITIVE BIASES FOR LANGUAGE

SEÁN G. ROBERTS

Language and Cognition Department, Max Planck Institute for Psycholinguistics
Nijmegen 6525 XD, The Netherlands
sean.roberts@mpi.nl

BILL THOMPSON, KENNY SMITH

Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, EH8 9AD, UK
bill@ling.ed.ac.uk, kenny@ling.ed.ac.uk

Models of cultural evolution demonstrate that the link between individual biases and population-level phenomena can be obscured by the process of cultural transmission (Kirby, Dowman, & Griffiths, 2007). However, recent extensions to these models predict that linguistic diversity will not emerge and that learners should evolve to expect little linguistic variation in their input (Smith & Thompson, 2012). We demonstrate that this result derives from assumptions that privilege certain kinds of social interaction by exploring a range of alternative social models. We find several evolutionary routes to linguistic diversity, and show that social interaction not only influences the kinds of biases which could evolve to support language, but also the effects those biases have on a linguistic system. Given the same starting situation, the evolution of biases for language learning and the distribution of linguistic variation are affected by the kinds of social interaction that a population privileges.

1. Introduction

The interaction between individual cognitive biases for learning, the amount of linguistic diversity in a population and how that diversity is used to support social interactions forms a complex, adaptive system. It's clear that there is a genetic basis for the ability to learn a language, but recent studies have demonstrated that the structures and distributions of linguistic features are also affected by cultural transmission (Kirby, Cornish, & Smith, 2008; Dunn, Greenhill, Levinson, & Gray, 2011). This may obscure the relation between properties of individual learners and population-level cultural phenomena (Kirby et al., 2007), making it difficult to infer from linguistic structures the existence of isomorphic cognitive biases. Consequently, in order to make predictions about the cognitive underpinnings of language, we must also understand the interactions between individual cognition and cultural evolution in populations. Computational models have ad-

dressed this issue (e.g. Smith et al., 2003; Nowak et al., 2001; Niyogi, 2006; Smith & Thompson, 2012, hereafter S&T). However, many assume a mature system of communication is one where users have converged on monadic conventions, or that the only relevant factors are properties of individual cognition.

Here we explore the consequences of relaxing that premise in favour of alternatives that better reflect the diversity of human social interaction. In particular we are interested in scenarios that embody a high degree of socially conditioned linguistic variation. Our focus is bilingualism, which we view as a socially constructed property of individuals: bilinguals learn to condition linguistic structures on social variables determined by the constitution of a population. Scenarios such as this may appear to be at odds with evolutionary reasoning. If communicative coordination is associated with a fitness payoff, we might expect populations to converge on monadic conventions, supported by innate biases to expect little variation. Indeed, this expectation is borne out by existing models of cultural evolution (S&T). However, in reality most humans are exposed to a large amount of linguistic diversity and acquire communicative competence in multiple languages. Also, while it may be traditional to view second language acquisition as a demanding task, empirical evidence shows that children are adept at learning multiple languages simultaneously (Byers-Heinlein & Werker, 2009).We extend S&T's model to explore a range of social contexts, including ones that privilege monolingualism, bilingualism, linguistic similarity (parity) and linguistic difference (exogamy). We find various conditions that lead to the evolution of biases supporting bilingualism. We show that assumptions about social interaction not only influence the *kinds* of cognitive biases that could evolve to support language, but also the nature of their effects on population-level culture.

2. Model definition

We adopt as our framework the iterated learning model whereby learners acquire a behaviour by observing similar behaviours in others who acquired their behaviours in the same way (Smith et al., 2003). We adopt and extend the Bayesian model developed by Burkett and Griffiths (2010) and extended by S&T in which learners can learn multiple languages from multiple teachers. The model involves discrete generations, each with a finite population of N agents who receive input from a previous generation, infer a hypothesis about how that data was produced, and then use that hypothesis to produce data for a subsequent generation.

Learning proceeds as follows: on the basis of a set of observations $d = \{d_1, d_2, \ldots, d_b\}$, each learner infers the ambient frequencies of two possible languages l_0 & l_1, and induces an hypothesis, $h = (P(l_0), P(l_1))$, where $P(l_i)$ represents the learner's estimate of the frequency of language l_i. As shorthand we can characterise h by $h_0 = P(l_0)$, since $P(l_1) = 1 - P(l_0)$ (the mean distribution of h in the population is labelled θ). Learners make inferences in a Bayesian rational framework, using Bayes' rule to compute the posterior distri-

bution $P(h|d) \sim P(d|h)P(h)$. The likelihood computations are straightforward: observations d are made up of b interchangeable utterances, each of which can take one of two forms, u_0 & u_1. These forms are typically diagnostic of one or the other language, so that $P(u_i|l_i) = 1-\varepsilon$ and $P(u_i|l_{j\neq i}) = \varepsilon$, where ε is small and represents errors in production. The likelihood function is simply the product of the probabilities for each utterance: $P(d|h) = \prod_{d_i} P(d_i|l_0)h + P(d_i|l_1)1-h$. Productions are based on these likelihoods: when a learner produces an utterance for the next generation, it samples a language from its hypothesis, and samples an utterance from that language according to the function above.

Each learner has a prior bias with two properties: One favours the use of each language in a particular proportion (G_0), and one controls the amount of variation they expect (α). During inference, hypotheses h are drawn from a Dirichlet process prior with base distribution G_0 and concentration parameter α. Computationally, we implement inference using a Gibbs sampler based on the Chinese restaurant process representation of the DP. G_0 specifies a distribution over the two possible language types. The concentration parameter α is a positive real, and regulates the influence of G_0 during inductive inference: as $\alpha \to \infty$, learners will induce hypotheses strongly determined by their prior preferences, so that $h \approx G_0$; as $\alpha \to 0$, h is determined largely by the learner's experiences. In our context we can interpret α as determining a learners expectations about linguistic diversity. High α leads learners to expect a wide distribution of languages in the population. Low α leads learners to expect homogeneity: linguistic variation is discounted in favour of monolingual hypotheses.

Learners inherit their prior biases genetically from 'parents' in the previous generation. The prior bias can mutate with probability μ, meaning that the distribution of priors evolves by natural selection. Reproductive success depends on the agents hypothesis and the fitness function. We test several fitness functions based on different conceptions of communicative success and social prestige.

S&T find that biological evolution via natural selection for communicative coordination leads to the emergence of low α. S&T's model rewards learners who converge on a common language. Assumptions of this kind are common in such models, and represent a sensible first pass at capturing the benefits of coordination in communication. However, we show below that this fitness metric directly privileges monolingualism, and so leads to linguistic homogeneity. In contrast, we show that populations of individuals with the same prior biases over languages, but with different fitness metrics, can lead to linguistic diversity.

Reproductive success is linked to the relationship between agents' hypotheses. We define this relationship using metric space notation $(\mathscr{H}, \rho)$, with $\mathscr{H} = \{(i, 1-i) \colon (i \in \mathbb{R}), (0 \leq i \leq 1))\}$ our set of possible hypotheses and ρ a metric on $\mathscr{H}$ which determines the fitness payoff between any two $h, h' \in \mathscr{H}$ as $\rho(h, h')$, where ρ reflects our various theoretical assumptions. The total fitness payoff for an agent is the sum of payoffs for the whole population. Below we define 5 metrics, each

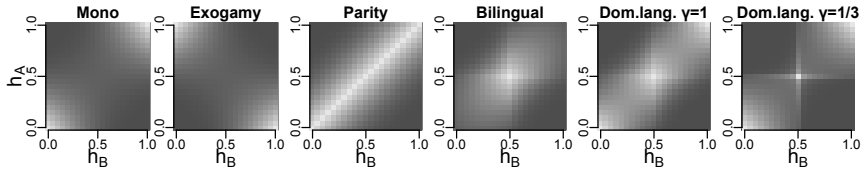

Figure 1. How the relationship between the hypotheses of two agents A and B maps onto fitness payoff for different metrics. Lighter shades represent better payoffs. This is part of the model input.

of which constructs a fitness landscape depicted in figure 1.

Type 1: Monolingual S&T's regime simply rewards convergence: fitness payoff is linked to communicative success, defined as being proportional to the probability that during a given encounter two learners use the same language. For S&T, then $\rho_m(h,h') = h \cdot h'$. We can visualise the fitness landscape defined by this metric as a heat map (see figure 1). As this shows, this assumptions strongly privileges monolingualism: the fittest pair of learners both speak only language l_0 or only language l_1 (i.e. $h_0 \approx h'_0 \approx 0$ and $h_0 \approx h'_0 \approx 1$), since these learners will tend overwhelmingly to converge on the same language. For some aspects of language, this kind of assumption is natural: coordination is at the heart of successful communication in many domains. However, the assumption means that it's impossible to be fully competent in both languages. This means that an agent with $h = 0.5$ is an analogue of a 'semilingual' individual who does not have native competence in any language (Bloomfield, 1927).This view of competence has been criticised, and is not wholly supported by the linguistic evidence (Martin-Jones & Romaine, 1986). More generally, the monolingual assumption may be appropriate for some scenarios, but does not reflect the diversity of human social interaction: many communities and societies privilege bilingualism and linguistic diversity, and in those cases fitness payoffs should reflect those systems.

Type 2: Bilingual Our first alternative is to explicitly privilege bilingualism: here the biggest fitness payoff goes to a pair of learners who both have command of both languages in equal proportion. Prestigious bilingualism is attested in many communities, and is often linked with the power to communicate between groups (De Mejía, 2002). Formally, we define our metric to be: $\rho_b(h,h') = 2(h \cdot h')(1 - |h_0 - 0.5|)(1 - |h'_0 - 0.5|)$. Here we are simply weighting the fitness payoff by the learners' combined distance from the entirely bilingual state ($h_0 = h'_0 = 0.5$). As in the 'monolingual' case, it pays to converge, but here there is only one hypothesis that yields the highest fitness payoff.

Type 3: Parity The monolingual and Bilingual regimes each privilege a particular subset of hypotheses on theoretical grounds, and so make reasonably transparent evolutionary predictions ($\rho_m \rightarrow$ low α, $\rho_b \rightarrow$ high α) . We can relax this premise and focus only on coordination by rewarding arbitrary parity. Under this regime maximum fitness payoff requires only that learners share a hypothesis:

any hypothesis is as good as any other, so long as the learners' hypotheses are matched. Here our metric is simply: $\rho_p(h,h') = 1 - |h_0 - h'_0|$.. This removes any obvious bias towards homogeneity or heterogeneity in the linguistic community. Human learners are highly sensitive to the distribution of linguistic variants they experience, and often try to match the behaviour of their interlocutor (Smith & Wonnacott, 2010).

Type 4: Linguistic Exogamy Some societies restrict marriage to members of different linguistic communities (e.g. Jackson, 1983), and these communities are often multilingual (Hill, 1978). In simple terms, learners receive higher fitness payoffs from interactions with linguistically foreign individuals, which is the inverse of the monolingual function: $\rho_{ex}(h,h') = 1 - (h \cdot h')$. As figure 1 shows, ρ_{ex} privileges interactions between maximally divergent hypotheses. However, the hypothesis with the best unilateral payoff (0.5) is not the optimal for an individual (0 or 1). The evolutionary predictions for this regime are unclear: we might expect populations to eventually contain monolingual speakers of both languages in roughly equal number. For this to happen in well-mixed cultural populations, learners may require a strong expectation for linguistic homogeneity (low values of α). However, previous models suggest that populations of learners with homogeneity biases tend to end up speaking only one language.

Type 5: Dominant Language Finally, we model the situation where learners can know a second language without detriment to their knowledge of their first language. The 'dominant language' metric assumes that fitness is proportional to communicative success, but a speaker always understands its dominant language, and understands its non-dominant language in proportion to the balance of its hypothesis. $\rho_d(h,h') = 1$ if $h > 0.5$ and $h' > 0.5$; $\rho_d(h,h') = 1$ if $h <= 0.5$ and $h' <= 0.5$; otherwise $\rho_d(h,h') = |h - h'|^{\gamma}$. This means that a learner will always get the maximum payoff for interacting with another learner who has a hypothesis with a tendency towards the same language. That is, they always understand their 'stronger' language. However, if their partner has a hypothesis with a tendency towards the opposite language, then the payoff is related to the difference between the hypotheses according to γ. When $\gamma = 1$, then the relationship is linear. Lower values of γ make the relationship exponential. The variable γ, therefore, specifies how much competence is required in a second language to receive a good fitness payoff from interacting with any other speaker. An individual with a hypothesis in the middle of the range can receive the maximum payoff from all other hypotheses. However, as γ decreases, the range where this is effective becomes increasingly narrower, making it a fragile state.

3. Results

We ran agent-based simulations to explore the co-evolution of cognitive biases (α) and linguistic systems (h) under our social models. In these simulations: $N = 100$; $\varepsilon = .05$; each learner is exposed to $b = 4$ utterances; Gibbs sampling was run for

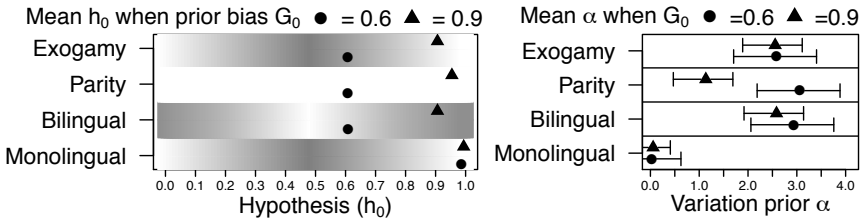

Figure 2. Emergent values of h_0 (left) and α (right) for different fitness metrics when the prior bias over hypotheses G_0 is weak (0.6) or strong (0.9). The background of the h_0 graph is shaded to indicate the maximum fitness payoff to an individual given that hypothesis (lighter shades = better payoff).

5 cycles at each learning event; simulations were run for 500 generations (α and h_0 converged after about 200). α mutates with probability $\mu = 0.01$. If a mutation occurs, α is drawn from a Gaussian distribution with the parent α as its mean and variance $\sigma^2 = 0.1$. We explored two settings for the prior bias over languages: a weak bias for l_0 ($G_0 = (0.6, 0.4)$) and a strong bias for l_0($G_0 = (0.9, 0.1)$).

The results are shown in figure 2. As in S&T, under the Monolingual metric α remains low (agents expect little diversity) and h_0 reflects an amplified prior over languages G_0 (agents use only l_0), regardless of the strength of the prior. h_0 also reflects the optimal fitness payoff. However, the alternative metrics behave differently. The Bilingual metric leads to high α (agents expect high diversity) and h_0 converges to G_0. h_0 does not reflect the optimal fitness payoff. The results for the Exogamy metric are the same, even though the payoff landscape is very different. Under the Parity metric, the value of α is high, but affected by the prior over languages (stronger bias leads to lower α). h_0 converges to the prior over languages, though is slightly amplified under the stronger prior. There was more variation in the emergent values of α under the alternative metrics than the monolingual metrics.

Figure 3 shows the results of manipulating the ease of comprehending a second language. As γ increases, there is a qualitative shift in the results of the simulations. With $\gamma > 0.7$ (comprehending a second language is easy), high α evolves (a 'bilingual' expectation) and the distribution of languages converges to the prior. However, with $\gamma < 0.7$ (comprehending a second language is harder), low α evolves and the distribution of languages is exaggerated (l_1 comes to dominate, non-convergence).

4. Conclusion

Our model demonstrated that linguistic diversity is dependent on individual cognitive biases, individual learning, social interaction and cultural evolution.

Under the monolingual social model, there are only two hypotheses that give

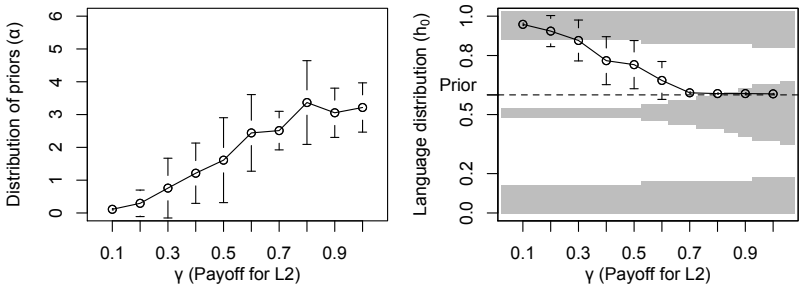

Figure 3. Results under the dominant language metric. The prior over hypotheses was 0.6. Gray areas indicate hypotheses that allow bigger fitness payoffs. There is a qualitative change around $\gamma = 0.7$.

unilateral optimal fitness (h=0 and h=1). This makes it easier for individuals to converge on a hypothesis, meaning there is less variation. This allows a low α to emerge, which leads to the agents' inference being more influenced by the data. In this case, the distribution of hypotheses climbs the fitness landscape until everyone has the same hypothesis. The alternative types of social interaction favour diversity (some transparently - e.g. bilingualism - others indirectly - e.g. exogamy). These regimes will always lead to learners that preserve diversity (high α). Therefore, the agent's inference is more affected by their prior. The resulting distribution reflects the prior, even when this does not reflect the optimal fitness payoff. That is, the social interactions select against rich-get-richer type learning that kills variation (even when a heavily biased prior means there isn't much to be gained, in fitness terms, by maintaining diversity), which is normally assumed in these kinds of evolutionary models.

Under the 'dominant language' model, we manipulated the amount of competence in a second language required to receive a fitness payoff. This variable interpolated between the two types of result. Even when moderately high competence was required for a fitness payoff, bilingualism emerged and learners evolved to expect variation in their input.

These results differ from the results of other models. First, bilingual biases can evolve and linguistic diversity can be stable. Secondly, our model shows that the constraints of social interaction, as well as individual learning biases and cultural evolution, can shape the emergent properties of linguistic populations. This suggests that a full explanation of language evolution must involve how language is used in interaction to shape social relationships. Future work could allow the relationship between social interaction and fitness to change over time.

Acknowledgments

SR was partly supported by an ESRC grant ES/G010277/1. BT was supported by an EPSRC studentship. Thanks to Simon Kirby and Amy Perfors for comments.

References

Bloomfield, L. (1927). Literate and illiterate speech. *Am. Speech*, *2*(10), 432–439.

Burkett, D., & Griffiths, T. (2010). Iterated learning of multiple languages from multiple teachers. In A. Smith, M. Schouwstra, B. de Boer, & K. Smith (Eds.), *The evolution of language: Proceedings of EvoLang 2010* (p. 58-65). World Scientific.

Byers-Heinlein, K., & Werker, J. F. (2009). Monolingual, bilingual, trilingual: infants' language experience influences the development of a word-learning heuristic. *Developmental Science*, *12*(5), 815-823.

De Mejía, A. (2002). *Power, prestige, and bilingualism: International perspectives on elite bilingual education* (Vol. 35). Multilingual Matters Ltd.

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*(7345), 79-82.

Hill, J. (1978). Language contact systems and human adaptations. *Journal of Anthropological Research*, *34*(1), 1–26.

Jackson, J. (1983). *The fish people: linguistic exogamy and tukanoan identity in northwest amazonia.* Cambridge University Press.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS*, *105*(31), 10681-10686.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *PNAS*, *104*(12), 5241-5245.

Martin-Jones, M., & Romaine, S. (1986). Semilingualism: A half-baked theory of communicative competence. *Applied Linguistics*, *7*(1), 26–38.

Niyogi, P. (2006). *The computational nature of language learning and evolution.* MIT press Cambridge, MA.

Nowak, M., Komarova, N., & Niyogi, P. (2001). Evolution of universal grammar. *Science*, *291*(5501), 114–118.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*(4), 371–386.

Smith, K., & Thompson, B. (2012). Iterated learning in populations: Learning and evolving expectations about linguistic homogeneity. In T. C. Scott-Phillips, M. Tamariz, E. A. Cartmill, & J. R. Hurford (Eds.), *The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9)* (p. 227-233). World Scientific.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449.