Bayesian Language Games

Bas Cornelissen

August 9, 2017

Few studies, it seems, have tried to bridge the gap between iterated learning and naming games. In this chapter I argue that Bayesian models of iterated learning can naturally be connected to naming games in the form of a new, Bayesian language game. This model of cultural evolution gives rise to a stable, lineage-specific language that reflects innate biases, but not faithfully so. With a proposed population structure, the game interpolates between an iterated learning model and a naming game and moreover incorporates a wide range of strategies. The model, in short, brings a unified perspective on two agent-based modelling paradigms and addresses some of the desiderata formulated in chapter ??.

Naming games and iterated learning are the central traditions of agent-based modelling of language evolution (Smith 2014; Grifoni, Ulizia, and Ferri 2016; Jaeger et al. 2009). In Jaeger et al. (2009) they even form the axes defining the space of agent-based simulations: naming games horizontally and iterated learning vertically. But the coordinate system looks rather empty. Although horizontal and vertical models have often been combined, the interaction between the two traditions seems extremely limited. A naive citation count makes this disturbingly clear.[1] A paper by Luc Steels (2016) with the inclusive title *Agent-based models for the emergence and evolution of languages* cites a grand total of zero iterated learning papers. Some years earlier, Steels (2011) scores 3/120 in a review called *Modelling the cultural evolution of language*. At least the traditions meet in mutual neglect: *The cultural evolution of language* (Tamariz and Kirby 2016) scores[2] 1/73 and Kirby, Griffiths, and Smith (2014) 2/60. But then again, the latter paper is called *Iterated learning and the evolution of language*.

A case of incommensurable paradigms? In this chapter, I will argue the opposite. Far from being incommensurable, Bayesian models of iterated learning and naming games naturally meet in a model I will call the *Bayesian language game*. This will be the extension of a Bayesian *naming* game, to be introduced first. Several closely related models can be found in the literature, but, to the best of my knowledge, have never been used to connect the two traditions. I review related work at the end of this chapter and would like to start where we left off in the previous chapter: the convergence proof of the naming game.

[1] I counted references to papers coming from the group of either Kirby (IL) or Steels (NG). All serious papers in either tradition cite extensively from the work of the respective groups.

[2] Admittedly, Tamariz and Kirby (2016) *does* cite the experimental semiotics literature. But then again, it does *not* include Steels under the heading 'naming games' (table 1), under which we do find some papers from Kirby's group.

The Bayesian naming game

The *Bayesian naming game* can be seen as an extension of the naming game studied by De Vylder and Tuyls (2006). We make similar simplifications and assume all N agents already know words $w_1, \dots, w_K$, and need to negotiate which of these words to use for the single object at hand. Each agent has an internal language θ, a distribution over the K words, based on which it produces words x using some production strategy. I use this naming game interpretation throughout the chapter, but it should be noted that the language θ has also be interpreted as a distribution over various linguistic variants. As Reali and Griffiths (2010) explain, "learning a language involves keeping track of the frequencies of variants of a linguistic form at various levels of representation, including phonology, morphology, and syntax". Ferdinand and Zuidema (2009) represent languages in a similar fashion.

The queue-learners in De Vylder and Tuyls (2006) 'learned' their language by computing the relative frequencies of observed words. Here, the Bayesian naming game takes a different turn. Following Bayesian iterated learning models, it assumes that agents are Bayesian reasoners updating their language using Bayes' rule. In other words, they use Bayesian updating as an *alignment strategy*. After observing utterances x, the agent infers the posterior distribution over languages

$$p(\theta \mid \mathrm{x}) \propto p(\mathrm{x} \mid \theta) \cdot p(\theta), \tag{1}$$

and *samples* a language accordingly (other strategies are discussed later). Just like iterated learning models, the biases of the agent enter the model explicitly in the form of a prior $p(\theta)$. But there is an important difference. Agents engage in multiple encounters and every time a hearer interact, its *beliefs* about the language it should use, have to be updated. The the posterior beliefs $p(\theta \mid \mathrm{x}_t)$ inferred during interaction t should thus serve as the prior beliefs $p_{t+1}(\theta)$ in round $t+1$. Strictly speaking, I use 'prior' as a technical term for the distribution $p_t(\theta)$. It can be *interpreted* as the 'beliefs' of the agent. In the first round, the prior encodes the *(innate) biases*, but later in the game, it encodes both innate biases and *past experience*. For simplicity, I consistently speak of *innate* biases, but a more more precise reading would be "everything that the learner brings to the task independent of the data" (Kirby, Smith, and Brighton 2004). The distinction becomes relevant in chapter **??**.

In general terms, round t in the Bayesian naming game has the following script.

- A hearer H and speaker S are randomly selected from the population.
- The speaker samples a language θ_t from her prior distribution $p_{S,t}(\theta)$. This is the posterior $p_S(\theta \mid \mathrm{x}_{t'})$ inferred during the last interaction t' she engaged in as a hearer. The selected language defines a distribution $p(x \mid \theta_t)$ over words. She samples b words $\mathrm{x}_t = (x_1, \dots, x_b)$ from that distribution and communicates those to the hearer.
- The hearer updates his beliefs $p_{H,t+1}(\theta) := p_H(\theta \mid \mathrm{x}_t)$ to the posterior, which is proportional to $p(\mathrm{x}_t \mid \theta) \cdot p_{H,t}(\theta)$. All other agents A, including the speaker, maintain their current beliefs: $p_{A,t+1}(\theta) := p_{A,t}(\theta)$.

This script outlines a general framework for Bayesian naming games. This chapter discusses one specific instantiation, a *Dirichlet-categorical naming game*, but it should be

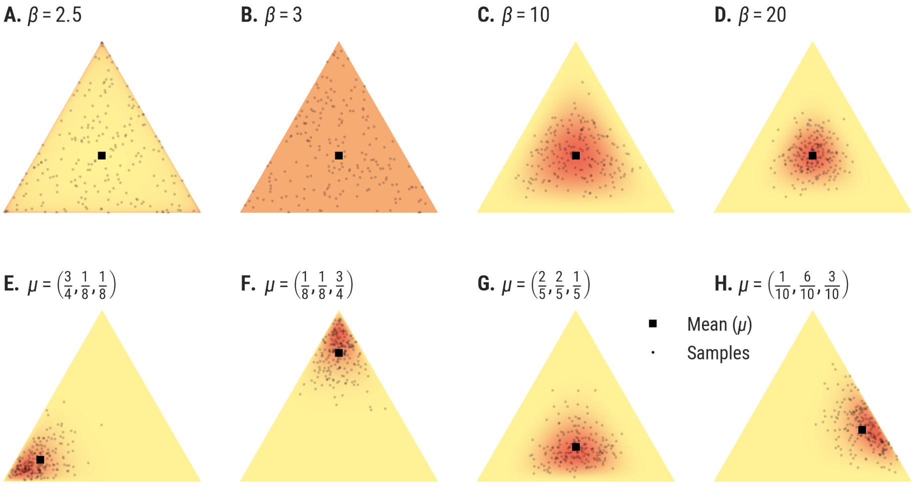


Figure 1 The Dirichlet distribution for various parameter settings. The Dirichlet can be parametrised by a point μ in the simplex and a scalar β. The mean of the distribution is determined by μ and β influences the variance. The first row (A–D) demonstrates the effect of β while fixing $\mu = (1/3, 1/3, 1/3)$; the second row (E–H) the effect of μ while keeping $\beta = 15$ fixed. Note that with $\beta \cdot \mu = (1, 1, 1)$ (subfigure B) one gets a uniform distribution over the simplex.

FIG02 Figure produced using code by Thomas Boggs at gist.github.com/tboggs/8778945

noted that the proposed framework is more general.

On a practical note, a mathematical development of the model is included in appendix **??**. The treatment in the main text is informal and to keep the notation uncluttered, deliberately sloppy: I do not decorate variables with the corresponding agent, time or index, unless strictly necessary.

The Dirichlet-categorical naming game

Following De Vylder and Tuyls (2006), the internal language θ of an agent is a categorical distributions over words, but what should the prior distribution over languages look like? The obvious candidate is the *Dirichlet distribution*, because it is the *conjugate prior* of the categorical. This means that the posterior distribution has the same parametric form as the prior, i.e. the posterior will also be a Dirichlet. If the prior at time t is parametrised by some parameter vector α_t, the *hyperparameter*, then posterior inference amounts to determining the new hyperparameter α_{t+1}, which can often be done analytically. So in terms of the Bayesian naming game, hearers only need to change the hyperparameter after an interaction to update their beliefs.

The Dirichlet distribution is defined over the *entire* simplex — not just over the finite subset Δ_Q, as with the multinomial in De Vylder and Tuyls (2006) — and thus assigns a probability to *every* language, *every* distribution over K words. It is parametrised by a vector $\alpha = (\alpha_1, \ldots, \alpha_K)$, but it is often convenient to split this into a normalised vector μ and a scalar parameter $\beta > 0$ and use $\alpha = \beta \cdot \mu$. The vector μ, since it sums to 1, lies in the simplex and determines the mean of the distribution. β is a kind of inverse variance, with larger values of β resulting in smaller variance. Figure 1 illustrates different parameterisations oft

With this conjugate prior, posterior inference amounts to updating the hyperparameter α — but how? Suppose the hearer receives words $x_t = (x_1, \ldots, x_b)$ — here b is the *bottleneck size* — and let $c_t = (c_1, \ldots, c_K)$ denote the corresponding vector of counts,

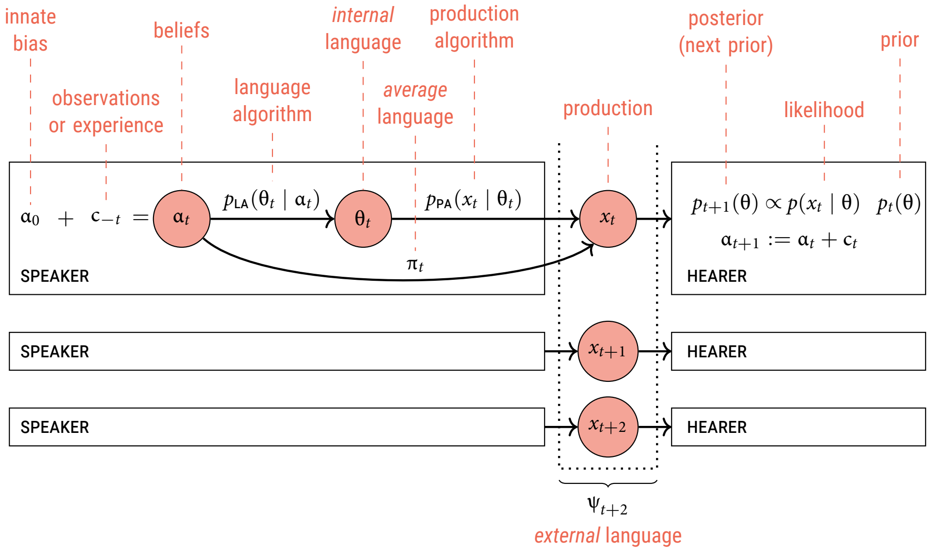


FIGURE 2 Illustration of the Bayesian naming game with all relevant concepts. See main text for details.

such that c_k is the number of k's in x. If α_t is the previous hyperparameter, then the posterior of the hearer is

$$p(\theta \mid \mathrm{x}_t) = \text{Dirichlet}(\theta \mid \alpha_t + \mathrm{c}_t), \tag{2}$$

and the belief update amounts nothing more than $\alpha_{t+1} := \alpha_t + \mathrm{c}_t$. The *Dirichlet-categorical* (DC) naming game we have defined can now be summarised as

$$\text{SPEAKER} \quad \begin{cases} \theta_t \mid \alpha_{t-1} & \sim \text{Dirichlet}\big(\alpha_{t-1}\big) \\ x_i \mid \theta_t & \sim \text{Categorical}\big(\theta_t\big), \quad i = 1, \ldots, b. \end{cases} \tag{3}$$

$$\text{HEARER} \quad \alpha_{t+1} := \alpha_t + \mathrm{c}_t \tag{4}$$

Note that the speaker still *samples* both languages and productions. Other strategies are discussed later.

PRIORS, BELIEFS, INNATE BIASES, AND PAST EXPERIENCE An additional benefit of the DC naming game is that it transparently represents several important concepts, which is visualized in figure 2. First of all, we can separate beliefs from the prior. I will call the hyperparameter α_t the *beliefs*, since those are updated after every interaction, and the distribution Dirichlet(α_t) the *prior*. Recall that the prior at $t = 0$ encodes the innate biases, but in later encounters captures past experience as well. Unraveling successive updates of the beliefs brings this fact to the fore:

$$\alpha_{t+1} \;=\; \alpha_0 + \mathrm{c}_1 + \mathrm{c}_2 + \cdots + \mathrm{c}_t \;=\; \alpha_0 + \mathrm{c}_{-t}, \tag{5}$$

where c_{-t} is the vector of counts of all observations before and including round t.[3] That is, c_{-t} captures all *past experience*, whereas α_0 captures "everything that the learner brings to the task independent of the data" (Kirby, Smith, and Brighton 2004). Equation 5 thus makes explicit that the beliefs in round t are the sum of innate biases α_0 and

[3] We set $c_t = 0$ if the hearer did not participate in round t.

past experience $\mathbf{c}_{-t}$. It also shows that the innate biases act as so called *pseudo-counts* of *pseudo-observations*. It is as if a newborn agent has already observed utterances with word counts given by α_0, before engaging in any interactions. The point of Griffiths and Kalish (2007), that the *prior* should not be seen as the innate bias, is even more to the point here. Alternatively, it can indeed be seen to regulate the amount of evidence needed to adopt a certain language. If α_t for example contains nothing but 20 observations of word w_3, the agent will need a lot of evidence before it will prefer to choose another word — irrespective of whether it concerns *pseudo* or *actual* observations.

INTERNAL, EXTERNAL, EXPECTED AND AGGREGATE LANGUAGES *Languages*, here, are always distributions over words $w_1, \dots, w_K$, but care should be taken to distinguish several different distributions. First of all, in every round the speaker chooses a language θ_t from which she generates words. This is the *internal* language (I-language). It is distinct from the *external* language (E-language) which consists of all utterances. The external language can be estimated with the relative frequencies of the words

$$\psi_t := \frac{1}{b \cdot t} \cdot \sum_{\tau=1}^{t} c_\tau \tag{6}$$

In that way the external language ψ_t also becomes a distribution over words.

Finally, we need to introduce the *expected* language and the *aggregate* language for practical reasons. Every agent entertains a full distribution over internal languages (the Dirichlet) and we cannot know beforehand which language it will use in a particular round. To probe the internal state of an agent we turn to its *expected language* $\pi_t = (\pi_1, \dots, \pi_K)$, the language it is expected to use in the next round. That is, we consider the marginal distribution

$$\pi_k := \; p\big(x = w_k \mid \alpha_{t-1}\big) = \int_\Delta p(x, \theta \mid \alpha_{t-1}) \; d\theta \;=\; \hat{\alpha}_k, \tag{7}$$

where $\hat{\alpha}_k$ is the k'th entry of $\hat{\alpha}_t = \alpha_t / \sum_j \alpha_j^{(t)}$, which is the normalised version of α_t. The conjugacy of the Dirichlet gives this marginal distribution a simple form: the expected language π_t is proportional to the beliefs α_t of the agent. Note that the expected language at $t = 0$ is the language completely determined by the innate biases: $\pi_0 = \hat{\alpha}_0$. Accordingly, we often identify the bias with $\hat{\alpha}_0$. Finally, the average of the expected languages of all agents in the population is called the *aggregate language*

$$\bar{\pi}_t := \frac{1}{N} \sum_{i=1}^{N} \pi_{A_i, t}, \tag{8}$$

consistent with our terminology in chapter **??**.

Phenomenology of the DC naming game

THREE-STAGE EVOLUTION Figure 3 show two typical runs of the Dirichlet-categorical naming game. Subfigure A shows the expected languages of all the agents (π_A, thin

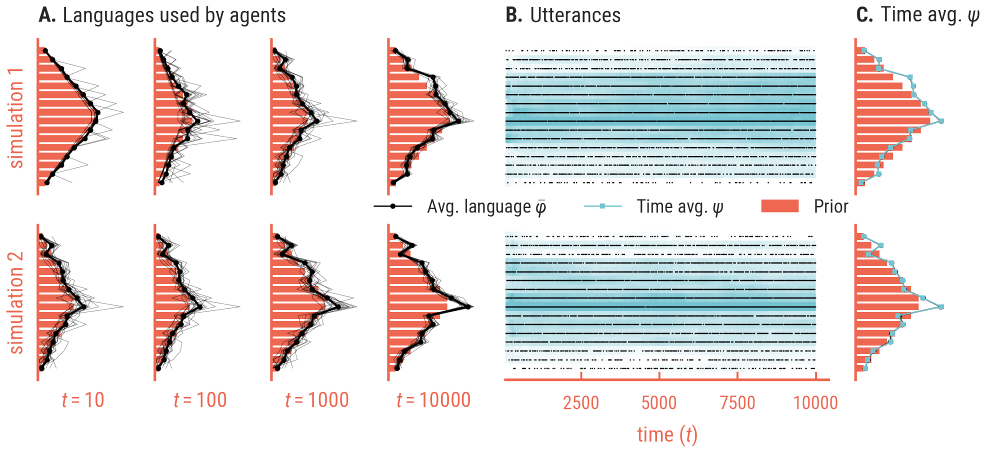


FIGURE 3 Two runs of the Bayesian Naming Game. A. The distributions of all agents (thin black lines) first diverge but eventually stabilise. They always reflect the prior (orange), B. Utterances (dots) at every time plotted over a moving average of 2000 time steps. C. The relative frequency of all utterances reflects the language adopted in the population. See main text for more details.

FIG05 $K = 16, N = 15, b = 1, \beta = 18, \eta = \zeta = 1, \gamma = \infty$

lines) and the aggregate language ($\bar{\pi}$, thick lines) after 10, 100, 1000 and 10000 encounters. The cultural evolution can be divided in three stages, which I will metaphorically call 'infancy', 'puberty' and 'adulthood'. In 'infancy', the agents have engaged in few encounters and the innate biases (α_0, orange) have a strong effect on the language they use. These 'infants' are fast learners: a single observation can drastically alter their beliefs. But they have not yet accumulated enough evidence to develop a consistent, more or less stable language. After a few hundred iterations, during 'puberty' this starts to change. By now all agents use much more stable, but different languages. Still, they are susceptible to new observations. This susceptibility slowly dies out during 'adulthood', when agents align their languages until, after ten thousand encounters, they have effectively negotiated a shared language. The resulting shared language is shaped by the cultural evolution. Different lineages thus adopt different languages, which the two simulations in figure 3 illustrate. Both lineages clearly reflect innate biases. So rather than a *convergence to the prior*, we observe a *reflection of the bias*.

Subfigure B and C focus on overt linguistic behaviour. The dots in subfigure B indicates which words were uttered, and the blue shades in the background show the external language over the last 2000 utterances.[4] The external language ψ_T is shown in subfigure C, together with the bias and $\bar{\pi}$ from subfigure B. The latter is hardly visible, since external language and the aggregate language seem to agree. This is not surprising: once the population has settled on a shared language, words are used in exactly the corresponding proportions. Note that the first two phases ('infancy' and 'puberty') are not so clear from subfigure B and C, although close inspection does reveal larger variability in the initial part of the game. The next two experiments present further evidence for (1) the three-stage evolution and (2) the reflection of the bias.

MEASURING THE DYNAMICS First, we need better ways to measure the dynamics of the game. The statistics used in chapter **??**, such as the number of unique words or the total number of words, are meaningless once the vocabulary is fixed.[5] To measure coherence, we use the (generalised) Jensen-Shannon divergence (JSD). The JSD can quantify the similarity of all the expected languages $\pi_{A_1}, \ldots \pi_{A_N}$ simultaneously (see appendix **??**

[5] When writing this, I realise that variants can of course be defined. In fact, I had done so 'before', in chapter **??**. Future work could transfer those measures to the Bayesian naming game.

[4] Note that the colours are 'normalised' in every column, such that the in every column the least frequent word is white and the most frequent ones the darkest blue.

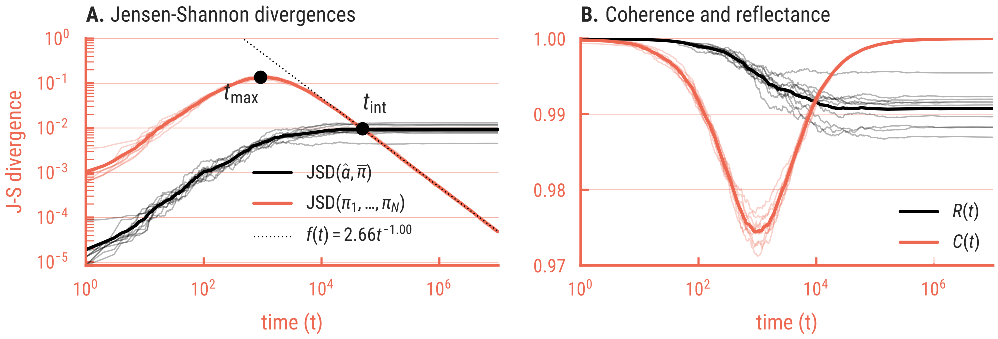


FIGURE 4 The Dirichlet-categorical name converges to a stable, coherent language. A. The distance between expected languages vanishes, but the aggregate language deviates from the bias. B. Coherence initially drops, but then increases to 1. The black line illustrates the *reflection of the bias*.

BNG02 $K = 40, N = 20, \eta = \zeta = 1, \gamma = \infty, \beta = 40$

for details). Normalising the JSD, the coherence measure becomes

$$C(t) = 1 - \frac{\text{JSD}\left(\pi^{(t)}_{A_1}, \ldots, \pi^{(t)}_{A_N}\right)}{\log_2(N)}, \tag{9}$$

such that $C(t) = 1$ indicates perfect coherence and lower values larger incoherence. Another question of interest is how the innate biases α_0 are *reflected* in the expected languages. To that end we measure the divergence between the aggregate language and the (shared) innate bias, which I will call the *reflectance*

$$R(t) = 1 - \text{JSD}(\hat{\alpha}_0, \bar{\pi}_t). \tag{10}$$

When $R = 1$ reflectance is perfect and the aggregate language coincides with the bias; lower values indicate poorer reflection of the prior.

CONVERGENCE The first results suggest that the population will always reach coherence. Note that contrary to Bayesian iterated learning, it is not straightforward to obtain such results analytically: We face the same difficulties as De Vylder and Tuyls (2006). Convergence was thus analysed in an experiment that measured how the coherence and reflectance change over time. The results are shown in figure 4B. The distance coherence (orange) initially decreases (during 'infancy') until it reaches a maximum (in 'puberty') and then starts to increase again (during 'adulthood'), indicating that the population reaches coherence. Subfigure A shows the divergences directly, and suggests that convergence is reliable, since $\text{JSD}\left(\pi^{(t)}_{A_1}, \ldots, \pi^{(t)}_{A_N}\right)$ is eventually well approximated by a function of the form $a \cdot t^{-1}$, This is illustrated by the dotted line, obtained using linear regression on doubly logarithmic coordinates.

However, the stable language is *not* identical to the bias. Rather, the effect of the bias diminishes as can be seen from the reflectance (figure 4B, black). The reflectance decreases, signalling a *divergence* from the bias, until it stabilises below $R = 1$. The final reflectance is fairly consistent across runs and seems to be determined by the strength β of the (see below). In summary, in every run of cultural evolution, in every lineage, the population develops a different, stable and shared language that reflects the innate biases, but diverges from it within certain bounds.

SCALING Figure 4 highlights two 'critical' points t_{max} and t_{int}, namely the maximum of $\text{JSD}(\pi_1, \ldots, \pi_N)$ and the intersection of that with $\text{JSD}(\bar{\pi}, \hat{\alpha}_0)$ respectively. These points

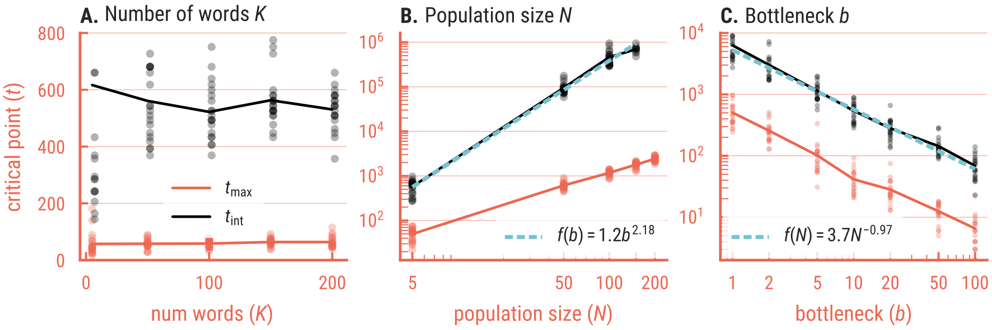


Figure 5 Effects of the language, population and bottleneck size on convergence time, probed by the critical points $t_{\max}$ and t_{int}. See main text for details.

BNG03 Parameters are fixed at $N = 5$, $K = 10$ and $b = 10$, if they are not varied. $\eta = \zeta = 1$, $\gamma = \infty$, $\beta = 100$.

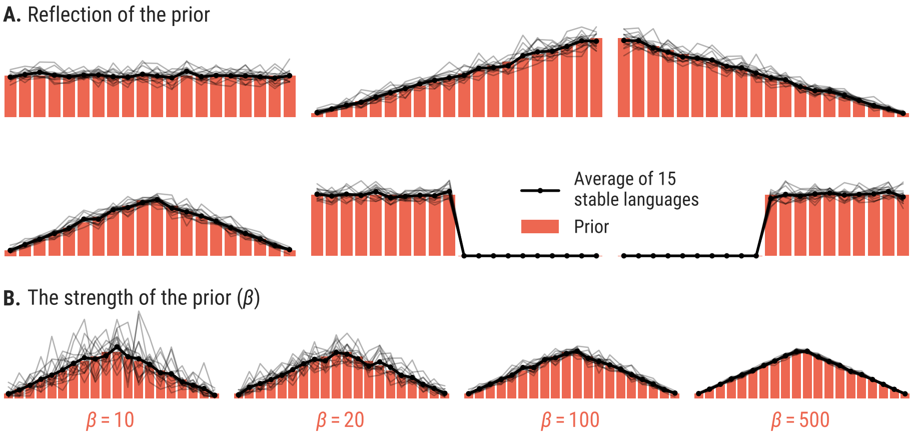


Figure 6 A. Different runs of evolutionary history result in different stable languages (thin black lines) that all reflect the prior (orange) in the sense that cultural evolution reproduces the prior *on average* over many runs. This is illustrated with six differently shaped priors. B. How well the languages reflect the prior is regulated by the strength of the prior (β).

BNG04/07 $K = 20$, $N = 10$, $b = 10$, $\zeta = \eta = 1$, $\gamma = \infty$, $\beta = 100$

seem to provide reliable indications of the convergence time and can thus be used to analyse how it is influenced by different parameters (cf. Baronchelli, Loreto, and Steels 2008; Baronchelli 2017). Figure 5 shows the effects of the language size (K), the population size (N) and the bottleneck size (b) on the convergence time, measured by the location of $t_{\max}$ and t_{int}. The number of words does not seem to have a strong effect on the convergence time. This is somewhat surprising and it might be worth investigating further. The population size does have a clear effect and seems to follow a power law $t_{\text{int}} \propto N^{2.18}$ (estimated using linear regression). The minimal naming game exhibits similar power-law scaling, although with a different exponent of 1.5 (see eq. **??**). This means that convergence time increases quickly as the population grows. But, as subfigure c shows, growing convergence times can be countered by increasing the bottleneck. This, too, exhibits power-law behaviour $t_{\text{int}} \propto b^{-.97}$, which is suspiciously close to $t_{\text{int}} \propto b^{-1}$. This is not surprising since larger bottlenecks allow for a more faithful transmission of the language, leading to faster convergence. Finally it should be stressed that the numerical results are rather rough estimates, since the explored range is very limited. Future work might extend the range to obtain more reliable numbers, or search for analytic results.

REFLECTION OF THE BIAS What mechanism underlies the ‘reflection of the bias’ in the final language? One possibility is that every converged language is a ‘draw’ from some distribution around the bias. That would mean that cultural evolution reproduces innate biases, but only *on average*. To test this, the game was repeated 15 times with six differently shaped biases. Indeed, each of the 15 lineages developed a distinct language, but the average of all lineages closely aligns with the innate bias (see figure 6A). The next question might be how much the emerging languages can deviate from the bias. This, it seems, is determined by the strength of the bias, β. Recall that the bias enters the model as the parameter of a Dirichlet distribution and α_0 can thus be factorised as $\alpha = \beta \cdot \mu$ where β is an inverse variance for the corresponding Dirichlet. Higher values of β result in smaller variance. This translates directly to the distance the resulting languages can have from the bias, as illustrated by figure 6B. These results corroborate the idea that in this model, cultural evolution effectively samples a language from a distribution around the bias. Interestingly, the resulting pattern is a common one in linguistics: *wide constrained variation* (Regier, Kemp, and Kay 2015). Colour terms, to name one example, vary across languages, but within certain constraints Regier, Kemp, and Kay (2015).

Language and production strategies

LANGUAGE STRATEGIES Chapter **??** discussed two strategies for selecting languages in iterated learning models: sampling a language or using the maximum of the posterior (MAP). The same strategies can be introduced in Bayesian naming games, using a parameter η to interpolate between them. In the Dirichlet-categorical model, the exponentiated distribution even has a simple analytical form (see eq. **??**):

$$\begin{aligned} p_{\text{LA}}(\theta \mid \mathrm{x}, \alpha) &\propto \big[p\big(\theta \mid \mathrm{x}, \alpha\big)\big]^{\eta} && (11) \\ &= \text{Dirichlet}(\theta \mid \eta \cdot (\alpha - 1) + 1). && (12) \end{aligned}$$

Exponentiation shifts the distribution towards its mode, the point with highest probability, and moreover shrinks the variance, as illustrated in figure 7. We assume that agents only use the exponentiated posterior during *production*, and use the normal (un-exponentiated) posterior as the prior in the next round. In other words, a hearer updates its beliefs to $\alpha_{t+1} := \alpha_t + c_t$, and *not* to $\eta(\alpha_t - 1) + 1 + c_t$. This means that agents will use the (internal) language they are most confident about, but remember how uncertain they were about other languages. After all, if an agent were to use the exponentiated posterior as the prior in the next round, it would effectively assume that the language it last encountered will from now on be used by all other agents. For that reason η really determines a production strategy, and not a *learning strategy* (the name commonly used in IL). I have called it the *language* strategy to distinguish it from the actual production strategy (see below, and also figure 2).

PRODUCTION STRATEGIES In a similar fashion, different strategies for picking *words* can be defined by sampling from

$$p_{\text{PA}}(x \mid \theta, \alpha) \propto \big[\, p(x \mid \theta, \alpha)\,\big]^{\zeta}. \qquad (13)$$

One reason for introducing the MAP-strategy for selecting utterances (i.e., $\zeta = \infty$) is that it mirrors the production strategy used in naming games. There, agents typically

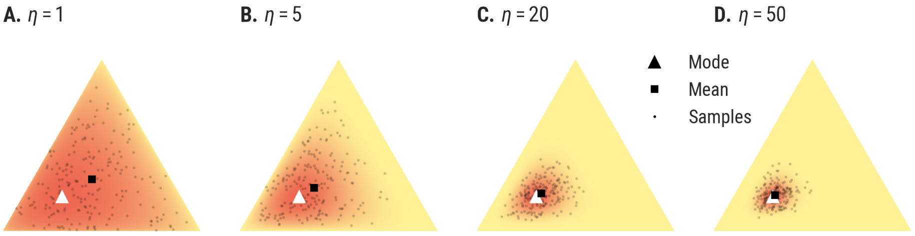


FIGURE 7 Exaggerating (or exponentiating) a Dirichlet distribution shrinks the variance and as η grows, the mean approaches the mode.

FIG02

produce the word with the highest score. But if we assume, following Griffiths and Kalish (2007), that agents are Bayesian and have accurate knowledge of the production strategy, they should infer a different posterior distribution: $p(\theta \mid x) \propto p_{\text{PA}}(x \mid \theta) \cdot p(\theta)$. This distribution is no longer a Dirichlet distribution (see appendix **??**) and a result, posterior inference cannot take the form of updating α. This significantly complicates the game and partly for that reason we assume that agents update their posterior *without* taking into account ζ. *For $\zeta > 1$ agents are therefore not (perfect) Bayesian reasoners.* This, I would argue, is not too problematic, since the parameter ζ is primarily introduced to reproduce the naming game, which itself does not use Bayesian agents. Moreover, technical considerations suggest that Bayesian agents that *do* take into account ζ would after a single observation deem all languages to be absurd, if they do not assign the highest probability to the observed word. That also seems unrealistic.[6]

In short, a round in the Dirichlet-categorical naming, with language and production strategies parametrised by η and ζ, takes the following form

$$\text{SPEAKER} \quad \begin{cases} \theta_t \mid \alpha_{t-1} & \sim \text{Dirichlet}\big(\ \eta(\alpha_{t-1} - 1) + 1\ \big) \\ x_i \mid \theta_t & \sim \text{Categorical}\big(\theta^{\zeta} / \Sigma(\theta^{\zeta})\big), \quad i = 1, \ldots, b. \end{cases} \tag{14}$$

$$\text{HEARER} \qquad \alpha_{t+1} := \alpha_t + c_t \tag{15}$$

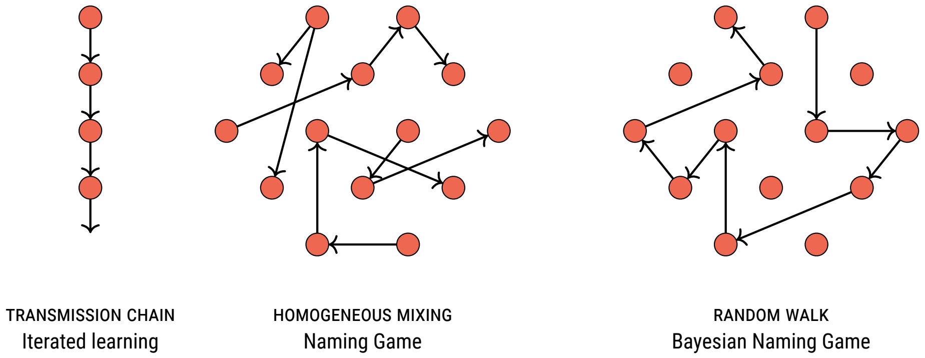
where $\Sigma(\theta^{\zeta}) := \theta_1^{\zeta} + \ldots \theta_K^{\zeta}$ denotes the sum of the entries of a vector.

In conclusion, Bayesian naming games can use different language and production strategies by importing parameters η and ζ from iterated learning and naming games respectively. We will evaluate all these strategies empirically and, in some cases, analytically. But it is better to do that later, in tandem with a new population structure that connects the Bayesian naming game to Bayesian models of iterated learning.

Bayesian language games

The Bayesian naming game was directly inspired by Bayesian models of iterated learning. The strategies, we have just seen, can also be connected to strategies used in the naming game. We now take the analogies one step further and explicitly connect the two paradigms. I reserve the name *Bayesian naming game* for the game studied above, and refer to the extension that we will define here as the *Bayesian language game*, since

[6] But then again, this probably happens because agents do not take into account that the language comes from multiple sources (cf Ferdinand and Zuidema 2009; Smith 2009) and is discussed later.



TRANSMISSION CHAIN
Iterated learning

HOMOGENEOUS MIXING
Naming Game

RANDOM WALK
Bayesian Naming Game

FIGURE 8 The proposed transmission model, a random walk through the population, combines the transmission chains used in iterated learning with the homogeneous mixing from the naming game.

it includes iterated learning-type models. To connect iterated learning to the Bayesian naming game, the population model has to be changed. I propose to add two ingredients: random walks and a life expectancy. The model will do a random walk through a population of fixed size. If agents 'die' after every interaction, the random walk becomes a transmission chain used in iterated learning. If the agents live forever, the random walk resembles homogeneous mixing from the naming game. Random walks might not be a very realistic model of linguistic interaction (although similar to homogeneous mixing), but formulating the most realistic model is not our main motivation either. Rather, connecting the two paradigms aims to highlight what they have in common and where they diverge.

INGREDIENT 1: RANDOM WALKS Transmission chains and homogeneous mixing are naturally combined into a random walk (see figure 8). Starting with a random first agent, in every round only one new agent is selected. The previous hearer becomes the next speaker. In that way a path through the population is formed that, when unraveled, mirrors a transmission chain. This trick is also used by Whalen and Griffiths (2017) in the context of arbitrary graphs. Note that this walk does not impose any restrictions on which agents can interact. Over time all agents are visited equally often and with equal probability.[7] The underlying social network is fully connected and in that sense there is homogeneous mixing. Note that using a random walk in the *minimal* naming game would be rather pointless: only the first agent will invent a word, which then spreads the population. This is caused by the extreme form of lateral inhibition, and the Bayesian language game seems unaffected by this.[8]

INGREDIENT 2: VARY THE LIFE EXPECTANCY Although a random walk forms a chain, it is not a typical *transmission chain* since agents can join in several times. For iterated learning, this issue is particularly pressing — you would not want the great-grandmothers to reappear as the children of their great-granddaughters. Fortunately, this is easily remedied by the second ingredient: death. If speakers were to die after every encounter, and if their places were taken by newborns, the random walk *does* reduce to a transmission chain. Conversely, if agents live forever one retrieves the naming game. And for intermediate life expectancies, one gets a gradual turnover of the population with both horizontal interactions (between agents that have lived for a while) and vertical inter-

[8] I have not been able to isolate systematic differences between homogenous mixing and random walks in the Bayesian language game, although future work could investigate this more systematically

[7] More precisely, the random walk is a Markov chain over the population with uniform stationary distribution.

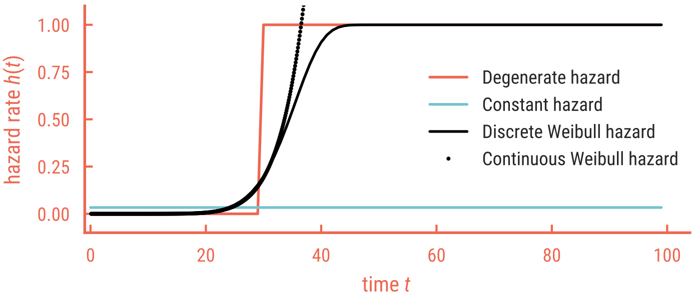


FIGURE 9 Different hazard functions. The more realistic (continuous/discrete) Weibull hazard is better approximated by a degenerate than a constant hazard function.

actions (between newborns and older agents) — a bit like the real world.

Birth-death processes are fairly common in the language evolution literature. To cite just two examples, de Boer and Vogt (1999) and Smith et al. (2002) model population turnover by removing one random agent in every round, and replacing it with a new agent. The problem with this approach is that it implies a rather unrealistic model of life-expectancy. To see why, one has to look at the so called *hazard rate*: probability that an agent will die in a given round, given that it is not dead yet (Rogríguez 2007). In the mentioned studies, this quantity is constant: $1/N$. Constant hazard rates do arise naturally, for example in radioactive decay, but not in human mortality rates. Those are much higher amongst elderly (and infants) and therefore not constant. For that reason, demographers have adopted different models often building on either the *Weibull* or *Gompertz* distribution (Juckett and Rosenberg 1993).

In appendix **??** I have outlined a *discrete Weibull* model of life expectancy. It has one parameter γ, which is the average life-expectancy. Since mathematical analyses might benefit from an even simpler model, I alternatively propose to use a *degenerate hazard* function that assigns all agents an identical, fixed life-span. This seems a better approximation of the Weibull than a model with constant hazard-rate (see figure 9). In the simulations below, I have indeed used a degenerate model with a fixed life-expectancy of γ, i.e. every agent dies after γ interactions as a speaker.[9]

Characterising Bayesian language games

The *Bayesian language game* is simply the Bayesian naming game extended with the random walk and population turnover outlined above. It can reproduce various different models, depending on three parameters:

- **Language strategy** η. Determines to what extend the agents favour more likely languages. $\eta = 1$ yiels samplers, $\eta = \infty$ maximisers.
- **Production strategy** ζ. Regulates the tendency to produce more likely productions; $\zeta = 1$ for samplers and $\zeta = \infty$ for maximisers.
- **Life expectancy** γ. The average life expectancy of an agent in terms of the number of rounds it can play as a speaker. For $\gamma = 1$ for iterated learning; $\gamma = \infty$ for a naming game.

Of course, the population size, number of words and bottleneck size are also of interest, but η, ζ and γ most directly determine the type of game. The next question is simple:

[9] Here too, I have to leave it to future work to *systematically* assess the impact of the different models of population turnover.

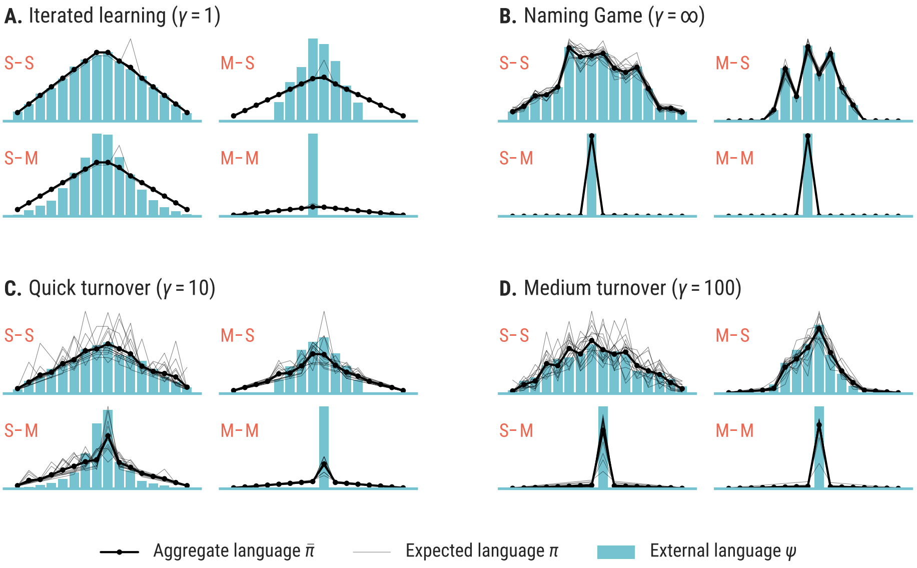


FIGURE 10 Typical outcomes of the Dirichlet-Categorical language game for the extreme strategies (sample–sample, MAP–sample, sample–MAP, MAP–MAP) in populations with immediate turnover (A, iterated learning, $\gamma = 1$), no turnover (B, naming game, $\gamma = \infty$) and two intermediate turnovers (C and D). See the main text for a discussion.

FIG08 $K = 16, N = 15, b = 1, T = 10000$

how? What kind of behaviour can we expect for different parameter settings? To find out, an experiment was set up to explore a larger part of the parameter space (η, ζ, γ) in a systematic fashion. The parameters appear to interpolate relatively smoothly between the extreme cases where η, ζ and γ are either 1 or ∞. The extreme cases are, I believe, most clearly illustrated by the outcomes of single runs. The main text discusses those, and I refer to appendix **??** for a more systematic exploration of the parameter space confirming the findings discussed here.

The central figure is 10. It shows one run of the Dirichlet-categorical language game for four different life expectancies: $\gamma = 1$ (iterated learning), $\gamma = \infty$ (naming game) and the intermediate $\gamma = 10$ and $\gamma = 100$. For every γ the four 'extreme' language-production strategies (η, ζ) are shown: sample-sample, sample-MAP, MAP–sample and MAP–MAP. Note that the blue bars show the external language, not the prior. All runs use the same prior: the by now familiar 'pyramid'. I first discuss the effect of population turnover (γ) and then turn to the different strategies.

VARYING LIFE EXPECTANCY: BETWEEN IL AND NG'S The results for iterated learning (subfigure A), are easily misinterpreted. At every point, only one agent has some past experience: the current speaker. That speaker corresponds to the thin line deviating from the aggregate language. The expected language of all other agents is exactly their bias. Through the lens of the Bayesian language game, we see an almost perfectly homogeneous population, which explains why the aggregate language $\bar{\pi}$ is the same for all strategies. This also means that the coherence is near-maximal, but this could be seen as an artefact. After all, considering the external language reveals an important discrepancy for non-sampling strategies. The internal language of agents, as seen from the aggregate language, is in strong disagreement with the external language (orange). That means that no agent, not even the speaker, has a faithful internal representation of the language actually used. With more experience the discrepancy disappears: in

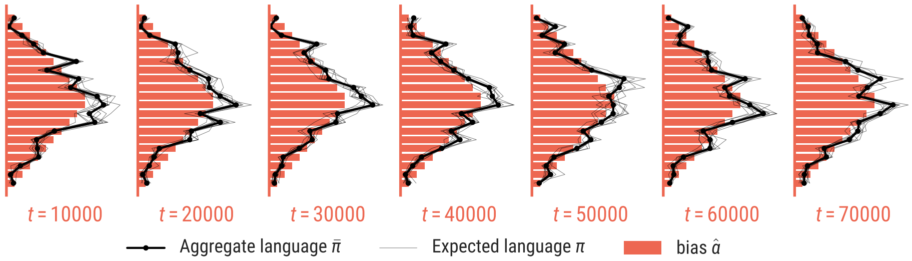


FIGURE 11 Gradual language change in the Bayesian language game for a particular choice of parameters. The effect seems brittle: slightly different parameter settings can give the kind of behaviour shown when $\gamma = \infty$.

FIG11 $K = 20, N = 10, b = 2, \beta = 40, \gamma = 700, \eta = 2, \zeta = 1$, deterministic hazard.

naming game (subfigure B) the external and internal languages are in fair agreement. More experience can also result from larger bottlenecks, which would result in only the speaker having a better representation of the language.

Another interesting observation is suggested by the two mixed strategies, MAP–sample and sample–MAP. Both exaggerate the bias, but in different ways. Maximising only the language appears to *prune* low-probability languages, whereas maximising only productions seems to *exponentiate* the language. This would explain the shape of the limiting language under a MAP–sample strategy in the naming game. That appears to consistently deviate from an exponentiated distribution and indeed more closely resembles a pruned distribution. Needless to say, more work is needed to confirm this 'pruning-vs-exponentiating' hypothesis. The most striking difference between the Bayesian iterated learning model and the naming game is the 'predictability' of cultural effects with the former. Even maximising strategies appear to result in languages that are determined by the bias, and some simple operation (possibly pruning or exponentiating). They are seemingly uninfluenced by the contingencies of the cultural process, in sharp contrast to the Bayesian naming game. This suggests that even maximising strategies in Bayesian iterated learning result exhibit fairly "uninteresting" (cf. Dediu 2009) behaviour.

The Bayesian naming game arguably exhibits more "interesting" behaviour, since the resulting languages are clearly shaped by a the contingencies of a rough, stochastic process of cultural evolution. The intermediate life expectancies seem to interpolate between IL and NG behaviour. The longer the agents live, the — yes — 'stronger and stabler' the cultural effects become, and the more languages can move away from the biases. For intermediate languages, variability can be large since new agents can always be introduced. An interesting question if the Bayesian language game can also reproduce gradual language change while maintaining a fair stability. Preliminary experiments suggest this is the case (see figure 11, although they also suggest that the effect is brittle in the sense that for example increasing γ quickly seems to result in behaviour more similar to $\gamma = \infty$.

EXTREME STRATEGIES FOR THE BAYESIAN NAMING GAME The Bayesian *naming* game implements a kind of lateral inhibition in the form of Bayesian updating. So do other strategies correspond to the different alignment strategies in the naming game? I discuss all strategies below, also $\gamma < \infty$/

- **Sample–sample (s–s, $\eta = \zeta = \infty$).** This is the 'default' strategy in the Bayesian naming game and corresponds to the sampler-strategy in iterated learning. This strategy exhibits lateral inhibition, in the sense that the 'score' of an observed word increases, while the score of other words decrease. By *score* the probability of the word under the expected language $s_t(x) := p(x \mid \alpha_t)$ is meant. After observing x_t it can be shown (see eq. **??**) to change to

$$s_{t+1}(y) = \frac{\Sigma(\alpha_t)}{\Sigma(\alpha_t) + 1} \cdot s_t(y) + \frac{[\![y = x]\!]}{\Sigma(\alpha_t) + 1}, \tag{16}$$

where $[\![$ condition $]\!]$ is the indicator function evaluating to 1 if the condition holds, and 0 otherwise. The update differs from the basic lateral inhibition strategies (e.g. Wellens 2012). First, the inhibition works by scaling rather than subtraction of a fixed parameter δ_{inh}. Second, the effect of the updates decreases with time since $\Sigma(\alpha_t)$ increases over time. Note that this proves that the expected language of one particular agent will converge, since the updates vanish. The simulations earlier this chapter suggest that all agents moreover converge to the same language, which reflects the bias. This is also what we see in figure 10B (S–S).

- **MAP–sample (M–S, $\eta = \infty, \zeta = 1$).** This strategy is used by maximisers in iterated learning. The lateral inhibition mechanism takes a very similar form as in the sample-sample strategy. In particular, the updates eventually also vanish, proving 'individual' convergence. Simulations suggest that coherence always emerges and the stable language appears to reflect an amplified or exaggerated version the bias. The exaggeration is apparent from figure 10B (M–S), where the resulting language is more peaked than the prior. The prior is not shown, but is the 'pyramid' also visible in subfigure A (S–S).
- **Sample–MAP (S–M, $\eta = \infty, \zeta = 1$).** This strategy is hardest to analyse, since the scores $p(x \mid \alpha)$ do not seem to have a simple expression (see appendix **??** for details). However, when bias is flat the strategy reduces to the case analysed by De Vylder and Tuyls (2006), which implies convergence to a single-word language. Indeed figure 10B (S–M) confirms that idea.
- **MAP–MAP (M–M, $\eta = \zeta = \infty$).** The MAP–MAP strategy corresponds to the frequency strategy from chapter **??**. An agent with this strategy uses the language with highest probability, the mode, and then utters the largest component from the mode. This amounts to producing $x_t = \arg\max_k \alpha_k$, the word with the highest counts, including pseudo-counts. The only words these agents will every use are the maxima of the bias. Consistent with chapter **??**, we find convergence to a single-word language in 10B (M–M).

Conclusions

This chapter proposed a Bayesian naming game based on a Dirichlet-categorical model. In the standard version of the game ($\eta = \zeta = 1$) the population reaches coherence in a typical three-stage process, metaphorically called 'infancy', 'puberty' and 'adulthood'. The resulting language reflects the bias, but is clearly shaped by the contingencies of cultural evolution. The model thus gives rise to lineage-specific, stable languages. In sum, it answer many of the desiderata formulated in chapter **??**. Concretely, it explicitly represents biases **??**; incorporates strategies from both the iterated learning and naming game literature **??**; seems susceptible to mathematical analysis **??**, as further discussed in chapter **??**; exhibits nontrivial cultural effects **??**; and results in a stable language (**??**).

The Bayesian *naming* game was extended to the Bayesian *language* game by introducing language- and production strategies (η and ζ) and a population model consisting of (1) a random walk and (2) a life expectancy γ for every agent. For $\gamma = 1$ this produced an iterated-learning model, for $\gamma = \infty$ in a naming game. A characterisation of the parameter space suggested several conclusions. First, that agents in an iterated learning model never faithfully represent the language actually used. And second, that the effect of maximising languages or productions are different and correspond to something like pruning or exponentiating the bias respectively. This in turn indicates that for those strategies are also relatively 'uninteresting' in iterated learning models, in the sense that that the outcome seems to be predictably determined by the bias. This is not the case for the Bayesian naming game, where the cultural process leaves a non-trivial on the language. That does not mean that the process is completely unpredictable, since the resulting language appears to be a draw from some distribution around the prior, allowing for only limited variability.

An interesting further question concerns 'stable' language change. Initial results suggest this can occur, but is somewhat brittle and does therefore not fulfil the desideratum of robustness (**??**). In general, the characterisation however does suggest a fair robustness. All small values of γ result in behaviour very similar to $\gamma = 0$ (iterated learning), and all large values are similar to $\gamma = \infty$. The same goes for the strategies: there seems to be a relatively smooth transition between the extreme cases. That means that understanding the extreme cases gives a fair sketch of the kind of behaviour that can be expected. In short, those are that sampling strategies result in (external) languages reflecting the biases, either perfectly (iterated learning) or imperfectly, when mediated by culture (naming game); mixed strategies exaggerate the biases, but differently when languages or productions are maximised (and again perfectly or imperfectly); and pure maximising strategies result in degenerate distributions. The longer the life span, the closer the external and internal languages align and the greater the language stability.

Of all the desiderata formulated in chapter **??**, only one remains. This is not to say that the models presented perfectly addressed all points, merely that they did so sufficiently for the purposes of this thesis — I discuss it's shortcomings in the final chapter. In the second part, I address the remaining desideratum: empirical testability. Let me end this chapter with some remarks on related work.

RELATED WORK The Bayesian iterated learning model is closely related to various models proposed in the literature. In the naming game literature, De Vylder and Tuyls (2006) is the closest analogue I have been able to find. The Dirichlet-categorical naming game nearly has their model as a special case, with a flat prior and a MAP language-strategy ($\eta = \infty$). The queue-agents can be roughly approximated by a fixed life-expectancy corresponding the length of the queue, but the analogy is not perfect. An interesting question is whether their results can be extended to the continuous case here. Even more closely related is the model by Reali and Griffiths (2010). In fact, it is the exact same Dirichlet-categorical model, but only studied in the iterated learning context.[10] Interestingly, they show that the model is in that case equivalent to the

[10] I unfortunately only became aware of this while writing up the results and time does not allow me to include an in-depth discussion.

Wright-Fisher model of genetic drift. Needless to say, this is an area ripe for future research. The first sketches of a 'Bayesian' naming game can also be discerned in Kirby, Tamariz, et al. (2015), who consider a population of two Bayesian agents interacting without population turnover. Ferdinand and Zuidema (2009) similarly represent languages (hypotheses) as categorical distributions, but use a different prior. A prior 'extending' the Dirichlet prior used here, the so called Dirichlet Process, has figured in Burkett and Griffiths (2010) and Kirby, Tamariz, et al. (2015). However, all these studies had different goals and none of them explicitly explored the parallels with naming games, hence I do not further discuss them here.

Bibliography

Baronchelli, Andrea (2017). "A gentle introduction to the minimal Naming Game". In: pp. 1–24. DOI: 10.1075/bjl.30.08bar. arXiv: 1701.07419.

Baronchelli, Andrea, Vittorio Loreto, and Luc Steels (2008). "In-depth Analysis of the Naming Game: The Homogeneous Mixing Case". In: *International Journal of Modern Physics C* 19.05, pp. 785–812. DOI: 10.1142/S0129183108012522.

Burkett, David and Thomas L. Griffiths (2010). "Iterated learning of multiple languages from multiple teachers". In: *The evolution of language: Proceedings of EvoLang*, pp. 58–65.

De Vylder, Bart and Karl Tuyls (2006). "How to reach linguistic consensus: A proof of convergence for the naming game". In: *Journal of Theoretical Biology* 242.4, pp. 818–831. DOI: 10.1016/j.jtbi.2006.05.024.

De Boer, Bart and Paul Vogt (1999). "Emergence of Speech Sounds in Changing Populations". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 1674, pp. 664–673. DOI: 10.1007/3-540-48304-7_87.

Dediu, Dan (2009). "Genetic biasing through cultural transmission: Do simple Bayesian models of language evolution generalise?" In: *Journal of Theoretical Biology* 259.3, pp. 552–561. DOI: 10.1016/j.jtbi.2009.04.004.

Ferdinand, Vanessa and Willem Zuidema (2009). "Thomas' Theorem meets Bayes' Rule: a Model of the Iterated Learning of Language". In: *Proceedings of the 31st Annual Conference of the Cognitive Science Society.* Austin, Texas, pp. 1786–1791.

Griffiths, Thomas L. and Michael L. Kalish (2007). "Language Evolution by Iterated Learning With Bayesian Agents". In: *Cognitive Science* 31.3, pp. 441–480. DOI: 10.1080/15326900701326576.

Grifoni, Patrizia, Arianna D Ulizia, and Fernando Ferri (2016). "Computational methods and grammars in language evolution : a survey". In: *Artificial Intelligence Review* 45.3, pp. 369–403. DOI: 10.1007/s10462-015-9449-3.

Jaeger, Herbert et al. (2009). "What Can Mathematical, Computational, and Robotic Models Tell Us about the Origins of Syntax?" In: *Biological Foundations and Origin* *of Syntax*. Ed. by Derek Bickerton and Eörs Szathmáry. The MIT Press, pp. 385–410. DOI: 10.7551/mitpress/9780262013567.003.0018.

Juckett, David A. and Barnett Rosenberg (1993). "Comparison of the Gomperz and Weibull Functions as Descriptors for Human Mortality Distributions and their Intersections". In: *Mechanisms of Ageing and Development* 69.1-2, pp. 1–31. DOI: http://doi.org/10.1016/0047-6374(93)90068-3.

Kirby, Simon, Tom Griffiths, and Kenny Smith (2014). "Iterated learning and the evolution of language". In: *Current Opinion in Neurobiology* 28, pp. 108–114. DOI: 10.1016/j.conb.2014.07.014.

Kirby, Simon, Kenny Smith, and Henry Brighton (2004). "From UG to Universals: Linguistic adaptation through iterated learning". In: *Studies in Language* 28.3, pp. 587–607. DOI: 10.1075/sl.28.3.09kir.

Kirby, Simon, Monica Tamariz, et al. (2015). "Compression and communication in the cultural evolution of linguistic structure". In: *Cognition* 141, pp. 87–102. DOI: 10.1016/j.cognition.2015.03.016.

Reali, Florencia and Thomas L. Griffiths (2010). "Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift." In: *Proceedings. Biological sciences / The Royal Society* 277.1680, pp. 429–36. DOI: 10.1098/rspb.2009.1513.

Regier, Terry, Charles Kemp, and Paul Kay (2015). "Word meanings across languages support efficient communication". In: *The handbook of language emergence*, pp. 237–263. DOI: 10.1002/9781118346136.ch11.

Rogríguez, Germán (2007). "Survival Models". In: *Lecture Notes on Generalized Linear Models*. Chap. 7.

Smith, Andrew D.M. (2014). "Models of language evolution and change". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 5.3, pp. 281–293. DOI: 10.1002/wcs.1285.

Smith, Kenny (2009). "Iterated learning in populations of Bayesian agents". In: *Cogsci Society Conference*, pp. 697–702.

Smith, Linda B et al. (2002). "Object name learning provides on-the-job training for attention." In: *Psychological science : a journal of the American Psychological Society / APS* 13.1, pp. 13–19. DOI: 10.1111/1467-9280.00403.

Steels, Luc (2011). "Modeling the cultural evolution of language". In: *Physics of Life Reviews* 8.4, pp. 339–356. DOI: 10.1016/j.plrev.2011.10.014.

— (2016). "Agent-based models for the emergence and evolution of grammar". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1701, p. 20150447. DOI: 10.1098/rstb.2015.0447.

Tamariz, Monica and Simon Kirby (2016). "The cultural evolution of language". In: *Current Opinion in Psychology* 8, pp. 37–43. DOI: 10.1016/j.copsyc.2015.09.003.

Wellens, Pieter (2012). "Adaptive Strategies in the Emergence of Lexical Systems". PhD thesis. Vrije Universiteit Brussel.

Whalen, Andrew and Thomas L. Griffiths (2017). "Adding population structure to models of language evolution by iterated learning". In: *Journal of Mathematical Psychology* 76, pp. 1–6. DOI: 10.1016/j.jmp.2016.10.008.