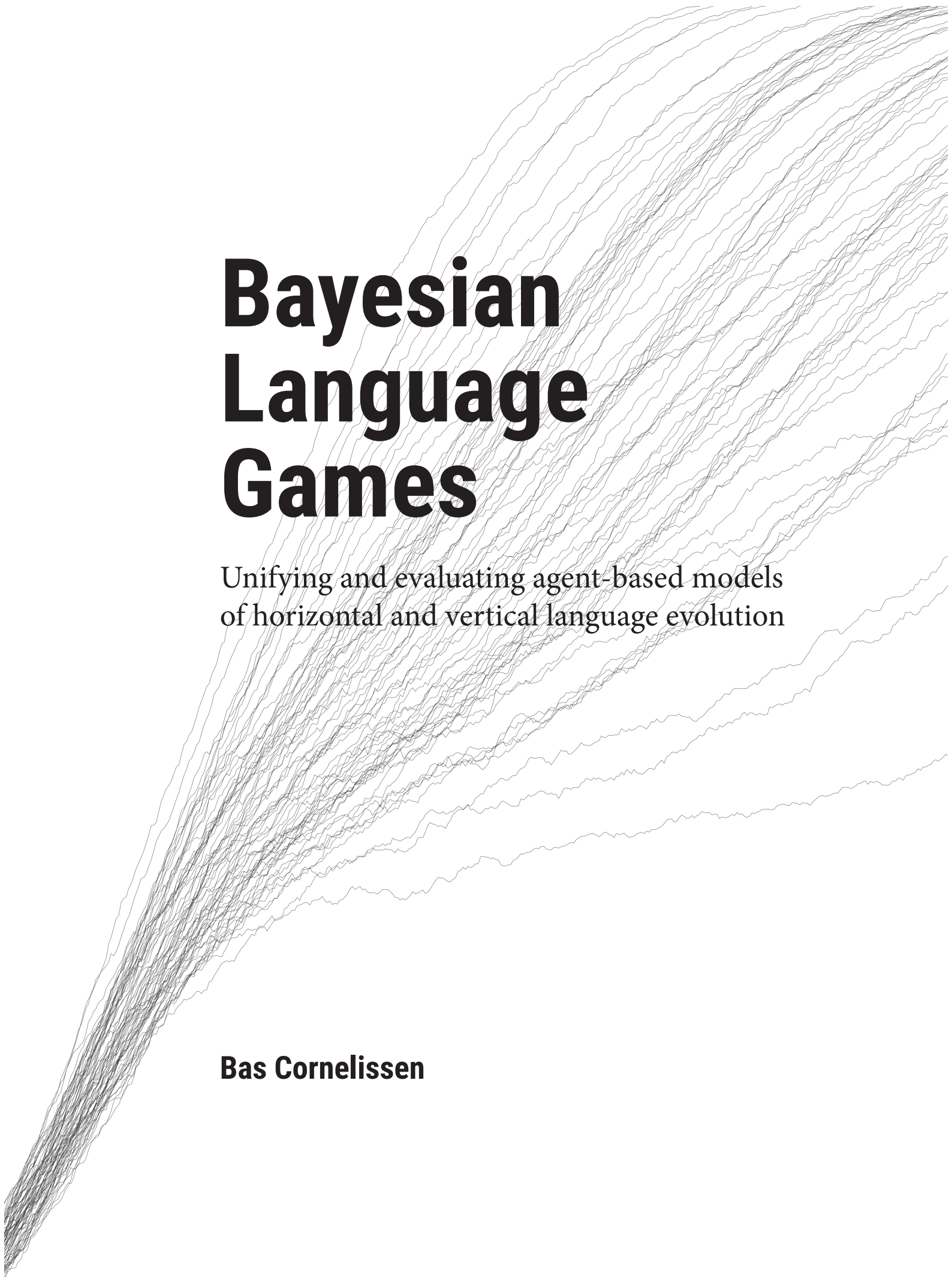


Bayesian Language Games

Unifying and evaluating agent-based models of horizontal and vertical language evolution

Bas Cornelissen

Bayesian Language Games

Unifying and evaluating agent-based models of horizontal and vertical language evolution

MSc Thesis *(Afstudeerscriptie)*

written by

Bas Cornelissen
(born 26 February 1992 in Utrecht)

under the supervision of **Willem Zuidema**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:	**Members of the Thesis Committee:**
30 August 2017	prof. dr. Frank Veltman
	dr. Wilker Aziz
	prof. dr. Benedikt Löwe

Institute for Logic, Language and Computation

Abstract. Human language is one of the most intricately structured communication system in the natural world. Over the last decades researchers in various fields have developed the idea that languages are primarily shaped by processes of cultural evolution, and that these processes can account for the structure of language. Computational models play an important role in their arguments. This thesis asks what those models can teach us about cultural language evolution. To that end, the first part of this thesis connects the two main branches of agent-based models, naming games and iterated learning, in a new *Bayesian language game*. The game gives a unified view on the field and suggests a characterisation of the behaviour exhibited by the main agent-based models of language evolution. It moreover addresses shortcomings of earlier models. We find lineage-specific languages reflecting the innate biases of the learners. The second part of this thesis aims to compare that behaviour with the evolution of actual language. Numeral systems are argued to be an ideal empirical test case for models of cultural language evolution. We revisit Hurford's pioneering work on the modelling of the emergence of numeral systems, and discuss some further results.

Preface. Dear reader, I will keep it short. As I write, my mother is preparing a delicious meal. And, as you will understand, I cannot keep my parents waiting. They haven't seen me much over the last few months, and I'm afraid they are not the only ones. Let me just say thanks to all those wonderful, warm, and loving people that make life so much fun. Oh and Jelle, of course, thanks for putting up with me ;-)

1 The cultural origins of language

Communication abounds in the natural world, but few, if any, communication systems are so intricately structured as human language. We can understand the meaning of a sentence that has never been uttered before, because we understand the words it consists of and the way in which they are combined — that is to say that language is by and large *compositional*. The words, morphemes, themselves also have an internal structure: they consist of phonemes, combined in a accordance with a combinatorial, phonological system specific to the language. The semantic and phonological combinatorics together give language what Hockett (1960) dubbed a 'duality of patterning' — a design feature not commonly found in communication systems of other species. Most animal communication systems are *holistic*: every vocalisation has one specific function, and no further internal structure (Zuidema 2013). There are, it seems, few interdependencies between vocalisations. But language is quite different: just about any two sentences will have many interdependencies in their phonology, morphology, hierarchical structure, and so on. It is in this sense that language has a distinct *systematicity* (Kirby 2017).

Now, the Big Question is this: where does it come from? And, indeed, why only us? Well, we find ourselves in good company. Robert Berwick and Noam Chomsky (2017) have been thinking about the same question, and leave us no doubt how we should *not* try to account for the structure of language: "by means of a cultural-biological interaction". No, "this latter effort fails in all respects". Some of that work is "trivially inadequate", or otherwise the "Kirby-type models", "the Kirby work" and "the Kirby line of research" do of course "not say anything about the initial origin of compositionality".

Good, good! More than enough reason to write a thesis about this "Kirby type work" — and in passing perhaps even cite more than two of his papers. But let's leave the polemics there. After all, what greater good do they really serve?

Berwick and Chomsky (2017) raise some fair concerns, for example regarding the testability and empirical validity of the Kirby-type theories. Or regarding their explanatory power: if your 'agents' already know, say, context-free grammars, and they 'evolve' a compositional language, does your model explain anything? Are you then, if anything, addressing the evolution of universal grammar, or just modelling language change? These are all valid concerns — or so I think, since I address some of them in this thesis, too. But reading Berwick and Chomsky's paper — I haven't read the book yet — one starts to understand what it means if two scientific traditions have lost their common ground and have started talking about radically different things, only superficially using the same words.

My apologies if I overwhelmed the reader with this, well, lack of introduction, but I want to get started sooner rather than later. One should know that the Kirby-type theories, in the broad sense, are really the work of a large group of linguists, cognitive scientists, biologists, anthropologists, computer scientists, statisticians, and even physicists, who over the last 30 years or so have tried to turn the question of language evolution upside down. The 'traditional' debates centred around the question whether language was an adaptation or not. Does language perhaps add to our reproductive fitness, or was it a key mutation event that catapulted language in the world? In either case, the origin of language was thought to be a fundamentally biological concern. We only had to find out how those brains evolved to accommodate language. Well, the

Kirby-type work turned that around: how do languages have to change, in order to fit in our brains (Christiansen and Chater 2016a)? Rather than evolving brains, let's think about evolving languages.

For this to even make sense, one has to drastically change the, let's say, Chomskyan conception of language. Language now becomes a complex adaptive system in its own right, that is subject to all kinds of pressures, at multiple different levels and timescales simultaneously (Kirby and Hurford 2002; Christiansen and Chater 2016a; Smith 2014; Kirby 2017; Steels 2016). On a biological level and evolutionary timescale the innate mechanisms that underly human cognition and language need to develop. At an individual level and much shorter timescale, humans acquire language, constrained by what their biological makeup can accommodate. These learning biases influence, on a cultural level and historic timescale, the cultural dynamics of language. Universals that emerge through processes of cultural evolution in turn shape the fitness landscape and indeed, interactions exist within in and across all levels. Rather than looking solely at the biology of language, the question is how it can interacts with acquisition and use. The idea developed is that "language has been adapted through cultural transmission over generations of language users to fit the cognitive biases in inherent in the mechanisms used for processing and acquisition" (Christiansen and Chater 2016b, p. 12).

Returning to Berwick and Chomsky (2017), I think they are quite right to point out that this work "does not really tackle questions about the evolution UG, but rather questions about how particular languages change over time, once the faculty of language itself is in place". I doubt Kirby would disagree. Indeed, this is the entire point. Since if it turns out that particular languages tend to change over time in a somewhat systematic way that *explains* their structure, then the explanatory burden is lifted from the faculty of language. And this is precisely what the Kirby-type theories argue for: that biology does not have to carry the entire explanatory load, but that a fair share of linguistic structure can be explained by a process of *cultural evolution*. As Kirby (2017) concludes, "we expect the language faculty to contain strong constraints only if they are domain general (e.g. arising from general principles of simplicity) and that any domain-specific constraints will be weak." If anything, the Kirby-type work is in the business of explaining away UG, rather than explaining it.

Now, the evolution of language is a notoriously hard problem — I belief this is where one is supposed to mention the *Société* — but how does one go about if language is moreover a complex adaptive system? Using mathematical and computational models is one solution. Modelling makes all assumptions absolutely transparent, allows one to verify their coherence and consistency and generate new hypotheses, or even predictions (Jaeger et al. 2009). Accordingly, models have figured prominently in the literature on cultural language evolution. And that is where the habitat of this thesis is to be found. *What can these models of language evolution actually learn us about language evolution?* That is the heart of this thesis, but as such, the question is too open ended. I therefore break up the question in two parts.

1. *What kind of* behaviour *can we expect from* agent-based *models of language evolution?* To answer this question, I aim to formulate a model that captures a substantial share of the agent-based modelling tradition, and try to characterise its behaviour. The 'substantial share' is easily identified, since the field falls apart

in two (strictly separated) traditions. **Chapter 2** introduces the 'vertical' tradition around *iterated learning*. This is the Kirby-type work that addresses how vertical transmission between generations can shape language. **Chapter 3** introduces the 'horizontal' tradition around *naming games*, which focusses on the self-organising power resulting from local interactions within a generation. Although the traditions are strictly separated, I will argue in **chapter 4** that their models are very similar. To do so, I take inspiration from Bayesian models of iterated learning and propose the *Bayesian language game*. By further changing the population model I can interpolate between both traditions and thus analyse their behaviour in a single unified framework. I moreover argue that it addresses some of the problems left open by Bayesian models of iterated learning.

2. *How does that behaviour relate to actual evolved language?* Once it is clear what kind of behaviour these agent-based models exhibit, one should ask what this learns us about language evolution. But how can one start answering such a question without an empirical test case? In **chapter 5** I therefore argue that numeral systems are a good test case, and explain in some detail what the structures are one should try to explain. In **chapter 6** I make a first start with simulating the emergence of numeral systems, which largely amounts to revisiting and extending the pioneering work of James Hurford.

The reader might want to note that the summaries at the start of every chapter further flesh out this outline.

The problem of language evolution challenges many disciplinary boundaries. This thesis alone borders at least several branches of linguistics, statistical physics, biology, probability theory and Bayesian (cognitive) modelling. When starting with the current work, I was new to pretty much all of these fields (except perhaps some courses in the latter two fields) and many important results might very well have escaped my attention. But some things I deliberately left out, or only touch on in passing. These include (1) models of biological evolution, (2) evolutionary game theory, (3) genetic algorithms, (4) evolution of UG, (5) empirical studies of transmission chains and (6) experimental semiotics. Neither will I further discuss broader debates, of which the Berwick and Chomsky paper is part. Instead, I try to address issues in the field of agent-based modelling of language evolution on its own terms. Indeed, the reader will notice that focus in the majority of this thesis is decidedly on the models, more so than on possible interpretations thereof. This partly reflects personal interest, but more importantly, I belief interpretations of models should, insofar as possible, built on a sound understanding of the model themselves. That is what I hope this thesis, if anything, contributes to.

SOURCE CODE AND DATA I have made the source code of all experiments publicly available via bascornelissen.nl/msc-thesis (including all figures and LaTeX files). A reference to the 'raw' data from all experiments can also be found there. The captions of many figures contain a code like **BNG06** or **FIG03**. This identifies the experiment or figure in the repository.

NOTATION Throughout the thesis I use several notational conventions. Vectors are written in boldface, as in $\mathbf{x} = (x_1, \ldots, x_K)$, and indexed by k if they have length K. Sometimes it will be convenient to abbreviate $x_0 := \sum_{k=1}^{K} x_k$. We often need to decorate variables with time-indications. The most consistent unambiguous solution, $\mathbf{x}^{(t)}, x_k^{(t)}$ etc., often clutters the notation. Therefore, vectors (boldface) simply get their indication in the subscript $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots)$ and I only use the superscript (t) when confusion can arise, as in $x_k^{(t)}$. If X is a random variable with distribution Dist parametrized by λ, then by $X \sim \text{Dist}(\lambda)$ we mean that $p_X(x) = \text{Dist}(x \mid \lambda)$ where the latter is the density (mass) function. We mostly drop the random variables and write $p(x)$ and $p(x \mid z)$ rather than $p_X(x)$ and $p_{X|Z}(x \mid z)$. Normalizing constants are often irrelevant and we write $p(x) \propto f(x)$ to indicate proportionality, i.e. that $p(x) = 1/C \cdot f(x)$ where C does not depend on x. With regard to sets, $\mathbb{N}, \mathbb{Z}, \mathbb{R}$ denote the natural numbers, integers and reals. If A and B are sets, their Cartesian product is $A \times B$ and B^A denotes the set of all functions $f : A \to B$. 'Spaces' typically have a calligraphic character, so x lies in $\mathcal{X}$ and parameters tend to be Greek. Finally, x is nearly always an observable variable (e.g. an utterance), z and unobservable variables (e.g. internal representation of a language), m a meaning and s a signal.

2 Iterated Learning

Could it be that structure in language emerges because it is transmitted from one generation to the next? Is cultural *transmission* the force shaping language? Early models of iterated learning suggested precisely that. Bayesian models improved the early work by separating the biases of the learners from the effects of transmission. But they also indicated that cultural evolution only allows the prior biases to surface, a result that sparked a small controversy. The 'convergence to the prior' was shown to break down in more complicated populations, again creating room for the shaping force of cultural evolution. This chapter introduces the iterated learning tradition and ends with a list of desiderata for models of cultural language evolution. The list serves as a guide to the remainder of this thesis.

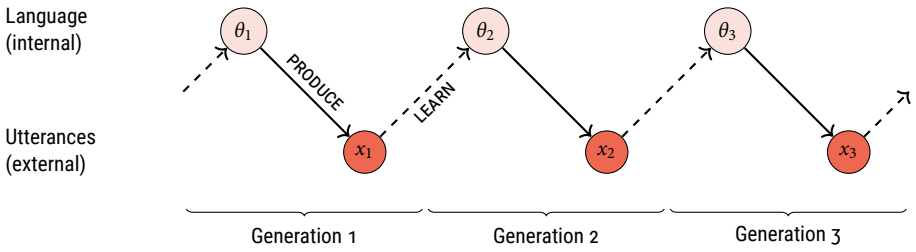


FIGURE 2.1 In the iterated learning model, the language produced by the previous generation serves as the primary linguistic data for the next.

Adapted from Kirby (2001).

Early iterated learning models

In the early years of this century, James Hurford, Simon Kirby, Kenny Smith and others, developed the idea that cultural transmission, in the form of *iterated learning* (IL), could be the source of structure in language. Early work in this tradition tried to isolate a "minimal set of assumptions and hypotheses with which linguistic structure can be explained" (Brighton 2002). The result was a simple model of cultural transmission between generations consisting of a single agent each. In the model, language alternates between an internal representation (I-language in Chomskyan parlance; Chomsky 1986, pp. 19–24) or an external representation in the form of actual utterances (E-language), as figure 2.1 illustrates. The first agent (the parent) is presented with several objects for which it produces some utterances. Those utterances form the primary linguistic data from which the second agent (the child) has to learn a language.[1] The child goes on to become the parent of the next generation, forms expressions for several (other) objects, which are observed by the next agent, and so on.

Every generation learns a language by observing the language of the previous generation, who themselves learned it from the generation before them. The target of learning is therefore the outcome of the same learning process and this gives rise an evolutionary dynamics on the cultural level: the fact that a language has to be learned over again shapes the language itself to become better learnable, hence better transmissible. And key to better transmission, many studies suggested, was the acquisition of some form of systematicity. Paraphrasing Hurford (2000), language appeared to be structured, because cultural transmission favours systematicity.

THE EMERGENCE OF COMPOSITIONALITY, I This conclusion was primarily based on computer simulations of the emergence of compositionality which I briefly want to discuss. Suppose, following Brighton (2002), that agents are positioned in an environment with various objects. The objects have F possible features, each taking V values, and thus correspond to points in a F-dimensional meaning space $\mathcal{M}$. The features might be color and shape, taking values triangular, rectangular or circular and orange, blue and black respectively. A language associates meanings $m \in \mathcal{M}$ to signals s in a space $\mathcal{S}$ of signals, typically strings over some alphabet. Certain languages are compositional, meaning that the signals can be decomposed in subsignals that each bear one aspect of the meaning. Compositional languages should be distinguished from *holistic* languages where meanings correspond to a signals without there being any underlying regularity.

1 The utterances alone are not enough, unless you assume the child can mind-read. Instead meaning-signal pairs are often communicated.

Consider the following language with alphabet $\{t, r, c, o, b, k\}$:

$$(\triangle, \color{#e8553e}\bullet) \mapsto to, \quad (\triangle, \color{#6ab0d4}\bullet) \mapsto tb \quad (\triangle, \bullet) \mapsto tk$$
$$(\square, \color{#e8553e}\bullet) \mapsto so, \quad (\square, \color{#6ab0d4}\bullet) \mapsto sb \quad (\square, \bullet) \mapsto sk$$
$$(\bigcirc, \color{#e8553e}\bullet) \mapsto co, \quad (\bigcirc, \color{#6ab0d4}\bullet) \mapsto cb \quad (\bigcirc, \bullet) \mapsto ck$$

This language is clearly compositional, since the first subsignal indicates the shape, (*t*riangle, *r*ectangle, *c*ircle) and the second subsignal the color (*o*range, *b*lue, blac*k*). In fact, that description is much more efficient:

$$\triangle \mapsto t, \quad \square \mapsto s, \quad \bigcirc \mapsto c$$
$$\color{#e8553e}\bullet \mapsto o, \quad \color{#6ab0d4}\bullet \mapsto b \quad \bullet \mapsto k$$

Rather than listing the signals corresponding to each of the $V^F = 3^3$ meanings (the worst case scenario for a holistic language), a compositional languages can be *compressed* to $F \cdot V = 2 \cdot 3$ rules listing to which *sub*signal every feature maps. That also means that one can faithfully reconstruct a compositional language from $F \cdot V$ signals, whereas it would need to observe *all* signals to reconstruct a holistic language (in the worst case). A compositional language is, in short, more *compressible* and as a result better *transmissible*.

TRANSMISSION BOTTLENECKS AND GENERALISATION In reality, children do not observe their entire language (e.g. all English sentences), but only a subset of it. They face a *transmission bottleneck*[2] better known as the *poverty of the stimulus*. If there is no such bottleneck all languages can be transmitted in their entirety, and faithfully so. The language can consequently not be changed by transmission and the initial language marks a *steady state*, maintained throughout all future generations. In the presence of a bottleneck, however, the learner is forced to *generalize* the observed data to a full language, in which case systematic errors can slowly accumulate.

The exact generalisation mechanism can take many different forms, such as (heuristic) grammar induction (Kirby 2001; Zuidema 2003), training a neural network (Kirby and Hurford 2002; Smith 2002) or constructing a finite state transducer (Brighton 2002). All these mechanisms try to discern some structure (e.g. compositionality) in the language. Sometimes, that allows the child to produce signals for unobserved meanings. But in other cases, the child is forced to invent a new signal. Incidentally, the new signal introduces a structure previously absent in the language. The next generation is then more likely to infer a language that reproduces that structure. As time passes, the differences between successive generations shrink and the language becomes more and more stable: the transmission bottleneck forced the language to become better transmissible. This is how the poverty of the stimulus solves the poverty of the stimulus (cf. Zuidema 2003).

A variety of different models confirmed this account. To name a symbolic and connectionist example, Kirby (2001) showed that a bottleneck caused the emergence of a stable, compositional language in agents representing language with definite-clause grammars. In another study, Kirby and Hurford (2002) found that the number of training instances passed between neural-network agents acted as a bottleneck, with

2 In fact, various different bottlenecks have been put forward; see Cornish (2011, ch. 4) for an overview and a discussion of the empirical findings regarding the presence of such a bottleneck.

a medium-sized training set leading to structured meaning-signal mappings. The fact that these, and many other, different models gave rise to similar behaviour is in itself striking. But it also makes it difficult to decipher what exactly is going on.

The shape-color example gives a hint, since there is a language much more compressible than a compositional one: the degenerate language that expresses *every* meaning with the same signal. The fact that none of the early studies seem to have produced degenerate languages, suggests that a bias against those must have been present (Cornish 2011). Or, conversely, that the learning algorithms implicitly pressured towards compositional languages. This opens up the possibility that "cultural evolution does no more than transparently map properties of the biology of an individual to properties of language" (Kirby 2017). Kirby points out that there are reasons to doubt this conclusion: The size of the bottleneck and the structure of the domain for example influence the simulations. Nonetheless, it became clear that in order to make claims about the shaping force of cultural evolution, one needs to know 1) what the *implicit biases* in the model are, 2) what the biases of the agents are and 3) how those interact with the cultural process.

Iterated learning with Bayesian agents

In 2005, Thomas Griffiths and Michael Kalish reinterpreted the iterated learning model in a population of Bayesian agents. One reason for doing so is that it connects the iterated learning model to a rich Bayesian modelling tradition in cognitive science (see e.g. Perfors et al. 2011; Goodman and Tenenbaum 2016; Griffiths, Kemp, and Tenenbaum 2008) and the formal models of human behaviour that have been proposed there. The Bayesian model of Griffiths and Kalish also solved the issues arising from implicit biases, since it *explicitly* encodes the biases of the learners. Moreover, the authors managed to characterise the long-term behaviour of the model — *convergence to the prior* — which sparked a small controversy. In the years that followed, the Bayesian paradigm appears to have surfaced as the primary approach to modelling iterated learning (Kirby, Griffiths, and Smith 2014; Kirby 2017). For that reason, and for its role in the next chapter, I want to go through the model in detail.

Recall from figure 2.1 that in iterated learning, a language alternates between a 'latent' internal representation θ and an 'overt' external representation x. Agents use a *production* and *language algorithm* (PA and LA) to move between these representations.[3] The idea put forward by Griffiths and Kalish (2007) is to model these production and language algorithms with probability distributions. An agent using language θ_t has a distribution $p_{\text{PA}}(x_t \mid \theta_t)$ over productions describing how to select a utterance. Conversely, it has a distribution $p_{\text{LA}}(\theta_t \mid x_{t-1})$ from which the agent picks a language after observing data x_{t-1} produced by the previous agent. Note that, as figure 2.1 illustrates, these are the only dependencies. Productions are conditionally independent from previous productions and the same goes for languages. This seems reasonable as an agent cannot use the previous production when making a new one (only its representation thereof) and clearly an agent cannot use the *unobservable* language of the previous agent directly. In short, iterated learning becomes a stochastic process on the random variables x_t and θ_t, which are conditionally independent from previous x_t's and

3 The language algorithm is usually called a *learning* algorithm. Since that terminology causes some confusion in chapter 4, I use the term *language algorithm*.

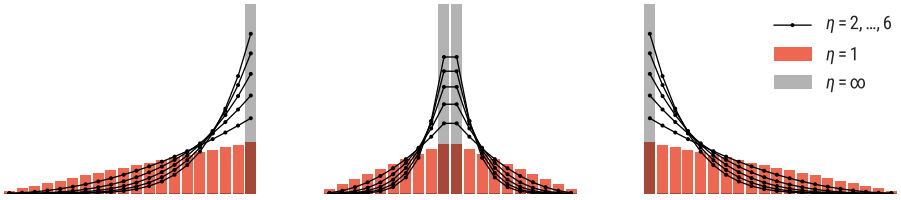


FIGURE 2.2 Exponentiating a distribution moves the probability mass towards the mode. Illustrated for three different distributions.

FIG03

θ_t's respectively.

What makes these agents 'Bayesian' is that their language algorithm reuses the the production algorithm and the prior beliefs of the agents using Bayes' rule. When confronted with data x_t, the agents infer the *posterior* distribution

$$p(\theta_t \mid x_{t-1}) \propto p_{\text{PA}}(x_{t-1} \mid \theta_t) \cdot p(\theta_t), \tag{2.1}$$

which captures how likely every language θ_t is in light of the observed the data. The posterior distribution balances two factors. First — and this is where the production algorithm is reused — how probable the agent itself regards the observed data to be, if it were to use language θ_t. This is the *likelihood* term $p(x_t \mid \theta_t)$. And second, how likely the language is in the first place: the *prior* $p(\theta_t)$.

Interestingly, *before* Griffiths and Kalish published their Bayesian interpretation, Kirby, Smith, and Brighton (2004) also noted that the language acquisition can be seen as Bayesian inference. The prior, they state, corresponds to Universal Grammar or the Language Acquisition Device: "everything the learner brings to the task *independent of the data*" (italics in original). However, Griffiths and Kalish (2007) stress that the prior "should not be interpreted as reflecting innate constraints *specific* to language acquisition" (my italics). The prior is, in other words, not necessarily domain specific, but aggregates all factors that influence language acquisition, including learned biases. Therefore, "the prior is better seen as determining the amount of evidence that a learner would need to see in order to adopt a particular language". Nevertheless many later papers use the prior primarily to capture innate learning biases (e.g. Kirby, Griffiths, and Smith 2014; Kirby 2017).

So how does a Bayesian agent adopt a particular language? Kirby, Smith, and Brighton (2004) assume agents pick the language with the highest probability under the posterior, the *maximum a posteriori* (MAP) estimate $\arg\max_\theta p(\theta \mid x)$. Griffiths and Kalish (2005), however used a different strategy where agents *sample* a language from their posterior, i.e. they are probability matching. The two strategies can be seen as extreme cases of a more general strategy: sampling from a *exponentiated* (or 'exaggerated') version of the posterior (Kirby, Dowman, and Griffiths 2007):

$$p_\eta(\theta_t \mid x_{n-1}) \propto p(\theta_t \mid x_{n-1})^\eta, \qquad \eta \geq 1. \tag{2.2}$$

For $\eta = 1$ this is the same as the sampling strategy, but as η increases, more and more of the probability mass is moved towards the maximum of the distribution (the mode) until sampling becomes indistinguishable from the MAP strategy (see figure 2.2). The language *algorithm* thus takes the posterior distribution and applies the language *strategy* (sample or maximise) to adopt a language.

THE EMERGENCE OF COMPOSITIONALITY, II It might be helpful to go through a concrete example. Griffiths and Kalish (2005) introduced a 'binary' language, which figured in several later studies (Griffiths, Canini, et al. 2007; Burkett and Griffiths 2010; Kirby, Tamariz, et al. 2015). It is a special case of the shape-color example introduced earlier, with two colours and two shapes (so $F = V = 2$). The language was introduced to study the emergence of compositionality. If we simplify the encoding, it is easier to see what the compositional languages are. Write 0 for a triangle, 1 for a square, 0 for black and 1 for orange, such that (□, ●) for instance becomes 10 and (△, ●) becomes 01. Using alphabet $\{a, b\}$ there are 4 compositional languages given by the feature-subsignal mappings

$$
\begin{aligned}
&(1) \qquad 0 \mapsto a, \quad 1 \mapsto a \\
&(2) \qquad 0 \mapsto a, \quad 1 \mapsto b \\
&(3) \qquad 0 \mapsto b, \quad 1 \mapsto a \\
&(4) \qquad 0 \mapsto b, \quad 1 \mapsto b
\end{aligned}
$$

In this scenario there are 4 meanings (▲, ▲, ■, ■) and $4^4 = 256$ ways to map four meanings to four signals $\{aa, ab, ba, bb\}$. This gives 256 languages of which 4 compositional and 252 holistic.

Not all languages are equally likely. A hierarchical prior that puts a fraction α of the probability mass on the compositional languages:

$$
p(\theta) = \begin{cases} \frac{\alpha}{4} & \text{if } \theta \text{ is compositional} \\ \frac{1-\alpha}{256} & \text{otherwise} \end{cases} \tag{2.3}
$$

Once an language θ has been fixed, the agent is presented with new meaning m for which it then produces a signal s by sampling from the distribution

$$
p(s \mid m, z) = \begin{cases} 1 - \varepsilon & \text{if } m \mapsto s \text{ in language } \theta \\ \varepsilon/3 & \text{otherwise} \end{cases} \tag{2.4}
$$

This means the agent will pick the signal s corresponding to m under language θ most of the time, but has a small probability ε of making an error and uniformly picking one of the other signals. Together with a completely independent distribution $p(m)$, typically a uniform one, this specifies the production algorithm

$$
p_{\text{PA}}(x \mid \theta) = p(s \mid m, \theta) \cdot p(m), \qquad x = (m, s) \tag{2.5}
$$

If $\mathrm{x} = ((m_1, s_1), \ldots (m_b, s_b))$ is the list of the utterances produced by the previous agent, then the posterior distribution is

$$
p(\theta \mid \mathrm{x}) \propto p(\theta) \cdot \prod_{i=1}^{b} p_{\text{PA}}(x_i \mid \theta), \tag{2.6}
$$

and, as usual, the language algorithm takes the form $p_{\text{LA}}(\theta \mid \mathrm{x}) \propto p(\theta \mid \mathrm{x})^{\eta}$.

Figure 2.3 illustrates the resulting simulation in a population of samplers ($\eta = 1$). It shows which language was used in every generation (left): one of the 252 holistic languages (H) or a compositional language (C1–4). The compositional languages seem

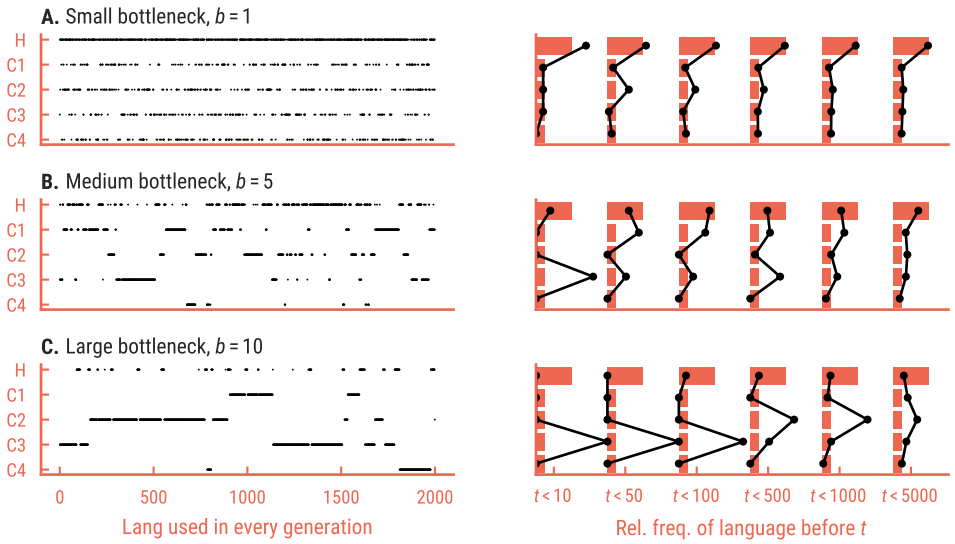


FIGURE 2.3 Emergence of compositionality in the Bayesian iterated learning model of Griffiths and Kalish (2007). On the left, the language used in every generation with H one of 252 holistic languages and C1–4 the compositional languages. On the right the relative frequency of every language up to a certain time t. These relative frequencies converge to the prior (orange). Larger bottlenecks (subfigures A–C) slow down convergence.

GK01 WebPPL simulation with $\alpha = 0.5$, $\varepsilon = 0.001$ and samplers ($\eta = 1$).

to be used much more frequently, which is confirmed by the plots on the right. There we see the relative frequency of every language up to several points t in the simulation. These plots indicate that the relative frequencies converge to the prior, shown in orange. Since the compositional languages have a higher prior probability than each of the holistic languages, they are more frequent. The convergence rate towards the prior is much faster when the bottleneck is small ($b = 1$, subfigure A) than when it is large ($b = 10$, subfigure C). It is clear why this happens: the more data is transmitted, the greater the probability that the child can reconstruct the language. The result is that languages will be stable throughout multiple generations, as seen from the lines in figure 2.3C. Nevertheless, even with a large bottleneck the relative frequencies seem to converge to the prior, be it very slowly. We will discuss all these findings in more detail later. First, what discuss the observed 'convergence to the prior'.

Convergence to the prior

Let me briefly summarise what we have seen so far. Bayesian agents observe utterances x_{t-1} produced by the previous agent, and then use Bayes' rule to infer a language. This language is θ is drawn from $p_\eta(\theta_t \mid x_{t-1}) = p(\theta_t \mid= x)^\eta$, where η interpolates between a sampling- and MAP-strategy for $\eta = 1$ and $\eta = \infty$ respectively. All this results in a chain of the form

$$x_0 \longrightarrow \theta_1 \longrightarrow x_1 \longrightarrow \theta_2 \longrightarrow x_2 \cdots . \tag{2.7}$$

Griffiths and Kalish (2005) noted that several Markov chains can be discerned in eq. 2.7, of which the long-term behaviour is well-studied: They often converge to a so called stationary distribution. This characterised the long-term behaviour of the iterated learning model.

Appendix A introduces the relevant convergence results for Markov Chains; I only summarise them here. Consider a system with a set of possible states S. If the vari-

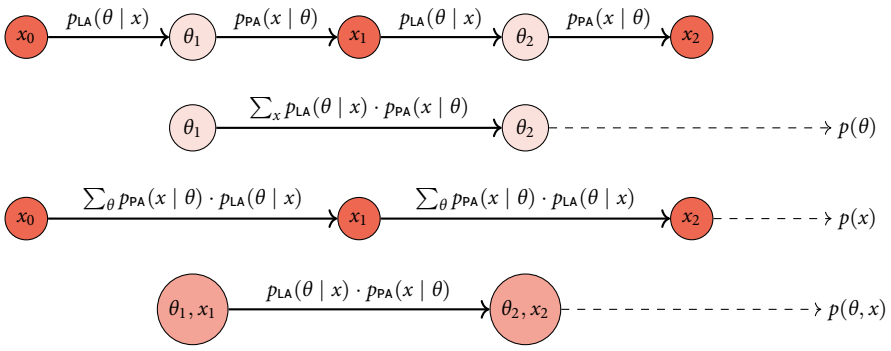


FIGURE 2.4 Different Markov chains hidden in the Bayesian iterated learning model, and to which stationary distribution they converge (right).

Figure adapted from Griffiths and Kalish (2007).

ables $x_0, x_1, x_2, \ldots$ indicate the state at every time step, they form a Markov chain if the probability of moving to another state only depends on the last state: $p(x_t \mid x_0, \ldots, x_{t-1}) = p(x_t \mid x_{t-1})$. If the number of states is finite, these *transition probabilities* can be collected in the transition matrix T. Suppose the initial distribution over states is given by vector π, then the next distribution is $p(x_1 = i) = (\mathrm{T}\pi)_i$ and after t steps, $p(x_t = i) = (\mathrm{T}^t\pi)_i$. These probabilities can converge to the so called *stationary distribution* $\pi^\star$ which must be an eigenvector of T since $\mathrm{T}\pi^\star = \pi^\star$. If the Markov chain is *ergodic* it is guaranteed to have a unique stationary distribution to which it converges: $p(x_t = i) \to \pi^\star_i$ as $t \to \infty$. Ergodicity, briefly, ensures that the chain keeps revisiting the entire state space and has a positive probability of reaching any other state from any given state in a finite number of steps. How often it visits every state is given by the stationary distribution, in the sense that the relative frequencies of visited states converges to the stationary distribution.

PROOF OF THE CONVERGENCE TO THE PRIOR Griffiths and Kalish (2005) noted that by marginalising out the productions x_t in eq. 2.7 one obtains the following Markov chain (see also figure 2.4):

$$p(\theta_t \mid \theta_{t-1}) = \sum_{x_{t-1}} p_{LA}(\theta_t \mid x_{t-1}) \cdot p_{PA}(x_{t-1} \mid \theta_{t-1}). \tag{2.8}$$

We hitherto assumed that the transition probabilities remain constant over time, that is, we looked at time-homogeneous chains. The Markov chain in eq. 2.8 is only homogeneous if all agents use the same production and language algorithms. In particular, they should all use the same prior. We will later discuss the validity of this assumption. If these assumptions hold and the chain is moreover ergodic, *then* the long-term behaviour of iterated learning is known: convergence to the stationary distribution, independent of the initial distribution.

The stationary distribution $\pi^\star$ of this distribution happens to be the prior $q(\theta) := p(\theta)$. To show this, one has to see that

$$q(\theta_{t+1}) = \sum_{\theta_t} p(\theta_{t+1} \mid \theta_t) \cdot q(\theta_t) \tag{2.9}$$

I have written q for the prior to highlight that we do not know whether $q(\theta_{t+1})$ is a marginal distribution of $p(\theta_{t+1}, \theta_t)$. In that case, the equality would hold trivially. Oth-

erwise, the following derivation shows that eq. 2.9 holds:

$$\sum_{\theta_t} q(\theta_t) \cdot p(\theta_{t+1} \mid \theta_t) = \sum_{\theta_t} q(\theta_t) \cdot \sum_{x_t} p_{\text{LA}}(\theta_{t+1} \mid x_t) \cdot p_{\text{PA}}(x_t \mid \theta_t) \tag{2.10}$$

$$= \sum_{x_t} p_{\text{LA}}(\theta_{t+1} \mid x_t) \cdot \sum_{\theta_t} q(\theta_t) \cdot p_{\text{PA}}(x_t \mid \theta_t) \tag{2.11}$$

$$= \sum_{x_t} p_{\text{LA}}(\theta_{t+1} \mid x_t) \cdot p_{\text{PA}}(x_t)$$

$$\overset{(\star)}{=} \sum_{x_t} \frac{p_{\text{PA}}(x_t \mid \theta_{t+1}) \cdot q(\theta_{t+1})}{p_{\text{PA}}(x_t)} \cdot p_{\text{PA}}(x_t)$$

$$= q(\theta_{t+1}) \sum_{x_t} p_{\text{PA}}(x_t \mid \theta_{t+1})$$

$$= q(\theta_{t+1})$$

where $(\star)$ holds by definition of $p_{\text{LA}}(\theta_{t+1} \mid x_t)$ and because we use samplers ($\eta = 1$). For maximisers, the proof breaks down at this point.

Similar results hold for the other Markov chains hidden in the iterated learning model (see figure 2.4). When averaging over interpretations rather than productions, one obtains a Markov chain on the productions:

$$p(x_{t+1} \mid x_t) = \sum_{\theta_{t+1}} p(x_{t+1} \mid \theta_{t+1}) \cdot p(\theta_{t+1} \mid x_t). \tag{2.12}$$

A proof analogous to eq. 2.10 shows that this chain converges to the *prior predictive distribution* $p(x) = \sum_\theta p_{\text{PA}}(x \mid \theta) \cdot p(\theta)$ Finally, one could consider a Markov chain over the state space of language-utterance pairs $(\theta, x) \in \Theta \times \mathcal{X}$ with transition probabilities

$$p(\theta_{t+1}, x_{t+1} \mid \theta_t, x_t) = p(\theta_{t+1} \mid x_t) \cdot p(x_t \mid \theta_t). \tag{2.13}$$

This chain has the joint $p(\theta, x) = p_{\text{PA}}(x \mid \theta) \cdot p(\theta)$ as its stationary distribution. Interestingly, this shows that Bayesian iterated learning implements a *Gibbs sampler*.

Gibbs samplers are often used in Bayesian statistics, whenever it is not possible to work with complicated distributions analytically. *Monte Carlo methods* are work-arounds that collect many samples from the distribution, and approximate the distribution using those samples. To obtain samples, one constructs a Markov chain whose stationary distribution is the distribution of interest. Over time, the visited states will be (correlated) samples from the target distribution. This is the basic idea behind many *Markov Chain Monte Carlo* (MCMC) methods and Gibbs sampling is one of those. It can be used to approximate a joint distribution $p(\theta, x)$ if it is easy to sample from the conditional distributions $p(\theta \mid x)$ and $p(x \mid \theta)$. In every iteration, it fixes one of the variables, say θ_t and samples a new x_{t+1} from $p(x_{t+1} \mid \theta_t)$. Then it fixes x_t and samples θ_{t+1} from $p(\theta_{t+1} \mid x_{t+1})$, and so on. This results in a new sample (θ_{t+1}, x_{t+1}) from the joint after every 'sweep' through the variables. Indeed, this procedure exactly mirrors Bayesian iterated learning with sampling agents, and it follows that the chain in eq. 2.13 converges to $p(\theta, x)$ (see Griffiths and Kalish 2007 for a longer discussion).

CONVERGENCE TO THE MAXIMUM OF THE PRIOR? What kind of behaviour should one expect in populations of maximisers? This turns out to be a much harder question. There are, to the best of my knowledge, two analytical results — we will return to empirical evaluations in chapter 4 — both suggesting that in populations of maximisers the behaviour is largely determined by the prior, but in a less direct way. First of all, Kirby, Dowman, and Griffiths (2007) analyses the stationary distribution for maximisers ($\eta > 1$) using a constrained set of languages that spread the probability mass uniformly over a (sub)set of utterances.[4] In other words, $p(x \mid \theta)$ is either 0 or equal to a $f(x)$, where the latter does not depend on θ. In that case, the stationary distribution is proportional to $p(z)^{\eta}$. This implies that cultural evolution results in an exaggerated version of the prior (cf. figure 2.2).

A similar conclusion follows from the second result, due to Griffiths and Kalish (2007). They note that maximisers (now $\eta = \infty$) implement a version *Expectation-Maximisation* (EM). This is an iterative algorithm used in models with hidden variables to estimate parameters that are increasingly close to the maximum likelihood estimates, or, in our case, MAP estimates. The trick is to use the current parameters to estimate the *expected* likelihood of the observed and hidden variables, and then update the parameters so that they *maximize* that likelihood. When computing the expectation analytically is intractible, it can be approximated by drawing several samples. The case using a single sample is called *stochastic EM*. Now, suppose, in EM jargon, there are no observed variables, x_t is the latent variable and θ_t the parameter, then stochastic EM in this model amounts to Bayesian iterated learning in a population of maximisers (see Griffiths and Kalish (2007) for details). This characterisation is not as clear-cut as with samplers, but suggests that the stationary distribution over languages will roughly be centred on the maxima of the prior (Griffiths and Kalish 2007).

Convergent controversy

The *convergence to the prior* was the first general result about the long-term behaviour of the iterated learning model. For populations of samplers, the result was crystal clear: starting from any initial distribution, the probability that an agent down the chain would be using language θ is given by the prior probability $p(\theta)$. And this is precisely what we observed in figure 2.3, which shows the the emergence of compositionality — or rather, the emergence of the prior. The model is an ergodic Markov chain, and over time the probability that a certain language will be used therefore converges to its probability under the stationary distribution, which is the prior. Compositional languages have high probability under that prior, and consequently emerge. Maximisers are much harder to analyse. The probability that a language is used by maximisers seems to be largely determined by the maxima of the prior. Now, what are the implications of all this for cultural language evolution?

4 This constraint on languages has a purely mathematical motivation: it is precisely what is needed to factorise the normalising constant in the posterior.

BOTTLENECKS AND WEAK BIASES Iterated learning was inspired by the idea that language is a compromise between "the biases of learners, and other constraints acting on language during their transmission" (Smith 2009), originally in the form of a transmission bottleneck. But in the Bayesian models, the bottleneck hardly plays any role.

Griffiths and Kalish (2007) conclude that "the emergence of languages with particular properties does not require a bottleneck" (p. 466). Larger bottlenecks do slow down convergence since they imply more faithful transmission and this increases language stability. The Markov chain's walk through the state space consequently slows down, which, somewhat paradoxically, also slows down convergence. But in the long run bottlenecks play no role — at least for samplers. This seems to undermine the idea that compressible languages emerge *because of* cultural transmission. Should we conclude, then, that languages are not shaped by cultural evolution, but primarily by innate constraints? Griffiths and Kalish (2007) conclude that their results "do not indicate which of these explanations is more plausible" (p. 475). There's something for everyone: if the prior captures innate biases, "iterated learning acts as an engine by which these constraints result in universals" (p. 475), but if you prefer the transmission process to actually change the priors, then you "can take heart from our results for learners who use MAP estimation".

Kirby, Dowman, and Griffiths (2007) follow the latter advise. Their paper discusses an iterated learning model with maximisers that have a prior bias towards regular languages. Bottleneck effects can occur in populations of maximisers (Griffiths and Kalish 2007) and the authors accordingly conclude that as the bottleneck tightens in their model, "regularity is increasingly favoured". But there is something peculiar about this conclusion: It seems to hold only because their prior favoured regularity. Had their prior favoured irregularity, irregularity would have been increasingly favoured under a tighter bottleneck.[5] In the Bayesian model, transmission at most amplifies pre-existing biases, which of course can be seen as an effect of cultural transmission. Another conclusion of Kirby, Dowman, and Griffiths (2007) is therefore that processes of cultural evolution can "completely obscure" the *strength* of the bias. A small tendency to favour languages with higher prior probability (i.e. $\eta > 1$) amplifies weak biases and results in strong universals. The strength of the bias has no role, only the ordering of the languages. All in all, it suggests a rather toothless process of cultural evolution. Several researchers therefore started tweaking the assumptions of the model to find out how robust the results are.

POPULATION STRUCTURE AND HETEROGENOUS POPULATIONS The population structure was one of the first things addressed. It should be noted that (Griffiths and Kalish 2007) generalised their findings to somewhat different scenario, with finite generations evolving in (discrete or) continuous time (cf. Nowak, Komarova, and Niyogi 2001). In that case the *proportion* $p_t(\theta)$ of the population speaking language θ at time t converges to the prior $p(\theta)$, as can easily be seen. If $\mathrm{p}_t = (p_t(\theta) : \theta \in \Theta)$ and T the transition matrix, these proportions change as

$$\mathrm{p}_{t+1} = \mathrm{T}\mathrm{p}_t, \tag{2.14}$$

which describes a linear dynamical system with a unique stable equilibrium. The same derivations as eq. 2.10 show the prior is that equilibrium. However, Niyogi and Berwick (2009) argue that this is an unrealistic model of language evolution as it precludes the possibility of bifurcations. Moreover, language stability cannot be maintained: even if only 0.01% of the population uses a different language, it will spread to a larger share of the population (the prior admitting). As a remedy Niyogi and Berwick (2009) propose

5 I found their PNAS paper is a bit sketchy on the details of their simulations, but these conclusions follow directly from Griffiths and Kalish (2007) and as far as I can see apply equally to Kirby, Dowman, and Griffiths (2007).

an alternative model where agents learn from a mixture of the languages used in the previous generation, not just one. This leads to markedly different nonlinear behaviour with bifurcations and possibility multiple equilibria, which they argue accurately describes historical developments (namely, that English is no longer a 'verb-second' language).

That the behaviour changes in different populations structures was confirmed in several other studies. Smith (2009) similarly considered infinite generations of agents learning from multiple parents. He reports that this precludes convergence to the prior and introduces a dependency on the initial distribution of languages in the population. Ferdinand and Zuidema (2009) draw the same conclusion, but also drop the assumption that all agents share the same innate biases, i.e. that the population is *homogeneous*. In heterogeneous population the convergence to the prior breaks down. Dediu (2009) finds that the strong differences between samplers and maximisers disappears in populations with a different structure or heterogeneity.

The agents in studies such as Ferdinand and Zuidema (2009) are not Bayesian agents in the strict sense that agents assume to be learning from a single language, while in fact the data comes from several sources. Burkett and Griffiths (2010) address this issue in a hierarchical model where agents take into account that they are possibly learning from multiple languages. Accordingly, the convergence to the prior reappears. Very recently, Whalen and Griffiths (2017) extended this to populations with arbitrary network structures, although it should be stressed that agents still learned from a single teacher. Nevertheless, the emerging consensus appears to be that in slightly more complicated population structures (with possibly imperfect Bayesian reasoners) the convergence to the prior can break down and nontrivial cultural effects appear.

LINEAGES AND CUMULATIVE CULTURAL EVOLUTION It is somewhat surprising that the population structure received most criticism, since that aspect of the Bayesian model is perfectly in line with the original iterated learning model. Some other parts, I would argue, are not. First of all, the type of convergence — in language or in probability of using a language — is markedly different. In early iterated learning studies, the population converged to a stable language which could be transmitted faithfully along many generations. In the Bayesian models, nothing of this sort happens. In the simulation of the emergence of compositionality (figure 2.3) one clearly sees that successive generations can acquire radically different languages: picture English-speaking parents, themselves born to Basque parents, whose children miraculously learned Hungarian.

Transmission in the Bayesian model generally not faithful — indeed, this is necessary for 'convergence' to occur at all. That seems particularly problematic for a model of cultural evolution. Even if transmission shapes languages, it has to be somewhat faithful if one expects any kind of cultural *evolution*. Tomasello (1999) points out that faithful transmission is important because it enables a so called *cultural ratchet*, where cultural innovations are passed on and improved upon by later generations. Cultural evolution, as a result, is *cumulative* and products of cultural evolution consequently reflect their full historical development. If it is not already uneasy that the defining property of a Markov chain is being memoryless, ergodicity certainly conflicts with the idea of cumulative cultural evolution. In an ergodic Markov chain, every 'lineage' is guaranteed to revisit all possible languages infinitely often. That amounts to an infi-

nite reinvention of the wheel — pretty much the exact opposite of cumulative cultural evolution.

Conclusions

The first iterated learning models suggested that languages primarily pick up systematicity during cultural transmission. Simulations showed how compositional structure accumulated in initially unstructured languages when a bottleneck pressured the languages to become more compressible. However, the learning algorithms that generalise a few observations to a full language implement all kinds of implicit biases, and possibly provide an implicit pressure towards compositional structures. To make general claims about the interaction of cultural processes and innate biases, the two need to be separated clearly. Bayesian iterated learning models did precisely that, but were also shown to *converge to the prior*. That meant that the probability that a certain language would be used, is after a while completely determined by the biases of the learners, independent of the initial conditions. In populations of maximisers, the relation is less transparent and the shape of the prior (its maxima in particular) largely appears to determine the outcome of cultural evolution.

The Bayesian iterated learning model moved the explanatory load from the cultural process to the prior biases of the learners. However, the strong conclusions were in several studies shown to break down in more complicated populations. The Bayesian model moreover results in an arguably unrealistic model of cultural evolution, with no language stability, nor any cumulative effects. Despite these shortcomings, the field made significant progress due to the work of Griffiths and Kalish. Explicitly encoding the biases of the learners made studies of the interactions between the 'nature' and 'nurture' of language much more principled, and moreover resulted in cognitively better motivated agents. The focus on analytic results regarding the long-term behaviour brought further transparency to the somewhat opaque conclusions suggested by simulations alone — irrespective of whether the results are ultimately convincing.

In sum, combining the criticism and benefits, I would draw up the following list of desiderata for a model that aims to show that cultural processes can shape the evolution of language (in arbitrary order):

(D1) **Explicate biases.** The biases of the agents should be explicitly specified in the model.
(D2) **Strategies.** The model should explore a wide range of strategies, such as sampling or MAP strategies.
(D3) **Analysable.** The model should be amenable to analytical scrutiny, and it should ideally be possible to draw general conclusions about long-term behaviour.
(D4) **Nontrivial cultural effects.** The model should exhibit non-trivial cultural effects, which might for example result in lineage-specific evolution: different runs resulting in different outcomes of cultural evolution.
(D5) **Robustness to population structure.** The model should exhibit behaviour that is fairly robust to changes in population structure.

(D6) **Language stability.** The model should result in a 'reasonable' degree of language stability. Reasonable, since languages are never perfectly stable (see also Kirby 2001).

(D7) **Empirically testable.** The model should give an empirically plausible, mechanistic explanation of cultural evolution, which is further testable against empirical linguistic findings (predating the lab).

The list is no doubt incomplete, but mainly serves as a guide to what I will address in this thesis, most notably in chapter 4.

3 Naming Games

How can a population negotiate a shared language without central coordination? This is the terrain of naming games, the second class of agent-based models. In local, horizontal interactions, agents 'align' their language until they reach coherence. We discuss several alignment strategies, some of which return in later chapters, and conclude with a proof suggesting that a stable, single-word language always emerges. The model used therein is the stepping stone for the next chapter, where we connect naming games to Bayesian models of iterated learning.

Naming games (NG) or language games were pioneered in the 90s by Luc Steels and colleagues. The view of language that motivated their work was similar to the views expressed in the iterated learning literature. As Steels (1995) puts it, "language is an autonomous adaptive system, which forms itself in a self-organising process". However, language games approach the adaptive system from a different angle than iterated learning. The development of linguistic structure is not primarily driven by transmission, as Kirby and others proposed, but "by the need to optimise *communicative success*" (p. 319, my italics). The central question takes the form (Steels 2011): how can a convention of some sort (lexical, grammatical, or otherwise) emerge in and spread through a population as a result of local communicative interactions, that is, without central coordination? So if iterated learning is a model of *vertical* language evolution, then the naming games model *horizontal* language evolution.

One of the first studies to explore this, Steels (1995), used a game in which (software) agents negotiated a spatial vocabulary. Equipped with a primitive perceptual apparatus, the agents learned to identify each other by name or spatial position in a shared simulated environment. Later research extended this approach to embodied robotic agents, grounding their 'language' in the physical world. These *grounded naming games* (Steels 2012; Steels 2015) introduce additional complexities pertaining to the perceptual and motor systems of the robots. We focus on non-grounded games, which can be divided into two branches. The first is centred around the *minimal naming game*, studied extensively using methods from statistical physics. The second extended the first naming games to more complex and possibly realistic linguistic scenarios. This chapter discusses and compares both branches. Of particular interest is the kind of dynamics one can expect from these models. We therefore conclude with the proof by De Vylder and Tuyls (2006) suggesting that naming games always converge to a stable, single-word language.

The basic naming game

Picture a group of people encountering a colourless green object for which they do not have a name. Of even worse, suppose they don't have a shared language at all. Confused, I suppose, they furiously shout out names for the object. But can they gradually align their vocabularies by carefully attending to what the others are saying, until they have agreed on a word for the object — *gavagai*, perhaps?

Frivolities aside, this is the essence of the naming game. It imagines a population of N agents in a shared environment filled with objects, which the agents try to name. At the start of the game, there is no agreement whatsoever about the names of the objects. Every agent has an inventory of names for the objects (a lexicon), which is adjusted after every round with the goal of increasing communicative success. In every round, two randomly selected agents interact, one as speaker, one as the hearer, according to the following script (Wellens 2012):

1. The speaker selects one of the objects which is to serve as the *topic* of the interaction. She[6] produces a name for the object, either by using one of the names she already knew, or by inventing a new name.

6 'Gender' is only introduced to convieniently disambiguate the intended agent: the speaker (she) or the *h*earer (he). This even puts the 'men' in the role of listener — which I belief is sometimes regarded to be the appropriate role.

A. Failed communication

SPEAKER	HEARER	⟹	SPEAKER	HEARER
Gavagai	Spam		**Gavagai**	Spam
Cofveve	Foo		Cofveve	Foo
Spam			Spam	**Gavagai**

B. Successful communication

SPEAKER	HEARER	⟹	SPEAKER	HEARER
Gavagai	**Spam**		**Spam**	**Spam**
Cofveve	Foo			
Spam				

FIGURE 3.1 The updates of the minimal naming game illustrated. If communication fails, the hearer adds the word uttered by the speaker (bold) to its vocabulary. After a success, both empty their vocabularies and keep only the communicated word.

Figure inspired by Wellens (2012).

2. The hearer receives the word, interprets it and points to the object he believes was intended.
3. The speaker indicates whether she agrees or disagrees, in that way signalling whether communication was successful.
4. Both the speaker and hearer can update their inventories.

The script is a broad outline and concrete implementations are more specific. How, for example, does the speaker select a word in step 1? The typical assumption is that the speaker uses her own experience as a proxy of the hearer's inventory and opts for a signal she would likely interpret correctly herself. This is a so called *obverter* strategy (Oliphant and Batali 1996). Or more importantly, how do the speaker and hearer update their lexicons after the encounter? Here, the sky is the limit. Does the speaker update her lexicon, or the hearer, or both? What happens after successful communications, what after failure? In years of research, one particular script emerged, which is discussed below. It also became clear that whichever update strategy is used, it must improve the *alignment* between the lexicons Steels (2011). That means that the probability that a future encounter will be successful is increased. Such strategies thus reinforce successfully communicated words and this often installs a winner-takes all dynamics which, in the end, leads to a (unique) shared convention. This is best seen in the so called *minimal naming game*.

The minimal strategy

The *minimal naming game* was introduced by statistical physicist Andrea Baronchelli (2006) and simplifies earlier naming game in several respects (Baronchelli, Felici, et al. 2006). First, it assumes that homonymy cannot occur. Homonymy can only be introduced when a speaker invents a *new* word for an object that happens to have been used already to name another object. If the space of possible new words is large enough, we can safely assume that invented words are unique and homonymy will be absent. Secondly, one can assume, without loss of generality, that there is only one object. If there is no homonymy, the update in step 4 will never affect words used for a different object. The competition between the synonyms for a particular object is thus completely independent from other objects. As a result, the dynamics of a naming game with multiple objects is fully determined by the dynamics of a game with a single object.

In the minimal naming game, the inventory of every agent is a list of words. In step 1, the speaker select one word uniformly at random from her inventory. The update in

step 4 distinguishes two cases.

- **Success.** If the hearer knows the word, communication is successful. Both hearer and speaker remove all *other* words from their inventories, yielding two perfectly aligned inventories with one single word.
- **Failure.** If the hearer does not know the word, communication fails and the hearer adds the word to his lexicon.

Figure 3.1 illustrates how the inventories of agents change after failed and successful communication. The dynamics of the games can be studied by collecting several statistics (cf. Baronchelli 2017; Wellens 2012), typically with a certain resolution (e.g. after every 10 rounds). Concretely, we measure the following:

- **(Probability of) communicative success** $p_s(t)$. The probability that an interaction at time t is successful. These probabilities are estimated by averaging this binary variable over many runs.
- **Total word count** $N_{\text{total}}(t)$ The total number of words used in the population at time t. Some authors prefer to divide it by the population size to get the average number of words per agent.
- **Unique word count** $N_{\text{unique}}(t)$**.** The number of unique words used in the population at time t.

Due to the stochasticity of the games, individual runs vary substantially and can obscure underlying regularities. Conversely, the behaviour of a single run can suggest regularities that do not generalise. For that reason, we study the average behaviour of the games, obtained by averaging over many simulation runs.

PHENOMENOLOGY The minimal naming game goes through three distinct phases, as illustrated in figure 3.2. In the first phase, most interacting agents will have empty vocabularies and thus invent new words. This results in a sharp increase of the number of unique words N_{unique} in the population. In the second phase, no new words are invented, but the invented words spread through the population. Alignment is still low and words will rare rarely be eliminated, so N_{total} keeps growing. In the third phase, after the peak of N_{total}, this changes. Interactions are increasingly likely to be successful, leading to a sharp increase in communicative success and a drop in N_{total} as more and more words are eliminated. This also results in the characteristic S-shaped curve of p_{success}. Eventually the population reaches coherence in the absorbing state where all agents share one unique word and reach perfect communicative success ($N_{\text{unique}} = 1$, $N_{\text{total}} = N$ and $p_{\text{success}} = 1$).

The game has two important properties, that one might call *effectiveness* and *efficiency*. The resulting communication system is *effective* because agents learn to communicate successfully, and *efficient* in the sense that agents do not memorise more words than strictly necessary (one, in this case). A simple argument shows that the minimal naming game almost always reaches an efficient and effective stable state (Baronchelli, Felici, et al. 2006). At any point in the game, there is a positive probability of reaching coherence in $2(N-1)$ steps: pick one speaker and let her speak to all other $N-1$ agents twice. The first time, a hearer might still have to adopt the word, but after the second interaction only one word will remain in his inventory. If p is the probability

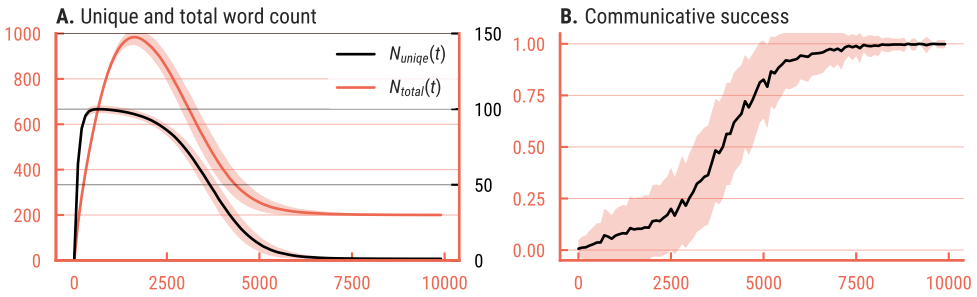


FIGURE 3.2 The dynamics of the minimal naming game. An sharp transition leads to convergence and the emergence of consensus.

MNG01 Results shown for $N = 200$; avg. of 300 runs, 1 std. shaded.

of this (unlikely) sequence of interactions, the probability that it has not occurred after $k \cdot 2(N-1)$ steps is less than $(1-p)^k$, which decreases exponentially in k. With probability 1, the population will thus reach coherence as $k \to \infty$. The argument is somewhat unsatisfactory as it does not reveal anything about the dynamics: how fast is the convergence, for example?

SCALING RELATIONS AND NETWORK STRUCTURE To obtain a better insight in the dynamics, one can adopt a methodology commonly used in statistical physics and look at *scaling relations*. The question is then how certain quantities, like convergence time, *scale* with the size of the system, i.e. the number of agents. To that end, two critical points are identified: the time t_{conv} where the game reaches coherence and the time $t_{\max}$ at which point $N_{\text{total}}(t)$ reaches its maximum. It turns out that these quantities depend on the population size N in a power-law fashion (Baronchelli, Felici, et al. 2006; Loreto et al. 2011):

$$t_{\text{conv}},\ t_{\max},\ N_{\text{total}}(t_{\max}) \propto N^{\alpha} \qquad \text{where } \alpha \approx 1.5 \tag{3.1}$$

Now note that $N_{\text{total}}(t_{\max})/N$ is the maximum number of words each agent has to store on average — the maximum memory load, perhaps. Baronchelli (2017) concludes that "the cognitive effort an agent has to take, in terms of maximum inventory size, *depends on the system size* and, in particular, diverges as the population gets larger" (Baronchelli 2017, italics in original). Although interesting, I would be hesitant to concede that linguistic activity in a small language community requires less cognitive effort than the same activity in a larger community.

Besides the scaling effects, the role of the network structure of the population has been studied extensively (see Baronchelli 2017, for an overview). In the classical naming game any two agents can interact — there is *homogeneous mixing* — corresponding to a fully connected social network. Varying the topology (to e.g. more realistic small-world networks, Dall'Asta et al. 2006) strongly influences the dynamics. This is reflected by different scaling relations, but not by convergence per se: the population still negotiates a unique word — as long as the networks remains connected, of course.

TABLE 3.1 Parameter settings for four different strategies, whose behaviour is shown in figure 3.3. Note that equivalent parametrisations also exist; see main text for details.

	δ_{inc}	δ_{inh}	δ_{dec}	s_{init}	s_{max}
MINIMAL STRATEGY	0	1	0	1	1
LAT. INHIBITION STRATEGY 1	1	1	0	1	∞
LAT. INHIBITION STRATEGY 2	0.1	0.5	0.1	0.5	1
LAT. INHIBITION STRATEGY 3	0.1	0.2	0.2	0.5	1
FREQUENCY STRATEGY	1	0	0	1	∞

Lateral inhibition strategies

The minimal strategy is somewhat opportunistic in that it forgets all other words after a successful encounter. It has been suggested that subtler alignment mechanisms might yield faster convergence times: so called *lateral inhibition strategies* (see Wellens 2012 ch. 2, for an overview). The name is ultimately derived from biology, where excited neurons can be found to *inhibit* neighbouring neurons. Similarly, lateral inhibition strategies decrease the chance of using competing words again. To that end, they assign a *score* to every word. If a word is communicated successfully, its score is increased, and the scores of competitors are decreased or *inhibited*. The production mechanism must also accounts for the scores, typically by producing the highest-scoring word.

The (basic) lateral inhibition strategy was first formulated in Steels and Belpaeme (2005) and is described by five nonnegative parameters (Wellens 2012)[7]

$$\delta_{\text{inc}}, \quad \delta_{\text{inh}}, \quad \delta_{\text{dec}}, \quad s_{\text{init}}, \quad s_{\text{max}}. \tag{3.2}$$

After a success, both agents increase the score of the communicated word by δ_{inc} and decrease scores of competitors by δ_{inh}. After a failure, the hearer adopts the word with score s_{init} and the speaker decreases the score by δ_{dec}. Whenever a score drops below (or equals) 0 the word is removed, and scores can never grow larger than s_{max}. Other inhibition strategies have also been used and will be discussed in chapter 7.

The minimal strategy is a special case of the lateral inhibition strategy, for $\delta_{\text{inc}} = \delta_{\text{dec}} = 0$ and $\delta_{\text{inh}} = s_{\text{init}} = 1$ (see also table 3.1). With those parameters new words get score 1 and this score is never further increased. It *can* be inhibited, by 1, which leads to to immediate removal. In this strategy, the scores thus play a purely administrative role. A strategy where scores play a larger role, is the *frequency strategy* which counts how often every word has been encountered. This strategy however exhibits no form of lateral inhibition. The minimal strategy and frequency strategy thus mark two extremes: the former has the strongest possible form of lateral inhibition, the latter none. Between these endpoints lie the proper lateral inhibition strategies.

I want to discuss three fairly different LI strategies here: LI strategy 1 is a strategy that returns in chapter 6; strategy 2 is taken from Wellens (2012); and strategy 3 is a variation thereof. The parameters are listed in table 3.1 and figure 3.3 shows the dynamics. First of all note that the dynamics of N_{unique} can strongly differ for different strategies (subfigure A). If for example $\delta_{\text{inh}} = \delta_{\text{dec}}$ as in LI strategy 3, many more words can be invented. But eventually this strategy gives rise to an efficient language. So do all other strategies, except that the frequency strategy results in a maximally inefficient languages where all agents know all words. Since agents only use the most frequent word, perfect communicative accuracy is still attained, as is the case for the other strategies.

7 Wellens (2012) only uses δ's in $(0, 1)$, but this general formulation allows the inclusion of the frequency strategy.

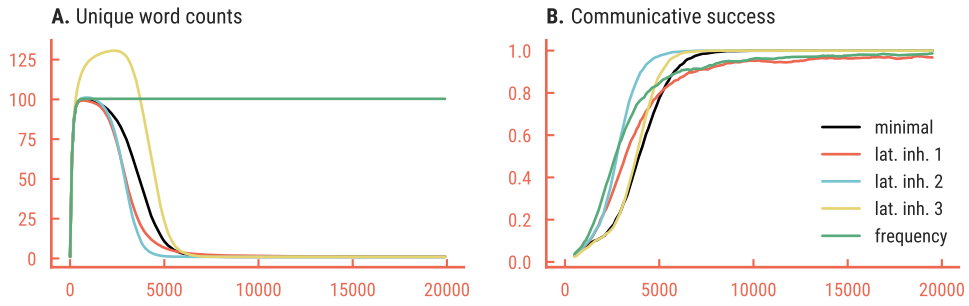


FIGURE 3.3 Comparison of the four naming game strategies in table 3.1. The the unique word count and communicative success show that all strategies reach communicative success. The stable language for the frequency strategy is not efficient.

LINGO1 Results shown for $N = 200$; avg. of 300 runs. p_{success} is a rolling average over a centered window of 1000 iterations.

These are just five strategies, but what does the rest of the strategy space look like? In appendix B I systematically explore a larger part of the space, following Wellens (2012). I indeed find that δ_{inh} interpolates between the minimal and frequency strategy. Further, relatively large δ_{inc} can lead to temporary stabilisation at a non-equilibrium state, until inhibition takes over the stable state is reached. However, I should note that I do not replicate Wellens's finding that the frequency converges faster than the minimal strategy (see also figure 3.3), and have not been able to reconstruct why. Although the behaviour might vary initially, the long-term behaviour is unaffected: convergence to a single-word language.

In sum, all strategies discussed allow the population to solves the naming problem and leads to effective communication within the population. Any form of lateral inhibition dampens competing words, a result of which agents eventually forget all but one word. The frequency strategy is the only discussed strategy that is not *efficient* in this sense. For different parameter settings communicative success can increase earlier, later or even stabilise temporarily, but will eventually be reached nonetheless. Indeed, it seems that "adding a scoring mechanism yields only marginal improvements in terms of communicative and alignment success" (Wellens 2012, p. 23)[8] Why, one wonders, is the convergence so robust?

Proof of convergence

To the best of my knowledge, De Vylder and Tuyls (2006) provided the only analytical result indicating that non-minimal naming games converge to a shared, single-word language. The results apply to a variant of the game, which makes similar simplifications as the minimal naming game: there is no homonymy and only a single object. It moreover starts 'later' in the game, when all agents have already engaged in an interaction and no new words are invented. At this point, there are K unique words $w_1, \ldots w_K$ in the game and the authors assume none of these is ever removed — very much like the frequency strategy. Similarly, speakers use observed frequencies to determine which word they will produce. For production strategies that reinforce or amplify the most frequent word, the authors are able to prove convergence to a single-word language. However, their proof applies to a *deterministic* model, the *sampling-response model*, and De Vylder and Tuyls use simulations — not a proof — to argue that their results

8 That is, for the basic naming game, since Wellens (2012) finds that in more complicated games, subtle update mechanisms *can* be beneficial.

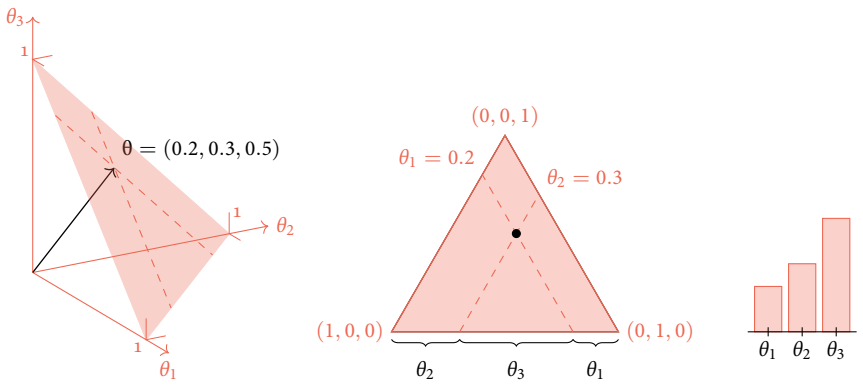


FIGURE 3.4 A discrete distribution θ over three values corresponds to a point in the 2-simplex, a triangular slice of $\mathbb{R}^3$ (left). The simplex can be embedded in the plane (middle), so that every point in the triangle determines a distribution (right).

generalise to the actual *stochastic, turn-based model*. I will present the deterministic, *sampling-response* model in some detail, partly because it is the stepping stone for the next chapter.

PRELIMINARIES First of all, we need to introduce the *simplex*: the space of discrete probability distributions. A probability distribution over K words is described by a vector $\theta = (\theta_1, \ldots, \theta_K)$ such that all θ_k are positive, and they together sum to 1, i.e. $\sum_k \theta_k = 1$. Note that the last entry, θ_K, is determined by the others and constraint $\sum_k \theta_k = 1$. Probability vectors therefore lie in a $(K-1)$-dimensional slice of $\mathbb{R}^K$. This slice is known as the $(K-1)$-*simplex* Δ^{K-1}, or simply Δ if no confusion can arise. The 2-simplex corresponds to a triangle, as illustrated in figure 3.4.

The model proposed by De Vylder and Tuyls (2006) considers a population of N agents who keep a queue of the last Q words they have observed.[9] A speaker will utter a word based on the relative frequencies of the words in her queue. Formally, we write $c = (c_1, \ldots, c_K)$ for the vector of *counts*, i.e. c_k the the number of k's in the queue. The counts correspond to (relative) frequencies $\theta = (\theta_1, \ldots, \theta_K)$ where $\theta_k = c_k/Q$. The point $\theta = (0.2, 0.3, 0.5)$ in figure 3.4 for example depicts the frequencies of $K = 3$ words in a queue of length $Q = 10$ with 2 occurrences of w_1, 3 of w_2 and 5 of w_3. By 'frequencies' θ we from now on mean *relative* frequencies and we also call θ the *language* of an agent. The frequencies lie in a discrete subset Δ_Q of the simplex which depends on the size of the queue Q (see figure 3.5).

Given a language, a *response function* r determines with what probability each word is uttered. Consider for example the response function r that puts all mass on the most frequent word. In our example with $\theta = (0.2, 0.3, 0.5)$ this means that $r(\theta) = (0, 0, 1)$, so the probability of uttering w_3 is $p(x = w_3 \mid \theta) = 1$. More generally, $r : \Delta \to \Delta$ maps the language θ_A of agent A to a *word distribution* $\pi_A := r(z_A)$, such that the probability of uttering word $x = w_k$ is

$$p(x = w_k \mid A) = \pi_{A,k}, \qquad \text{where } \pi_{A,k} = \big[r(\theta_A)\big]_k \tag{3.3}$$

9 The notation of De Vylder and Tuyls (2006) maps to ours as follows: $n \rightsquigarrow K$, $K \rightsquigarrow Q$, $m_i \rightsquigarrow c_i$, $x_i \rightsquigarrow q_i$, $s(k) \rightsquigarrow \pi_k$ $\Sigma \rightsquigarrow \Delta$, $\sigma \rightsquigarrow \theta$ (mostly), and $\tau \rightsquigarrow \bar{\pi}$.

THE SAMPLING-RESPONSE MODEL It is not easy to analyse this game directly. Consider how the language θ of a hearer changes during an interaction. The only thing that matters is the probability of hearing a word, not which speaker uttered it. We obtain

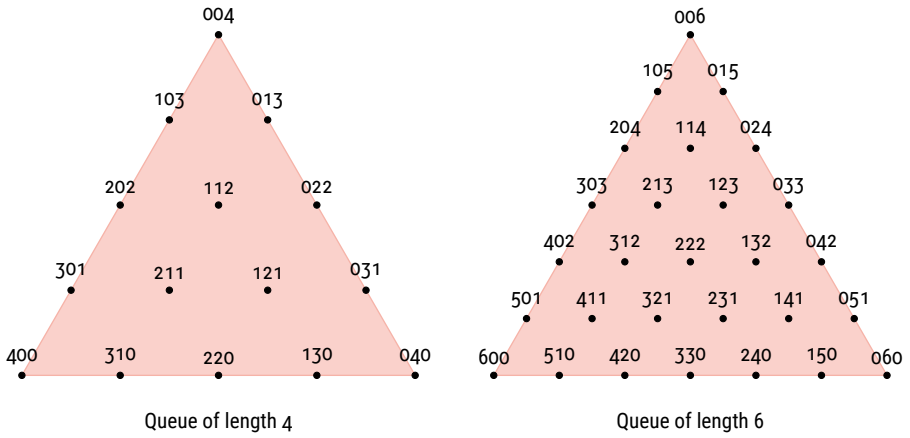


FIGURE 3.5 All possible frequencies of 3 words in a queue of length 4 (left) and 6 (right) form a discrete subset of the simplex. The corresponding relative frequencies are the 'languages' used by agents in the sampling-response model. Frequencies (a, b, c) are labeled abc.

Figure inspired by (De Vylder and Tuyls 2006).

those probabilities by averaging over all possible speakers (for simplicity, agents are allowed to speak to themselves),

$$p(x = w_k) = \bar{\pi}_k, \qquad \text{where } \bar{\pi} = \frac{1}{N}\sum_{A=1}^{N} \pi_A, \tag{3.4}$$

and call this average word distribution the *aggregate languages* as it aggregates the languages of all agents. Since the language of the hearer changes in every round, the aggregate language $\bar{\pi}$ also varies from round to round. To obtain a analysable model, De Vylder and Tuyls (2006) nonetheless assume it temporarily remains constant. In the resulting *sampling-response* model all agents interact synchronously in successive *episodes*. During an episode all agents simultaneously receive Q utterances, drawn from the aggregate language $\bar{\pi}$. One episode therefore corresponds to multiple rounds of the original turn-based game, enough to ensure that all agents have 'flushed' their queues, i.e. acted as a hearer Q times. Indeed, the analogy is not perfect, but deliberately so.

Importantly, *the sampling-response model is deterministic* and analysing how an agent's language changes during an episode becomes much easier. Concretely, if $\bar{\pi}$ is the aggregate language during an episode t, the probability of observing frequencies θ_t is the probability of observing the corresponding counts c_t amongst Q independent draws from $\bar{\pi}_t$. A multinomial probability, that is, so the sampling-response model takes the form

$$\theta_t \mid \bar{\pi}_t \sim \text{Multinomial}\big(c_t \mid \bar{\pi}_t\big), \qquad c_t = Q \cdot \theta_t, \tag{3.5}$$

$$x_t \mid \theta_t \sim \text{Categorical}(r(\theta_t)) \tag{3.6}$$

We can use this to compute the word distribution of agent A *after* episode t:

$$\begin{aligned} \pi_{A,k}^{(t+1)} &= p(x_{t+1} = w_k \mid A, \bar{\pi}_t) && (3.7)\\ &= \sum_{\theta \in \Delta_Q} p(x_{t+1} = w_k \mid \theta) \cdot p(\theta \mid \bar{\pi}_t) && (3.8)\\ &= \sum_{\theta \in \Delta_Q} [r(\theta)]_k \cdot \text{Multinomial}(c \mid \bar{\pi}_t). && (3.9) \end{aligned}$$

Note that the word distribution does not depend on A. This implies that the next aggregate language is also

$$\bar{\pi}_{t+1} = \sum_{\theta \in \Delta_Q} r(\theta) \cdot \text{Multinomial}(c \mid \bar{\pi}_t). \tag{3.10}$$

This defines a deterministic transition $\bar{\pi}_t \mapsto \bar{\pi}_{t+1}$ from the aggregate language in one episode to the next.

CONVERGENCE In summary, in episode t of the sampling-response model, all agents simultaneously hear Q words drawn from the aggregate language $\bar{\pi}_t$. At the end of the episode, all agents have updated their language, resulting in a new aggregate language $\bar{\pi}_{t+1}$. The new language is a deterministic function of $\bar{\pi}_t$, but also depends on the response function r. De Vylder and Tuyls (2006) showed that under certain conditions $\bar{\pi}_t$ converges to an aggregate language with only a single word:

$$\lim_{t \to \infty} \bar{\pi}_t = (0, \ldots, 0, 1, 0, \ldots, 0). \tag{3.11}$$

Perhaps the most important condition was that the response function must be *amplifying*. That means, roughly, that the response function increases the probability of producing the most frequent word (with respect to its frequency). We have already seen the prime example of an amplifying function: the function that *exponentiates* a distribution (see figure 2.2):

$$r_\eta(\theta) = \frac{1}{\sum_{k=1}^{K} \theta_k^\zeta} \cdot \left(\theta_1^\zeta, \ldots, \theta_K^\zeta\right), \qquad \zeta > 1 \tag{3.12}$$

With this response function, the population would eventually adopt a language with only one word.

Conclusions

Naming games try to understand how self-organisation can lead to the emergence of a shared vocabulary. To reach coherence, agents have to align their vocabularies after every encounter, for which various strategies can be used. The strategies discussed in this chapter — the minimal, frequency, and lateral inhibition strategies — all lead to the emergence of a consensus. As long as the alignment strategy implements some kind of competition damping, for example in the form of lateral inhibition, the resulting language is effective and agents remember no more words than strictly required. The frequency strategy was the exception, where agents remember all words but nevertheless reach full communicative success.

The naming games discussed in this chapter are the simplest, but, it seems, most important models in the literature. By dropping various assumptions, different games have been obtained. One could for example allow homonymy, in which case coherence can then still be reached by damping competing homonyms and synonyms. This introduces a kind of *mapping-uncertainty* (Wellens 2012) (which word corresponds to which object?) that will return in chapter 6. This problem is stronger in so called

Guessing games where the speaker is not allowed to indicate the object the hearer, who has to *guess* the intended object. Rather than simplifying the basic naming game, it can also be extended. Recently, Steels (2015) for example proposed the *syntax game*, where agents communicate n-ary relationships rather than words. Both the script and underlying mechanism are in the end very simimlar to the original naming game. This also seems to hold for work that has adopted *fluid construction grammars* to extend the representational capacities, hoping to move closer towards natural language (see e.g. Steels (2016) for an overview). Since this thesis concerns itself with the dynamics of the underlying game, such extensions have been left out.

The underlying, long-term dynamics of naming games seems rather clear. Where Bayesian iterated learning found a convergence to the prior, in naming games one finds convergence to single-word, coherent stable language, if the alignment mechanism somehow amplifies the highest-scoring word. So much both the proof of De Vylder and Tuyls (2006) and experimental results suggest. The proof applies only to a deterministic variant of the naming game, and it remains an open problem to show convergence for stochastic naming games. The wide range of experimental results suggests this should be possible — at least as long as the rules of the game are respected. As soon as the rules are changed, convergence can break. Baronchelli, Dall'Asta, et al. (2007) for example introduced a parameter β regulating the probability with which agents update their inventories. They find that for values of β below some critical point, multiple words can survive in the population.

The desiderata formulated in the last chapter were clearly motivated by iterated learning models, and might not be directly relevant for naming games. Nevertheless, it should be pointed out that naming games give rise to stable languages (D6), are to some extend analysable (D3) and appear to be robust to population structure (D5). They include various strategies (D2), but the resulting behaviour is always similar terms of long-term behaviour. One desideratum the naming game does clearly not fulfil is the explicit representation of the learning biases. In fact, agents have hardly any cognitive makeup, but this is addressed in the next chapter.

4 Bayesian Language Games

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 2.

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.[10] 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[11] 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.

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 $w_1, \ldots, w_K$, 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

$$p(\theta \mid x) \propto p(x \mid \theta) \cdot p(\theta), \tag{4.1}$$

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

11 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.

10 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.

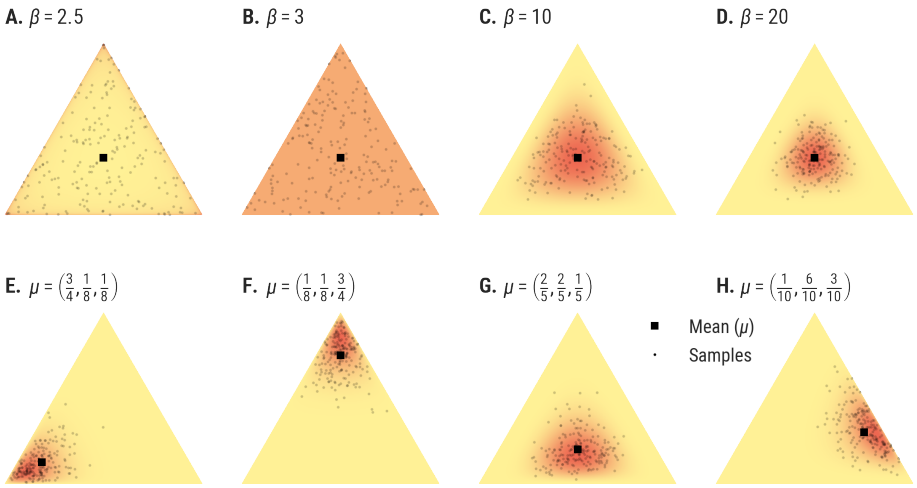


FIGURE 4.1 The Dirichlet distribution for various parameter settings. The Dirichlet can be parametrised by a point μ in the simplex and a scalar β. The mean of the distribution is determined by μ and β influences the variance. The first row (A–D) demonstrates the effect of β while fixing $\mu = (1/3, 1/3, 1/3)$; 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

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 $p(\theta \mid x_t)$ 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 6.

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 $p_S(\theta \mid x_{t'})$ 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 $x_t = (x_1, \ldots, x_b)$ from that distribution and communicates those to the hearer.
- The hearer updates his beliefs $p_{H,t+1}(\theta) := p_H(\theta \mid x_t)$ to the posterior, which is proportional to $p(x_t \mid \theta) \cdot p_{H,t}(\theta)$. 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 noted that the proposed framework is more general.

On a practical note, a mathematical development of the model is included in appendix C. 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 4.1 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, 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

$$p(\theta \mid x_t) = \text{Dirichlet}(\theta \mid \alpha_t + c_t), \tag{4.2}$$

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

$$\text{SPEAKER} \quad \begin{cases} \theta_t \mid \alpha_{t-1} & \sim \text{Dirichlet}(\alpha_{t-1}) \\ x_i \mid \theta_t & \sim \text{Categorical}(\theta_t), \quad i = 1, \ldots, b. \end{cases} \tag{4.3}$$

$$\text{HEARER} \quad \alpha_{t+1} := \alpha_t + c_t \tag{4.4}$$

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 4.2. 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 $\text{Dirichlet}(\alpha_t)$ 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:

$$\alpha_{t+1} = \alpha_0 + c_1 + c_2 + \cdots + c_t = \alpha_0 + c_{-t}, \tag{4.5}$$

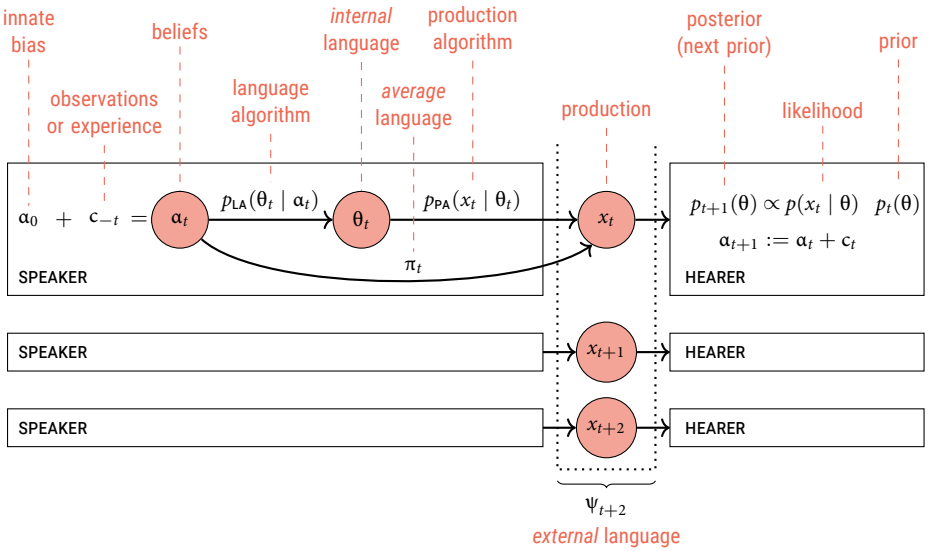


FIGURE 4.2 Illustration of the Bayesian naming game with all relevant concepts. See main text for details.

where c_{-t} is the vector of counts of all observations before and including round t.[12] 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 4.5 thus makes explicit that the beliefs in round t are the sum of innate biases α_0 and past experience 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 $w_1, \ldots, w_K$, 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

$$\psi_t := \frac{1}{b \cdot t} \cdot \sum_{\tau=1}^{t} c_\tau \tag{4.6}$$

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* $\pi_t =$

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

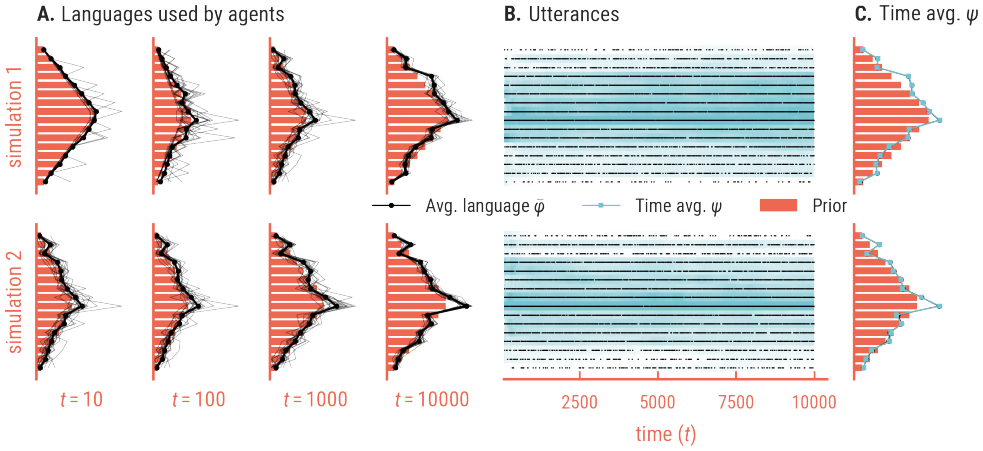


FIGURE 4.3 Two runs of the Bayesian Naming Game. A. The distributions of all agents (thin black lines) first diverge but eventually stabilise. They always reflect the prior (orange), B. Utterances (dots) at every time plotted over a moving average of 2000 time steps. C. The relative frequency of all utterances reflects the language adopted in the population. See main text for more details.

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

$(\pi_1, \ldots, \pi_K)$, the language it is expected to use in the next round. That is, we consider the marginal distribution

$$\pi_k := p(x = w_k \mid \alpha_{t-1}) = \int_\Delta p(x, \theta \mid \alpha_{t-1})\, d\theta = \hat{\alpha}_k, \tag{4.7}$$

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*

$$\bar{\pi}_t := \frac{1}{N} \sum_{i=1}^{N} \pi_{A_i, t}, \tag{4.8}$$

consistent with our terminology in chapter 3.

Phenomenology of the DC naming game

THREE-STAGE EVOLUTION Figure 4.3 show two typical runs of the Dirichlet-categorical naming game. Subfigure A shows the expected languages of all the agents (π_A, thin 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 4.3 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.[13] 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 3, such as the number of unique words or the total number of words, are meaningless once the vocabulary is fixed.[14] To measure coherence, we use the (generalised) Jensen-Shannon divergence (JSD). The JSD can quantify the similarity of all the expected languages $\pi_{A_1}, \dots \pi_{A_N}$ simultaneously (see appendix C for details). Normalising the JSD, the coherence measure becomes

$$C(t) = 1 - \frac{\text{JSD}\left(\pi^{(t)}_{A_1}, \dots, \pi^{(t)}_{A_N}\right)}{\log_2(N)}, \tag{4.9}$$

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*

$$R(t) = 1 - \text{JSD}(\hat{\alpha}_0, \bar{\pi}_t). \tag{4.10}$$

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 4.4B. 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 $\text{JSD}\left(\pi^{(t)}_{A_1}, \dots, \pi^{(t)}_{A_N}\right)$ 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.

14 When writing this, I realise that variants can of course be defined. In fact, I had done so 'before', in chapter 6. Future work could transfer those measures to the Bayesian naming game.

13 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.

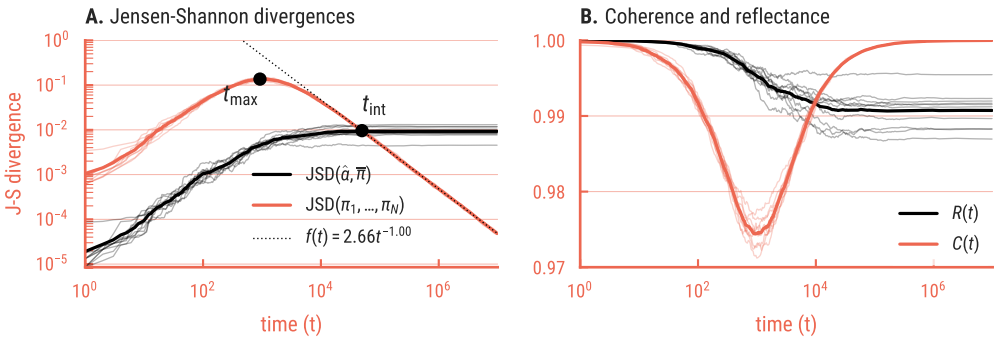


FIGURE 4.4 The Dirichlet-categorical name converges to a stable, coherent language. A. The distance between expected languages vanishes, but the aggregate language deviates from the bias. B. Coherence initially drops, but then increases to 1. The black line illustrates the *reflection of the bias*.

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

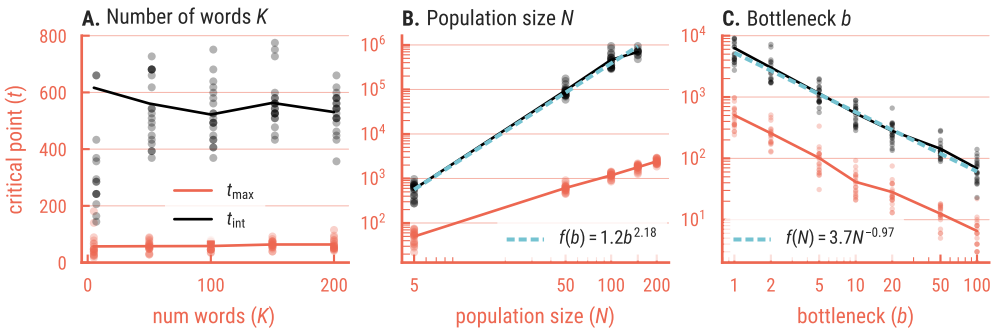


FIGURE 4.5 Effects of the language, population and bottleneck size on convergence time, probed by the critical points t_{max} and t_{int}. See main text for details.

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

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 4.4B, 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 4.4 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 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 4.5 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 $t_{int} \propto N^{2.18}$ (estimated using linear regression). The minimal naming game exhibits similar power-law scaling, although with a different exponent of 1.5 (see eq. 3.1). 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 $t_{int} \propto b^{-.97}$, which is suspiciously close to $t_{int} \propto b^{-1}$. This is not surprising since larger bottlenecks allow

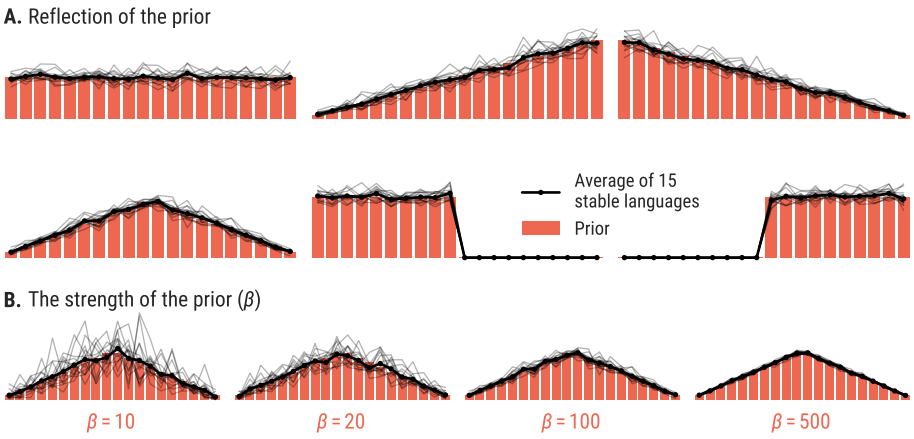


FIGURE 4.6 A. Different runs of evolutionary history result in different stable languages (thin black lines) that all reflect the prior (orange) in the sense that cultural evolution reproduces the prior *on average* over many runs. This is illustrated with six differently shaped priors. B. How well the languages reflect the prior is regulated by the strength of the prior (β).

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

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 4.6A). 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 4.6B. 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 2 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

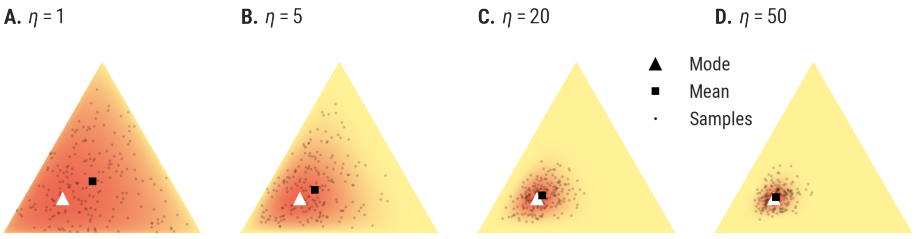


FIGURE 4.7 Exaggerating (or exponentiating) a Dirichlet distribution shrinks the variance and as η grows, the mean approaches the mode.

FIG02

exponentiated distribution even has a simple analytical form (see eq. 7.32):

$$p_{\text{LA}}(\theta \mid x, \alpha) \propto \left[p(\theta \mid x, \alpha)\right]^{\eta} \tag{4.11}$$

$$= \text{Dirichlet}(\theta \mid \eta \cdot (\alpha - 1) + 1). \tag{4.12}$$

Exponentiation shifts the distribution towards its mode, the point with highest probability, and moreover shrinks the variance, as illustrated in figure 4.7. 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 4.2).

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

$$p_{\text{PA}}(x \mid \theta, \alpha) \propto \left[\, p(x \mid \theta, \alpha) \,\right]^{\zeta}. \tag{4.13}$$

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 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 C) 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.[15]

15 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.

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

$$\text{SPEAKER} \quad \begin{cases} \theta_t \mid \alpha_{t-1} & \sim \text{Dirichlet}\big(\ \eta(\alpha_{t-1} - 1) + 1\ \big) \\ x_i \mid \theta_t & \sim \text{Categorical}\big(\theta^\zeta / \Sigma(\theta^\zeta)\big), \quad i = 1, \ldots, b. \end{cases} \tag{4.14}$$

$$\text{HEARER} \qquad \alpha_{t+1} := \alpha_t + c_t \tag{4.15}$$

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 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 4.8). 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.[16] 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.[17]

17 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

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

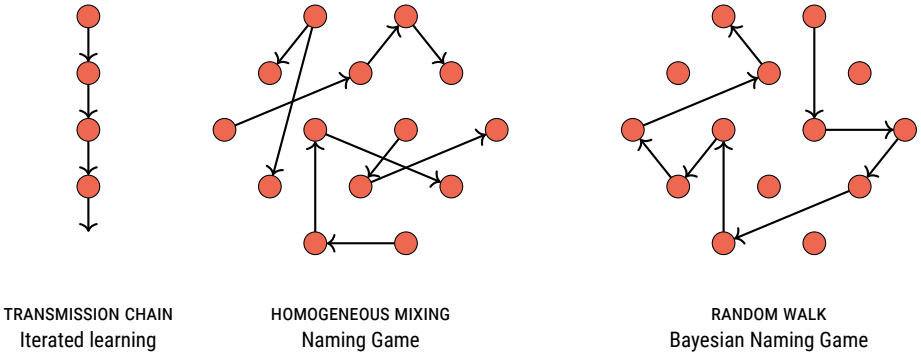


FIGURE 4.8 The proposed transmission model, a random walk through the population, combines the transmission chains used in iterated learning with the homogeneous mixing from the naming game.

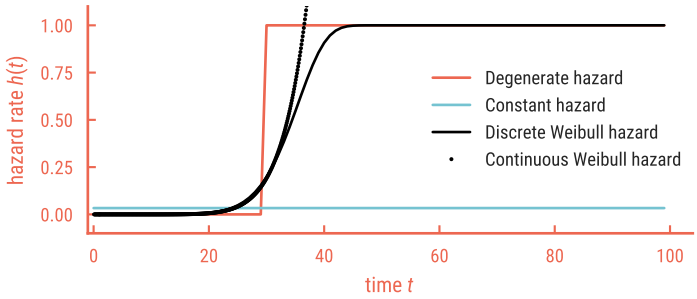


FIGURE 4.9 Different hazard functions. The more realistic (continuous/discrete) Weibull hazard is better approximated by a degenerate than a constant hazard function.

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 interactions (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 E 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 4.9). 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.[18]

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: 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 D for a more systematic exploration of the parameter space confirming the findings discussed here.

The central figure is 4.10. 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

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

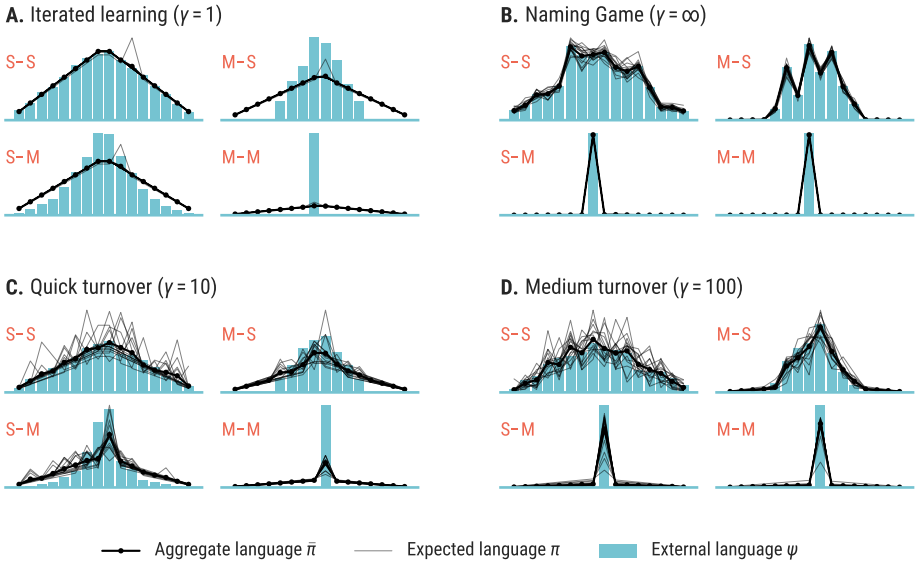


FIGURE 4.10 Typical outcomes of the Dirichlet-Categorical language game for the extreme strategies (sample–sample, MAP–sample, sample–MAP, MAP–MAP) in populations with immediate turnover (A, iterated learning, $\gamma = 1$), no turnover (B, naming game, $\gamma = \infty$) and two intermediate turnovers (C and D). See the main text for a discussion.

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

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 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 al-

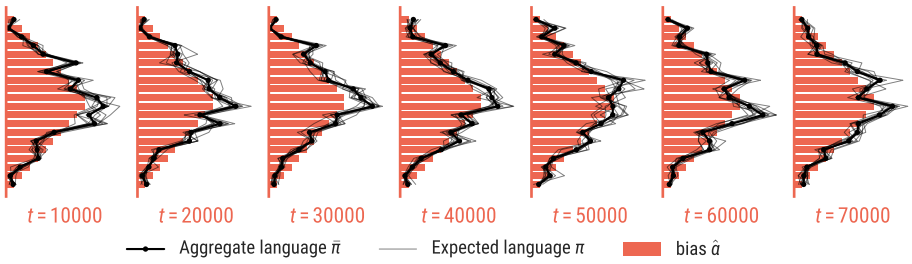


FIGURE 4.11 Gradual language change in the Bayesian language game for a particular choice of parameters. The effect seems brittle: slightly different parameter settings can give the kind of behaviour shown when $\gamma = \infty$.

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

ways 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 4.11, 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. 7.42) to change to

$$s_{t+1}(y) = \frac{\Sigma(\alpha_t)}{\Sigma(\alpha_t) + 1} \cdot s_t(y) + \frac{[\![y = x]\!]}{\Sigma(\alpha_t) + 1}, \tag{4.16}$$

  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 4.10B (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 4.10B (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 C 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 4.10B (S–M) confirms that idea.
- **MAP–MAP (M–M, $\eta = \zeta = \infty$).** The MAP–MAP strategy corresponds to the frequency strategy from chapter 3. 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 3, we find convergence to a single-word language in 4.10B (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 2. Concretely, it explicitly represents biases (D1); incorporates strategies from both the iterated learning and naming game literature (D2); seems susceptible to mathematical analysis (D3), as further discussed in chapter 7; exhibits nontrivial cultural effects (D4); and results in a stable language ((D6)).

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 ((D5)). 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 2, 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.[19] Interestingly, they show that the model is in that case equivalent to the 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.

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

5 Numeral systems

Models of cultural language evolution are seen to be ecologically valid, primarily because their conclusions are reproducible in equivalent laboratory experiments. More direct comparisons seem vital, and this chapter proposes numeral systems as a test case. Various reasons are given: reconstructions of their development have been proposed, lots of empirical data are available, the design space is vast, cognitive mechanisms well-studied and finally, numerals are simple enough to be easily modelled. The next chapter addresses early attempts at modelling the cultural evolution of numeral systems.

Numerals, as Bernard Comrie once put it, seem to be one of the rare cases where "present-day languages provide direct insight into the evolution of language" (Comrie 2013). They suggest an evolution in multiple stages, starting with words for the numbers 1–4, then building up a counting sequence, which gave rise to 'serialised' additive constructions and finally to the fully recursive systems that prevail today. He was not the first to note this. The multistage evolution of numerals was one of the central ideas developed in James Hurford's *Language and Number*. Numeral systems, he writes "show marks of successive phases of invention in the building up of the whole." (p. 78). As a result, "one can 'read' the history of a system, just like the history of an old building, from the contrasting styles of its pieces, from the foundations up" (p. 83). If it possible to reconstruct the evolution of numeral systems, doesn't that make it the ideal test case for models of language evolution? That is exactly what I suggest in this chapter.

Balancing expressivity and simplicity

A good empirical test case is indispensable. As Dediu et al. (2013) observe, "much work in agent-based modelling has proceeded in the absence of empirical linguistic data, input from linguists, or psychological considerations regarding learning, memory and processing" (p. 330). The absence might be a side effect of a growing body of laboratory experiments with cultural evolution (see Tamariz (2017) for a recent review), providing evidence for the ecological validity of the models (Smith 2014). To give one famous example, Kirby, Cornish, and Smith (2008) presented human participants, organised in a transmission chain, with an artificial language learning task. The first participant learned randomly generated names for a subset of a larger collection objects, which differed in colour, shape and movement. In a testing phase, the subject was asked to reproduce the names of *all* the objects, including unseen ones. A subset of the reproduced names was used to train the next subject. That subect again reproduced names for all objects, which were presented to another subject, and so on. Transmitting only a subset of the meaning-symbol pairs gave rise to a transmission bottleneck. This led to increase in learnability in the form of a systematic underspecification. The resulting language for example only distinguishes shape, but ignores colour and movement. Underspecification comes at the cost of expressivity, but when adding a pressure for expressivity, compositional languages emerged (see figure 5.1). Initially, this pressure was imposed by artificially removing homonymy (Kirby, Cornish, and Smith 2008), later by introducing communication in a chain of dyads (Kirby, Tamariz, et al. 2015). In either case, the experimental findings are surprisingly consistent with early iterated learning models and are therefore said to "provide empirical support for computational and mathematical models of iterated learning" (Kirby, Cornish, and Smith 2008).

The idea that language balances competing pressures for expressivity and learnability connects iterated learning to the work of Terry Regier and colleagues. In a series of papers, they argue that languages are optimised for efficient communication, implying a similar balance between expressivity and simplicity. They found evidence for this in colours terms (Regier, Kay, and Khetarpal 2007), kinship relations (Kemp and Regier 2012), spatial relations (Khetarpal et al. 2013; Regier, Kemp, and Kay 2015), and,

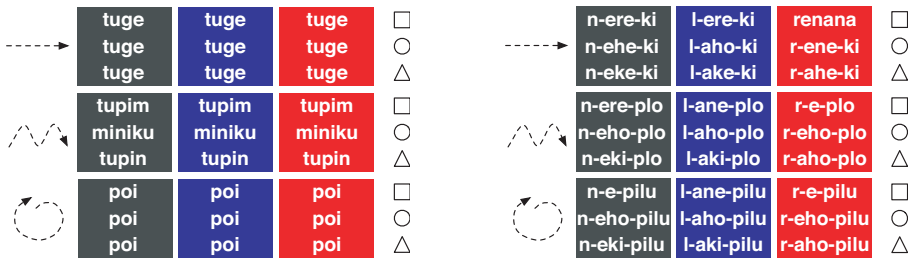


FIGURE 5.1 Transmission pressures for learnable languages, resulting in systematic underspecification (left). Introducing a pressure for expressivity results in compositional structure (right).

Figure reproduced from Kirby, Cornish, and Smith (2008) without permission.

indeed, numeral systems (Xu and Regier 2014). It was soon recognised that iterated learning might explain the observed near-optimality (Levinson 2012). Xu, Griffiths, and Dowman (2010) and Carstensen et al. (2015) accordingly set up iterated learning experiments where human subjects reproduced colour terms, concluding that "colour-naming universals may come from the learning and perceptual biases of human learners, brought out through the process of cultural transmission" (Xu, Griffiths, and Dowman 2010). Interestingly, perceptual biases in the form of just-noticeable differences, were also sufficient to reproduce realistic colour-term patterns in a simulation using a variant of the naming game (Baronchelli, Gong, et al. 2010).

However, colour terms might not provide the most convincing case for the argument that *compositional* structure in language results from pressures for learnability and expressivity (e.g. Kirby, Tamariz, et al. 2015) — colour terms have, to the best of my knowledge, typically no compositional structure. In that respect numeral system are much more promising. Few, if any, structures seem to balance expressivity and simplicity so effectively Most languages can accurately name a vast range of numbers using a small lexicon and some simple recursive rules. In English, for example, the thirteen words *one*, *two*, *three*, *four*, *five*, *six*, *seven*, *eight*, *nine*, *ten*, *hundred*, *thousand*, and *million* can be used to name nearly all numbers up to one billion (10^9). This might make numerals atypical linguistic structures, but that does not mean their structure cannot be explained by typical linguistic processes such as grammaticalisation (von Mengden 2008).

This chapter, then, has two goals. First, it argues that numeral systems are a good test case for models of language evolution. One argument can be found in the discussion above — numerals balance simplicity and expressivity, which should be reproducible using iterated learning — and further arguments are developed later in this chapter. Second, the chapter outlines the basic structure of numeral systems, a prerequisite for all that follows. In light of the above discussion, the *internal*, arithmetical structure of numerals is of primary interest and not, say, the grammatical role of numerals (e.g. why is *five blue balls* grammatical, and *blue five balls* not?). Restrictions of this type are necessary, since the linguistic literature on numerals is vast. In one survey, Harald Hammerström identifies 13 500 references to primary sources, over 100 monographs and hundreds of articles. I am practically oblivious to all this literature and will rely on several more general, secondary sources, which fortunately sketch a relatively clear picture of the world's numeral systems.

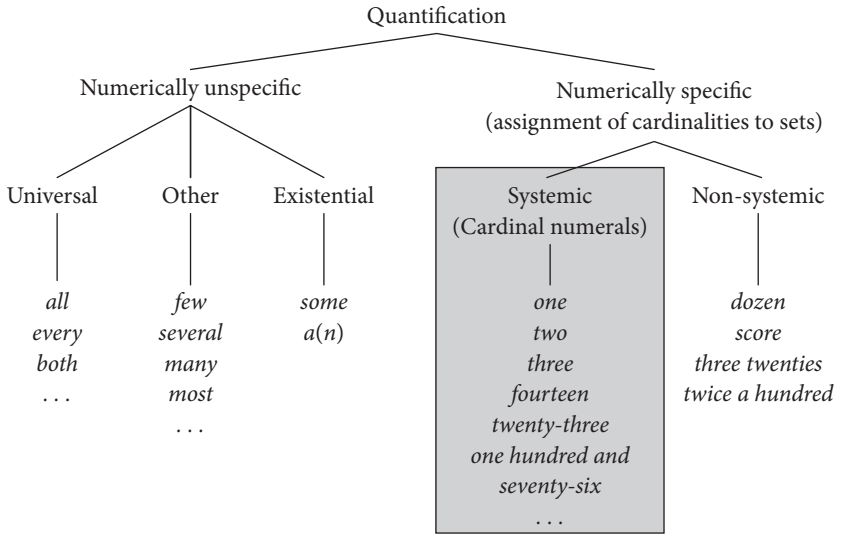


FIGURE 5.2 Numerals are numerically specific, systematic quantifiers.

Reproduced from von Mengden (2008) without permission.

An introduction to numeral systems

DEFINING NUMERALS So what exactly are numerals? Numerals are expressions for numbers, but we need to be more preicse. The the expressions of interest are *cardinals* like *one, two, three*. These express quantity and should should be distinguished from *ordinals* such as *first, second, third* expressing order. von Mengden (2008) presents a categorisation that clearly distinguishes cardinals from other quantifiers; it is reproduced in figure 5.2. At the highest level he distinguishes *numerically specific* from numerically *unspecific* quantifiers. The latter include vague quantifiers such as *some* and *many* but also universals such as *all*. Other words, like *score* and *dozen* have specific muneric values, but are typically not considered to be numerals, for the simple reason they are not part of a *system* in they way *one, two, three,* and so on, are. They for instance do not normally occur in the counting sequence.

To a first approximation, numerals are systematic, numerically specific expressions, but considering a more refined definition is instructive. Hammarström (2009) defines numerals as "(1) spoken, (2) normed expressions that are used to denote the (3) exact number of objects for an (4) open class of objects in an (5) open class of social situations with (6) the whole speech community." These clauses exclude certain other numerical expressions. Symbolic systems such as Roman and Arabic number symbols are excluded by (1); non-standard expressions like *three-times-five-and-two* for *seventeen* are excluded by (2); and (3) excludes numerically unspecific expressions. (4) excludes counting systems that are exclusively used to count a restricted class of objects, such as the Wuvulu system for counting coconuts (Hammarström 2009). (6) excludes specialised (mathematical or technical) jargon, and (5) excludes body-tally systems. Those systems use a fixed sequence of body parts to which speakers point when indicating a number. The sequences can be quite elaborate: some extended body counting systems in the highlands of New Guinea use a sequence of 23 body parts[20] (Comrie 2013). As Hammerström explains, "body tallying has to be done on a physically present person

20 Starting on the left side of the body: little finger, ring finger, middle finger, index finger, thumb, wrist, middle of forearm, inside of elbow, middle of upper arm, shoulder, collarbone, hole above breastbone, and then continuing in reverse order at the other side of the body

and to understand what number is referred to, the process must be watched". This makes body-tally systems markedly different from the other numeral systems.

The constraints (1)–(6) can hardly be called restrictive. Hammarström (2009) estimates there are at least 3500 numeral systems. Accordingly, Numeralbank,[21] a databank of numeral systems, contains around 4200 numeral systems. These numeral systems can be divided in at least two categories: *restricted* and *recursive systems* Comrie (2011) and Xu and Regier (2014). Restricted systems have little internal structure (e.g. all numbers are lexicalised or use only additive constructions) and typically cannot express numbers higher than 20. These systems are extremely rare. Most numeral systems are recursive, meaning they are organised around one or several bases and use multiplication and addition to recursively express a vast range of numbers. English uses the bases 10, 100 and 1000 (and possibly more), and happens to be a representative example: The vast majority of the world's languages use a decimal system, followed by vigesimal (base 20) and quinary (base 5) systems (Comrie 2013). Yet many other systems exist. There are languages without base, or languages using base 3, 4, 5, 6, 8, 12 or 15, perhaps supplemented by higher bases like 40, 60 and 80 (see Hammarström 2009, for a survey).

SIMPLE NUMERALS, COMPLEX NUMERALS AND BASES Let's spell out the structure of numeral systems in more detail, by considering the English system. The first ten numbers are expressed by mono-morphemic forms which we will call *simple numerals* (von Mengden 2008). Forms like *hundred* or *thousand* are also simple numerals. They can be combined to form *complex numerals* such as *two thousand three hundred sixty five*. The transition between simple and complex numerals can be a smooth one, as evident from French: *onze (11), douze (12), treize (13), quatorze (14), quinze (15), seize (16), dix-sept (17), dix-huit (18), dix-neuf (19)* (Calude and Verkerk 2016). Simple numerals can be further subdivided in *atoms* and *bases*. The atoms are the numerals from *one* up to *nine*; *ten, hundred* and *thousand* are the bases (higher bases are discussed later). Even if bases are "the most salient single characteristic" of numeral systems (Hammarström 2009), defining what exactly counts as a base is tricky.

Comrie (2013) defines a base as a number b occurring in multiplicative expressions of the form $x \times b + y$, where he stresses that order is unimportant. Although the idea is clear, it is not precise enough: The expression *six hundred and four* is of the given form, but should not suggest that 6 is an English base. Greenberg (1978) opts for a more technical definition. From a linguistic point of view, 2×10 (*two tens*) and 10×2 (*ten twos*) are not identical. The order of the factors indicates which of the two is the 'constant' (*augend* in additive/*multiplicand* in multiplicative constructions) and which the 'variable' (*addend*/*multiplier*). Bases, according to Greenberg, are multiplicands occurring in a series like $2 \times 10, 3 \times 10, 4 \times 10, \ldots$ — they are *serialised multiplicands*. But even this definition has shortcomings: It ignores subtractive constructions of the form $x \times b - y$. Hammarström (2009) *does* account for those and defines a number b_i to be a base if (1) the next higher base b_{i+1} is a multiple of b_i and (2) "a proper majority" of the numbers between b_i and b_{i+1} are expressed as $n \times b_i + k$ or $n \times b_i - k$ for some $k < b_i$. For English, this definition designates 10, 100, 1000, 10^6 and 10^9 as bases. Indeed, definitions are difficult, but the idea should be clear.

21 Numeralbank is part of Glottobank and largely based on the work of Eugene Chan. He collected many numeral systems at mpi-lingweb.shh.mpg.de/numeral. For most languages, it contains the expressions for 1–30, 40, 50, 60, 70, 80, 90, 100, 200, 1000 and 2000.

ISOLATED, MIXED AND ADDITIVE BASES It will be convenient to introduce some terminology for 'special' base-like numbers, that are not bases in the strict sense. French is famous for containing remnants of a vigesimal system, expressing numbers above 80 using base 20, as in *quatre-vingt-dix-sept* 'four-twenty-ten-seven' ($97 = 4 \times 20 + 10 + 7$). However, using Hammarström's definition, this would not make 20 a French base, as only a minority of the numbers between 20 and 100 (the next base) are expressed using 20. Comrie (2011) calls such cases *isolated* bases.

The Welsh expression *deu naw* 'two nine' ($18 = 2 \times 9$) is another example of an isolated base, since Welsh has a fairly clear vigesimal system — with one notable exception: It uses 100 as a base, which is not a power of 20. When the bases in a system are not powers of a single base, the system has *mixed bases*. In the case of Welsh, 20 does qualify as a base, since most numbers between 20 and 99 are expressed using multiples of 20. 70 is for example expressed as *deg a thrigain* 'ten on three-twenty' (3×10). Such mixed vigesimal-decimal system are very common (Comrie 2013), but other mixes are also attested. Supyire uses a particularly complex mix of base 20, 80 and 400, while expressing numbers below 20 using additive constructions involving 5 and 10. Comrie (2011) gives the expression for 799 as

(1) *kàmpwóò ŋ̀kwuu sicyɛɛré 'ná béé-tàànre ná ké 'ná báár-ìcyɛ̀ɛ̀rɛ̀* Supyire
fourhundred eighty four and twenty-three and ten and five-four
$799 = 400 + (80 \times 4) + (20 \times 3) + 10 + (5 + 4)$

Additive constructions for numbers below the lowest base, as with 10 and 5 in Supyire, occur in more languages. Mixtec languages for example use a vigesimal system with single words for 10 and 15. The same pattern is found in Biblical Welsh (Hurford 1975). Georgian expresses the number 11 to 19 using addition with 10, which is reminiscent of the French *quatre vingt dix sept*, where 10 is used in a similar additive construction. In these examples, 5, 10 or 15 are not proper bases, but one might call such numbers *additive bases*, if they are smaller than the first base and occur in multiple additive constructions — that is, if they are *serialized augends* (Greenberg 1978). It should be stressed that neither additive nor isolated bases are bases according to Hammerström's definition, while mixed bases are.

EXPONENTIATION AND MATHEMATICAL BASES. The notion of 'base' is used in mathematics, in a similar, but different way. A decimal system in the mathematical sense uses bases $10^0, 10^1, 10^2, 10^3, 10^4, 10^5, \ldots$, whereas a decimal system in the linguistic sense would use a finite subset such as $10^2, 10^3, 10^6, 10^9$. The defining property of mathematical bases is that they are exponents. It has been argued that exponentiation also plays a defining role in numeral systems (e.g. Hurford 1975), but this is controversial (see Comrie 1999; Comrie 2013). First, nothing signals that *hundred* and *thousand* correspond to the first and second power of 10. The use of portmanteau forms for high bases is in fact widespread: *million* originates from Italian *millione*, the augmentative of *mille* (thousand), something like 'a big thousand.' Second, even if the sequence *billion*, *trillion*, *quadrillion*, and so on, is somewhat productive, the relation to the corresponding powers is opaque (n-illion meaning 10^{3n+3}). Third, such high powers are, with the notable exception of at least Sanskrit (see below), often a recent invention used mostly in technical context. Finally, exponentiation is not very consistent across

languages. Illustrating the last few points, the UK adopted the so called *short scale system* in favour of the *long scale system* as recently as 1974. The short scale system uses *million* for 10^6, *billion* for 10^9, *trillion* for 10^{12}, and so on; the long scale system uses *million* for 10^6, *milliard* for 10^9, *billion* for 10^{12}, *billiard* for 10^{15}, and so on. Mandarin uses a completely different sequence of powers: 10^4, 10^8, 10^{12}, 10^{16}. Sanskrit even has a monomorphemic series of bases for all powers of 10 up to 10^{11} (Comrie 2011). The simple and safe solution, in sum, is to define the bases of a numeral system by listing all of them.

SUBTRACTION, DIVISION AND FRACTIONS Arithmetical operations other than addition and multiplication are also used. One finds subtraction in the Latin expression *un-de-viginti* 'one-from-twenty' ($19 = 20 \leftarrow 1$), where I wrote $a \leftarrow b$ for $b - a$ to respect the order of the constituents. In Biblical Welsh, one finds expressions like *onid pedwar deugain* 'minus four two-twenty' ($36 = -4 + 2 \times 20$) and Ket expresses 70, 80 and 90 as $100 - 30$, $100 - 20$, $100 - 10$ respectively. Comrie (2011) gives the example

(2) *qus'am ʌɣam dəŋas' bən's'aŋ ²kiʔ* Ket
one left.over thirty without hundred
$71 = 1 + (30 \leftarrow 100)$

Some languages even use division, although it might be more accurate to speak of multiplication by fractions (Comrie 2011). Welsh expresses 50 as *half cant* 'half hundred' ($50 = 1/2 \times 100$). Danish offers more examples, expressing 50 as *halvtreds*, which is derived from *halv-tred-sinds-tyve* 'half-third-times-twenty' ($50 = 2\,1/2 \times 20$). Here half-third can be interpreted as the third half: $2\,1/2$ after $1/2$ and $1\,1/2$, but Comrie (2011) lists it as an example of *overcounting*.

OVERCOUNTING AND OVERRUNNING. The English equivalent of overcounting would be the expression*three on the way to to fifty* for 43. More formally, if b is some base, an example of overcounting is of the form *a towards* $(x+1) \times b$ for $x \times b + a$. One could read an example of overcounting in *half-tred*: half on the way to three, but the interpretation 'the third half' also seems likely. Clearer examples are cited by Hurford (1975). Some Mayan languages express 41 as *huntuyoxkal*, which translates to *the first of the third score* ($1 + 3 \times 20$). *Overrunning* is a different phenomenon, where one uses addition when multiplication might be expected, or multiplication if a higher base would be expected. The English equivalent would be *tenteen* for 20 or *tenty* for 100. Comrie (1992) gives several examples, starting with Polabian

(3) *visem-nocti, diva(t)-nocti, disa(t)-nocti* Polabian
eight-ten, nine-ten, ten-ten
18, 19, 20

The French *soixante dix neuf* ($6 \times 10 + 10 + 9$) would be another example, and indeed some varieties of French have adopted the simpler *septante neuf*. A clear example of multiplicative overrunning can be found in Old Islandic

(4) *otto tiger, níu tiger, tío tiger, ellefo tiger* Old Islandic
eight ten, nine ten, ten ten, eleven ten
8×10, 9×10, 10×10, 11×10

ORDER Languages can differ in how they order the constituents of numerals. The English *twenty five* (20 + 5) becomes *vijfentwintig* 'five-and-twenty' (5 + 20) in Dutch, and German similarly 'reverses' the order. As Calude and Verkerk (2016) point out, the ordering (base–atom, as in English or atom–base as in Dutch) has not always received enough attention. But there are interesting regularities. If languages use both the base–atom and atom–base order, the system always uses atom–base for the smallest numbers and at some point switches to base–atom. The reverse never happens (Greenberg 1978). Greenberg proposed the cognitive explanation that for large numbers the base term is much more informative and salient. Calude and Verkerk (2016) find that in Indo-European languages, the change in order, if it happens, practically always happens below 20.

CONTINUITY OF THE COUNTING SEQUENCE Another property of numeral systems is that they never have gaps.[22] If a system can express numbers up to L, it has expressions for *all* numbers $1, \ldots, L$. This perhaps unsurprising observation has some importance for philosophical discussions concerning the nature of numbers. Hurford (1987) dedicates a full chapter to three possible explanations of the continuity. First, in extreme form, the *referential-pragmatic hypothesis* holds that cardinalities are properties of collections ('threeness') that are fairly directly perceptible (*subitised*, see below). Continuity follows from the claim that n is more likely to be expressed than $n + 1$. Second, the *conceptual-verbal hypothesis* assumes we are born with the concepts ONE, NUMBER and SUCCESSOR. As a result children cannot but construct a continuous sequence, like "little Peano's." Third, the *ritual hypothesis* assumes numbers are the result of reciting a ritualised sequence; the meaning is grounded in the ritual of counting. All of these positions are problematic and Hurford therefore suggests a synthesis: Small numbers up to 3–4 are subitised and we learn the notion of successor only after exposure to a conventional counting sequence.

THE PACKING STRATEGY The main (near-)universal property of numeral systems formulated in Hurford (1975) is the so called *packing strategy*. Conceptually, it is a simple principle: "When forming an expression for a high number, pick the highest-valued expression available as a starting point, and then build on that" (Hurford 1987, p. 243). Even though it is regularly cited it has not received much attention in the literature. One reason might be that the packing strategy was originally formulated as a fairly technical set of constraints on a phrase-structure grammar. But it is in fact a simple principle, *both* technically and conceptually. That becomes clear if you reformulate the principle *outside* the specific framework of Hurford (1975). In appendix F I reduce the packing strategy to the claim that *complex numerals use the largest multiple of the largest base possible.* A slightly more general formulation would be *the difference between a and b in $a + b$ and $a \times b$ should be maximised.* For example, Mixtec uses both 15 and 10 as base and could thus express 19 as either $10+9$ or $15+4$. The packing strategy correctly picks the latter, as it uses the larger of the two possible bases 10 and 15. Counterexamples also exist: *twenty-three hundred* does *not* conform to the packing strategy since *two thousand three hundred* expresses the same number using a larger base.

22 But see Zhou and Bowern (2015) for *possible* gaps in some *restricted* Australian systems.

A DECIMAL WORLD The previous discussion should not suggest a world where any two languages will use a radically different numeral systems, with mixes of subtraction, overcounting, division, overrunning and what not. No, that is certainly false. In fact, "we live in a decimal world. […] Bases other than 10 or 20 are extremely rare in the modern world" (Comrie 2013). It is no big mystery why the particular base 10 is so prevalent — "no contemporary linguist has ever thought it necessary to spell this explanation out, let alone argue against it" (p. 39–40). Indeed, boydy parts are a common source for certain number names: "Nouns for 'hand' probably provide the most widespread source for numerals for 'five' in the languages of the world" (Heine and Kuteva 2002, p. 166).

But the fact that we live in a decimal world also has another reason, clearly illustrated by the vigesimal systems, the most common after decimal systems. Vigesimal systems were dominant in Mesoamerica before the European invasions, but by now a mixed decimal-vigesimal system is mostly used. The systems typically include a word for 100 derived from Spanish *ciento*. Comrie (2013) signals a "worldwide historical trend for the dominant decimal system to encroach on and replace other systems" and concludes that "non-decimal numeral systems are even more endangered than the languages in which they occur." But why are numeral systems particularly prone to replacement? One explanation is that "numerals, much more so than most other parts of a language, are very culture-bound, being tied to the educational system in modern societies, to trading relations even in the earlier and less modernised societies" (Comrie 1999).

If the development of numeral system can follow a rough, unpredictable path, shaped out by all sorts of historical contingencies, all accounts of cultural evolution of numeral systems should be careful to check that abstracting away from those is justified. One account has a fair argument for this: that most recursive numeral systems share a similar basic structure.

The evolution of numeral systems

Not many studies have accounted of the evolution of numeral systems. But the accounts I have found, most notably of von Mengden (2008), Hurford (1987), Hurford (2007), and Comrie (1999) suggest a broadly similar picture. I recount the particularly lucid version of von Mengden (2008). It focusses on the decimal numeral system as found in most Indo-European languages, but appears to be equally applicable to other recursive systems. The reconstruction is based on the 'growth marks' (Hurford 1987) left by successive stages in the development of the numeral systems. It therefore rests on two crucial assumptions: first, that numerals are an ordered sequence, and second, that the sequence is continuous. Together, they suggests that properties of lower parts of the number sequence diachronically *precede* properties of higher parts (von Mengden 2008).

1. SUBITISING AND SIMPLE NUMERALS The simplest numeral system, if a system at all, would consist of only simple numerals. The few languages with only simple numerals reach no higher than 5 (Greenberg 1978, p. 256), but often only to 3 or 4 (von Mengden 2008). There is a simple explanation for the discontinuity around 4: Small quantities

up to 3–4 are directly perceptible, a phenomenon called *subitising*. Even newborns can fairly accurately discriminate different sized sets of 1, 2 or 3 items, while above 4 items their performance drops below chance level (Feigenson, Dehaene, and Spelke 2004). Importantly, subitising does not require counting, but quantities are recognised automatically. Even though the exact nature and the limits of subitising remain contested (Dehaene 2011; Feigenson, Dehaene, and Spelke 2004), it is clear that the lowest quantities are relatively directly perceptible.

Hurford (2001) argued that subitising is evidenced by languages, where numerals up to 3–4 are treated specially. For example, (exact) grammatical number distinctions beyond singular, dual and trial do not exist. Idiosyncrasies in words for 1–4 provide further evidence. Hurford cites the many irregular and suppletive forms for the first 4 or so ordinals in several languages. Similarly, small numbers more often have distinct case or gender forms, and sometimes a different word order. It should be noted that the sheer frequency of low numbers (Dehaene and Mehler 1992) could provide an alternative explanation for some of these effects. Nevertheless, languages with only subitizable cardinalities could form the fist step towards numeral systems. But they differ from numeral systems in two respects (von Mengden 2008). First, they are not organized in a sequence and second, they cannot be said to be systematic.

2. COUNTING AND THE EMERGENCE OF NUMERACY

The obvious starting point for a conventional counting sequence is not verbal but gestural: a fixed sequence of body parts (von Mengden 2008). The number *n* can then be indicated by highlighting the *n*'th body part in the sequence, while at the same time naming the body parts. At some point the gestures might become redundant and the names themselves form the counting sequence. Evidence for this idea can still be found in the use of body parts as atoms (Heine and Kuteva 2002, p. 166). von Mengden (2008) thus concludes that "we can safely assume that body-part expressions are the main source for cardinal numbers" (p. 299). Once the body-part expressions are standardised as cardinals, they often loose their original meaning. Consequently, the ordering of the words becomes arbitrary, as the association with a sequence of body parts is lost. As remembering long sequences is difficult, sporadic complex forms might emerge to express higher numbers. von Mengden (2008) argues that these expressions would have an underlying arithmetic structure (like complex numerals) but this would likely be rendered opaque. The resulting mono-morphemic forms would behave like atoms, necessitating a more transparent, recursive system.

3. SERIALISATION

A more transparent system would be one using *serialisation* Greenberg (1978). This means that one numeral is combined with an entire sequence of consecutive simpler numerals. The Medieval Latin numeral *decem* (10) is thus serialized in *un-decim (11), duo-decim (12), tre-decim (13), quattuor-decim (14), quinque-decim (15), se-decim (16), decem et septem (17), decem et octo (18), decem et nouem (19)* (von Mengden 2008, p. 301). Crucially, or so von Mengden argues, the compositional structure of the serialised expressions remains transparent. But how would serialisation emerge? One explanation would be that once the conventional sequence ends, another words is used for the whole collection. Consider people counting a pile of objects. Whenever they run out of counting words, the sequence ends, they group the objects and start

again (Hurford 2007). The groups will be given a name, perhaps *a ten*, or closer to the body counting: *a hand* or *two hands*. This word will start to function as a base. Initially this would be an *additive* base; multiplicative serialisation and actual bases would have emerged later. This is suggested by *1-deletion*, that languages tend not to use multiplication by 1 (Hurford 1987, p. 54) for the first base, as in *ten* instead of *onety*.

4. FUNCTIONAL ELEMENTS At this point, the numeral system is already fairly developed. Numerals will have acquired internal, arithmetical structure, which is perhaps sufficient for the purposes of this thesis. Tut the system will undergo further changes, in processes such as *grammaticalisation*. Grammaticalisation theory aims to describe how lexical forms can gradually evolve in grammatical forms (Heine and Kuteva 2002). During that process some of their semantic, morphosyntactic, and phonetic properties are lost, and the lexical forms take on another, more grammatical role. Consider numeral bases, which are often expressed by different morphemes, such as the English *ten, -ty* and *-teen*, a phenomenon called *base suppletion*(Hurford 1987, p. 56). von Mengden (2008) argues that this is a clear example of grammaticalisation, in which the base *ten*gradually took up a more grammatical role. The suffix *-ty* still means 10, but also signals that it occurs in a multiplicative construction. Similarly, *-teen* signals an additive construction. The resulting expressions such as *nineteen* or *ninety* are therefore more grammatical than that from which they are supposedly derived (something like *nine and ten* and *nine tens*).

Conclusions

Numeral systems, it seems, emerged in a multistage process where it gradually picked up more complex, recursive structure. This process is likely rooted in practices such as body-tallying (von Mengden 2008) or the practice of reciting a counting sequence (Hurford 2007). Grouping objects might be natural first step towards multiplicative constructions. Processes such as grammaticalisation further shaped the systems, resulting in functional affixes such as *-ty* and *-ten*. It does not seem too far fetched to extend this account to include subtraction, overrunning and the like. The origin of the arithmetic structure of numerals, after all, seems to lie in concrete counting practices.

Although the typology of numerals support this account, it is clearly speculative, and leaves open many questions. This is a terrain where modelling could proof useful. For example, Von Mengden suggests that in stage 2, only serialised complex expressions remain transparent, and that earlier, sporadic compounds are rendered opaque. Hurford (1987), on the other hand, seems to suggest that various competing expressions, all transparent, would be used simultaneously. A process of social negotiation would then lead to the standardisation of a single expression for every number. He supported this idea with simulations, which I discuss in detail in the next chapter.

But this would be an example of models helping our understanding of numerals, while the point of this chapter was to argue the reverse: that the numerals could help understand the models of language evolution. Since I touched upon various arguments throughout the chapter introducing, let me list all arguments explicitly.

- There is a fairly well-supported account of how numeral systems might have developed in several successive stages. This, one could say, is the empirical benchmark.
- There is a wealth of empirical data regarding numeral systems in the form of NumeralBank, comparable to the World Colour Survey (WCS).
- The design space of numeral systems is vast. The many arithmetic operations attested (subtraction, division, overrunning and -counting, besides addition and multiplication) and use of mixed bases all bear testimony to this.
- Numeral systems seem to be the school-book example of a balance between expressivity and simplicity, one of the predictions made by iterated learning. Models should be able to reproduce this.
- The cognitive mechanisms of number cognition have been studied extensively, although I have not discussed this in detail here. See for example Dehaene (2011) (but also Hurford 1987.

The next chapter discusses some first results in this direction.

6 Emergent numeral systems

Numeral systems seem to be an ideal testbed for models of language evolution. They have a clear hierarchical, recursive structure, there is plenty of variation, and their structure suggests how they evolved. So what can the models of language evolution discussed in the first part of this thesis tell about the emergence of numeral systems? This chapter addresses that question. To that end we first revisit and reinterpret the work of James Hurford, who addressed the same question over 30 years ago. This reveals that naming games can be sensitive to biases in the domain, which can be distinguished from biases of the agents. Finally, an attempt to simulate the evolution of numeral systems directly is presented and discussed.

Numeral systems strongly restrict the set of allowed expressions for numbers. One would be surprised to find a book for *three fours and seven euros* — that number is called *nineteen*. The degree of standardisation, to put it differently, is remarkably high for numerals. Why so? Hurford (1987) wonders whether the standardisation could be the result of a process of social interaction. In his account of the evolution of numeral systems, linguistic innovators play an important role. These rare individuals occasionally invent new linguistic constructions, such as additive or multiplicative constructions. Between phases of linguistic innovation the invented rules spread the population and do not change substantially, until the next innovation comes along. This might be an idealisation, but even if one prefers gradual phases of innovation rather than an individual innovator, the question remains how a linguistic innovation eventually become standardised.

Hurford addresses the standardisation of a base, the most salient characteristic of a recursive numeral system. Using two[23] agent-based simulations, he argues that the standardisation of a shared base could be the result of a process of social negotiation. In line with current terminology, I have baptised the simulations the *additive* and *multiplicative Base Game* (BG). The simulations represent two successive 'stages' in evolutionary history, following the invention of additive and multiplicative constructions respectively. This chapter starts by revisiting Hurford's work, then moves on to some further simulations and concludes with a discussion.

Hurford's base games

In both base games, populations of N agents "spend their time uttering numeral expressions to each other": constructions like $7+6$ or $3\times 10+5$. Interaction between agents are assumed to be homogeneous. Initially, all agents will use different expressions for the same numbers, but over time the same *base* should be used to form expressions. Since a base is not defined outside a numeral system, Hurford takes the base to be the largest of the constituents: the base of both $a\times b$ and $a\times b+c$ is $\max(a,b)$. So 3×7, $4\times 7+3$ and $4+7$ all use base 7. Note that the summand c in a multiplicative construction can thus be larger than the base. We return to these assumptions in the discussion.

Hurford assumes speakers try to use expressions that a hearer is likely to know. To that end agents track the frequencies or scores $s(b)$ of all bases b they encounter.[24] A simple criterion decides which bases are *favoured*, i.e. which bases an agent prefers to use. A base b is *favoured* if

$$s(b)>0 \qquad \text{and} \qquad s(b)\geq \xi\cdot \max_{b'} s(b'), \tag{6.1}$$

for $1\geq\xi>0$.[25]

Just like ζ in the the Bayesian language games, the parameter ξ determines the production strategy: how much the agent tends towards using the most frequent bases. For $\xi=1$ the base with positive and maximal frequency are favoured. The smaller ξ gets, the lower the threshold for being favoured. This slows down convergence, and I have used $\xi=1$ throughout this chapter. Appendix G includes a brief analysis of the effects of ξ. The set of bases favoured by agent A is denoted $F(A)$.

25 Hurford does not require $s(b)>0$, but this simplifies the discussion and does not alter the behaviour of the game.

24 In this chapter b does not denote the bottleneck, but a base.

23 Hurford discusses several other variants. I restrict myself to the two his simulations which appear to be most important for his argument.

ADDITIVE BASE GAME It will be convenient to reformulate Hurford's models in a more formal fashion. Suppose agents know simple numerals for the numbers $\mathcal{S} = \{1, \ldots, B = 2K\}$, which they can combine in additive constructions. We assume $B = 2K$ to be even as it greatly simplifies the discussion. Since agents generally prefer to use the simple numerals for numbers below B, they only form complex numerals for the numbers $\mathcal{N} = \{B + 1, \ldots, B + B\}$, which I call the *domain*. This implies that the numerals $n \in \mathcal{S}$ for which $n + n < B + 1$ cannot be used as a base. For example, if $B = 10$, agents know simple numerals $\{1, \ldots, 10\}$, communicate about the domain $\{11, \ldots, 20\}$ using bases in $\{6, \ldots, 10\}$. Note that base 10 is most *expressive* since it can be used to form expressions for all numbers in the domain. With base 9 you can only get to 18, with 8 up to 16, and so on. To make this precise, note that there are K numbers that could be used as a base, namely

$$b_1 := K + 1, \quad b_2 := K + 2, \quad \ldots \quad b_K := K + K = B \tag{6.2}$$

For the set of numbers n that are *expressible by base* b we write

$$E(b) = \{n \in \mathcal{N} : n \leq b + b\}, \tag{6.3}$$

Conversely $E^{-1}(n) = \{b : n \in E(b)\}$ denotes the set of bases that express n.

Hurford (1987) only considered the decimal case $B = 10$. He argued that because 10 is most expressive, it will soon become the most frequent base, resulting in the standardisation of the expressions $10 + 1, 10 + 2, \ldots, 10 + 10$. The additive base game was proposed to test the consistency of this account and follows the following script.

1. A number $n \in \mathcal{N}$ is chosen from the domain.
2. The speaker considers the set $C = F(S) \cap E^{-1}(n)$ of candidate bases: favoured bases that moreover express n. She randomly picks a base b from C if it is nonempty, or from $E^{-1}(n)$ otherwise. The speaker expresses n as $b + (n - b)$.
3. The hearer H receives the expression, determines the base used, and updates the score $s_H(b)$ of b by 1.

MULTIPLICATIVE BASE GAME The second game, the *multiplicative base game*, corresponds to the next stage in the evolution of numeral systems. By this time, an innovator has invented multiplicative constructions of the form $a \times b + c$, which have become available to the entire population. We assume agents prefer the simpler additive constructions for the numbers $B + 1$ up to $B + B$. Therefore, the domain is $\mathcal{N} = \{2B + 1, \ldots, B^2 + B\}$ and the numbers expressible by b are $E(b) = \{n \in \mathcal{N} : n \leq b^2 + B\}$. As before, agents track frequencies to determine which bases they favour. But there is an additional criterion. When an agent favours several bases, some numbers can be expressed in different ways. For example, if 10, 9, 8 and 7 are all favoured, 21 can be expressed as $2 \times 10 + 1$, $2 \times 9 + 3$, $2 \times 8 + 5$ and 3×7. In this case agents prefer *simpler* expressions — here 3×7 — as the result of a general simplicity bias (Hurford 1987). Step 2 of the script is changed to accommodate this:

2. The speaker considers all expressions of the form $n = a \times b + c$, where b is a base in C, or otherwise in $E^{-1}(n)$ when C is empty. If there are 'simple' expressions $n = a \times b$, she communicates one of those, or otherwise a random other one.

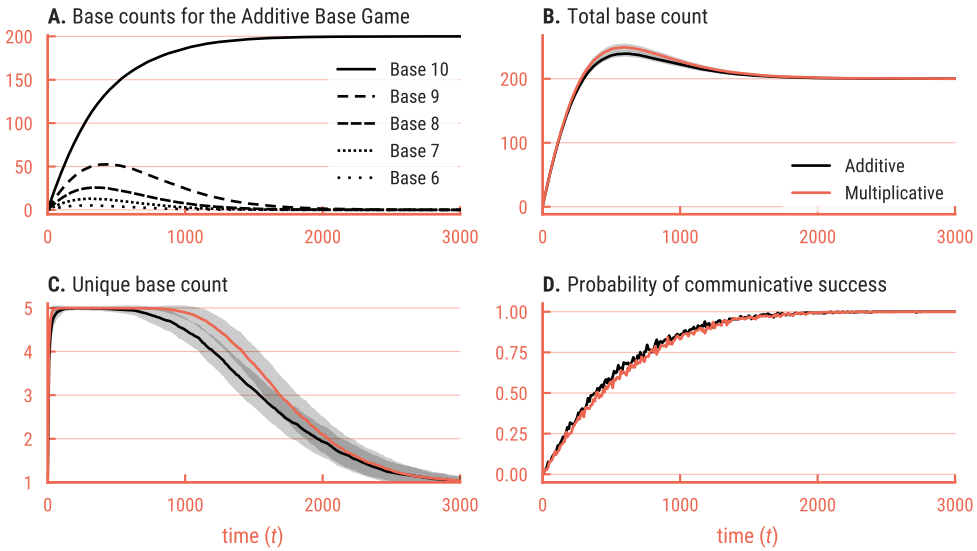


FIGURE 6.1 Comparison between the Additive Base Game (black) and the Multiplicative Base Game (orange). The dynamics of the two games are remarkably similar. Dynamics are visualized using A. the base counts of all possible bases for the Additive Base Game only (the Multiplicative case looks extremely similar); B. the total base counts; C. the unique base count; and D. the probability of successful communication. See main text for details.

HUR03 Results shown for $N = 200$, $B = 10$, $\xi = 1$; avg. of 600 runs; 1 std. shaded.

BASIC PHENOMENOLOGY Unlike Hurford (1987), we have twenty years of agent-based modelling research at our disposal. I therefore reanalyse the games using a more modern methodology and analyse the following statistics

- **(Probability of) communicative success** $p_s(t)$ We consider an interaction successful if the base used by the speaker is favoured by the hearer.
- **Base count** $N_{\text{base}}(b, t)$**.** The number of agents that favour base b at time t.
- **Total base count** $N_{\text{total}}(t)$**.** The total number of bases favoured by agents in the population, i.e. $N_{\text{total}}(t) = \sum_b N_{\text{base}}(b, t)$.
- **Unique base count** $N_{\text{unique}}(t)$**.** The number of unique bases favoured in the population at time t, i.e. $N_{\text{unique}}(t) = |\{b : N_{\text{base}}(b, t) > 0\}|$.

Figure 6.1 summarises the dynamics of the base games. Subfigure A illustrates the typical stages every simulation goes through. Initially, none of the bases is favoured and bases are used with roughly equal probability. This is followed by a phase in which different bases compete for a share of the population. The largest two bases are the two biggest competitors, but base 10 soon takes over, leading to the emergence of a shared, decimal system. Subfigure B shows that multiple bases compete directly for each agent's preference. This can be seen from the peak of N_{total}, which crosses the population size $N = 200$, indicating that agents at that point on average favour more than one base. Even when base 10 is already dominant, it can take a long time to eliminate all other bases from the population, as seen from C. But in D one sees that population eventually communicates sucessfully. The difference between the additive and multiplicative game, in terms of dynamics, appears to be small. Most importantly, Hurford's intuition about the emergence of a decimal system, is confirmed.

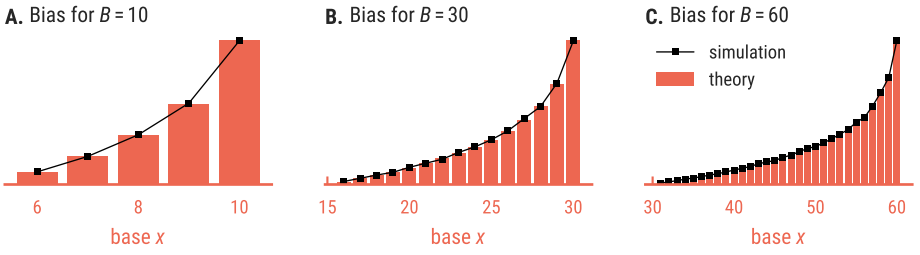


FIGURE 6.2 In the additive base game, the probability of using a base without any past experience (i.e., no preferences) is strongly skewed towards the highest base. The game has a strong implicit prior for using high bases.

FIG10 The 'simulation' is the rel. freq. of 100 000 samples.

Domain adaptivity in the base games

The similarities between the dynamics of the base games (figure 6.1) and the dynamics of the naming games (e.g. figure 3.3) are striking. Indeed, the base games seem to be naming games in disguise. Rather than negotiating a name for an object, agents negotiate a base: which of the 'words' $6, \ldots, 10$ should name the 'object' BASE. Like the Bayesian language game, the 'vocabulary' is fixed to $K = 5$ bases, and the strategy used corresponds directly to the *frequency strategy* (when $\xi = 1$).

But there are also some striking dissimilarities. Most notably, the same 'word' — base 10 — is adopted in every simulation: The base game seems to be biased towards adopting the highest base. Hurford recognises this, but does not explain what this implicit bias exactly is. But this can be done, for the additive base game, at least. It is a nice mathematical puzzle to show that the probability that a base will be used (by an agent with no past experience) is

$$p(b_j) = \frac{1}{K}\big(H_K - H_{K-j}\big), \tag{6.4}$$

where $H_n = 1/1 + \cdots + 1/n$ is the n'th harmonic number (see appendix G for a proof). The distribution is plotted in figure 6.2, for $B = 10, 30$ and 60. The figure clearly demonstrates that the additive base game has a strong *implicit bias* towards using the largest base.

IMPLICIT BIASES AND EXTERNAL CONSTRAINTS I have called the bias *implicit* since the bias seems to be the result of certain constraints built into the model. In this case, the constraints are arithmetic in nature and ensure that base 10 is much more expressive than base 6. The constraints moreover appear to be *external* to the agent — properties of the domain, rather than properties of the agent.[26] This raises the question whether the *implicit biases* arising from external constraints ("$13 > 6+6$") should be distinguished from the biases somehow internal to the agent ("I have ten fingers"). If the different biases influence the behaviour in qualitatively different ways, that is a clear indication that the two should be distinguished. The next experiment shows that this is indeed the case.

Consider a population where every agent has an internal bias towards using a certain base. For example base 10, because the ten fingers are particularly salient. We model this by initially assigning a score s_0 to one particular base b_0. These initial scores act as pseudo-observations, in the same way as the biases work in the Bayesian language

26 Readers objecting to mathematical constraints being somehow external to the agents should note that this does not undermine the main point that biasses can differ in kind.

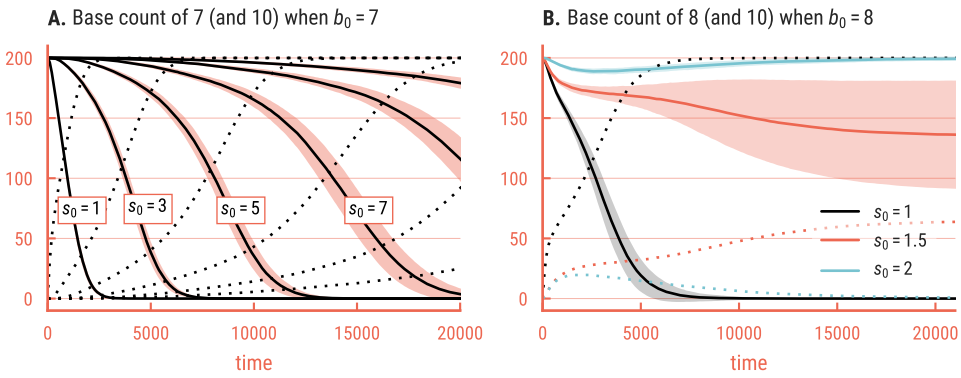


FIGURE 6.3 The additive base game in populations biased towards using base 7 (left) or base 8 (right), with varying initial score s_0 (higher scores indicate stronger bias). The figure illustrates that the biases implicit in the domain and the biases of the agents work differently: agents cannot overcome the former (see main text for details). Note: averages over 300 runs are shown and for $s_0 = 1.5$ individual runs convert to either base 10 or base 8.

HURO5 Results shown for $N = 200$, $\eta = 1$, $n_{\max} = 46$; avg. of 300 runs, 1 std. shaded.

game. Figure 6.3A reports an experiment with the additive base game where the population is biased towards using base $w_0 = 7$, with various different initial frequencies s_0. The initial state seems to be unstable, irrespective of the initial frequency of base 7 (although it takes longer before base 7 dies out if the initial frequency is higher). Apparently, a septenary system cannot be maintained. I presume this is so because base 7 can express only less than half of the numbers in the domain. The probability that an agent will use another, larger base is therefore greater than the probability that it will use base 7, independent of whether it favours base 7. This makes a septenary language unstable, and a larger base will eventually take over. The important point is that the biases of the agents *cannot overcome the constraints of the domain*. No matter how strong their initial bias for using base 7, nothing will change the fact that $7 + 7 < 15$, and the population will have to *adapt* to the constraints in the domain.

But the the internal biases and external constraints can interact in nontrivial ways. Figure 6.3B shows the same experiment, but now for a population biased towards base 8. The behaviour is remarkably different. Most notably, an octal system *can* be maintained, presumably since base 8 *does* express more than half the numbers in the domain. As a result, agents favouring base 8 are also more likely to use it. But there is a caveat: the initial frequency must be large enough. If $s_0 \geq 2$, the population always adopts a base-8 system. For $s_0 \leq 1$, base-8 slowly dies out and the decimal system takes over. But in between, $1 < s_0 < 2$ we observe a bifurcation. It should be noted that the plot shows the average over 300 runs. In every such run, the population seems to adopt *either* a decimal *or* an octal system.

These experiments demonstrate that the additive base game is not just the Bayesian naming game with a particular choice of bias. This initially seems plausible, since the implicit biases takes the form of a distribution over K bases $b_1, \ldots, b_K$ and even the strategy ($\xi = 1$) corresponds directly to a MAP–MAP strategy ($\eta = \zeta = \infty$) in the Dirichlet-categorical NG. But since biases in that game act as pseudo-counts, sufficient counter-evidence will make the bias disappear. Biases and experience are, in that respect, on completely equal footing. In the additive base game, this is not the case. No amount of counter-evidence can overcome the external constraints in the sense that they never disappear. We return to this point in the discussion.

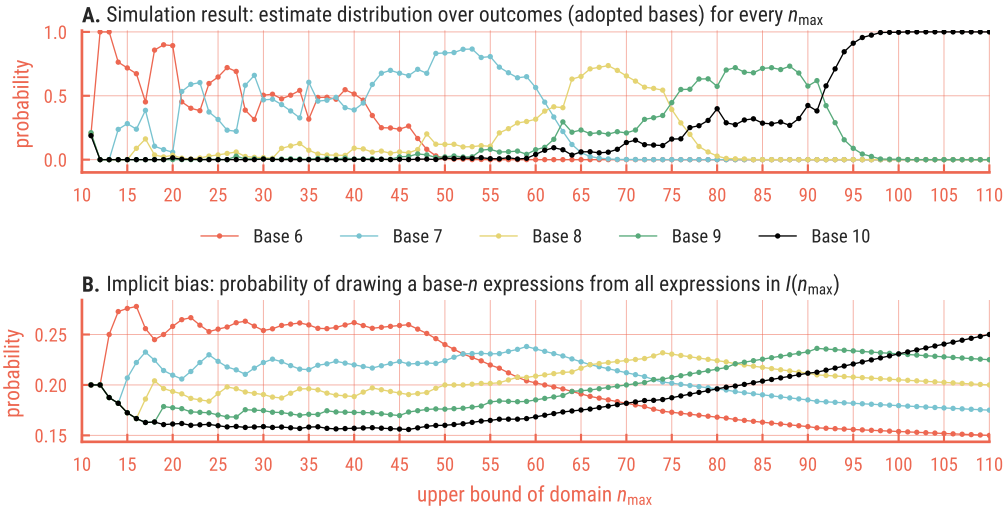


FIGURE 6.4 Domain adaptivity in the multiplicative base game. Figure A. shows the distribution over outcomes (adopted bases) for every the domain $I(n_{max}) = \{11, \ldots, n_{max}\}$. Note that the plot shows 99 distributions, one for each n_{max}. The game appears to exaggerate certain biases implicit in the domain. Figure B. shows an approximation of these biases: the probability that an expression randomly drawn from all expressions for numbers in $I(n_{max})$ uses base b. Details are in the main text.

HUR07 Results shown for $N = 200$, $\eta = 1$. Each of the 99 distributions is the average over 600 runs.

DOMAIN ADAPTIVITY IN THE MULTIPLICATIVE BASE GAME The driving force behind the implicit bias in the additive base game is the difference in *expressivity* between bases, with the largest base having a strong expressive advantage over the other bases. One might wonder: are there other implicit biases, that come to the fore when bases are *equally expressive*? This is best studied in the multiplicative base game, which allows for larger domains. The idea is to restrict the the domain such that no base has an expressive advantage. On a domain with upper bound $n_{max} = 6^2 + 10$, the limit of a multiplicative base-6 system, all all bases have equal expressivity. If there are no other biases, the probability that base is adopted should be equal for all bases. But initial experiments suggested that something very different was happening: the probability that a base would be adopted, depends on the size of the domain, i.e. on n_{max}. The following experiment therefore takes a more general approach. For every $n_{max} \in \{11, \ldots, 110\}$, 600 simulations of 5000 iterations were repeated, with domain $I(n_{max}) = \{11, \ldots, n_{max}\}$.[27] In each simulation, one particular base might be adopted. The experiment aims to quantify the probability that a base is adopted on a certain domain, by averaging over 600 runs.[28] I will call the distribution over adopted bases the *distribution over outcomes*.

Figure 6.4A shows the results. Every point on the x-axis corresponds to a domain $I(n_{max})$. *Above* that point, the distribution over outcomes is plotted. The figure thus shows $110 - 11 = 99$ distributions. It indicates that different bases are more likely to be adopted on different domains $I(n_{max})$. When agents communicate about $I(18)$, base 6 is adopted most frequently, when communicating about $I(70)$, base 8 is most likely to emerge, etc. The pattern is somewhat jumpy, but probably not because of random noise. The estimates seem to be reliable, since averaging only 150 runs yields a similar pattern as the averages of 600 runs shown here. There seems to be a *structural* reason for the jumpiness, a bias implicit in the structure of the domain, to which the language adapts.

In the additive base game, we specified the bias explicitly. In this case, we will approximate it with the relative frequency of base-b expressions. How many ways there are to express a number n using base b? Consider $n = 26$ and $b = 6$. The only possible

28 More precisely, we recorded the final distribution over bases, since it sometimes happened that the population had not yet converged after 5000 iterations.

27 In Hurford's game, agents communicate about $\{21, \ldots, 110\}$, but for this experiment the domain was extended.

expressions in the multiplicative base game are $4 \times 6 + 2$ and $3 \times 6 + 8$. More generally, the only two possible base b-expressions are:

$$a \times b + c \quad \text{and} \quad (a-1) \times b + (b+c). \tag{6.5}$$

After all, for any lower factor $a-2$, the remainder would be at least $b+b+c > 2b > B$, and therefore inexpressible in the game. So how many base-b expressions does a number n have? We denote this quantity by $N_b(n)$. If $n > b^2 + B$, it simply cannot be expressed and $N_b(n) = 0$. Now let $c := n \bmod b$. If $b + c > B$ then eq. 6.5 shows that $N_n(b)$ must be 1. The same holds when $c = 0$, since preference is then given to the the simpler $x \times b + 0$. In all other cases, there are 2 expressions, so in sum,

$$N_b(n) = \begin{cases} 0 & \text{if } n > b^2 + B \\ 1 & \text{if } b + c > B \text{ or if } c = 0\ , \\ 2 & \text{otherwise.} \end{cases} \qquad c := n \bmod b \tag{6.6}$$

The total number of base-b expressions for numbers in the interval $I(n_{\max})$ is the sum $N_b(I(n_{\max})) := \sum_{n=n_{\min}}^{n_{\max}} N_b(n)$, and the relative frequency of base-b expressions amongst all expressions in the interval $I(n_{\max})$ is

$$f(b, n_{\max}) := \frac{N_b(\ I(n_{\max})\ )}{\sum_{b'} N_{b'}(\ I(n_{\max})\ )}. \tag{6.7}$$

Figure 6.4B shows the relative frequencies for all bases and 99 values of $n_{\max}$, corresponding to subfigure A. The figure suggests that the relative frequency $f(b, n_{\max})$ is a good first approximation of the implicit bias implicit in the domain, and the multiplicative base game seems to have exaggerated this bias. The match is far from perfect, since we have not accounted for all complexities of the game. First, the peaks in the implicit bias (subfigure B) have shifted to the left in the simulation results (subfigure A). Consider the implicit bias towards base 6 in $I(90)$. In an actual game with this domain, the base would never be adopted, since more than half of that domain is not expressible by base 6. We have not accounted for that, but the game, of course, does and accordingly 'shifts' the implicit bias to the left (e.g. 6 has lower probability there). Second, the game seems to exaggerate the implicit bias in the sense that bases with very low relative frequency are never adopted, and most probability mass is moved to the most frequent base. This is in line with earlier findings about frequency strategies in chapter 4, but does not seem to correspond to exponentiation alone.

Even though the approximation of the implicit biases is far from perfect, the experiment strongly suggests that there can be are other biases at play, besides the expressive advantage. In this case the sheer number of expressions for a given number appears to give rise to a complex bias that makes different bases more likely to be adopted on different domains. The language, numeral system, in this sense exhibits a kind of *domain adaptivity*. The cultural process moreover seems to exaggerate the structure of the implicit bias, and the underlying reason might be the amplifying character of the frequency strategy. In the final section of this chapter we further discuss the implicit biases.

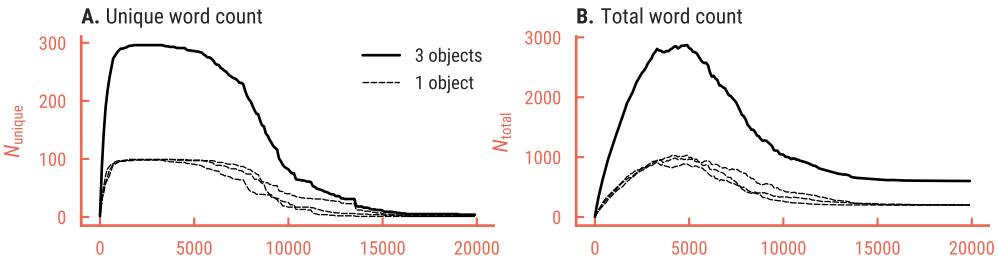


FIGURE 6.5 A naming game with three objects is just the sum of three independent single-word NG's when there is no homonymy. Dashed lines show the statistics per object; solid lines for the 'total' game: the sum of the dashed lines.

FIG07 Results shown for 1 run; $N = 200$, using the current strategy.

Counting games

Hurford's models only addressed the standardisation of a base and assumed that (1) the atoms of the system were in place and (2) that the additive or multiplicative constructions had been invented. These two assumptions roughly correspond to the first three stages in the evolution of numeral systems as outlined in the previous chapter. In this section, rather than assuming these stages, we attempt to reproduce them. The first stage, the emergence of the first simple numerals, seems to be the type of problem naming games excel at. The lowest numbers subitised and could therefore be named directly. So, let's suppose there is no homonymy, and agents need to negotiate, say, 3 numbers. We could run the simulation, but there is really no need to do so: we already know that this works. Playing a naming game with 3 objects in the absence of homonymy is the same as playing 3 independent single-object games at the same time (see figure 6.5).

Can we extend this approach to the next stage, where a larger counting sequence of simple numerals emerges, by simply asking agents to negotiate more words? Yes, but it raises a problem: nothing about the words signals that they are ordered (although their semantics of course does). In fact, using a naming game to model the emergence of simple numerals commits to a certain conception of number. Recall from the previous chapter that the *referential-pragmatic hypothesis* holds that quantity is directly perceptible and can therefore be named unproblematically, as in a naming game. For the lowest numbers up to around 4 this might be realistic, but for larger numerals, it is not. The game we develop for that reason aims to align more closely with the *ritual hypothesis* (and Hurford's synthesis). It assumes that numerals are grounded in a practice of counting (Hurford 2007): reciting a the sequence *one, two, three*, and so on. As discussed in the previous chapter, this sequence might initially be a sequence of for example body parts. The result is that the word *eight* does not mean 8 because it refers to some object EIGHT, but because it is the 8th word in a conventionalised counting sequence. In this case, the semantics of numerals in a counting sequence are implicitly defined by their position in the sequence, by their order. This implies that numerals are *necessarily* ordered and continuous (uninterrupted), as some argue they should be (von Mengden 2008).

The goal then is to adapt the naming game so that agents negotiate a counting sequence. To do so, first consider a naming game *with* homonymy. In that case agents communicate object–word pairs (o, w) where the same word can occur in different pairs. For the population to reach coherence, the alignment strategy has to dampen

competing pairs: all pairs using the same object, but a different word, *or the same word, but a different object*. All this is very similar to the original naming game, when words are replaced by pairs and the inhibition rules is extended accordingly. We will use[29] lateral inhibition strategy 1 from 3.1, that is, with

$$\delta_{\text{inc}} = \delta_{\text{inh}} = \delta_{\text{init}} = 1, \qquad \delta_{\text{dec}} = 0 \tag{6.8}$$

Next, we replace object-word pairs by pairs of consecutive numerals: (one, two), (two, three), and so on. To illustrate this, suppose an agent knows the following pairs:

$$(\textsc{start}, \text{one}), (\textsc{start}, \text{two}), (\text{two}, \text{three}), (\text{two}, \text{four}), (\text{one}, \text{two}),$$

where START is a purely administrative symbol indicating the start of the counting sequence. The agent can form the following sequences:

START, one, two, three,	START, one, two, four,
START, two, three,	START, two, four

Other agents might be able to form different sequences, but the idea is that after many interactions, a consensus will emerge where only one sequence remains in use.

In every interaction the agents must somehow communicate a *sequence* of words, rather than single words. This can be done in various ways, and I experimented with three slightly different scripts (L denotes the *limit* of the system):

1. **Unbounded counting game.** Starting from $x_0 = \textsc{start}$, the speaker iteratively chooses the next pair (x_i, x_{i+1}) with the highest score, *generating as many new pairs as necessary*, until it reaches a randomly drawn number $n \leq L$. The sequence $(\textsc{start}, x_1, x_2, \ldots, x_L)$ is presented to the hearer, who updates the score of every pair (x_i, x_{i+1}) according to the alignment strategy. (With $L = 1$ this is a normal naming game.)
2. **Instructional counting game.** Similar to the unbounded version, but now the speaker can invent at most 1 new pair, *whose score remains 0*, and always tries to count up to L (rather than to $n \leq L$). In this game, agents can therefore only acquire new numerals if another agent instructed them how to count on.
3. **Joint counting game.** Starting with START, the speaker utters the first word. Both agents perform the usual updates. *Only if the hearer knows the word*, i.e. if communication was successful, they continue with the next word. The round goes on in this fashion until communication breaks down or they reach L.

To measure the dynamics of these games, besides the usual statistics the *(counting sequence) length* $\ell(t)$ is measured: the length of the initial segment of the counting sequence about which the entire population agrees at time t. It should also be noted that a complication arises when pairs are removed. If one thinks of the pairs as describing a tree with START at the root, the removal of a pair can render an entire branch inaccessible. To keep statistics like N_{total} (which now counts pairs) informative, all inaccessible pairs are removed before collecting the statistics. Finally, the communicative success is no longer boolean but averaged over all words in a round.

29 This is actually an unfortunate choice, since figure 3.3 suggests convergence is relatively slow. When doing this work, I was however not aware of the existence of lateral inhibition strategies and the like.

The dynamics of the three games is visualised in figure 6.6. All three games allow the population to negotiate a shared counting sequence, but in quite different ways. In

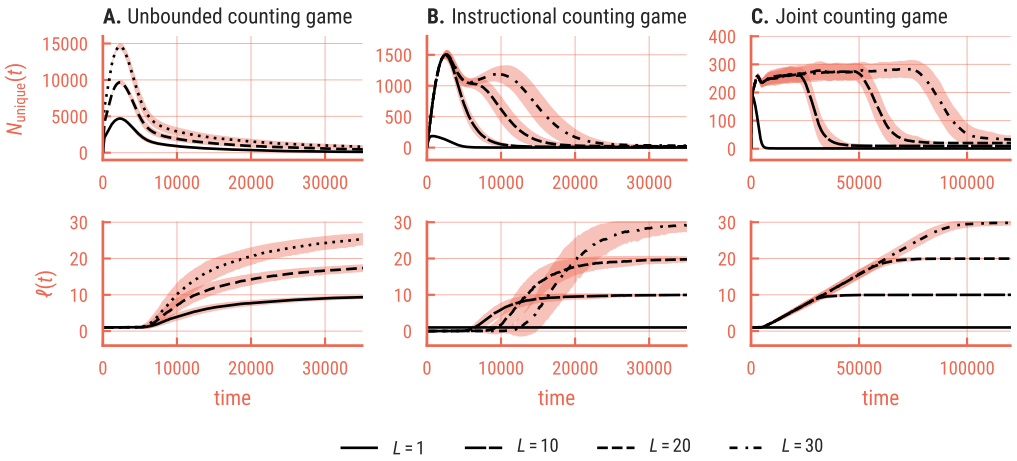


FIGURE 6.6 Dynamics of the three counting games measured by the number of unique pairs and the initial segment length. In all cases the population develops a counting sequence, but the dynamics are strikingly different. See main text for details.

Results shown for $N = 200$; avg. of 600 runs, 1 std. shaded.

the unbounded counting game (A) convergence time does not depend on the limit L: negotiating 10, 20 or 30 simple numerals takes equally long. Closer inspection reveals that the population has 'found' a simple trick. My implementation happens to represent words by successive integers (rather than, say, strings). A typical[30] counting sequence at the end of the game turns out to be of the form

$$\text{START},\ \underbrace{212, 213, 214}_{\text{fragment 1}},\ \underbrace{350, 351, 352, 353}_{\text{fragment 2}},\ \underbrace{658, 659, 660}_{\text{fragment 3}}.$$

where the numbers are 'words' and every fragment is generated by a single speaker. It seems that the population has 'adapted' to a higher L by communicating *fragments* rather than pairs. Since fragments become larger when L does, convergence times remain approximately stable.

In the instructional game this cannot happen, since in every encounter at most one word can be invented. But the game does exhibit some funny behaviour. If a population for example has to count up to $L = 10$, they can do so earlier than when they have to count up to $L = 30$. This can be seen from the length $\ell(t)$ in figure 6.6B and is, of course, mildly absurd. Interestingly, the joint counting game results in perfectly linear behaviour: the population is always exactly on the same page. After all, a speaker can only count on if the hearer agreed about the sequence so far.

SIMPLE NUMERALS, AND BEYOND? Populations of counting agents can negotiate a counting sequence, so can they go on and negotiate a recursive numeral system? The idea was again to again take inspiration from the evolutionary account in chapter 5 and see if a practice of 'grouping' could give rise to serialisation, or, simply put, additive constructions: ten stones and another 3 (Hurford 2007). When counting agents could at some point decide to group what they had counted so far. This means that agents have to score possible group sizes — bases, really — and use a base depending on its score. In other words, agents would have to play a naming game to negotiate the group size, the base, while simultaneously playing a counting game to negotiate a sequence of atoms. These two problems are interdependent: the length of the counting sequence

30 I have checked all this by computing the distribution of fragment lengths for every L.

is determined by the base, or, conversely, the base is the numeral where the counting sequence of atoms ends.

What I typically observed in all experiments with games of this type — I considered many variations on this theme — was that negotiating a single base is much easier than negotiating a long counting sequence. The population would thus adopt the first base it could count to (typically 2, as I excluded group size 1). To prevent this from happening one should, in one way or another, get the population to postpone the negotiation of the base, until the counting sequence has developed sufficiently. But this amounts to implementing an implicit bias towards a certain base. I did not continue along these lines, since it appeared to add little to our understanding of the the evolution of numeral systems. After all, if the model is in the end bringing some implicit biases to the fore, one might as well take the Bayesian naming game and encode the biases explicitly as a prior.

Conclusions

This chapter explored the first experiments trying to align agent-based simulations of the cultural evolution of numeral systems with their actual evolution, as discussed in chapter 5. The starting point was the pioneering work of James Hurford, who introduced the base games where the population negotiated a shared base with additive and multiplicative constructions. Detailed analyses suggested that the base games implements a strong implicit biases towards the highest base, resulting from the expressive advantage that base has: it can express all the numbers in the domain, whereas the smallest base can only express a few numbers. When this advantages is removed by restricting the domain such that all bases have equal expressivity, more subtle frequency patterns in the domain appear to determine the outcome of the cultural process. Further analyses revealed that the biases implicit in the domain affect the behaviour of the game differently than biases of agents. The latter can be overcome by counter evidence, the former cannot.

The findings highlight one simple point: naming games can be driven by implicit biases. This is reminiscent of early iterated learning models, where the biases leading to the emergence of compositionality were often hard to isolate. In iterated learning literature (Kirby, Dowman, and Griffiths 2007; Kirby, Dowman, and Griffiths 2007) explicit biases are often advocated, as is done in Bayesian models. The additive base game provide a compelling argument for doing the same with naming games. Explicating the biases, as we did in eq. 6.4, makes transparent to what extend the outcome (a decimal system) is determined by the assumptions in the model. However, we have also seen that hard constraints influence the behaviour of the model differently than the biases in for example the Bayesian naming games. This suggests that Bayesian models not only introduced explicit biases in the iterated learning literature, but might also have changed the way the biases work compared to early iterated learning models.

Although the games exhibit a certain domain adaptivity, one must be careful not to jump to the conclusion that numeral systems are shaped by their domain. After all, the adaptivity in the multiplicative base game is rather artificial.[31] Base 6 has higher expressivity on, say, $I(18)$ than base 10, only because the game allows the summand c

31 It might be interesting to note that in some trial experiments where, as the result of some bugs in the implementation, prime numbers had a slight frequency advantage, leading to the emergence of prime bases. Clearly these kind of biases are not driving the evolution of numeral systems.

in an expression $a \times b + c$ to be larger than the base b, as seen from eq. 6.5.[32] Such *overrunning* is rarely found in actual numeral systems, as discussed in the previous chapter. That adaptivity of the multiplicative base game (figure 6.4) is therefore an artefact of overrunning. It further implies that synonymy only disappears if the domain is not restricted, that is, if base 10 *does* have an expressive advantage. On smaller domains a population might evolve a base 6 system, where both the expressions $3 \times 6 + 8$ and $4 \times 6 + 2$ would be perfectly 'standard'. This is not only undesirable for a model of standardisation, it also has consequences for the packing strategy, which Hurford's simulations partly aimed to explain. The first of these expressions, after all, violates the packing strategy.

MECHANISMS AND INTERPRETATIONS Although the base games indicate that social processes can lead to standardisation of a base, that is not to say that the model is *realistic*. Roughly, the models suggest that standardisation arises because people prefer to use the most frequently observed bases. Interestingly Hurford himself suggests a very different explanation before introducint his simulations, arguing that standardisation would arise because "the obvious communicative advantages of standardised, canonical forms, and because no communicative advantage is lost by such a standardisation, due to the rather special nature of numerals/numbers" (p. 273). Something along those lines indeed seems plausible, but such explanations are not supported by the models. I think this highlights a serious difficulty for the use of models: the mechanism and the interpretation need not be aligned. That is, the mathematical explanation of why a model exhibits the behaviour it does, is often not in line with the interpretation given to the model, since that is typically based only on the behaviour observed in simulations. And insofar the interpretation and the mechanism are misaligned, modelling does not contribute much to showing the consistency of an informal account either. One way to resolve such discrepancies is by evaluating models against empirical data: by testing the underlying mechanisms and the predictions they make.

When it comes to that, the base games do not await a very bright future. They for example predict that higher bases are more likely to be adopted, while in fact bases higher than 20 are rare (Hammarström 2009). Moreover, numeral systems use multiple bases to counter expressive restrictions, so the disadvantage of a base-6 system in the base games is in reality ameliorated by introducing the larger base 36.

32 I am not sure if Hurford (1987) was aware of this. His definition of a base is explicit and implies that "7 would be the the base in both (3×7) and $((4 \times 7) + 3)$" (p. 295). He also mentions expressions such as $3 \times 4 + 9$ (p. 294), which leave me to conclude that he did not rule out overrunning.

7 Conclusions

What can models learn us about language evolution? That is the question this thesis set out to address. Without any introduction, it suggests a fairly philosophical thesis, but we found nothing of the sort — deliberately so. This thesis aimed to delineate the space of possible model behaviour and in that way get some understanding of what we can expect to learn from the models. Of course, the space is restricted to 'relevant' models, interesting enough to have been studied for several decades. Those, by and large, come in two flavours: the horizontal naming games and vertical iterated learning tradition. The main contribution of the thesis is an attempt to unify these two traditions in a new model: the Bayesian language game.

BAYESIAN NAMING GAME This Bayesian language game is a simple extension of the Bayesian *naming* game, also proposed in this thesis. In the Bayesian naming game, all agents haven an internal language θ from which they draw words in every encounter. After observing utterances, they update their beliefs using Bayesian updating. Concretely, we studied a Dirichlet-categorical variant of this framework, meaning that the prior beliefs of the agent are modelled by a Dirichlet distribution. Although mathematically simple, the game exhibits a rich behaviour. First of all, the population converges to a shared, stable language (D6). Second, this language *reflects the bias*: it is clearly shaped by both the innate biases and the cultural process, in a non-trivial way. That means that different lineages develop different languages (D4) Third, the game develops in three stages, metaphorically called 'infancy', 'puberty' and 'adulthood'. In the first two stages the language primarily develops its characteristic, contingent shape, and during 'adulthood' it stabilises.

The Bayesian naming game addresses the lack of language stability in the Bayesian iterated learning models, but it also incorporates one of the main innovations of the Bayesian models: the explicit representation of innate biases (D1). Further, in this game the innate biases can be transparently separated from past linguistic experience. Consequently, the 'total' beliefs of an agent are, quite literally, the sum of innate biases and past experience. Another innovation of the Bayesian iterated learning models was the study of different strategies agents use for selecting languages: using the most likely language under the posterior (MAP or maximising) or sampling a language. This can be directly translated to the Bayesian naming game in the form of a parameter η, together with a production strategy, parametrised by ζ, derived from the naming game literature. After all, the game is primarily a *naming* game and was accordingly shown to implement a kind of lateral inhibition, one of the alignment strategies used in that field.

BAYESIAN LANGUAGE GAME The Bayesian *language* game extended the Bayesian *naming* game by changing the population structure so that it can either take the form of a naming game, or the form of an iterated learning model. The language game performs a random walk through a population of fixed size. It moreover adds a life expectancy, and when an agent dies it is replaced by a newborn agent. The random walk, when unraveled, forms a chain (iterated learning), but since it is random, it simultaneously approximates homogeneous mixing (naming game). The crucial parameter that interpolates between naming games and iterated learning is the life expectancy γ. The Bayesian language game is therefore parametrised by three crucial parameters: η for

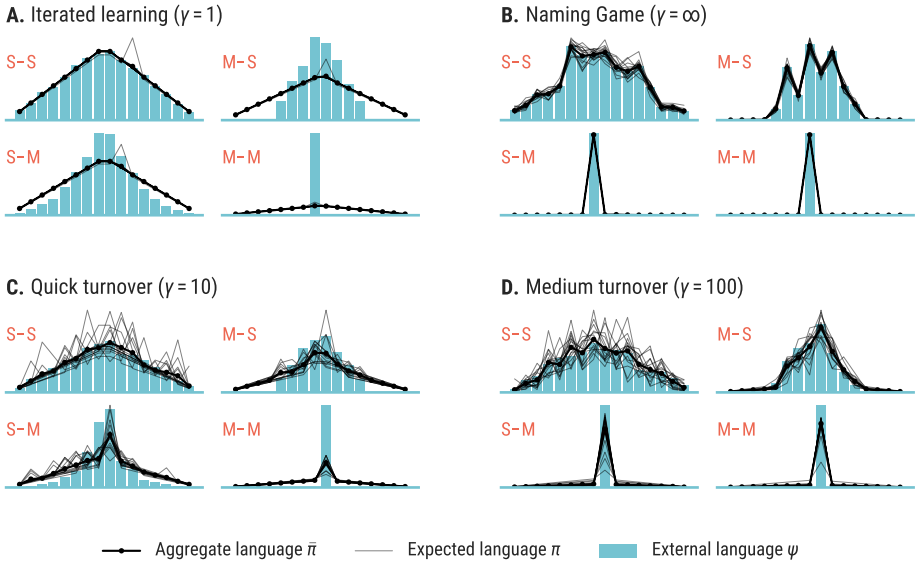


FIGURE 7.1 Typical outcomes of the Dirichlet-Categorical language name for the extreme strategies (sample–sample, MAP–sample, sample–MAP, MAP–MAP) in populations with immediate turnover (A, iterated learning, $\gamma = 1$), no turnover (B, naming game, $\gamma = \infty$) and two intermediate turnovers (C and D).

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

the language strategy, ζ for the production strategy and γ for the life expectancy. This, then, is the 'space of agent-based models' whose behaviour we aimed to characterise: the parameter space (η, ζ, γ) of the Bayesian language game.

The eight extreme cases, with $\eta, \zeta, \gamma \in \{1, \infty\}$ are of special interest. With $\gamma = 1$ the language game reduces to iterated learning, and with $\gamma = \infty$ it reduces to a naming game. The pure sampling strategy (sample–sample) corresponds to $\eta = \zeta = 1$; the mixed sampling-maximising strategies to $\eta = 1, \zeta = \infty$ (sample–MAP) and $\eta = \infty, \zeta = 1$ (MAP–sample); and the pure maximising strategy to $\eta = \zeta = \infty$ (MAP–MAP). A systematic search through the space suggested that the behaviour of the game would relatively smoothly change between these eight extreme cases. Characterising their behaviour therefore goes a long way to charting the behaviour one can expect from agent-based language models. The result of cultural evolution in the extreme cases (and two intermediate population turnover rates) are shown again in figure 7.1. The main conclusions it suggests are:

- Bayesian naming games results in languages reflecting the bias, but shaped by cultural evolution.
- In iterated learning models, no agent faithfully represents the external language.
- Pure sampling strategies stay close to the bias, either perfectly (iterated learning), or imperfectly (naming game)
- Mixed strategies differently 'exaggerate' the bias: maximising languages results in some kind of pruning, maximising productions in some kind of exponentiation.
- Pure maximising strategies result in degenerate languages.
- The language spoken in an iterated learning model seems predictably determined by the biases — even for maximising strategies.
- Intermediate life expectancies interpolate between these two extremes.[33]

That, in short, that was the answer this thesis formulated on the first subquestion ad-

33 With one notable exception: there might be stable language change in between, but possibly only for a very limited parameter range.

dressed in this thesis: what kind of behaviour can we expect from agent-based models of language evolution? Well, see figure 7.1, and interpolate!

EMPIRICAL VALIDITY That might be a good start, but what does it tell us about language evolution? In all fairness, I do not think I have formulated anything close to a clear answer to this question. Perhaps because this is a not exactly an easy question.

In the second part of the thesis I identified numeral systems as an interesting test case for models of language evolution. I gave various reasons, pertaining to the amount of linguistic data available, the sheer size of the design space (which makes for an interesting search problem), the balance between expressivity and simplicity in numerals and, most importantly, that the structure of numerals themselves suggest a reconstruction of their origins. The last chapter tried to connect that reconstruction to agent-based models of language evolution. That led to some interesting findings, such as the fact that the biases of agents and the constraints of the domain interact in different ways. Or the sensitivity of naming games to subtle frequency biases in the underlying domain.

In fact, that finding is perfectly illustrative of the problem: all agents seem to be doing in these models is a more or less sophisticated form of frequency administration. Administration surely helped the rise of human civilisation, but perhaps not this kind of administration. *Perhaps*. The point is that adapting to frequencies describes a *mechanism* that amounts to an explanation of the observed phenomenon — but is it the right one, is it justified? Frequency 'administration' could explain the standardisation of bases, as we have seen in the last chapter. But one could also explain that by using arguments of communicative efficiency, as Hurford (1987) also proposed. Which one to prefer? Or take a different example: the idea that compositionality emerged because it balances expressivity and compressibility. The explanations suggested for one of the purest compositional structures, that of numerals, point in a quite different direction: a counting practice, grouping objects, or, later, processes of grammaticalisation. Which one is it?

As soon as one leaves the world of models and enters the world of language, in this case numeral systems, it becomes clear that you have moved to a different, lower level of description. One could conclude that the higher level description is 'wrong', but that would be too simple. Descriptions at different levels are rarely perfectly aligned — that's the whole point of having different levels of descriptions in the first place. But what, then, is the conclusion? Perhaps that it is time to see if the levels of description align at all. That is, to see if the models can explain actual linguistic data, outside the computer or the lab.

I did not like Berwick and Chomsky's paper, as the reader could no doubt tell from the introduction. Just as pretty much any other field of science, the 'Kirby-type work' has problems and good, sharp critiques are refreshing. But the style — demeaning, territorial, grotesque. I mean, the academic equivalences thereof. Science, the knowledge commons, those are, in the end, beautiful, praiseworthy things and you would hope for a more more communal spirit.[34] In any case, there is one point at which I could not disagree with Berwick and Chomsky (2017): when they write that, as far as they know, the "whisper down the lane" properties (slang for 'iterated learning') have not been applied to "the empirical phenomena of language change that linguists have actu-

34 Ah, how could I have missed that? The second sense: "(of conflict) between different communities, especially those having different religions or ethnic origins: violent communal riots." That explains everything.

ally measured" outside the lab or simulations. It is just that I would prefer not to read that as a fatal blow, but as a programme.

I will leave it at that. The remainder of this chapter lists the main contributions, some points of discussion and future work.

Main contributions

The main two contributions of this thesis are (1) the Bayesian Naming Game, an attempt at unifying the naming game with the Bayesian model of iterated learning, and (2) the identification of numeral systems as a testbed for models of cultural language evolution. Various other additions to existing literature have been made throughout the thesis. The ones I believe to be of most interest are listed below.

- **Lineage-specific languages can reflect a prior.** A demonstration that longer lifespans can lead to lineage-specific, shared languages that reflect the prior, but do not mirror it exactly. This suggests that cultural evolution *can* change languages in nontrivial ways, while still being subject to innate constrains.
- **Random walks as a transmission model.** A new model of cultural transmission in the form of a random walk through a population of fixed size. This combines the transmission chains with homogeneous mixing.
- **Connecting naming games to iterated learning** The thesis tried to connect two agent based modelling paradigm. Most specifically this results in a direct mathematical parallel between the model of De Vylder and Tuyls (2006) and Bayesian iterated learning models.
- **Numerals as a testbed for language evolution.** I have argued that numeral systems are a promising test case to relate models of language evolution with actual language. Numeral systems (i) allow one to model a subsystem of natural language directly; (ii) there is plenty of linguistic variation (i.e. the search of possible systems); (iii) the cognitive capacity for numerosity has been studied extensively; and (iv) there are benchmarks in the form of a reasonable reconstruction of the origins of numeral systems, which is moreover supplemented by recent empirical studies.
- **Realistic population turnover using Weibull distributions.** Proposed a more realistic death-process for iterated learning models based on the Weibull distribution, often used to model life-expectancy. Earlier iterated learning studies have often assumed constant hazard rates, which are unrealistic.
- **Revisited Hurford's base games.** I have shown that Hurfords early simulations of the standardisation of a base are variants of the Naming Game.
- **Domain-adaptivity in the Base Game.** Demonstrated that the Base Game adapts to the domain in the sense that the behaviour depends on frequency patterns determined by domain.
- **Counting Game.** A new type of naming game in which agents negotiate a counting sequence. In this model simple numerals derive their meaning solely from their position in the sequence.

- **Packing strategy without grammar.** A reformulation of the packing strategy independent of the phrase-structure grammar proposed in Hurford (1975) in terms of the semantic (arithmetic) structure of the numerals. This is particularly relevant for empirical studies of this proposed universal.

Future work

Throughout this thesis, possibilities for future work have been identified. Let me list the most practical points first, and then move on to the more interesting questions.

- **Other lateral inhibition strategies.** This thesis only compared Bayesian updating has only to the *basic* lateral inhibition strategy in Wellens (2012). Wellens also mentions the so-called *interpolated* LI strategy, which can be interpreted as the Rescorla-Wagner or Widrow-Hoff rule. In hindsight, that seems closer to Bayesian updating (in form, at least) and future work could address the exact relation.
- **Measures for the Bayesian naming game**. The measures used to analyse the Bayesian naming game could be improved by developing appropriate variants of measures used in naming games (communicative success, number of (unique)). The methodology of chapter 6 could be a starting point.
- **Scaling relations.** The change in convergence time in the Bayesian naming game under varying population, vocabulary and bottleneck size could be investigated more thoroughly, either empirically or analytically.
- **Models of life-expectancy.** Chapter 4 argued for more realistic models of life-expectancy without actually analysing whether they make a difference in the simulations. A systematic analysis would be valuable.
- **Random walks vs. homogeneous mixing.** Preliminary experiments did not reveal any systematic differences between random walks and homogeneous mixing in the Bayesian naming game, but this should be checked in a principled fashion.
- **Mathematical problems.** The mathematical analyses gave rise to several problems. The first concerns the characterisation of the posterior distribution of an agent with a maximising word strategy $\zeta > 1$; see appendix C. Another concerns the bias in Hurford's additive base game: has this distribution (proportional to the difference of two harmonic numbers) been studied before?

OTHER BAYESIAN NAMING GAMES This thesis mainly developed the Dirichlet-categorical naming game as the simplest instantiation of a more general framework. Future work could study other models inside the same framework. An interesting possibility would be to drop the conjugacy assumption, and consider more general probabilistic models. In fact, the Dirichlet-categorical model surfaced from early experiments with Bayesian agents represented as so called *probabilistic programs*. Probabilistic programming (see Ghahramani 2015 for an overview) provides a remarkable representational flexibility by using programming languages to specify the models, and their universal approximate inference algorithms (sampling or variational methods) significantly speed up development time. The downside in our context would be that poor inference could

have unexpected results when repeated for tens of thousands of rounds and thus trouble the analyses.

CONNECTIONS TO BIOLOGICAL EVOLUTION Reali and Griffiths (2010) were the first to connect Bayesian models of iterated learning to biological evolution, i.e. the Wright-Fisher model. Since the Bayesian Naming Game is in many ways identical to their model, it seems worthwhile to explore further connections with models developed in population genetics.

NONPARAMETRIC EXTENSION OF THE BAYESIAN NAMING GAME The Bayesian Naming Game made the simplifying assumption that the number of words in the population is fixed, an assumption made earlier in De Vylder and Tuyls (2006). The model would stay closer to the original naming games if this assumption could be dropped. Invention after all plays a crucial role in the Naming Game (Steels 2011), but more generally it has been argued that inventors play an important role in language change and evolution (Hurford 1987). Fortunately, there is a straightforward extension of the Bayesian Naming Game which does not fix the number of categories — is not *parametric* in that sense. This model would use an infinite analogue of the Dirichlet distribution, a Chinese restaurant process, that allows for the use of an arbitrary number of words. Every agent then faces the choice of either using an old word or inventing a new one. This extension has been discussed in the iterated learning literature before Reali and Griffiths (2010) and Burkett and Griffiths (2010).

Combined with population turnover, the non-parametric model seems to solve another issue: that words never disappear from the population completely. In classical lateral inhibition games, words scoring below a certain threshold are removed. This mechanism allows the emergence of *efficient* vocabularies. Population turnover could replace that mechanism in a non-parametric variant of the Naming Game: low-frequency words might simply not be acquired by newborn agents, and consequently die out. If this prediction is indeed true, it suggests a very exciting possibility. Suppose one splits the population in two parts at a certain point in time, and let both halves continue the game separately. Given the stochasticity of the game, it seems likely that different words would die out, effectively leading to different lineages. Future work could investigate such 'linguistic speciation' in a non-parametric extension of the Bayesian Naming Game.

Note that this also highlights a problem with the notion of 'lineage specificity' as I have used it. 'Lineages' were understood to mean 'runs', since it has not been shown that one can, say, split the population and develop two distinct branches. For the Bayesian naming game, the later seems moreover unlikely, since the language develops its characteristic shape primarily in the early phase of the game, before agents reach 'adulthood'. After that, the population only converges to a completely stable language, which is not realistic either (Kirby 2001). Increasing population turnover addresses this, and the nonparametric suggestion outlined above might solve the former.

MATHEMATICAL ANALYSIS OF THE BAYESIAN NAMING GAME Simulations indicate that coherence emerges in the Bayesian Naming Game, but I have not provided a formal proof of this. In fact, there are two separate questions: (i) do all agents converge to the same

distribution? And (ii) can one characterise this distribution in terms of the parameters η, ζ and γ? It seems unlikely that this is a simple problem, since even the special case of the naming game has so far resisted analytical scrutiny (De Vylder and Tuyls (2006) only provide analytical results for a deterministic variant of the game). One possible line of attack would investigate whether the limiting distribution is indeed a sample from some distribution around the prior. If so, a characterization of that distribution would be likely to provide valuable further insights into many models of cultural language evolution simultaneously.

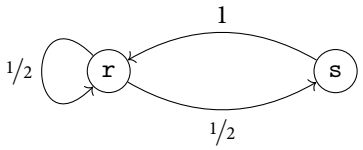
And the list continues. But that's it for now.

Appendices

A Converging Markov Chains

Convergence results for Markov chains are key to understanding the long-term behaviour of Bayesian iterated learning models. This appendix introduces those results for ergodic Markov chains.

RAINY AND SUNNY DAYS Let's consider a 'simple', non-linguistic scenario: the weather. Suppose sunny (`s`) and rainy (`r`) days are equally probable, but that every sunny day is deterministically followed by a rainy day:[35]



(7.1)

This model results in walks through the state space of the form `r`, `r`, `s`, `r`, `s`, `r`, `s`, `r`, `r`, `r`, `s`, `r`, `r`, `s`, and so on. The numbers along the edges are the *transition probabilities* of moving from one state to the next. Importantly, these probabilities do not change over time, a property known as *time-homogeneity*, and depend on the current state *only*. What happened in the past is irrelevant. This 'memorylessness' happens to be a defining characteristic of Markov chains. If $x_0, x_1, x_2, \ldots$ denote the states (rainy or sunny) at day $0, 1, 2, \ldots$, these random variables form a Markov chain if

$$p(x_t \mid x_0, \ldots, x_{t-1}) = p(x_t \mid x_{t-1}), \qquad t \geq 1, \tag{7.2}$$

that is, if the probability of being in state x_t *only* depends on the very last state x_{t-1} the system was in. What about the very first state? The probability that $x_0 = i$ is separately by a *initial distribution* π, just a probability vector if the number of states is finite. In our example, the state space $S = \{\mathtt{r}, \mathtt{s}\}$ is indeed finite. The transition probabilities can then be collected in a *transition matrix*

$$\mathrm{T} = \begin{bmatrix} 1/2 & 1 \\ 1/2 & 0 \end{bmatrix} = [t_{i \to j}]_{j,i \in S} \tag{7.3}$$

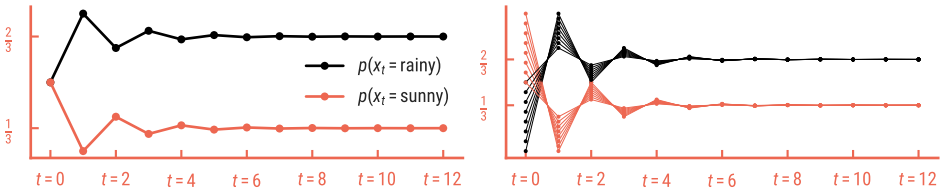
where entry $t_{i \to j}$ at position (j, i) is the probability of transitioning from state i to j. The probability that $X_0 = i$ is given by the initial distribution, but what is the probability we are in i at a later time, say $t = 1$? To find out we compute the marginal probability

$$p(X_1 = j) = \sum_{i \in S} p(X_1 = j \mid X_0 = i) \cdot p(X_0 = i) = \sum_{j \in S} t_{i \to j} \cdot \pi_i = (\mathrm{T}\pi)_j$$

In other words, the marginal distribution over states at time $t = 1$ is given by the vector $\mathrm{T}\pi$. Repeating this trick, we find that at time t this probability is given by $\mathrm{T}^t\pi$, where T^t is the t'th power of the transition matrix.

35 This 'gappy process' is adapted from Mathias Madsen's notes on *Random Processes and Ergodicity* (Madsen 2015). The presentation of the formalism is based on Norris (1997).

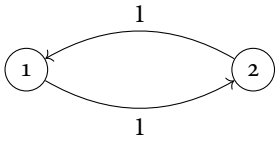
LIMITING AND STATIONARY DISTRIBUTIONS Suppose we start with initial distribution $\pi = (1/2, 1/2)$. Then the marginal distribution at $t = 1$ is $T\pi = (3/4, 1/4)$, in the next time step $T^2\pi = (5/8, 3/8)$, then $(11/16, 5/16)$, $(21/32, 11/32)$, and so on. These distributions converge to the *limiting distribution* $\pi^\star = (2/3, 1/3)$:



After a while, it will rain with probability $2/3$ and will be sunny with probability $1/3$, regardless of the initial condition, or so the plot on the right suggests. Looking at the transition diagram 7.1, that makes sense: every sunny day necessarily comes with one extra rainy day. Note that a 'converged' chain still hops between states (i.e., it will be either rainy or sunny), only the *probability* with which it is one of these stabilizes.

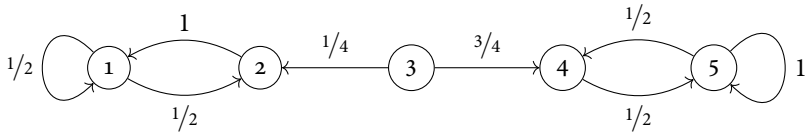
Now, once probabilities of being in state `r` and `s` are exactly $3/2$ and $1/3$ respectively, that should remain true in the next time step — they have converged, after all. And yes, that is as true as it gets. The distribution $\pi^\star$ is therefore also called the *stationary distribution*. The stationary distribution is left unchanged by the transition matrix, so $T\pi^\star = \pi^\star$, which makes it an *eigenvector* of the transition matrix with the eigenvalue 1. In our simple example, the limiting distribution (the limit of $T^t\pi$ as $t \to \infty$) and the stationary distribution (eigenvector of T) are the same. That is not true for all Markov chains.

APERIODICITY AND IRREDUCIBILITY Take this Markov Chain:



Starting from $X_0 = 1$, it will be in state 1 at all even times and in state 2 at odd times — this chain is *periodic*. As a result, $p(X_t = i)$ alternates between 0 and 1 and does not converge over time. But even though it has no limiting distribution, it does have a stationary distribution $\pi^\star = (1/2, 1/2)$, as is easily verified.

Here more more common thing that can go 'wrong': sinks can lead to multiple stationary distributions. Consider two connected copies of our weather model:



If the chain at some point reaches node 2 it will keep jumping between 1 and 2 afterwards. $\{1, 2\}$ thus acts like a 'sink' from which there is no escape. The set $\{4, 5\}$ is another such sink. Both sinks having their 'own' stationary distributions and depending on the initial condition, the chain converges to one of them — or a mixture. Concretely, write δ_i for degenerate and deterministic distribution with all mass on state i. If we start with the initial distribution $\pi_0 = \delta_1$, the chain converges to the stationary distribution $(2/3, 1/3, 0, 0, 0)$. But if we start on the other side with $\pi_0 = \delta_5$, then it converges to $(0, 0, 0, 1/3, 2/3)$. Finally, starting in the middle with $\pi_0 = \delta_3$ results in a mixture of both: $(2/12, 1/12, 0, 3/12, 6/12)$. All of these are stationary distributions. (And so are all the convex combinations of $(2/3, 1/3, 0, 0, 0)$ and $(0, 0, 0, 1/3, 2/3)$.)

In this case, there is no unique stationary distribution because the graph has several sinks from which one cannot reach the other parts of the graph. If *every* state is reachable from every other state with positive probability in a finite number of steps, then it is called *irreducible*. Such a Markov chain has no sinks and every state will almost surely be visited again and again.[36] It (almost) never stops visiting the *entire* state space.

ERGODICITY Markov chains that are both irreducible and aperiodic are said to be *ergodic*. Together, the two properties are sufficient to ensure convergence of a Markov chain to a unique stationary distribution, just as in our weather model:

Theorem 1 (1.8.3 in Norris 1997). *Let $(x_0, x_1, \dots)$ be an ergodic Markov chain with initial distribution π, transition matrix T and stationary distribution $\pi^\star$. Then*

$$p(x_n = i) \longrightarrow \pi^\star_j \quad as \quad n \longrightarrow \infty \tag{7.4}$$

for all $i \in S$.

As we have seen, an ergodic Markov chain over time traverses the entire state space in an aperiodic fashion. One might wonder if any regularity underlies the states it visits. Is it more likely to visit high-probability states under the stationary distribution, for example? Indeed, it is. The relative frequencies of visited states in fact converge to the stationary distribution. The important point is that this connects a distribution *over time* — the visited states — with a distribution *over the state space* — the stationary distribution. *Ergodic theory* studies this relation between time- and space-averages and I want to state one result here. To do so we define these averages, for any bounded function $f : S \to \mathbb{R}$, as

$$f_{\text{time}} := \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) \quad \text{and} \quad f_{\text{space}} := \sum_{i \in S} \pi^\star_i f(i) \tag{7.5}$$

The result is as follows.

Theorem 2 (Ergodic Theorem; 1.10.2 in Norris 1997). *Let $(x_n)_{n \geq 0}$ be as in theorem 1 and let $f : S \to \mathbb{R}$ be any bounded function. Then the time-average of f almost surely converges to the space-average of f:*

$$p\big(f_{time} \longrightarrow f_{space} \quad as \quad n \longrightarrow \infty\big) = 1. \tag{7.6}$$

36 This is generally not true for infinite state spaces, but introduced as an additional condition (*recurrence*)

I omit the proof, but want to draw a practical conclusion. As seen, one can find the stationary distribution as an eigenvector of (an estimate of) the transition matrix, but the previous theorem shows a more intuitive option: measuring relative frequencies. If $V_i(t)$ is the number of visits to state i before time t, then $f_i(t) = V_i(t)/t$ is the relative frequency. For an ergodic Markov chain, this converges to the stationary distribution, $f_i(t) \to \pi_i$ as $t \to \infty$, with probability 1. This is (a consequence of) the Ergodic theorem (Norris 1997, p. 1.10.2) when f_i is the indicator function. As the reader might notice, this fact has been used extensively.

B Lateral Inhibition Strategies

In chapter 3 we explored five different lateral inhibition strategies, and concluded that they always converge to an effective, shared language. Do these conclusions indeed generalize to the rest of the 6-dimensional, strategy space? The convergence proof suggests so, but does not apply to the naming games directly. In this appendix part of the parameter space is therefore explored systematically. The results indicate that effective languages eventually emerge for all strategies, although the dynamics before convergence can vary substantially.

Recall that the space of lateral inhibition strategies is defined by five nonnegative parameters

$$\delta_{\text{inc}}, \quad \delta_{\text{inh}}, \quad \delta_{\text{dec}}, \quad s_{\text{init}}, \quad s_{\text{max}}. \tag{7.7}$$

I have not been able to find a systematic analysis of the parameter space. Wellens (2012) does compare several strategies and suggest that the the value of the parameters determines the strategy. For example concluding that "a higher value [of δ_{inh}] improves alignment". I belief this is slightly inaccurate, since the strategies are invariant under scaling. In other words, it is the *relative* value of the parameters that matters. There are many more such equivalences. One could for example use any other $\delta_{\text{inh}} \geq s_{\text{init}}$ without altering the minimal strategy; or fix $s_{\text{max}} := s_{\text{init}}$ and use any $\delta_{\text{inc}} > 0$. Similarly, a different $\delta_{\text{inc}} > 0$ leaves the frequency strategy unchanged, since scores greater than s_{init} are of the form $s_{\text{init}} + k \cdot \delta_{\text{inc}}$ and essentially track the frequency k anyway.

We map two slices of the strategy space by fixing either the increments or inhibition parameter and varying the other (following Wellens 2012). Figure 7.2 reports the results. Fixing $\delta_{\text{inc}} = 1$ while varying δ_{inh} (figure 7.2A) reveals that the inhibition parameter δ_{inh} interpolates between the minimal strategy ($\delta_{\text{inh}} = 4^2$ or larger; purple) and and the frequency strategy ($\delta_{\text{inh}} = 0$; black). Both reach eventually reach perfect communicative success, but the stronger the lateral inhibition, the faster so. The number of unique words N_{unique} initially grows identically for all δ_{inh} as inhibition plays hardly any role at the start of the game. In the frequency strategy, no words are ever removed and the resulting vocabulary is therefore not *efficient*. It is hard to tell if the

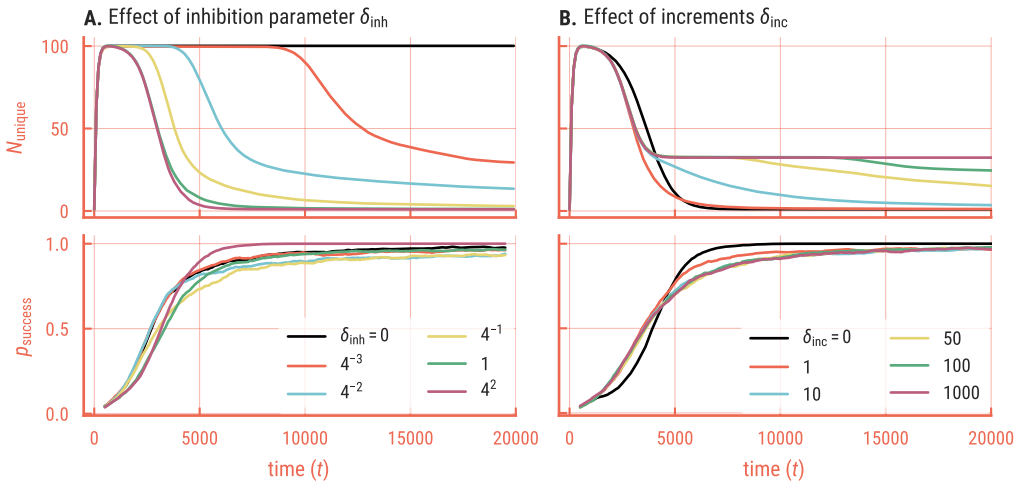


FIGURE 7.2 A. The effect of δ_{inh} keeping $\delta_{\text{inc}} = 1$ fixed. It interpolates between the minimal strategy and frequency strategy. B. the effect of δ_{inc} for $\delta_{\text{inh}} = 1$ fixed. For large δ_{inc}, the inhibition is rendered ineffective.

LINGO3 Results shown for $N = 200$, $\delta_{\text{dec}} = 0$, $s_{\text{init}} = 1$, $s_{\text{max}} = \infty$; avg. of 300 runs. p_{success} is moreover a rolling average over a centered window of 1000 iterations.

amount of lateral inhibition matters in the long-term. The plot seems to suggest that this is not the case, and even the slightest lateral inhibition will (after a significantly longer time) result in a one-word language.

The effect of the increment δ_{inc} is shown in figure 7.2B. One can see that the minimal strategy corresponds to $\delta_{\text{inc}} = 0$, but larger increments yield different dynamics. After the peak of N_{unique}, words with score $\delta_{\text{init}} = 1$ are quickly removed, as it takes a single inhibition. But words that have been heard multiple times have scores of at least $\delta_{\text{init}} + \delta_{\text{inc}}$ and need many more inhibitions to be removed. There appear to be around $N/6$ such words. The result is a temporary stabilisation of N_{unique}. Eventually inhibition takes over and competing words start disappearing. The (very) long-term behaviour thus appears to be the same as before: convergence to a single-word language.

C Mathematical details of Dirichlet-categorical NG

This appendix develops the Dirichlet-categorical naming game in a more rigorous fashion. Please refer to chapter 4 for extensive motivation.

First, recall our notational conventions. The most precise notation would be of the form

$$\alpha_A^{(t)} = \left(\alpha_{A,1}^{(t)}, \ldots, \alpha_{A,K}^{(t)}\right) \in \mathbb{R}^K, \tag{7.8}$$

and indicates the agent, the time and indices. I nearly always prefer a cleaner notation and often drop agents, or even time indices, whenever they are irrelevant. Also, vectors (boldface) get their time index in the subscript. Further recall that $\Sigma(\alpha) := \sum_k \alpha_k$ and that we write $[\![$ condition $]\!]$ for the indicator function evaluating to 1 if the condition holds and to 0 otherwise

Dirichlet and categorical distributions

The Dirichlet distribution is a continuous multivariate probability distribution defined over the interior of the $(K-1)$-simplex which we denote as $\Delta^{K-1} = \{x \in \mathbb{R}^K : \sum_k x_k = 1 \text{ and } 0 < x_i < 1\}$. We will only consider the $(K-1)$-simplex, so drop the superscript. Samples of a Dirichlet can thus be interpreted as K-dimensional probability-vectors. The Dirichlet is parametrised by a K-vector α and its density is given by

$$p(\theta \mid \alpha) = D(\alpha) \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}, \qquad D(\alpha) = \frac{\Gamma(\Sigma(\alpha))}{\prod_{k=1}^{K} \Gamma(\alpha_k)}, \tag{7.9}$$

where $D(\alpha)$ is the normalising constant, computed using the gamma function Γ, a continuous extension of the factorial with $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$. If $\theta \sim \text{Dirichlet}(\alpha)$, then it has the following properties (e.g. Bishop 2006)

$$\mathrm{E}[\theta_k] = \frac{\alpha_k}{\Sigma(\alpha)}, \qquad \mathrm{Var}[\theta_k] = \frac{\alpha_k(\Sigma(\alpha) - \alpha_k)}{\Sigma(\alpha)^2(\Sigma(\alpha)+1)}, \qquad \mathrm{Mode}[\theta_k] = \frac{\alpha_k - 1}{\Sigma(\alpha) - K}. \tag{7.10}$$

It is often convenient to parametrise the Dirichlet differently, as $\alpha := \beta \cdot \mu$ with $\beta \in \mathbb{R}$ and $\mu \in \Delta$. Here β is the concentration parameter, a kind of inverse variance, and μ determines the location of the distribution. This translates into

$$\mathrm{E}[\theta_k] = \mu_k, \qquad \mathrm{Var}[\theta_k] = \frac{\mu_k(1-\mu_k)}{\beta + 1} \tag{7.11}$$

from which we see that the mean is determined by μ and that larger β lead to smaller variance. We will use both parametrisations interchangeably.

The categorical distribution is a discrete probability distribution over K outcomes, described by a probability vector $\theta \in \Delta$. Recall that $c_k = \sum_i [\![x_i = k]\!]$ counts the number of k's in x. The joint distribution of b i.i.d. categorical variables $\mathrm{x} = (x_1, \ldots, x_b)$ is then given by

$$p(\mathrm{x} \mid \theta) = \prod_{i=1}^{b} \prod_{k=1}^{K} \theta_k^{[\![x_i = k]\!]} = \prod_{k=1}^{K} \theta_k^{\sum_i [\![x_k = k]\!]} = \prod_{k=1}^{K} \theta_k^{c_k}, \tag{7.12}$$

DIRICHLET-CATEGORICAL DISTRIBUTION To show that the Dirichlet is the *conjugate prior* of the categorical distribution, consider the following model

$$\theta \sim \text{Dirichlet}(\alpha) \tag{7.13}$$

$$x_1, \ldots, x_b \sim \text{Categorical}(\theta). \tag{7.14}$$

In this case, conjugacy means that the posterior distribution $p(\theta \mid \mathrm{x}, \alpha)$ is of the same parametric form as the prior $p(\theta \mid \alpha)$, namely a Dirichlet. More precisely, we have

$$p(\theta \mid \mathrm{x}, \alpha) \propto p(\theta \mid \alpha) \cdot p(\mathrm{x} \mid \theta) \tag{7.15}$$

$$\propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \cdot \prod_{k=1}^{K} \theta_k^{c_k} \tag{7.16}$$

$$= \prod_{k=1}^{K} \theta_k^{\alpha_k + c_k - 1}. \tag{7.17}$$

In the last line one can recognise a Dirichlet density with parameters $\alpha + \mathrm{c}$. We conclude that the the posterior is Dirichlet$(\alpha + \mathrm{c})$-distributed, or, more explicitly,

$$\theta \mid \mathrm{x}, \alpha \sim \text{Dirichlet}(\alpha_1 + c_1, \ldots, \alpha_K + c_K). \tag{7.18}$$

This result also illustrates the workings of the hyperparameter α. It is as if the model pretends to have observed α_k more instances of category k that it actually has. For that reason, the α_k's are often called *pseudo-counts*.

We can also derive the compound distribution $p(x \mid \alpha) = \int_\Delta p(x \mid \theta) \cdot p(\theta \mid \alpha)\ d\theta$ by marginalizing out all probability vectors θ. To do this, we have to use a trick, which exploits the fact that the Dirichlet distribution is normalized,

$$\int_\Delta D(\alpha) \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\ d\theta = 1. \tag{7.19}$$

Moving the normalising constant out of the integral, we see that

$$\int_\Delta \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\ d\theta = \frac{1}{D(\alpha)} \tag{7.20}$$

Using that trick we can compute the marginal probability of x as

$$p(\mathrm{x} \mid \alpha) = \int_\Delta \prod_{i=1}^{b} p(x_i \mid \theta) \cdot p(\theta \mid \alpha)\ d\theta \tag{7.21}$$

$$= \int_\Delta \prod_{i=1}^{b} \theta_{x_i} \cdot D(\alpha) \cdot \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\ d\theta \tag{7.22}$$

$$= D(\alpha) \int_\Delta \prod_{k=1}^{K} \theta_k^{\alpha_k + c_k - 1}\ d\theta \tag{7.23}$$

$$= \frac{D(\alpha)}{D(\alpha + \mathrm{c})} \tag{7.24}$$

$$= \frac{\Gamma(\Sigma(\alpha))}{\Gamma(\Sigma(\alpha + \mathrm{c}))} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + c_k)}{\Gamma(\alpha_k)} \tag{7.25}$$

Note that when $b = 1$, hence $\mathrm{x} = (x)$, the (almost defining) relation $\Gamma(n+1) = n\Gamma(n)$ can be exploited to further simplify the distribution. Concretely, note that $\Gamma(\Sigma(\alpha +$

$\mathrm{c})) = \Gamma(\Sigma(\alpha) + 1) = \Sigma(\alpha)\Gamma(\Sigma(\alpha))$. This simplifies the first term in equation 7.25, so we can simplify the marginal probability to

$$p(x \mid \alpha) = \frac{1}{\Sigma(\alpha)} \cdot \frac{\Gamma(\alpha_x + 1)}{\Gamma(\alpha_x)} \cdot \prod_{k \neq x} \frac{\Gamma(\alpha_k)}{\Gamma(\alpha_k)} \tag{7.26}$$

$$= \frac{1}{\Sigma(\alpha)} \cdot \frac{\alpha_x \Gamma(\alpha_x)}{\Gamma(\alpha_x)} = \frac{\alpha_x}{\Sigma(\alpha)}. \tag{7.27}$$

This also gives the posterior predictive distribution $p(y \mid \mathrm{x}, \alpha)$, since that is just $p(y \mid \alpha')$ for the updated parameters $\alpha' := \alpha + \mathrm{c}$:

$$p(y \mid \mathrm{x}, \alpha) = \frac{\alpha_y + c_y}{\Sigma(\alpha + \mathrm{c})} \tag{7.28}$$

This is a remarkably simple result, indeed: the probability of observing y is proportional to the number of times it has been observed already, including the pseudo-observations.

Exponentiated distributions

Different strategies can be used for selecting languages or words in the Bayesian Naming Game. This is done by exponentiating the distributions $p(\theta \mid \alpha)$ and $p(\mathrm{x} \mid \theta)$ by two parameters, η and ζ respectively. The resulting distributions again take simple form:

$$p(\theta \mid \alpha)^\eta \propto \left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\right)^\eta = \prod_{k=1}^{K} \theta_k^{[\eta(\alpha_k - 1) + 1] - 1} \tag{7.29}$$

$$\propto \mathrm{Dirichlet}(\theta \mid \eta(\alpha - 1) + 1). \tag{7.30}$$

The case of the categorical is obvious,

$$p(x \mid \theta)^\zeta = \frac{\theta_x^\zeta}{\Sigma(\theta^\zeta)}. \tag{7.31}$$

So we conclude that

$$p_{\mathrm{LA}}(\theta \mid \alpha) = \mathrm{Dirichlet}(\theta \mid \eta(\alpha - 1) + 1) \tag{7.32}$$

$$p_{\mathrm{PA}}(x \mid \theta) = \mathrm{Categorical}(x \mid \theta^\zeta / \Sigma(\theta^\zeta)) \tag{7.33}$$

THE DIFFICULT CASE $\zeta = \infty$ Whenever $\zeta \neq 1$, Bayesian agents are facing a different inference problem and should infer the posterior distribution $p(\theta \mid \mathrm{x}, \alpha) \propto p(\mathrm{x} \mid \theta)^\zeta \cdot p(\theta \mid \alpha)$. However, this is no longer a Dirichlet distribution. To see this, we compute the joint

$$p(\theta, \mathrm{x} \mid \alpha) = p(\theta \mid \alpha) \cdot \prod_{i=1}^{b} p(x_i \mid \theta)^\zeta \tag{7.34}$$

$$= D(\alpha) \cdot \prod_k \theta_k^{\alpha_k - 1} \cdot \frac{1}{\Sigma(\theta^\zeta)} \cdot \prod_k \theta_k^{\zeta \cdot c_k} \tag{7.35}$$

$$= \frac{1}{\Sigma(\theta^\zeta)} \cdot D(\alpha) \cdot \prod_k \theta_k^{\alpha_k + \zeta c_k - 1} \tag{7.36}$$

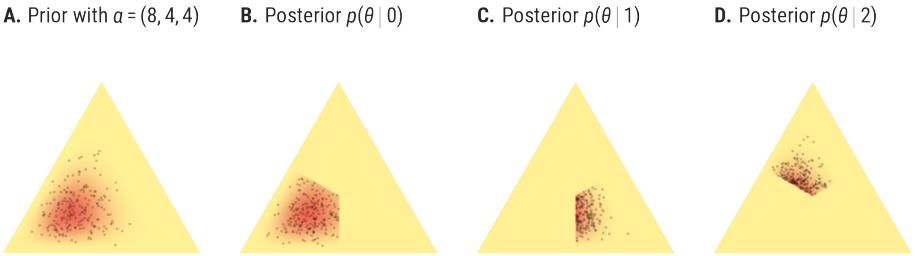


FIGURE 7.3 The posterior distribution $p(\theta \mid x)$ for various x if $\zeta = \infty$, that is, if agents always pick the most likely word. The posterior restricts the prior to the area of the simplex where $\arg\max_k \theta_k = x$.

FIG02

Although this is reminiscent of the Dirichlet density, it is not proportional to it, since the first term depends on θ. Consequently, deriving a closed-form expression of the posterior seems hard, as it involves solving the integral

$$\int_\Delta \frac{1}{\Sigma(\theta^\zeta)} \prod_k \theta_k^{\alpha_k + \zeta c_k - 1} \, d\theta. \tag{7.37}$$

The reciprocal of the sum hindered any progress on this point and all suggestions would be more than welcome.[37]

This might be an interesting problem in its own, partly because we *can* relatively easily identify the extreme cases. When $\zeta = 1$ we trivially get the normal posterior, but when $\zeta = \infty$ we can also get an idea of the posterior. After observing x, the language θ used to generate it can only be one where x gets maximum probability. In other words, θ must have been in $\Delta_x := \{\theta \in \Delta : \theta_x \geq \theta_k \text{ for all } k\}$. Consequently,

$$p(\theta \mid x, \alpha) = p(\theta \mid \theta \in \Delta_x, \alpha) \tag{7.38}$$

$$= \frac{[\![\theta \in \Delta_x]\!] \cdot \text{Dirichlet}(\theta \mid \alpha)}{\int_{\Delta_x} \text{Dirichlet}(\theta \mid \alpha) \, d\theta} \tag{7.39}$$

That is, the posterior is proportional to the prior, restricted to the section Δ_x of the simplex where the largest component is x. I have not tried integrating the Dirichlet over Δ_x yet, other than the symmetric case, i.e. $\alpha = \beta/K \cdot 1$, when the integral is simply $1/K$. In any case, one immediately sees that this cannot be a Dirichlet distribution: the posterior is discontinuous at the boundary of Δ_x, or at least at the part of it that lies inside Δ.

All this is illustrated in figure 7.3. It should be clear from that figure that agents who update their beliefs like this very quickly run into serious problems. After observing $x = 0$ all mass is restricted to Δ_0 and if the agent next observes $x = 1$, it has the problem that $p(1 \mid \alpha) = \int_{\Delta_0} p(\theta, 1) \, d\theta = 0$. This is one of the reasons for assuming that agents use exaggerated distributions only during production, and do *not* account for them during posterior inference.

Measuring the distance between languages

Most measures introduced to analyse the Dirichlet-categorical naming game, measure distances between languages. As a distance measure for distributions, the Jensen-Shannon divergence (JSD) can be used. The JSD is a symmetric version of the more

37 I also posted the problem at math.stackexchange.com/q/2360468.

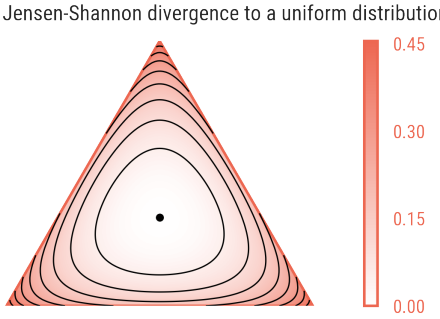


FIGURE 7.4 The divergence between distributions in the 2-simplex and the uniform distribution $(1/3, 1/3, 1/3)$ (indicated by a dot) under the Jensen-Shannon divergence. Points on the solid lines have the same distance to the uniform

FIG02 Figure inspired by a blogpost of Lior Pachter liorpachter.wordpress.com/tag/jensen-shannon-metric/

common Kullback-Leibler divergence and measures the similarity between probability distributions. Figure 7.4 illustrates the JSD of different discrete distributions to the uniform distribution. Formally, if $\pi_1, \ldots, \pi_N$ are probability distributions, their divergence is

$$\mathrm{JSD}(\pi_1, \ldots, \pi_N) := H\Big(\frac{1}{N}\sum_{i=1}^{N}\pi_i\Big) - \frac{1}{N}\sum_{i=1}^{N} H(\pi_i), \tag{7.40}$$

where H is the Shannon-entropy, a measure for the uncertainty in a distribution. As one can see, the JSD measures the difference between the entropy of the average distribution and the average entropy. If the divergence is zero, the distributions are identical since the JSD is the square of a metric (Endres and Schindelin 2003; Briët and Harremoës 2009). The divergence is moreover bounded,

$$0 \leq \mathrm{JSD}(\pi_1, \ldots, \pi_N) \leq \log_2(N) \tag{7.41}$$

so the normalised divergence, between 0 and 1, is obtained by dividing by $\log_2(N)$.

Bayesian updating and lateral inhibition

Intuitively, Bayesian updating implements a kind of lateral inhibition — but how exactly? We derive the 'update' rules in the Dirichlet-categorical naming game. Recall that every word is assigned a score $s(x) = p(x \mid \alpha)$. The question is how score of word y changes after observing x. That is, what is $s_{t+1}(x)$ in terms of $s_t(x)$? For a sampler, this follows directly from eq. 7.28:

$$s_{t+1}(y) = p(y \mid x, \alpha) = \frac{\alpha_y + [\![y = x]\!]}{\Sigma(\alpha) + 1} \tag{7.42}$$

$$= \frac{\alpha_y}{\Sigma(\alpha)} \cdot \frac{\Sigma(\alpha)}{\Sigma(\alpha) + 1} + \frac{[\![y = x]\!]}{\Sigma(\alpha) + 1} \tag{7.43}$$

$$= s_t(y) \cdot \frac{\Sigma(\alpha)}{\Sigma(\alpha) + 1} + \frac{[\![y = x]\!]}{\Sigma(\alpha) + 1} \tag{7.44}$$

For the MAP language strategy ($\eta = \infty$), the agent always chooses the mode ν of the distribution Dirichlet(α), i.e.,

$$\nu = \frac{\alpha - 1}{\Sigma(\alpha) - K} \tag{7.45}$$

The score $s(y) = p(y \mid x, \alpha)$ is therefore the y'th component of the mode. A similar argument as above shows that

$$s_{t+1}(y) = s_t(y) \cdot \frac{\Sigma(\alpha) - K}{\Sigma(\alpha) - K + 1} + \frac{[\![y = x]\!]}{\Sigma(\alpha) - K + 1}. \tag{7.46}$$

This is a similar lateral inhibition mechanism as the one used by samplers.

D Parameter space of the DC language game

The Bayesian language game has a language strategy parameter η, a production strategy parameter ζ and a parameter γ for the life expectancy. How do those influence the resulting behaviour of the model? This appendix reports an experiment that systematically analysed the behaviour in a larger part of the space.

The experiment measures three new quantities besides coherence and reflectance The first concerns the amount of *synonymy* in the language. If a language assigns all words the same probability, the synonymy is maximal, but if one word takes all probability mass, there is no synonymy. Synonymy is the inverse notion of efficiency in the naming games and formally defined as the relative Shannon entropy of the aggregate language,

$$S(t) := \frac{H(\bar{\pi}_t)}{\log_2(N)}, \tag{7.47}$$

where H is the entropy. $S(t) = 1$ indicates maximal synonymy, $S(t) = 0$ the complete absence of synonymy. Second, the *discrepancy* between the internal and external language is measured:

$$D(t) := \text{JSD}(\bar{\pi}_t, \psi_t), \tag{7.48}$$

where the aggregate language functioned as a proxy of all internal languages. Third, we measure the *variability* of the aggregate language as its standard deviation over time,

$$V(t) = \text{std}(\bar{\pi}_0, \ldots, \bar{\pi}_t). \tag{7.49}$$

If the languages used in the populations were relatively stable throughout the game, the variability should be low.

The experiment simulated 20 runs of the Dirichlet-categorical language game for every combination of the parameters $\eta \in \{1, 2, 5, 50, \infty\}$, $\zeta \in \{1, 1.5, 2, 5, \infty\}$ and $\gamma \in \{1, 10, 50, 10, 100, 1000, \infty\}$.[38] Every run had a duration of 100 000 iterations and

[38] To be completely clear: that amounts to 350 million rounds in 3500 independent simulations, using 175 different parameter settings.

all used the same, relatively flat, but nonuniform, prior. Trial experiments suggested that these parameter values sufficiently illustrate games in different parts of the parameter space. For example, $\eta > 100$ and $\zeta > 100$ already yield behaviour comparable to the infinite case and were left out. The coherence,[39] reflectance, synonymy, discrepancy, and variability were measured at the end of every game. Figure 7.5 shows all the results. Interpreting the results is tricky, and it helps to keep figure 4.10 in mind: The corners of every heat map in figure 7.5 mark the extreme strategies ($\eta, \zeta \in \{1, \infty\}$), which were also shown in figure 4.10. The life expectancies we considered earlier ($\gamma = 1, 10, 100$ and ∞) can also be found in figure 4.10. Now, the main conclusion is be that *our earlier findings are largely confirmed by the current experiment.* For intermediate parameter values, we observe a relatively smooth transition between the extreme cases $\eta, \zeta, \gamma \in \{1, \infty\}$.

We first discuss the last row, which corresponds to a naming game. The reflectance (column A) is much higher for samplers, and decreases quickly as soon as agents start to maximise their productions slightly (i.e. $\zeta > 1$). The synonymy (column B) suggests why: maximising production strategies result in a one-word language, that is, one with no synonymy. This explains why the reflectance is low for high ζ: the bias allows much more synonymy. The reflectance and synonymy also suggest that the language strategy (η) is far less influential than the production strategy (ζ). The fact that the change in vertical direction is smaller than the change in horizontal direction, and this seems to generalise to other life-spans as well. Looking at agents with shorter lifespans, we further see that the reflectance and synonymy increase as γ approaches 1 (iterated learning). As before, the reason is that the internal language of an agents with a short life span is nearly completely determined by the bias. Accordingly, reflectance is high and since the bias is fairly flat, so is synonymy. But, note that the discrepancy between the external language and the internal increases sharply for more maximising strategies. Finally the last column shows an increase in variability with intermediate lifespans. This could indicate continuous language change when agents have an intermediate lifespan of around $\gamma = 1000$ interactions, but more research is needed before such conclusions are warranted.

[39] Coherence is not shown, because all simulations appear to have reached coherence. This is an artefact of the measure used, which should thus be improved: populations with few observations, as in iterated learning, look perfectly coherent to our measure, because the shared bias fully determines their language.

E A discrete Weibull model of population turnover

Population turnover is commonly modelled by replacing one random agent in every round. Such a constant mortality rate is not very realistic, and this appendix proposes an alternative, discrete Weibull model. It is reparametrised such that the mean life expectancy is the only parameter.

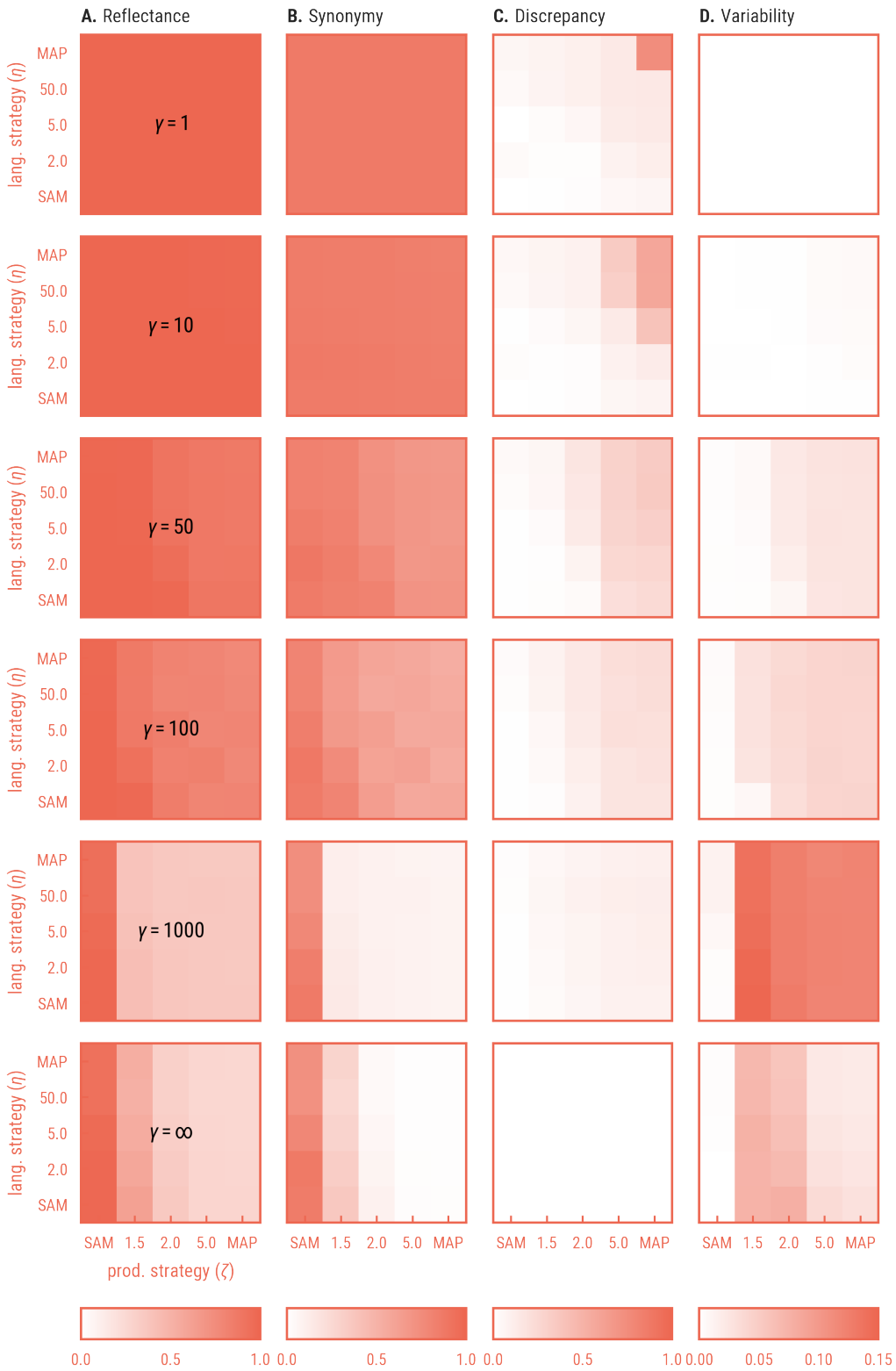


FIGURE 7.5 The behaviour of the Dirichlet-categorical language game across the parameter space (γ, η, ζ). Rows corresponds to life expectancies (γ); columns show the coherence, reflectance, synonymy and variability for every strategy (η, ζ). See figure 4.10 for the typical resulting languages in the extreme cases $\gamma, \eta, \zeta \in \{1, \infty\}$.

BNG06 Every cell is an average over 20 simulation runs. $K = 20$, $N = 10, b = 1, \gamma = \infty, \beta = 30$. Simulations used a deterministic hazard function.

How can we realistically model the mortality-rate in a population? This is a central question in *survival analysis* (e.g. Rogríguez 2007). In the context of a model of language evolution, if one agent dies in every iteration, the probability that any given agent dies at time t is thus γ — *given that it was still alive at time* $t-1$. If T is the random variable that measures the time of death of an agent, this means $p(T = t \mid T \geq t) = \gamma$. This conditional probability is known as the *hazard rate* $h(t)$, as it measures the rate of death (hazard) occurring at t.[40] In our example the hazard rate was constant. But, as explained in chapter 4, models with constant hazard rates are poor models of life-expectancy in human populations and demographers usually adopt either the *Weibull* or *Gompertz* distribution (Juckett and Rosenberg 1993). We here consider the simpler Weibull distribution, mainly because several discrete analogous have been proposed (Nakagawa and Osaki 1975; Stein and Dattero 1984; Almalki and Nadarajah 2014).

DISCRETE WEIBULL DISTRIBUTION The Weibull distribution (Weibull 1951) is a continuous distribution parametrized by a scale parameter $\kappa > 0$ and a shape parameter $\lambda > 0$. If $\kappa > 1$ the distribution is unimodal, meaning that most agents die around the same age, which is in turn determined by λ (see figure 7.6). For completeness, the density of a Weibull-distributed random variable T is

$$p(t \mid \lambda, \kappa) = \kappa/\lambda \cdot \left(t/\lambda\right)^{\kappa-1} \cdot \exp\left(-\left(t/\lambda\right)^{\kappa}\right). \tag{7.50}$$

Since language games are discrete time models, we use a discrete approximation known as the *Discrete Weibull* distribution[41] (Nakagawa and Osaki 1975), which preserves the so called *survival* function of the continuous distribution. The *survivor* function $S(t) = p(T \geq t)$ measures the probability of surviving to at least time t. The Weibull distribution, this function takes the form

$$S(t) = \exp\left(-\left(t/\lambda\right)^{k}\right), \tag{7.51}$$

and the Discrete Weibull is defined as the discrete distribution with the same survival function. This can be done, since the probability mass function is fully determined by the survival function:

$$p(T = t) = S(t) - S(t+1) = \exp\left(-\left(t/\lambda\right)^{\kappa}\right) - \exp\left(-\left(t+1/\lambda\right)^{\kappa}\right). \tag{7.52}$$

It should be stressed that the resulting distribution *approximates* the Weibull distribution, which for our purposes is sufficient.

The hazard rate of the Discrete Weibull distribution can be computed as

$$h(t) = p(T = t \mid T \geq t) = \frac{S(t) - S(t+1)}{S(t)} = 1 - \exp\left(\left(t/\lambda\right)^{\kappa} - \left(t+1/\lambda\right)^{\kappa}\right). \tag{7.53}$$

Recall that the hazard rate is the probability that an agent dies at time t, given that it hasn't died yet. Therefore, if we want to model a population where the probability of dying at time T approximately follows a Weibull distribution, we should in every round flip a coin with weight $h(t)$ to decide whether the agent dies.

41 I reparametrized the distribution by taking $\beta := \kappa$ and $q := \exp(-\lambda^{-\kappa})$, which is both computationally and conceptually more convenient.

40 It is usually defined for continuous T with density f as $h(t) = \frac{f(t)}{1-F(t)}$, in which case it is not a conditional probability but the *rate* of instantaneous hazard.

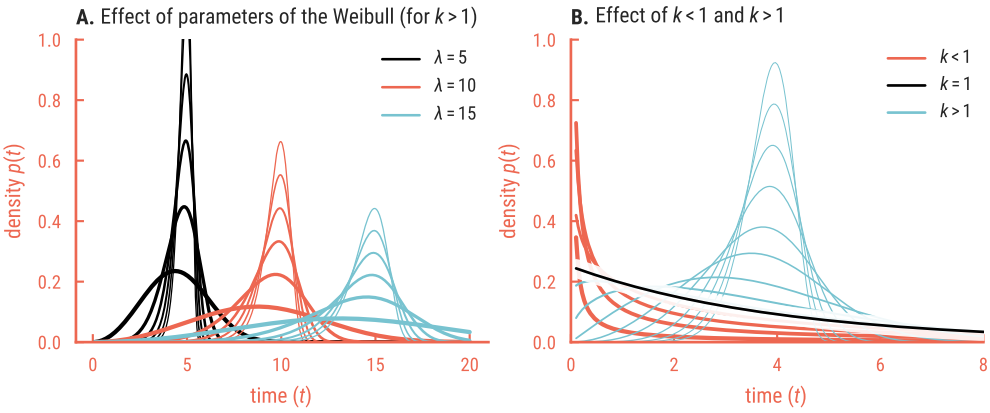


FIGURE 7.6 The Weibull distribution can model the probability that an agent dies at time t. A. Varying the parameters of a Weibull distribution illustrates that λ is a scale parameter and κ a shape paramater. B. If $\kappa > 1$ the Weibull is a unimodal distribution, whose variance decreases with higher κ (thinner lines), but for $\kappa < 1$ the distribution has no mode. When $\kappa = 1$ the Weibull reduces to a exponential distribution.

FIG04

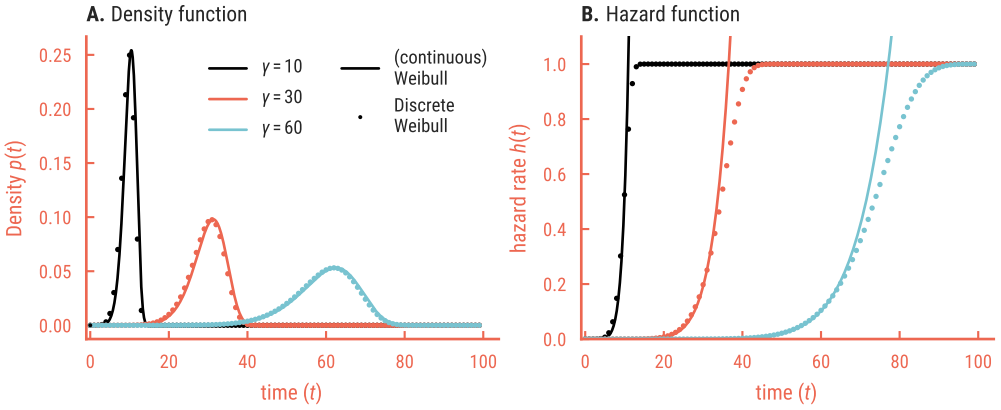


FIGURE 7.7 The single-parameter version of the continuous and discrete Weibull distribution. A. The distributions closely line up and γ is easily interpretable. B. The hazard rate increases with time, thus capturing ageing effects. Note that a continuous hazard rate $h(t)$ is not a distribution and exceeds 1.

FIG04

MODELLING POPULATION TURNOVER The Discrete Weibull gives us a more realistic model of population turnover, but its behaviour is regulated by two parameters: λ and κ. Ideally a *single* parameter would interpolate between immediate death (iterated learning) and immortality (naming game). This can be done by defining λ and κ in terms of a $\gamma \geq 1$:

$$\kappa(\gamma) := \log(\gamma) + \kappa_0 \qquad \text{and} \qquad \lambda(\gamma) := \frac{\gamma}{\Gamma\left(1 + {}^{1}/_{\kappa(\gamma)}\right)}, \qquad \gamma \geq 1 \tag{7.54}$$

where Γ is the gamma function. Figure 7.7 illustrates the effect of γ. Three reasons underly this reparametrization. First, the term $\Gamma\left(1+{}^{1}/_{\kappa(\gamma)}\right)$ makes γ interpretable: γ is the mean of the continuous distribution $\text{Weibull}\big(\lambda(\gamma), \kappa(\gamma)\big)$. Second, scaling $\kappa(\gamma)$ logarithmically with γ results in a realistic mortality distribution for all $\gamma \geq 1$. Third, the constant κ_0 guarantees that for $\gamma = 1$ the hazard rate is approximately 1, corresponding to instant death in iterated learning. I opt for[42] $\kappa_0 = 5$.

In sum, we have defined a discrete Weibull model, approximating the continuous Weibull, but parametrised by a single parameter γ, the average life expectancy. When used in combination with a random walk through a population of fixed size, γ thus interpolates between iterated learning ($\gamma = 1$) and a naming game ($\gamma = \infty$).

F Reformulating the packing strategy

The technical formulation of the packing strategy in (Hurford 1975) seems to have caused some confusion in the literature. This appendix reformulates the principle independent of the original generative framework, without compromising preciseness. This will bring some limitations of the packing strategy to the fore.

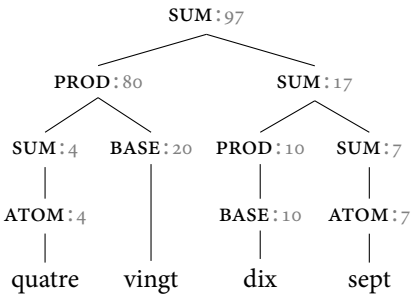
THE PACKING STRATEGY AS A CONSTRAINT ON TREES The packing strategy was introduced within the conceptual framework of generative grammar, as a 'significant generalisations' about number expressions and how they relate to numbers. Hurford (1975) analysed several numeral systems (English, French, Danish, Mixtec and Yoruba) using a phrase structure grammar which can be simplified to:[43]

$$\begin{aligned} \textsc{sum} &\longrightarrow \left\{ \begin{array}{l} \textsc{atom} \\ \textsc{prod}\ (\textsc{sum}) \end{array} \right\} \\ \textsc{prod} &\longrightarrow (\textsc{sum})\ \textsc{base} \end{aligned} \tag{7.55}$$

43 The original phrase structure rules constructed bases using exponentiation. This is controversial (see chapter 5) so I have use the most recent, simplified grammar from Hurford (2007). Note that the rewrite rule of PROD is different in Hurford (1987), where SUM is not optional. I have also changed notation and use SUM for NUM; PROD for PHRASE; and ATOM for DIGIT.

42 This results in the hazard rate $h(1 \mid \lambda = 1) > 1 - 10^{-8}$, which seems sufficiently close to 1.

where ATOM and BASE rewrite to one of the atoms and bases of the system respectively. It is easiest to think of this grammar as an *attribute grammar* (Knuth 1968) where every leaf (ATOM or BASE) has a fixed numeric value or *attribute*. Every nonterminal node corresponds to an operation that computes the value of the node from the values of its constituents. SUMs of course correspond to sums and PRODs to products. Here is the structure for French *quatre-vingt-dix-sept*, where I decorated nodes with their attributes in grey:

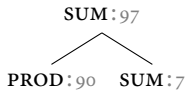


(7.56)

This is just one of the many structures with value 97 generated by the rules eq. 7.55. The packing strategy was introduced as a way to separate the wellformed from the ill-formed structures. It was therefore formulated as constraint on the structure of the trees, namely that:[44]

the sister constituent of a SUM must have the highest possible value. (7.57)

That is, the highest possible value while keeping the value of the parent constant. The sister constituent of a SUM can be a PROD or a BASE. Both can be found in eq. 7.56: at depth 3 we for example find a BASE with value 20 and a PROD with value 10. The reader might also have noticed that the node SUM : 17 *violates* the packing strategy. In a structure of the form



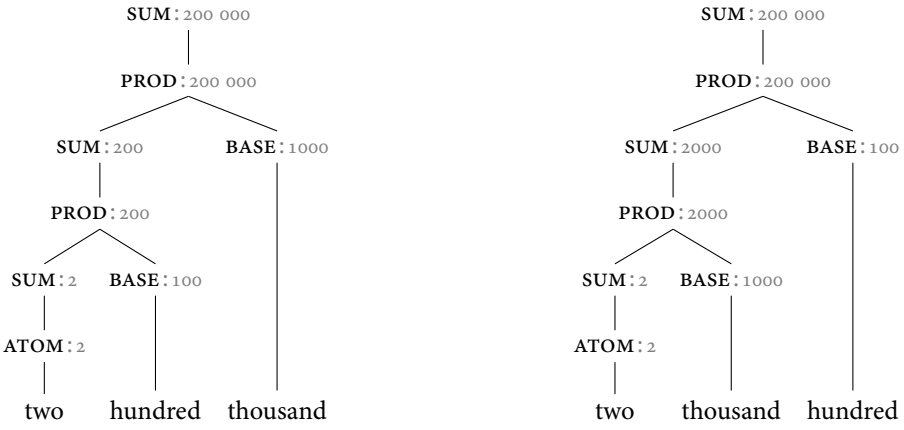
(7.58)

the node PROD : 90 is the sister constituent of a SUM and has a value higher than 80. We will discuss this problem later in more detail.

The packing strategy also accounts for the order of bases in large numerals, e.g. that

44 The formulation is from Hurford (1987) and Hurford (2007). The original also applied to bases constructed by exponentiation and is thus more complicated, as BASE nodes were nonterminals. Let A be a structure of category X (i.e. a PROD or a BASE) with value x and two constituents: a SUM and some node of another category Z (PROD or BASE). Then A is only wellformed if Z has the largest possible value $z \leq x$. That is, if there is no alternative Z' that also expands X with $\text{val}(Z) < \text{val}(Z')$.

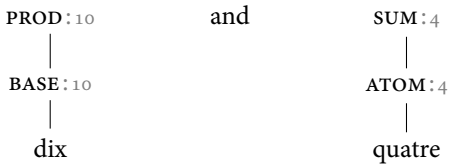
that *two hundred thousand* is wellformed, but *two thousand hundred* is not:



(7.59)

In the tree on the right, the sister node of SUM : 2000 is the node BASE : 100, and this violates the packing strategy, as it is also possible to form a tree where the corresponding sister has the higher value 1000. This is the tree shown on the left.

THE PACKING STRATEGY WITHOUT TREES Perhaps the tree representations overly complicated.[45] First they generate obvious redundancies in fragments like



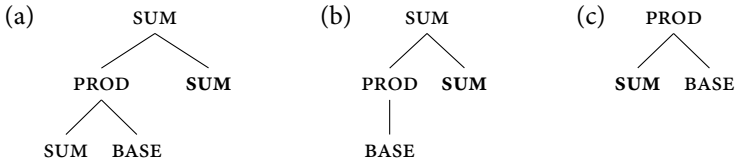
(7.60)

But more importantly, the same structures can be expressed using simple arithmetic formulae like $(4 \times 20) + (10 + 7)$ and $(2 \times 100) \times 1000$, as we have been doing throughout. More precisely, every tree corresponds to a formula built up from the values of the leaves using the binary operations addition and multiplication. Such formulae are not only simpler, they also have more expressive power. The phrase structure rule can for example only produce multiplicative constructions with a base. But as we have seen, languages sometimes contain multiplicative constructions with factors that are *not* considered bases properly: isolated or mixed bases. Similarly, additive constructions with two atoms are illformed by the packings strategy; a base has to figure in one of the constituents. The Welsh expression for 15 is an additive base, but not a base. This means that *correct* expressions of the form $15 + 2$ cannot be generated.[46] It is furthermore easy to extend the formulae with other (binary) operations like subtraction and division, whereas the phrase structure rules can only account for these using complicated extensions of the semantic interpretation which I will not reproduce here. Finally, the order of constituents, of the base and atom in particular, cannot be described in the formalism, which is problematic (Calude and Verkerk 2016). In short, the formulae are simpler, more expressive and stay closer to the semantic structure of number expressions.

46 Hurford (1975) does list 15 as a base, and thus circumvents this at the cost of using an arguably wrong notion of base.

45 I doubt whether Hurford would disagree; over the years he used ever looser variants of the grammar, and often opts for arithmetic formulae in the discussion (Hurford 1999; Hurford 2007).

So how can we express the packing strategy in terms of such formulae? Well, note that SUMs only occur as sister constituents (bold) in fragments of the form



which, when collapsing chains, correspond to formulae of the form

$$\text{(a)}\ \ (y \times b) + x, \qquad \text{(b)}\ \ b + x, \qquad \text{(c)}\ \ x \times b.$$

Here x is the sum of interest, y some other expresion and b a base. Sister constituents of SUMs are thus multiples of bases (considering $b = 1 \times b$ a multiple) and the packing strategy states that these should have the highest possible value. We can thus reformulate the packing strategy as:

Complex numerals use the largest multiple of the largest base possible. (7.61)

This directly suggests more general principles, such as:

The difference between a and b in $a + b$ and $a \times b$ should be maximised. (7.62)

This would also apply to multiplicative constructions like 5×6, which do not contain a base (in English). The principle would then correctly favour 3×10. Principle (7.62) could be taken as a good interpretation of the informal statement that "languages prefer to form numeral expressions by combining constituents whose arithmetical values are maximally apart, within the constraints defined by the syntax of the system" (Hurford 1987, p. 243). But this is *not* a literal reformulation of the packing strategy: it is slightly more general.

LIMITS OF THE PACKING STRATEGY One of the arguments for the importance of packing strategy was that it explained the peculiarities of French numerals Hurford (1975). As a final note, I would like to point out that Hurford's explanation is somewhat problematic. Recall that structure eq. 7.58 showed that French numerals do not satisfy the packing strategy. So how does Hurford use the packing strategy to explain why *soixante dix* (60 + 10) is wellformed, and 50 + 20 or 40 + 30 are not? Consider the following expressions for 70:

(a) 7×10 **(b)** $6 \times 10 + 10$ (c) $5 \times 10 + 20$
(d) $3 \times 20 + 10$ (e) $2 \times 20 + 3 \times 10$

The correct expression is (b), although some dialects use *septante* for (a). We can directly eliminate (c) since the packing strategy favours $(6 \times 10) + 10$ over $(5 \times 10) + 20$. But (b) is illformed, since 6×10 is illformed in the light of 3×20 (it is assumed that 20 is a base). To correct for this, two additional constraints are introduced (Hurford 1975, p. 101). The first states (in a complicated way) that 70×10, 80×10 and 90×10 are

illformed. This eliminates (a). The second states (in an even more complicated way) that all multiples or 20, except 4×20, are illformed. This eliminates (d) and (e), but also makes (b) *well*formed, as desired.

The Packing Strategy, in short, does not appear to *explain* much about the French numerals. On the contrary, it would predict a quite different, vigesimal system, which can only be remedied by introducing ad-hoc constraints. This is perhaps not surprising. The packing strategy predicts a completely regular numeral system, and it is hard to see how such a strategy in itself could account for irregularities like those encountered with French numerals. These conclusions do mean that the packing strategy might not be the important generalisation Hurford suggests it to be. The corresponding generalisations in Greenberg (1978) (roughly, 37 and 38) might even be of more empirical relevance. The latter captures that numeral systems are very predictable, or that "there is no 'surprise' in numeral larger than [a certain] base". If the French expression for 70 is irregular, so is the expression for 70 in $170 = 100 + 70$. Finally, to the best of my knowledge the packing strategy has never been *systematically* evaluated against a large collection of numeral systems either. This might be something to address in future work, which might benefit from the simplified formulation of the generalised packing strategy derived in this appendix.

G Base games

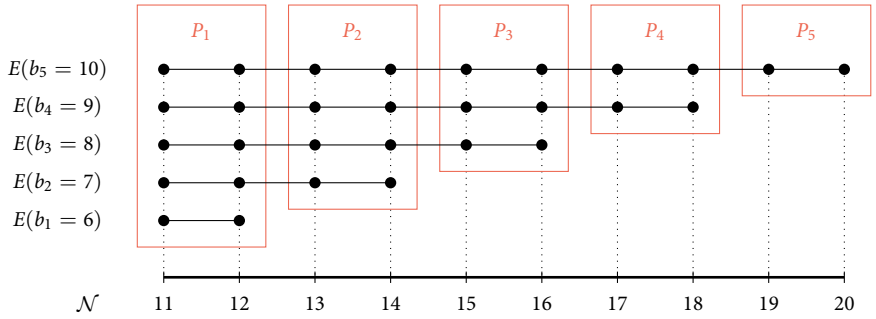
This appendix mathematically derives the implicit biases in the additive naming game and presents some further analyses of the parameters of the game.

Implicit biases in the additive naming game

The additive base game strongly favours the use of the highest base. To quantify that bias, we ask what the probability is that an agent, without any past experience, will use a certain base in the next round. 'Without past experience' is important, since we are interested in the *bias*, similar to the Bayesian naming game. The problem, I think, becomes much more intuitive if you keep the following diagram in mind:

The diagram illustrates the decimal case $B = 10$, so $K = 5$. Recall that $E(b)$ is the set of all numbers n in the domain that are expressible with base b; these sets are shown as horizontal lines with dots at every number n. We first compute the probability $p(n \in E(b_j))$ that we pick a number expressible by the j'th base, and then the probability that base b_j is selected *given* that the number is expressible. From the diagram one directly sees that

$$p\big(\, n \in E(b_j) \,\big) = \frac{|E(b_j)|}{B} = \frac{j}{K}, \tag{7.63}$$



which is the relative length of the black line corresponding to $E(b_j)$. Next, $p(b_j \mid n \in E(x_j))$ has to take into account that other bases might also express n. Given that $n \in E(b_j)$ there are j *equally likely* 'parts' $P_1, \ldots, P_j$ that n could have been in. The parts correspond the orange boxes. Now it is easily seen that the numbers n in box P_j can be expressed by $K - j + 1$ different bases. In the diagram, the numbers 13 and 14 are in part P_2 and can be expressed by $5 - 2 + 1 = 4$ different bases. But that means that

$$p\big(\, b_j \mid n \in E(b_j) \,\big) = \sum_{i=1}^{j} p(n \in P_i) \cdot p(b_j \mid n \in P_i) \tag{7.64}$$

$$= \frac{1}{j} \cdot \sum_{i=1}^{j} \frac{1}{K - i + 1} \tag{7.65}$$

$$= \frac{1}{j} \cdot \Big(\sum_{i=1}^{K} \frac{1}{i} - \sum_{i=1}^{K-j} \frac{1}{i} \Big) \tag{7.66}$$

$$= \frac{1}{j} \big(H_K - H_{K-j} \big), \tag{7.67}$$

where $H_n = \sum_{i=1}^{n} 1/i$ is known as the *n'th harmonic number* and we assume $H_0 := 0$. Putting everything together,

$$p(b_j) = p\big(\, n \in E(b_j) \,\big) \cdot p\big(\, b_j \mid n \in E(b_j) \,\big) \tag{7.68}$$

$$= \frac{1}{K} \big(\, H_K - H_{K-j} \,\big). \tag{7.69}$$

This implies a very strong bias towards using the highest base, which is further discussed in chapter 6.

As a final check, does eq. 7.69 really defines a distribution — is it normalised? The distribution in eq. 7.69 is normalised if and only if

$$\sum_{j=1}^{K} \frac{1}{K} (H_K - H_{K-j}) = 1 \tag{7.70}$$

Multiplying both sides by K and reordering, we see that this is equivalent to

$$\sum_{j=1}^{K-1} H_{K-j} = K \cdot (H_K - 1), \tag{7.71}$$

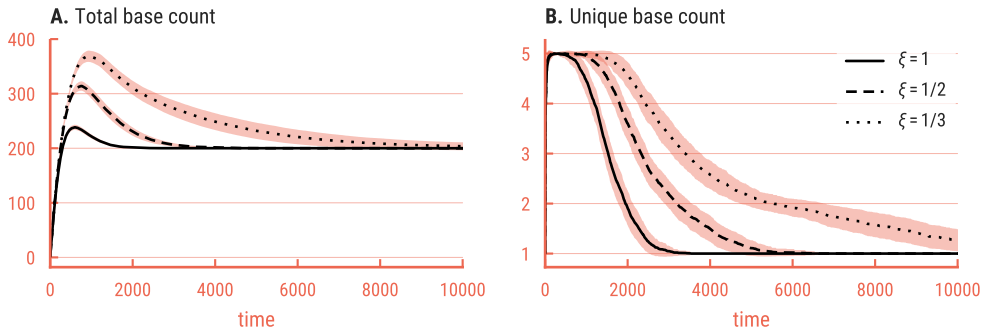


FIGURE 7.8 Effects of ξ, the parameter regulating the production strategy in the additive base game. Clearly, smaller values lead to slower convergence time.

HURO2 Results shown for $N = 200$, $B = 10$; avg. of 300 runs; 1 std. shaded.

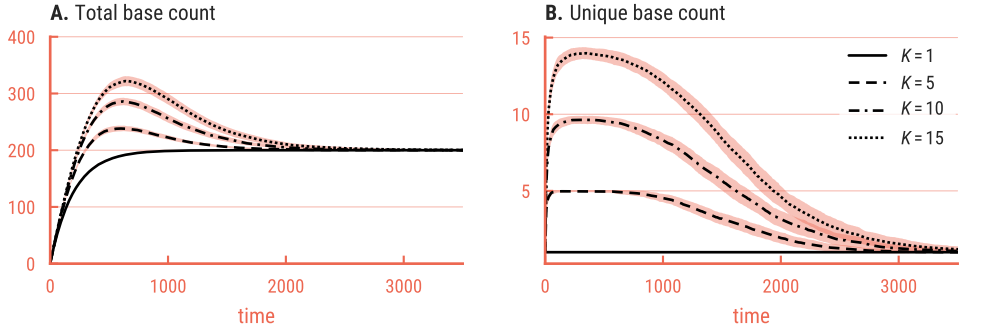


FIGURE 7.9 The effect of K, the number of bases, in the additive base game.

HURO1 Results shown for $N = 200$, $\xi = 1$; avg. of 300 runs; 1 std. shaded.

where we used the convention $H_{K-K} = H_0 := 0$ to sum only up to $K - 1$. Equation 7.71 is in fact true: it is one of the basic recurrence relations on the harmonic numbers. We can therefore conclude that the distribution in eq. 7.69 is really a distribution.

As a final note, I have note been able to find any reference of this distribution, but I haven't searched extensively either. But if the reader happens to recognise this distribution — I would be very interested to learn more about it.

Properties of the additive base game

EFFECT OF ξ Recall that the parameter ξ determined which bases were favoured; base b was *favoured* if

$$s(b) > 0 \qquad \text{and} \qquad s(b) \geq \xi \cdot \max_b s(b), \tag{7.72}$$

for $1 \geq \xi > 0$. The additive base game was simulated with $\xi \in \{1, 1/2, 1/3\}$ and the results are reported in figure 7.8. The main effect seems to be that lower values of ξ slow down convergence. This is not surprising: when $\xi = 1/3$, an agent only favours a single base if its frequency is 3 times as high as that of any other base. This is a much weaker preference for frequent bases than with $\xi = 1$, in which case agents always use the most frequent base.

EFFECTS OF THE NUMBER OF SIMPLE NUMERALS How does the number of simple numerals influence the game? Figure 7.9 summarises some experiments of the additive base game with $B \in \{1, 10, 20, 30\}$. Although these simple experiments do not yield strong

conclusions, the convergence time is not strongly influenced by B. That is, certainly not in a power law fashion, like population size. Rather, the convergence time does not seem to depend on K. One reason for this is that convergence time in the Dirichlet-categorical naming game was also not found to be strongly influenced by K.

Bibliography

Almalki, Saad J. and Saralees Nadarajah (2014). "A New Discrete Modified Weibull Distribution". In: *IEEE Transactions on Reliability* 63.1, pp. 68–80. DOI: 10.1109/TR.2014.2299691. arXiv: arXiv:1307.3925v1.

Baronchelli, Andrea (2006). "Statistical mechanics approach to language games". PhD thesis. Università di Roma "La Sapienza".

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

Baronchelli, Andrea, Luca Dall'Asta, et al. (2007). "Nonequilibrium phase transition in negotiation dynamics". In: *Physical Review E* 76.5, p. 051102. DOI: 10.1103/PhysRevE.76.051102. arXiv: 0611717 [cond-mat].

Baronchelli, Andrea, Maddalena Felici, et al. (2006). "Sharp transition towards shared vocabularies in multi-agent systems". In: *Journal of Statistical Mechanics: Theory and Experiment* 2006.06, P06014–P06014. DOI: 10.1088/1742-5468/2006/06/P06014. arXiv: 0509075 [physics].

Baronchelli, Andrea, Tao Gong, et al. (2010). "Modeling the emergence of universality in color naming patterns." In: *Proceedings of the National Academy of Sciences of the United States of America* 107.6, pp. 2403–2407. DOI: 10.1073/pnas.0908533107. arXiv: 0908.0775.

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.

Berwick, Robert C and Noam Chomsky (2017). "Why only us: Recent questions and answers". In: *Journal of Neurolinguistics* 43.B, pp. 166–177. DOI: 10.1016/j.jneuroling.2016.12.002.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine learning*. Information Science and Statistics. Springer.

Briët, Jop and Peter Harremoës (2009). "Properties of classical and quantum Jensen-Shannon divergence". In: *Physical Review A* 79.5, p. 052311. DOI: 10.1103/PhysRevA.79.052311. arXiv: 0806.4472.

Brighton, Henry (2002). "Compositional Syntax From Cultural Transmission". In: *Artificial Life* 8.1, pp. 25–54. DOI: 10.1162/106454602753694756.

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.

Calude, Andreea S. and Annemarie Verkerk (2016). "The typology and diachrony of higher numerals in Indo-European: a phylogenetic comparative study". In: *Journal of Language Evolution* 1.2, pp. 91–108. DOI: 10.1093/jole/lzw003.

Carstensen, Alexandra et al. (2015). "Language evolution in the lab tends toward informative communication". In: *Proceedings of the 37th Annual Meeting of the Cognitive Science Society.* 1.

Chomsky, Noam (1986). *Knowledge of language: Its nature, origin, and use*. Convergence. New York: Praeger.

Christiansen, Morten H. and Nick Chater (2016a). *Creating language. Integrating evolution, acquisition, and processing*. Cambridge, Massachusetts: MIT Press.

Christiansen, Morten H and Nick Chater (2016b). "The Now-or-Never bottleneck: A fundamental constraint on language". In: *The Behavioral and brain sciences* 39. DOI: 10.1017/S0140525X1500031X.

Comrie, Bernard (1999). "Haruai numerals and their implications for the history and typology of numeral systems". In: *Numeral Types and Changes Worldwide* 118, pp. 95–111.

— (2011). *Typology of Numeral Systems*.

— (2013). *Numeral Bases*.

Cornish, Hannah (2011). "Language Adapts: Exploring the Cultural Dynamics of Iterated Learning". PhD thesis. University of Edinburgh.

Culbertson, Jennifer and Simon Kirby (2016). "Simplicity and specificity in language: Domain-general biases have domain-specific effects". In: *Frontiers in Psychology* 6.JAN, pp. 1–11. DOI: 10.3389/fpsyg.2015.01964.

Dall'Asta, L et al. (2006). "Agreement dynamics on small-world networks". In: *Europhysics Letters (EPL)* 73.6, pp. 969–975. DOI: 10.1209/epl/i2005-10481-7. arXiv: 0603205 [cond-mat].

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.

Dediu, Dan et al. (2013). "Cultural Evolution of Language". In: *Cultural Evolution: Society, Technology, Language and Religion*. Ed. by Peter J. Richerson and Morten H. Christiansen. MIT Press. Chap. 16, pp. 304–332.

Dehaene, Stanislas (2011). *The number sense: How the mind creates mathematics*. Second. New York: Oxford University Press.

Dehaene, Stanislas and Jacques Mehler (1992). "Cross-linguistic regularities in the frequency of number words". In: *Cognition* 43.1, pp. 1–29.

Endres, D.M. and J.E. Schindelin (2003). "A new metric for probability distributions". In: *IEEE Transactions on Information Theory* 49.7, pp. 1858–1860. DOI: 10.1109/TIT.2003.813506.

Feigenson, Lisa, Stanislas Dehaene, and Elizabeth Spelke (2004). "Core systems of number". In: *Trends in Cognitive Sciences* 8.7, pp. 307–314. DOI: 10.1016/j.tics.2004.05.002.

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.

Ghahramani, Zoubin (2015). "Probabilistic machine learning and artificial intelligence". In: *Nature* 521.7553, pp. 452–459. DOI: 10.1038/nature14541.

Goodman, Noah D. and Joshua B. Tenenbaum (2016). *Probabilistic Models of Cognition*. v2.

Greenberg, Joseph H (1978). "Generalizations about numeral systems". In: *Universals of Human Language Volume 3 Word Structure*. Vol. 3, pp. 271–309.

Griffiths, Thomas L., Kevin R. Canini, et al. (2007). "Unifying rational models of categorization via the hierarchical Dirichlet process". In: *Proceedings of the 29th annual conference of the cognitive science society*, p. 323328.

Griffiths, Thomas L. and Michael L. Kalish (2005). "A Bayesian view of language evolution by iterated learning". In: *Proceedings of the 27th annual conference of the cognitive science society*, pp. 827–832.

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

Griffiths, Thomas L., Charles Kemp, and Joshua B Tenenbaum (2008). "Bayesian models of cognition". In:

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.

Hammarström, Harald (2009). "Rarities in Numeral Systems". In: *Rethinking universals: How rarities affect linguistic theory*. Ed. by Jan Wohlgemuth and Michael Cysouw. De Gruyter Mouton, pp. 11–60.

Hanke, Thomas (2010). "Additional rarities in the typology of numerals". In: *Rethinking Universals: How Rarities Affect Linguistic Theory*. Ed. by Jan Wohlgemuth and Michael Cysouw. Empirical Approaches to Language Typology. Berlin, New York: De Gruyter Mouton, pp. 61–90. DOI: 10.1515/9783110220933.

Heine, Bernd and Tania Kuteva (2002). *World Lexicon of Grammaticalization*. Cambridge University Press, p. 401. DOI: 10.1017/CBO9780511613463.

Hockett, Charles F. (1960). "The Origin of Speech". In: *Scientific American* 203.3, pp. 88–96. DOI: 10.1038/scientificamerican0960-88.

Hurford, James R. (1975). *The Linguistic Theory of Numerals*. Cambridge Studies in Linguistics. Cambridge, United Kingdom: Cambridge University Press.

— (1987). *Language and Number: The Emergence of a Cognitive System*. Oxford/New York: Basil Blackwell.

— (1999). *Artificially growing a numerical system*.

Hurford, James R (2000). "Social Transmission Favours Linguistic Generalization". In: *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*.

Hurford, James R. (2001). "Languages Treat 1 – 4 Specially". In: *Mind & Language* 16.1, pp. 69–75.
— (2007). "A performed practice explains a linguistic universal: Counting gives the Packing Strategy". In: *Lingua* 117.5, pp. 773–783. DOI: 10.1016/j.lingua.2006.03.002.
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.
Kemp, Charles and Terry Regier (2012). "Kinship Categories Across Languages Reflect General Communicative Principles". In: *Science* 336.6084, pp. 1049–1054. DOI: 10.1126/science.1218811.
Khetarpal, N et al. (2013). "Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses". In: *Proceedings of the 35th Annual Meeting of the Cognitive Science Society: Cooperative Minds: Social Interaction and Group Dynamics, July 31-August 3, 2013*, pp. 764–769.
Kirby, Simon (2001). "Spontaneous evolution of linguistic structure - An iterated learning model of the emergence of regularity and irregularity". In: *IEEE Transactions on Evolutionary Computation* 5.2, pp. 102–110. DOI: 10.1109/4235.918430.
— (2017). "Culture and biology in the origins of linguistic structure". In: *Psychonomic Bulletin & Review* 24.1, pp. 118–137. DOI: 10.3758/s13423-016-1166-7.
Kirby, Simon, Hannah Cornish, and Kenny Smith (2008). "Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language". In: *Proceedings of the National Academy of Sciences* 105.31, pp. 10681–10686. DOI: 10.1073/pnas.0707835105.
Kirby, Simon, Mike Dowman, and Thomas L. Griffiths (2007). "Innateness and culture in the evolution of language." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.12, pp. 5241–5245. DOI: 10.1073/pnas.0608222104.
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 and James R Hurford (2002). "The Emergence of Linguistic Structure: An Overview of the Iterated Learning Model". In: *Simulating the Evolution of Language*. London: Springer London, pp. 121–147. DOI: 10.1007/978-1-4471-0663-0_6.
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.
Knuth, Donald E. (1968). "Semantics of context-free languages". In: *Mathematical Systems Theory* 2.2, pp. 127–145. DOI: 10.1007/BF01692511.
Levinson, Stephen C. (2012). "Kinship and Human Thought". In: *Science* 336.6084, pp. 988–989. DOI: 10.1126/science.1222691.
Loreto, Vittorio et al. (2011). "Statistical physics of language dynamics". In: *Journal of Statistical Mechanics: Theory and Experiment* 2011.04, P04006. DOI: 10.1088/1742-5468/2011/04/P04006.
Madsen, Mathias Winther (2015). *Random Processes and Ergodicity*.
Nakagawa, Toshio and Shunji Osaki (1975). "The Discrete Weibull Distribution". In: *IEEE Transactions on Reliability* R-24.5, pp. 300–301. DOI: 10.1109/TR.1975.5214915.
Niyogi, Partha and Robert C Berwick (2009). "The proper treatment of language acquisition and change in a population setting". In: *PNAS* 106.25, pp. 10124–10129.
Norris, J. R. (1997). *Markov Chains*. Cambdrige: Cambridge University Press. DOI: doi.org/10.1017/CBO9780511810633.
Nowak, Martin a., Natalia L. Komarova, and Partha Niyogi (2001). "Evolution of universal grammar." In: *Science (New York, N.Y.)* 291.5501, pp. 114–8. DOI: 10.1126/science.291.5501.114.
Oliphant, Michael and John Batali (1996). "Learning and the Emergence of Coordinated Communication". In: *Center for research on language newsletter* 11.1, pp. 1–46. DOI: 10.1.1.27.2287.
Perfors, Amy et al. (2011). "A tutorial introduction to Bayesian models of cognitive development". In: *Cognition* 120.3, pp. 302–321. DOI: 10.1016/j.cognition.2010.11.015.
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, Paul Kay, and Naveen Khetarpal (2007). "Color naming reflects optimal partitions of color space." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.4, pp. 1436–41. DOI: 10.1073/pnas.0610341104.

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 (2002). "The cultural evolution of communication in a population of neural networks". In: *Connection Science* 14.1, pp. 65–84. DOI: 10.1080/09540090210164306.

— (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 (1995). "A Self-Organizing Spatial Vocabulary". In: *Artificial Life* 2.3, pp. 319–332. DOI: 10.1162/artl.1995.2.319.

— (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.

— (2012). "Introduction. Self-organization and selection in cultural language evolution". In: *Experiments in Cultural Language Evolution*, pp. 1–37. DOI: 10.1075/ais.3.02ste.

— (2015). *The Talking Heads experiment: Origins of words and meanings*. Ed. by Luc Steels and Remi van Trijp. Computational Models of Language Evolution. Language Science Press. DOI: 10.17169/langsci.b49.75.

— (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.

Steels, Luc and Tony Belpaeme (2005). "Coordinating perceptually grounded categories through language: A case study for colour". In: *Behavioral and Brain Sciences* 28.04, pp. 469–529. DOI: 10.1017/S0140525X05000087.

Stein, William E. and Ronald Dattero (1984). "A New Discrete Weibull Distribution". In: *IEEE Transactions on Reliability* R-33.2, pp. 196–197. DOI: 10.1109/TR.1984.5221777.

Tamariz, Monica (2017). "Experimental Studies on the Cultural Evolution of Language". In: *Annual Review of Linguistics* 3.1, pp. 389–407. DOI: 10.1146/annurev-linguistics-011516-033807.

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.

Tomasello, Michael (1999). *The cultural origins of human cognition*, p. 248.

Von Mengden, Ferdinand (2008). "The grammaticalization cline of cardinal numerals and numeral systems". In: *Rethinking Grammaticalization: new perspectives*. ii, pp. 289–308. DOI: 10.1075/tsl.76.14men.

Weibull, W. (1951). "A statistical distribution function of wide applicability". In: *Journal of applied mechanics* 18, pp. 293–297.

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.

Xu, Jing, Thomas L. Griffiths, and Mike Dowman (2010). "Replicating Color Term Universals through Human Iterated Learning". In: *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Portland, OR: Cognitive Science Society, pp. 352–357.

Xu, Yang and Terry Regier (2014). "Numerical Systems across Languages Support Efficient Communication: From Approximate Numerosity to Recursion". In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014)*, pp. 1802–1807. DOI: 10.1007/s13398-014-0173-7.2.

Zhou, Kevin and Claire Bowern (2015). "Quantifying uncertainty in the phylogenetics of Australian numeral systems". In: *Proceedings of the Royal Society B: Biological Sciences* 282.1815, p. 20151278. DOI: 10.1098/rspb.2015.1278.

Zuidema, Willem (2003). "How the Poverty of the Stimulus Solves the Poverty of the Stimulus". In: *Advances in Neural Information Processing Systems 15* 15, p. 51.

— (2013). "Language in Nature: On the Evolutionary Roots of a Cultural Phenomenon". In: *The Language Phenomenon: Human Communication from Milliseconds to Milennia*. Ed. by P.-M. Binder and Kenny Smith. Heidelberg: Springer. Chap. 8, pp. 163–189. DOI: 10.1007/978-3-642-36086-2_8.