Volume 249 Issue 3 7 December 2007 ISSN 0022-5193

Journal of Theoretical Biology

Available online at www.sciencedirect.com
ScienceDirect

Journal of Theoretical Biology 249 (2007) 606–616

Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Finite populations choose an optimal language

Christina Pawlowitsch*

Department of Economics, University of Vienna, Hohenstaufengasse 9, 1010 Vienna, Austria

Received 22 May 2007; received in revised form 9 August 2007; accepted 10 August 2007
Available online 28 August 2007

Abstract

This paper studies the evolution of a proto-language in a *finite* population under the frequency-dependent Moran process. A proto-language can be seen as a collection of concept-to-sign mappings. An efficient proto-language is a bijective mapping from objects of communication to used signs and vice versa. Based on the comparison of fixation probabilities, a method for deriving conditions of evolutionary stability in a finite population [Nowak et al., 2004. Emergence of cooperation and evolutionary stability in finite populations. Nature 428, 246–650], it is shown that efficient proto-languages are the only strategies that are *protected by selection*, which means that no mutant strategy can have a fixation probability that is greater than the inverse population size. In passing, the paper provides interesting results about the comparison of fixation probabilities as well as Maynard Smith's notion of evolutionary stability for finite populations [Maynard Smith, 1988. Can a mixed strategy be stable in a finite population? J. Theor. Biol. 130, 247–251] that are generally true for games with a symmetric payoff function.

Keywords: Language evolution; Evolutionary stability; Moran process; Fixation probabilities

1. Introduction

A central issue in modern linguistics is the search for *universals*, that is, properties that are shared by all languages. One of these design features of language is our ability to infer abstract concepts and to link them to arbitrary signs. A collection of such concept-to-sign mappings can be interpreted as a proto-language. Evolutionary game theory—which provides a formal framework for studying biological as well as cultural evolution of frequency-dependent phenomena—can be used to show how such a proto-language can evolve from a pre-linguistic environment (Nowak and Krakauer, 1999; Nowak et al., 1999; Trapa and Nowak, 2000; for a review of the more general purpose of integrating the computational and evolutionary aspects of language see, for example, Nowak et al., 2002; or Niyogi, 2006).

From a linguistic point of view, an interesting aspect of the evolution of concept-to-sign mappings is whether a simple replication mechanism will always lead to *bidirectionality*, that is, the property that whenever a particular sign is used to communicate a particular object, this sign will also evoke the image of this object (see, for example, Hurford, 1989). In an infinitely large population this is not necessarily the case. The deterministic replicator dynamics (Taylor and Jonker, 1978; Hofbauer and Sigmund, 1988; Nowak and Sigmund, 2004) can be blocked in a suboptimum state, where one object of communication is linked to two or more signs—synonymy, or where one sign is used for two or more objects—homonymy (Pawlowitsch, 2007). In the beginning of language, however, small population size most probably played a crucial role (see, for example, Enard et al., 2002). To get a good model of the onset of language—a model that actually serves the purpose of giving a plausible reconstruction of events—it is therefore important to turn to *finite* populations. There has been already an attempt to study an aspect of language evolution in a finite population. Komarova and Nowak (2003) have discussed the evolution of grammar in a population of finite size.

*Present address: Program for Evolutionary Dynamics, Harvard University, One Brattle Square, Cambridge, MA 02138.
Tel.: +1 617 496 4664; fax: +1 617 496 4629.
E-mail address: christina.pawlowitsch@univie.ac.at

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.08.009

In this paper I discuss the evolution of a lexical matrix in a finite population under a frequency-dependent Moran process in the style of Nowak et al. (2004). Evolutionary dynamics, according to this model, can be genetic or cultural. In addition to selection, drift is an intrinsic feature of this process. Eventually this process leads to fixation of a strategy throughout the whole population. But new variants can arise by mutations. One crucial aspect of the frequency-dependent Moran process in a finite population is that a single mutant strategy that has a (small) disadvantage in terms of relative fitness against the resident type can still generate a lineage and finally take over the whole population through the effects of drift. If selection is not present, and drift is the only evolutionary force at work, a single mutant strategy that appears in an otherwise monomorphic population has a chance to reach fixation of $1/N$, the inverse population size. Nowak et al. (2004) say that *selection opposes the replacement of the resident type* if the fixation probability of the mutant is lower than $1/N$. Here I shall explore what this implies for the evolution of proto-language. In particular I am interested in the emergence of bidirectionality.

2. The model

Let us consider a language game in the style of Nowak et al. (1999) or Trapa and Nowak (2000). Suppose there are n events that potentially become the object of communication, and that there are m available signs. A strategy in the role of the sender can be represented by a matrix

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{1m} \\ \vdots & & \vdots & & \\ p_{i1} & \cdots & p_{ij} & \cdots & p_{im} \\ \vdots & & \vdots & & \\ p_{n1} & \cdots & p_{nj} & \cdots & p_{nm} \end{pmatrix} \in \mathscr{P}_{n\times m}$$

and a strategy in the role of the receiver can be represented by a matrix

$$\begin{pmatrix} q_{11} & \cdots & q_{1i} & \cdots & q_{1n} \\ \vdots & & \vdots & & \\ q_{j1} & \cdots & q_{ji} & \cdots & q_{jn} \\ \vdots & & \vdots & & \\ q_{m1} & \cdots & q_{mi} & \cdots & q_{mn} \end{pmatrix} \in \mathscr{Q}_{m\times n},$$

where

$$\mathscr{P}_{n\times m} = \left\{ P \in R^+_{n\times m} : \sum_{j=1}^{m} p_{ij} = 1, \forall i \right\}, \tag{1}$$

$$\mathscr{Q}_{m\times n} = \left\{ Q \in R^+_{m\times n} : \sum_{i=1}^{n} q_{ji} = 1, \forall i \right\}. \tag{2}$$

The interpretation is that p_{ij} is the probability with which event i is linked to sign j, and q_{ji} is the probability with which sign j is linked to event i. If a sender P meets a receiver Q, the probability that they correctly communicate event i is

$$\sum_{j=1}^{m} p_{ij} q_{ji}.$$

We call the sum of these probabilities over all n events the *potential of communication*

$$\pi(P, Q) = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} q_{ji} = \mathrm{tr}(PQ). \tag{3}$$

For a given number of n objects of communication and a given number of m available signs, the maximally attainable potential of communication is $\min\{n, m\}$. We call a pair $(P, Q) \in \mathscr{P}_{n\times m} \times \mathscr{Q}_{m\times n}$ *efficient* if it attains the maximally available potential of communication. For simplicity we assume in the sequel that $m = n$. A pair $(P, Q) \in \mathscr{P}_{n\times n} \times \mathscr{Q}_{n\times n}$ is efficient if $\mathrm{tr}(PQ) = n$.

Assuming that communication is mutually beneficial, we identify the potential of communication with the payoff that both players, the sender as well as the receiver, get out of their interaction, that is,

$$\pi_P(P, Q) = \pi(P, Q) = \pi_Q(P, Q). \tag{4}$$

The strategy sets for a sender (1) and, respectively, a receiver (2) together with the payoff function (3) then constitute a two-player *asymmetric game* (the two players have different strategy sets) with a *symmetric payoff function* (the two players get the same payoff out of their interaction).

Language as a social phenomenon crucially hinges on the fact that every single individual potentially appears in both, the role of the sender as well as the role of the receiver. We assume that interaction is pairwise and that individuals adopt the role of the sender or the role of the receiver with equal probabilities. Formally this amounts to symmetrizing the asymmetric game, where a strategy of the symmetrized game is a pair of a sender and a receiver matrix

$$L = (P, Q) \in \mathscr{P}_{n\times m} \times \mathscr{Q}_{m\times n} \tag{5}$$

and the payoff function for (P_1, Q_1) interacting with (P_2, Q_2) is

$$f[(P_1, Q_1), (P_2, Q_2)] = \tfrac{1}{2}\mathrm{tr}(P_1 Q_2) + \tfrac{1}{2}\mathrm{tr}(P_2 Q_1). \tag{6}$$

Note that this payoff function is symmetric,

$$f[(P_1, Q_1), (P_2, Q_2)] = f[(P_2, Q_2), (P_1, Q_1)],$$

which, in this case, is a consequence of the symmetry of the payoff function in the asymmetric game (4) in combination with the symmetry of the weights $(\frac{1}{2}, \frac{1}{2})$ for being in the role of the sender or in the role of the receiver. Symmetric games with a symmetric payoff function are sometimes called *partnership games* or *doubly symmetric games*.

2.1. Infinitely large populations

The classical notion of a population game assumes an infinitely large population (see, for example, Hofbauer and Sigmund, 1998). A strategy is evolutionarily stable if it is a best reply to itself, and if in addition to that whenever there is an alternative best reply, the original strategy attains a strictly higher payoff against this alternative best reply than the alternative best reply attains against itself (Maynard Smith, 1982). If we look at the game described by Eqs. (5) and (6) as a game played in an infinitely large population, a strategy $(P, Q) \in \mathscr{P} \times \mathscr{Q}$ is an *evolutionarily stable* strategy if and only if it represents a bijective mapping from events to used signs and vice versa; that is, if and only if P is a permutation matrix and Q is the transpose of P (Trapa and Nowak, 2000). In the case of 3 objects of communication and 3 available signs an evolutionarily stable strategy is, for example,

$$\left[\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix},\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\right],$$

or

$$\left[\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix},\begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix}\right].$$

It is not difficult to see that sender–receiver pairs of this form are the only strict Nash strategies of this game (for a discussion of Nash strategies and best-response properties in terms of the P and Q matrices see the Appendix). The result about evolutionary stability then follows from the fact that for symmetrized asymmetric games, a strategy is evolutionarily stable if and only if it is a strict Nash strategy—which, in turn, follows from Selten's result that in asymmetric games, strict Nash strategies are the only evolutionarily stable strategies (Selten, 1980).

For games with a symmetric payoff function, *evolutionarily stable strategies* coincide with the *locally asymptotically stable* rest points of the replicator dynamics (Hofbauer and Sigmund, 1988, 1998). This is due to the fact that for this class of games, the replicator dynamics—which describes deterministic frequency-dependent evolution in an infinitely large population—induces an increase in average fitness along every non-stationary path. Yet the replicator dynamics will not always lead to an evolutionarily stable strategy. There are strategies of this game that are not evolutionarily stable (in the strict sense) but *weakly evolutionary stable* or *neutrally stable*. A strategy is neutrally stable if it is a Nash strategy, and if, in addition to that, whenever there is an alternative best reply to the original Nash strategy, this alternative best reply is not a better reply to itself than the original Nash strategy is to the alternative best reply (Maynard Smith, 1982). In the language of the P and Q matrices, a strategy (P, Q) is neutrally stable if and only if (i) neither P nor Q has a column with multiple maximal elements strictly between 0 and 1, and (ii) at least P or Q has no zero column (Pawlowitsch, 2007). This means that in a neutrally stable strategy, one sign can be used for two or more objects of communication—a column with multiple 1-entries in P, or one event can be inferred by two or more signs—a column with multiple 1-entries in Q. But there cannot be two or more objects that are in parallel linked to, or inferred by, two or more signs—two or more columns with multiple maximal elements strictly between 0 and 1; and there cannot be any sign that remains idle—a zero column in P—as long as there is an event that is never possibly understood—a zero column in Q. Neutrally stable strategies are Lyapunov stable in the replicator dynamics (Thomas, 1985; Bomze and Weibull, 1995). As a result, the replicator dynamics can be blocked in a state where ambiguity of concept-to-sign mappings remains in the population, and there is no convergence to an efficient proto-language (Pawlowitsch, 2007).

The deterministic replicator dynamics in an infinitely large population does not include any effects of drift. In a finite population, however, drift is automatically present as differences in relative fitness only translate into expected and not realized offspring.

3. Finite populations

One example for a game dynamical process in a finite population is the frequency-dependent Moran process as introduced in Nowak et al. (2004); see also Nowak (2006): One individual is selected proportional to its fitness and produces identical offspring, which replaces a randomly chosen individual from the population. The fitness is determined by a combination of a constant background fitness, which is the same for all individuals, and the payoff from the game.

Suppose first that there are only two competing languages.

Example 1.

$$L_1 = (P_1, Q_1) = \left[\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix},\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\right]$$

and

$$L_2 = (P_2, Q_2) = \left[\begin{pmatrix}\alpha&1-\alpha&0\\0&0&1\\0&0&1\end{pmatrix},\begin{pmatrix}1&0&0\\1&0&0\\0&1-\beta&\beta\end{pmatrix}\right],$$

where $\alpha, \beta \in (0, 1)$. Note that L_2 is a probabilistic strategy, where object 1 is linked to two signs, with probability α to sign 1, and with probability $1-\alpha$ to sign 2; and where sign 3 is associated with two objects, with probability $1-\beta$ with object 2, and with probability β to object 3.

Restricting attention to these two strategies, we can describe the payoff structure by a payoff matrix of the form

$$\begin{array}{c|c|c|} & L_1 & L_2 \\ \hline L_1 & a & b \\ \hline L_2 & c & d \\ \hline \end{array} \quad , \tag{7}$$

where $a=f(L_1,L_1)$, $b=f(L_1,L_2)$, $c=f(L_2,L_1)$, and $d=f(L_2,L_2)$. In our case we have

$$\begin{array}{c|c|c|} & L_1 & L_2 \\ \hline L_1 & 3 & 1+\frac{\alpha}{2}+\frac{\beta}{2} \\ \hline L_2 & 1+\frac{\alpha}{2}+\frac{\beta}{2} & 2 \\ \hline \end{array} \quad . \tag{8}$$

Note that $c=b$ as $f(L_2,L_1)=f(L_1,L_2)$.

Let N be the number of individuals. A state of the population is a vector $X=(X_1,X_2)$, where X_1 is the number of individuals playing L_1, and X_2 is the number of individuals playing L_2, with $X_1+X_2=N$, We use

$$F(L_1\,|\,X_1,X_2)=\frac{X_1-1}{N-1}f(L_1,L_1)+\frac{X_2}{N-1}f(L_1,L_2), \tag{9}$$

$$F(L_2\,|\,X_1,X_2)=\frac{X_1}{N-1}f(L_2,L_1)+\frac{X_2-1}{N-1}f(L_2,L_2) \tag{10}$$

to denote the fitness of L_1, and, respectively L_2, given that X_1 individuals speak L_1 and X_2 individuals speak L_2. For the specific form of the payoff function we have

$$F(L_1\,|\,X_1,X_2)=\frac{X_1-1}{N-1}3+\frac{X_2}{N-1}\left(1+\frac{\alpha}{2}+\frac{\beta}{2}\right), \tag{11}$$

$$F(L_2\,|\,X_1,X_2)=\frac{X_1}{N-1}\left(1+\frac{\alpha}{2}+\frac{\beta}{2}\right)+\frac{X_2-1}{N-1}2. \tag{12}$$

The frequency-dependent Moran process combines two evolutionary forces: selection and drift. Eventually this process leads to fixation of a single strategy throughout the whole population; new variation can only be brought in by mutations.

Let us assume that the population has reached the state where all individuals speak $L_2=(P_2,Q_2)$. To assess the stability of L_2 in an evolutionary sense Nowak et al. (2004) propose two criteria: (I) *selection opposes L_1 invading L_2* if a single L_1 mutant in a population consisting otherwise of L_2 has a lower fitness than the regular type L_2; and (II) *selection opposes L_1 replacing L_2* if a single L_1 mutant that appears in a population of L_2 has a chance to reach fixation that is lower than $1/N$.

(I) *The fitness of a single mutant*: *Selection opposes L_1 invading L_2* if

$$F(L_2\,|\,1,N-1)>F(L_1\,|\,1,N-1); \tag{13}$$

the fitness of a single mutant should be lower than the fitness of the regular type, given the state of the population *after* the mutation has appeared. For our linear fitness function we obtain:

$$\frac{1}{N-1}f(L_2,L_1)+\frac{N-2}{N-1}f(L_2,L_2)>\frac{N-1}{N-1}f(L_1,L_2). \tag{14}$$

In a letter to the *Journal of Theoretical Biology* from 1988 Maynard Smith uses condition (14), with a weak inequality sign, as a concept for *evolutionary stability* in the framework of a finite population game (Maynard Smith, 1988). Schaffer (1988) introduces the same condition as a general concept for evolutionary stability in a finite population—independent of the particular form of the fitness function.

For the sender–receiver game discussed here, condition (14) reduces to

$$f(L_2,L_2)>f(L_1,L_2), \tag{15}$$

which comes from the fact that $f(L_2,L_1)=f(L_1,L_2)$, and as a consequence the factor $(N-2)$ cancels out on both sides of the inequality.

Condition (15) says that L_2 has to be a unique best response to itself, that is, the condition that L_2 has to be a strict Nash strategy in the base game. Note that this is generally true for games with a symmetric payoff function.

Observation 1. For games with a symmetric payoff function, under the assumption of a linear fitness function, Maynard Smith's and Schaffer's ESS concept for finite populations,

$$F(L_2\,|\,1,N-1)\geqslant F(L_1\,|\,1,N-1),$$

reduces to the condition of a Nash strategy in the base game; with a strict inequality sign, it reduces to the condition of a strict Nash strategy.

In fact, this is not surprising. The concept of Nash equilibrium relies on the very idea of a deviation—or what we call here a mutation—under the hypothetical assumption that all other players do not change their strategy choices. In the framework of a classical population game, this *ceteris paribus* assumption translates into the assumption that a single player's strategy choice has a vanishing effect on the population's average strategy. In a finite population this is no longer true. Condition (14) compares the deviant's payoff to the payoff of a non-deviant, taking into account the effect of the deviation. If the weight of the deviant vanishes, we are back to the classical notion of a Nash-equilibrium strategy. But the same is true if the term that reflects the deviation cancels out for some other reason—here the symmetry of the payoff function.

For the two languages considered here, we have that

$$f(L_2,L_2)=2>1+\frac{\alpha}{2}+\frac{\beta}{2}=f(L_1,L_2); \tag{16}$$

L_2 is indeed *the unique best reply to itself*, and therefore a single L_1 mutant that appears in a population where all $N-1$ other individuals speak L_2 will attain a *strictly lower* fitness than the regular type L_2.

From (8), the payoff matrix, it is easily seen that L_1 is also a unique best reply to itself as $3>1+\alpha/2+\beta/2$. As a consequence, a single L_2 mutant that appears in a population where all other individuals speak L_1 will also attain a *strictly lower fitness* than the regular type L_1.

This means that on the basis of condition (14) we cannot distinguish L_1 from L_2; both have a strict advantage in terms of relative fitness against a single mutant that switches to the other strategy.

In the case of L_1 this is not surprising. L_1 is an efficient proto-language; it is a strict Nash strategy in the complete strategy space $\mathscr{P}_{3\times 3} \times \mathscr{Q}_{3\times 3}$. But L_2 is not; due to its synonymous use of the first and second sign, and its ambiguous interpretation of the third sign, it does not exploit the full potential of communication that is in principle available given that there are 3 objects of communication and 3 available signs. Note, though, that L_2 is a neutrally stable strategy if the game is played in an infinitely large population. The population as a whole would be better off if it could directly jump to the state where everybody used L_1. But given that we start from a state where everybody speaks L_2, L_2 has a strict advantage in relative fitness against a single L_1 mutant. For the deterministic replicator dynamics in an infinitely large population this is enough to be blocked in such an inefficient state. In a finite population, however, advantages in relative fitness are not the only thriving force of evolution. There is also drift. Even though a single mutant strategy has a lower fitness than the regular type, the frequency-dependent Moran process can still favor the fixation of this mutant strategy.

(II) *Fixation probabilities*: The frequency-dependent Moran process in the style of Nowak et al. (2004) allows us to introduce a parameter $\omega \in [0, 1]$ that measures the *intensity of selection*. Instead of $F(L_1 \mid X_1, X_2)$ and $F(L_2 \mid X_1, X_2)$ as given by Eqs. (9) and (10), we use the modified fitness functions

$$F_\omega(L_1 \mid X_1, X_2) = 1 - \omega + \omega F(L_1 \mid X_1, X_2), \tag{17}$$

$$F_\omega(L_2 \mid X_1, X_2) = 1 - \omega + \omega F(L_2 \mid X_1, X_2). \tag{18}$$

If $\omega = 0$, the payoffs of the game do not contribute to fitness at all, and we are in the case of *neutral evolution*. If $\omega = 1$ fitness is entirely determined by expected payoff; selection is strong. Note that for the deterministic replicator dynamics in an infinitely large population, the parameter ω reduces to a constant that does not alter the trajectories of the dynamics, but only the speed of convergence.

The probability that a single L_1 mutant becomes fixed in a population of L_2—that is, the probability that it generates a lineage that takes over the whole population—is given by

$$\rho_1 = 1 \Bigg/ \left(1 + \sum_{k=1}^{N-1} \prod_{X_1=1}^{k} \frac{F_\omega(L_2 \mid X_1, X_2)}{F_\omega(L_1 \mid X_1, X_2)}\right); \tag{19}$$

see, for example, Nowak et al. (2004), or Nowak (2006). In the case of neutral evolution, that is if $\omega = 0$, $\rho_1 = 1/N$. The idea of Nowak et al. (2004) for assessing the stability of L_2 in an evolutionary context is to compare the fixation probability of a single L_1 mutant under the frequency-dependent Moran process to this neutral threshold $1/N$. They say that *selection opposes L_1 replacing L_2 if $\rho_1 < 1/N$*. Similarly, we may say that *selection favors L_1 replacing L_2 if $\rho_1 > 1/N$*.

In general it can be quite laborious to calculate fixation probabilities. *Weak selection* characterizes the condition that the payoff of the game is just a small component of the fitness of a type, $\omega N \ll 1$. Nowak et al. (2004) derive conditions for the comparison of fixation probabilities in the case of weak selection for a game of the general form as represented by payoff matrix (7). They find that

$$\rho_1 > \frac{1}{N}$$
$$\Leftrightarrow a(N-2) + b(2N-1) > c(N+1) + d(2N-4). \tag{20}$$

For different values of N this gives

$N = 2$: $b > c$

$N = 3$: $a + 5b > 4c + 2d$

$N = 4$: $2a + 7b > 5c + 4d$

$\vdots$

In the limit for large N, we have

$$\rho_1 > \frac{1}{N} \quad \Leftrightarrow \quad a + 2b > c + 2d.$$

Condition (20) is not only valid for the Moran dynamics, but for a variety of other stochastic processes, such as the Wright–Fisher process (Imhof and Nowak, 2006) and pairwise-comparison processes (Traulsen et al., 2005; or Traulsen et al., 2006). Recently it has been shown that the limit condition for large N is valid for any process in the domain of Kingman's coalescent (Lessard and Ladret, 2007).

Observation 2. If $b = c$, that is, in the case of a symmetric payoff function, condition (20) reduces to

$$\rho_1 > \frac{1}{N} \quad \Leftrightarrow \quad a + b > 2d \tag{21}$$

for all $N \geqslant 3$. For $N = 2$ we always have neutrality, $\rho_1 = \frac{1}{2}$.

This means that a mutant strategy (P_1, Q_1) that appears in an otherwise monomorphic population (P_2, Q_2) can reach fixation with a probability greater than $1/N$ *even if it has a strict disadvantage in relative fitness against the resident type when it first appears in the population*—that is, even if, $b - d = f(L_1, L_2) - f(L_2, L_2) < 0$, but only *if this is outweighed by a payoff advantage that the mutant strategy has against itself relative to the payoff that the originally resident type gains from interaction with itself*, $a - d = f(L_1, L_1) - f(L_2, L_2) > f(L_2, L_2) - f(L_1, L_2) = d - b$.

For the game with the two languages considered here, from payoff matrix (8), we easily see that $a + b > 2d$ is fulfilled as $\alpha + \beta > 0$. Thus, a single L_1 mutant that appears in a population where all other individuals speak L_2 has a *higher probability to reach fixation than in the case of neutral evolution*. If, on the other hand, we consider L_1 as the resident type and L_2 as the mutant,

$$\rho_2 > \frac{1}{N} \Leftrightarrow d + c > 2a, \tag{22}$$

which is *not fulfilled* in our case, as we can, again, easily read off from payoff matrix (8),

$$d + c = 2 + 1 + \frac{\alpha}{2} + \frac{\beta}{2} < 6 = 2a.$$

Thus, the comparison of fixation probabilities—other than the comparison of the fitness of a single mutant—allows us to distinguish L_1 from L_2 in terms of its stability from an evolutionary point of view: L_1 has a fixation probability in L_2 that is *higher than in the case of neutral evolution*, but L_2 has a fixation probability in L_1 that is *lower than in the case of neutral evolution*. Selection favors L_1 replacing L_2, but it opposes L_2 replacing L_1.

Let us look at some other examples.

Example 2. Consider the case of two efficient proto-languages. Let

$$L_1 = (P_1, Q_1) = \left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right]$$

and

$$L_2 = (P_2, Q_2) = \left[\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}\right].$$

In this case the payoff matrix is

	L_1	L_2
L_1	3	1
L_2	1	3

.

From this we directly see that both are unique best replies to themselves,

$$f(L_1, L_1) = 3 > 1 = f(L_2, L_1),$$

$$f(L_2, L_2) = 3 > 1 = f(L_1, L_2),$$

which is, of course, true as both are strict Nash strategies in the complete strategy space, and therefore any single mutant that appears in an otherwise monomorphic population of the other language has a strictly lower fitness than the regular type. Assessing fixation probabilities, we find that

$$a + b = 4 < 6 = 2d \Rightarrow \rho_1 < 1/N,$$

$$d + c = 4 < 6 = 2a \Rightarrow \rho_2 < 1/N.$$

Therefore, selection opposes both, the fixation of L_1 in L_2 and vice versa.

Example 3. Next we consider the case of an efficient proto-language against a simple Nash language,

$$L_1 = (P_1, Q_1) = \left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right],$$

$$L_2 = (P_2, Q_2) = \left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right].$$

In this case, the payoff matrix is

	L_1	L_2
L_1	3	2
L_2	2	2

.

Of course, L_1 is a unique best reply to itself. L_2 is also best reply to itself, but not unique—L_1 is also a best reply to L_2,

$$f(L_2, L_2) = 2 = f(L_1, L_2).$$

As a consequence, a single L_1 mutant that appears in a population that consists otherwise of L_2 has the same fitness as the regular type L_2. However, as L_1 is a unique best reply to itself, L_1 will already have a higher fitness than L_2 as soon as there is a second L_1 speaker. Not surprisingly, then, the fixation probability of L_1 in L_2 is *higher* than in the case of neutral evolution,

$$a + b = 5 > 4 = 2d$$

but the fixation probability of L_2 in L_1 is *lower* than in the case of neutral evolution,

$$d + c = 4 < 6 = 2a.$$

Selection opposes L_2 replacing L_1, but selection favors L_1 replacing L_2.

Example 4. Finally we consider the case of two strategies that are neutrally stable in the complete strategy space and that display the same type of multiplicity in object-to-sign and sign-to-object mappings. Let

$$L_1 = (P_1, Q_1) = \left[\begin{pmatrix} 1-\alpha & \alpha & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1-\beta & \beta \end{pmatrix}\right],$$

$$L_2 = (P_2, Q_2) = \left[\begin{pmatrix} 1-\gamma & \gamma & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1-\delta & \delta \end{pmatrix}\right]$$

with all parameters strictly between 0 and 1, but $\alpha \neq \gamma$ and $\beta \neq \delta$. In this case the payoff matrix is

	L_1	L_2
L_1	2	2
L_2	2	2

and we are in the case of neutral evolution. Drift is the only evolutionary force at work; $\rho_1 = 1/N$ and $\rho_2 = 1/N$.

3.1. The complete strategy space

Eventually our aim is to uncover general patterns of the P and Q matrices that emerge from a finite population in the complete strategy space $\mathscr{P}_{n\times n}\times\mathscr{Q}_{n\times n}$. The examples above suggest that efficient proto-languages are the only strategies such that no mutant strategy has a fixation probability that is higher than the neutral threshold $1/N$. Indeed this can be shown. More precisely we can show the following implications.

Proposition 1. *Let* $(P,Q)\in\mathscr{P}_{n\times n}\times\mathscr{Q}_{n\times n}$. *Under the frequency-dependent Moran process with weak selection, for all* $N\geqslant 3$:

(a) *If P is a permutation matrix and Q is the transpose of P—that is, if (P,Q) is a strict Nash strategy in the complete strategy space $\mathscr{P}_{n\times n}\times\mathscr{Q}_{n\times n}$—then there is no $(P',Q')\in\mathscr{P}_{n\times n}\times\mathscr{Q}_{n\times n}$, $(P',Q')\neq(P,Q)$ such that $\rho'\geqslant\frac{1}{N}$.*

(b) *If (P,Q) is not of the form such that P is a permutation matrix and Q is the transpose of P, then there is some $(P',Q')\in\mathscr{P}_{n\times n}\times\mathscr{Q}_{n\times n}$, $(P',Q')\neq(P,Q)$ such that $\rho'>\frac{1}{N}$.*

A proof is given in the Appendix.

4. Interpretation and conclusions

The starting point for any evolutionary argument in a finite population is that *every mutant strategy has some positive probability to reach fixation*—simply by the chances of random drift. In the modeling framework used here, what we call *the payoff of the game* is just short hand, or a convention of language, for adding a fitness component to a birth-and-death process that introduces an element of frequency-dependent selection in addition to drift.

As with every model, our model of language evolution considers just one stylized aspect of reality, while abstracting from others. Of course, the overall fitness of an individual does not depend only on its communicative strategy, but on other cultural and biological traits.

Weak selection, where the payoff of the game is just a small component that is added to the background fitness, therefore, seems to be the natural case.

Since evolution in a finite population can produce any result, the best we can expect from adding a frequency-dependent fitness component to a birth-and-death process is an evaluation of the *direction of the effect* in which this modifies the outcome of the dynamics. As Nowak et al. (2004) say: *Selection opposes L_1 replacing L_2* if its fixation probability is lower than the neutral threshold $1/N$; and *selection favors L_1 replacing L_2* if its fixation probability is higher than $1/N$.

Under the deterministic replicator dynamics in a infinitely large population, a mutant strategy can only spread if it has at least the same payoff as the resident type—which is exactly what is reflected by the classical notion of an evolutionarily stable strategy (Maynard Smith, 1982). Indeed, Maynard Smith's (1988) and Schaffer's (1988) extension of evolutionary stability to finite populations—condition (14) with a weak inequality sign—captures exactly the same aspect: a mutant strategy should not do better than the originally resident type, given the post-entry state of the population. We have seen that for games with a symmetric payoff function, this condition reduces to the condition of a Nash-equilibrium strategy in the base game.

In Example 1, we have seen that a strategy (P,Q) that is of the form of a neutrally stable strategy (in the sense of a game played in an infinitely large population) satisfies Maynard Smith's (1988) and Schaffer's (1988) notion of stability for finite populations, when tested against a mutant that switches to an efficient proto-language (P',Q'). In fact, what defines a neutrally stable strategy of the game discussed here, is exactly that it has no alternative best replies in terms of an efficient language; otherwise there would be an alternative best reply that does better against itself than the originally resident type does against the alternative best reply (see Pawlowitsch, 2007). Under the deterministic replicator dynamics that operates in an infinitely large population this has the effect that evolution can be blocked in such an inefficient state.

In a finite population, however, a mutant strategy can spread even if it has a strict disadvantage in fitness relative to the resident type when it first appears in the population. As we have seen in Example 1, its probability to reach fixation can even be higher than the neutral threshold $1/N$. In general, Maynard Smith's (1988) and Schaffer's (1988) concept of evolutionary stability for finite populations is neither necessary nor sufficient for the notion of stability in terms of fixation probabilities as introduced by Nowak and his coauthors (see Nowak et al., 2004; or Nowak, 2006).

We have seen that for games with a symmetric payoff function, the condition of Nowak et al. (2004) reduces in a very nice way and does no longer depend on N, as soon as $N\geqslant 3$. Condition (21) states that a mutant strategy (P',Q') that appears in an otherwise monomorphic population (P,Q) will reach fixation with a probability greater than $1/N$ if and only if its disadvantage against the originally resident type $f(L',L)-f(L,L)$ is outweighed by an advantage that the mutant strategy has against itself relative to the payoff that the originally resident type has against itself $f(L',L')-f(L,L)$.

Proposition 1 tells us that under a frequency-dependent Moran process with weak selection, efficient proto-languages—that is, strategies that are strict Nash in the complete strategy space—are the only strategies such that no mutant strategy, out of the complete strategy space, has a fixation probability that is higher than $1/N$. In this sense we may say that in a finite population, efficient proto-languages are the only strategies that are both *protected* and *favored by selection*.

Of course, Proposition 1 is only applicable if $m=n$. If $m\neq n$ then there are no strategies that are strict Nash in the complete strategy space, and there will be much more drift between strategies that are selectively neutral irrespective of the state of the population. But still, in the complete strategy space, strategies will differ in terms of their communicative potential. For example, a neutrally stable strategy that links 3 objects to the same sign and leaves one sign unused, will be dominated by a neutrally stable strategy that keeps everything else constant, but that links only 2 objects to the same sign and uses the previously idle sign for the third object. A way to tackle a situation where $m \neq m$ is to define hierarchies of neutrally stable strategies. However, this does not concern the main conclusion of this paper, which is indeed inextricably linked to the assumption that $m = n$.

Intuitively one might guess that if we have the same number of objects of communication and the same number of signs, a simple feedback dynamics would always lead to a proto-language where every object is bijectively linked to one sign and vice versa (see for example, Lewis, 1969, and Wärneryd, 1993). However, in an infinitely large population this is not necessarily the case (Nowak and Krakauer, 1999; Pawlowitsch, 2007). What we learn from the results presented here is that finite populations are one possible solution to reconcile intuition with a formal argument. In a finite population, efficient proto-languages are the only strategies that are protected by selection.

The efficiency of a proto-language is not only interesting from a purely optimality-oriented point of view; it also has important implications for bidirectionality of concept-to-sign mappings. In the modeling framework used here, efficient proto-languages are the only strategies that display perfect bidirectionality: whenever a particular sign is used to communicate a particular concept, this sign will also evoke the image of this concept. Most linguistic theories take bidirectionality as an innate property of language, which is ultimately genetically determined. In an evolutionary framework we can address the question of how this universal design feature of language could evolve. Deterministic evolution in an infinitely large population does not necessarily lead to bidirectionality. Frequency-dependent selection in a finite population provides one possible foundation for bidirectionality.

Appendix A. Proofs and methods

A.1. Nash strategies and evolutionary stability

In evolutionary game theory, a strategy is usually called a *Nash strategy* if it is a best response to itself. For the game discussed here that is,

$$f[(P,Q),(P,Q)] \geqslant f[(P,Q),(P',Q')] \quad \text{for all } (P',Q') \in \mathscr{P}_{n\times m} \times \mathscr{Q}_{m\times n}.$$

And a strategy is called a *strict Nash strategy* if it is a unique best response to itself, that is, if the inequality above holds with a strict inequality sign.

Let

$$B(P) = \{Q \in \mathscr{Q}_{m\times n} : \mathrm{tr}(PQ) \geqslant \mathrm{tr}(PQ')\ \forall Q' \in \mathscr{Q}_{m\times n}\}$$

be the set of best responses to P—in the sense of the asymmetric game—and let

$$B(Q) = \{P \in \mathscr{P}_{n\times m} : \mathrm{tr}(PQ) \geqslant \mathrm{tr}(P'Q) \forall P' \in \mathscr{P}_{n\times m}\}$$

be the set of best responses to Q. As a general property of symmetrized games, in a Nash strategy the strategies played in two roles have to be best responses to each other. That is, (P,Q) is a Nash strategy of the symmetrized game if and only if $P \in B(Q)$ and $Q \in B(P)$. In a strict Nash strategy P has to be a unique best response to Q, and vice versa. Complementing this with a characterization of best-response properties in terms of the P and Q matrices gives us a good tool to characterize Nash strategies of the symmetrized game.

Lemma 1 (*Best-response properties*). *Let $P \in \mathscr{P}$ and $Q \in \mathscr{Q}$.*

(a) *For any $Q \in B(P)$*

$$\sum_{i \in \mathrm{argmax}_i(p_{ij^*})} q_{j^\star i} = 1 \quad \textit{and} \quad q_{j^\star i} = 0 \quad \forall i \notin \mathrm{argmax}_i(p_{ij^*});$$

(b) *for any $P \in B(Q)$*

$$\sum_{j \in \mathrm{argmax}_j(q_{ji^*})} p_{i^\star j} = 1 \quad \textit{and} \quad p_{i^\star j} = 0 \quad \forall j \notin \mathrm{argmax}_j(q_{ji^*}).$$

Of course, if $\bar{p}_{i^*j^*}$ is the *unique maximal element* in the j^*th column of $\bar{P}$, then, for any Q that is a best response to $\bar{P}$, $q_{j^\star i^\star} = 1$; and vice versa for the roles of P and Q reversed. Note also that by the contrapositive of Lemma 1,

(a) if $Q \in B(P)$, then

$$q_{j^\star i^\star} \neq 0 \Rightarrow p_{i^\star j^\star} = \max_i(p_{ij^*}) \Rightarrow p_{i^\star j^\star} \neq 0 \text{ or } p_{ij^*} = 0\ \forall i; \textit{and}$$

(b) if $P \in B(Q)$, then

$$p_{i^\star j^\star} \neq 0 \Rightarrow q_{j^\star i^\star} = \max_j(q_{ji^*}) \Rightarrow q_{j^\star i^\star} \neq 0 \text{ or } q_{ji^*} = 0\ \forall j.$$

For more details see Pawlowitsch (2007).

The proof of Proposition 1 makes extensive use of an earlier result on neutral stability. It is therefore useful to state this result briefly here.

Definition 1 (*Neutral stability*). A strategy $(P,Q) \in \mathscr{P}_{n\times m} \times \mathscr{Q}_{m\times n}$ is *neutrally stable* if and only if

(i) it is a Nash strategy, and if
(ii) whenever $f[(P,Q),(P,Q)] = f[(P',Q'),(P,Q)]$ for some $(P',Q') \in \mathscr{P}_{n\times m} \times \mathscr{Q}_{m\times n}$, then

$$f[(P,Q),(P',Q')] \geqslant f[(P',Q'),(P',Q')].$$

Lemma 2 (*Pawlowitsch, 2007*). *Let* $(P, Q) \in \mathcal{P}_{n\times m} \times \mathcal{Q}_{m\times n}$ *be a Nash strategy*. (P, Q) *is a neutrally stable strategy if and only if*

(i) *at least one of the two matrices, P or Q, has no zero-column, and*
(ii) *neither P nor Q has a column with multiple maximal elements that are strictly between* 0 *and* 1.

A.2. Proof of Proposition 1

Proof. Part (a) is easily seen from an indirect argument. Suppose that there is some $L' \in \mathcal{P}_{n\times n} \times \mathcal{Q}_{n\times n}$ such that $\rho' \geqslant \frac{1}{N}$. Then $f(L', L') + f(L', L) \geqslant 2f(L, L)$. However, this cannot be true as $f(L, L) > f(L', L)$, for all $(P', Q') \in \mathcal{P}_{n\times n} \times \mathcal{Q}_{n\times n}$ (L is a strict Nash strategy), and since (P, Q) exploits already the maximally available potential of communication,

$$f(L, L) = \mathrm{tr}(P, Q) = n \geqslant \mathrm{tr}(P', Q') = f(L', L')$$

for all $(P', Q') \in \mathcal{P}_{n\times n} \times \mathcal{Q}_{n\times n}$.

For part (b), the only interesting case is where $L = (P, Q)$ is a neutrally stable strategy (that is not an evolutionarily stable strategy). If $L = (P, Q)$ is not even a Nash strategy, that is, $Q \notin B(P)$ or $P \notin B(Q)$, or both, there will always be a $P' \in B(Q)$ and $Q' \in B(P)$ with $(P', Q') \neq (P, Q)$. Say, $Q \notin B(P)$, and consider $L' = (P, Q')$ where $Q' \in B(P)$. Then

$$\mathrm{tr}(PQ') + (\tfrac{1}{2}\mathrm{tr}(PQ) + \tfrac{1}{2}\mathrm{tr}(PQ')) > 2\mathrm{tr}(PQ),$$

and therefore,

$$f(L', L') + f(L', L) > 2f(L, L) \Rightarrow \rho' > \frac{1}{N}.$$

If $L = (P, Q)$ is a Nash strategy, but not a neutrally stable strategy, then there is some $(P', Q') \in \mathcal{P}_{n\times n} \times \mathcal{Q}_{n\times n}$ such that (1) $f[(P', Q'), (P, Q)] = f[(P, Q), (P, Q)]$, and (2) $f[(P', Q'), (P', Q')] > f[(P, Q), (P', Q')]$. By the symmetry of the payoff function, (2) can be written as $f[(P', Q'), (P', Q')] > f[(P', Q'), (P, Q)]$. Plugging (1) into the right-hand side of (2), we have that $f[(P', Q'), (P', Q')] > f[(P, Q), (P, Q)]$, which is equivalent to

$$\mathrm{tr}(P'Q') > \mathrm{tr}(PQ).$$

Furthermore, the condition that (P, Q) has to be a Nash strategy implies that $\mathrm{tr}(P'Q) \leqslant \mathrm{tr}(PQ)$ *for all* $P' \in \mathcal{P}_{n\times n}$. Otherwise, (P', Q) would be a better reply to (P, Q) than (P, Q) is to itself, $\mathrm{tr}(P'Q) + \mathrm{tr}(PQ) > \mathrm{tr}(PQ) + \mathrm{tr}(PQ)$, which cannot be true if (P, Q) is to be a Nash strategy. By an analogous argument we also have that $\mathrm{tr}(PQ') \leqslant \mathrm{tr}(PQ)$ *for all* $Q' \in \mathcal{Q}_{n\times n}$. From (1) we therefore have that $\mathrm{tr}(P'Q) = (PQ) = (PQ')$. As a consequence,

$$\mathrm{tr}(P'Q') + (\tfrac{1}{2}\mathrm{tr}(P'Q) + \tfrac{1}{2}\mathrm{tr}(PQ')) > 2\mathrm{tr}(PQ).$$

That is,

$$f(L', L') + f(L', L) > 2f(L, L) \Rightarrow \rho' > \frac{1}{N}.$$

Suppose then that $L = (P, Q)$ is a neutrally stable strategy (but not an evolutionarily stable strategy). From Pawlowitsch (2007) we know that if (P, Q) is a neutrally stable strategy, then (i) neither P nor Q can have a column with multiple maximal elements that are strictly between 0 and 1, and (ii) at least P or Q has no zero column. Suppose, without loss of generality, that P is the matrix that has no zero column. It then follows from the restrictions on P and Q and the fact that they have to be best responses to each other, that P has at least one *row* that has at least 2-entries strictly between 0 and 1, such that they are unique maximal elements of their respective *columns*.

To see why this is so, consider the following: Since P has no column with multiple maximal elements strictly between 0 and 1 and no zero-column, for every column in P we have that its maximum is (a) unique and equal to 1, (b) unique, but not equal to 1, or (c) equal to 1, but not unique. First note that P has to display at least one case of (b) or (c); otherwise P would be a permutation matrix, and by the properties of best responses, Q the transpose of P, which is exactly what is ruled out in this part of the proposition. Suppose then that P displays a case of (b), and let $p_{i^\star j^\star} \in (0, 1)$ be the unique maximal element of the $j^\star$th column of P. As Q is a best response to P, we have that $q_{j^\star i^\star} = 1$. This implies that $q_{j^\star i^\star}$ is *a maximal element* of the $i^\star$th column of Q. However, as $p_{i^\star j^\star} \neq 1$, but the sum over all $p_{i^\star j}$ such that $j \in \mathrm{argmax}_j(q_{ji^\star})$ has to be exactly equal to 1, $q_{j^\star i^\star} = 1$ cannot be the unique maximal element in the $i^\star$th column of Q, and there must be some $j \neq j^\star$ with $j \in \mathrm{argmax}_j(q_{ji^\star})$ such that $p_{i^\star j} \neq 0$. As Q is a best response to P, whenever $q_{ji^\star} \neq 0$, which is indeed the case for all $j \in \mathrm{argmax}_j(q_{ji^\star})$, then $p_{ji^\star}$ is a maximal element of its respective column in P. As P has no zero-column, the respective element in the $i^\star$th row of P is strictly between 0 and 1. As P has *no multiple* maximal elements that are strictly between 0 and 1, we have that for all $j \in \mathrm{argmax}_j(q_{ji^\star})$, $0 < p_{i^\star j} < 1$ is indeed the *unique maximal element* of its respective column in P, and we are done for this case. If P displays a case of (c), this means that there are at least two rows that have a 1 entry in the same column position and 0 otherwise. Provided that $m = n$, this implies that P has to have at least two columns with unique maxima strictly between 0 and 1. From (b) it then follows that they have to be in the same row.

Now, let $i^\star$ be that *row*. As Q is a best response to P, $q_{ji^\star}$ must be equal to 1 whenever its orthogonal element in P, $p_{i^\star j}$ is positive. As there are at least two such positive elements in the $i^\star$th *row* of P, the $i^\star$th *column* of Q must have at least two 1-entries. As on the other hand P is a best response to Q, the sum over all $p_{i^\star j}$ such that $j \in \mathrm{argmax}_j(q_{ji^\star})$ has to be exactly equal to 1.

A consequence of the multiple 1-entries in the $i^\star$th column of Q is that Q has at least (i) one zero column or (ii) two columns with unique maximal elements strictly between 0 and 1. We consider each of these two cases in turn.

(i) Suppose the elements of the $i^{\star\star}$th column in Q are all equal to 0. We now construct a potential mutant (P', Q'). For some $j^\star \in \mathrm{argmax}_j(q_{ji^\star})$ let $p'_{i^\star j^\star} = 1$ and $p'_{i^\star j} = 0$ for all $j \neq j^\star$. And for some $j^{\star\star} \in \mathrm{argmax}_j(q_{ji^\star})$, $j^{\star\star} \neq j^\star$, let $p'_{i^{\star\star}j^{\star\star}} = 1$, and $p'_{i^{\star\star}j} = 0$ for all $j \neq j^{\star\star}$. Otherwise $p'_{ij} = p_{ij}$. That is, we have constructed P' from P just by replacing its $i^\star$th row with a vector that has one 1 in its $j^\star$ position, and by replacing its $i^{\star\star}$th row with a vector that has one 1 in its $j^{\star\star}$ position. To construct Q' from Q it suffices to exchange its $j^{\star\star}$th row, such that $q'_{j^{\star\star}i^{\star\star}} = 1$ and $q'_{j^{\star\star}i} = 0$ for all $i \neq i^{\star\star}$.

It is then just a matter of payoff comparison to see that (P', Q') wins more against itself than it loses against (P, Q)—compared to the payoff of (P, Q) against itself: instead of. Summing over all $i \neq i^\star, i^{\star\star}$ and over all j we have

$$\sum_{i \backslash \{i^\star, i^{\star\star}\}} \sum_j p'_{ij} q'_{ji} = \sum_{i \backslash \{i^\star, i^{\star\star}\}} \sum_j p'_{ij} q_{ji} = \sum_{i \backslash \{i^\star, i^{\star\star}\}} \sum_j p_{ij} q'_{ji}$$
$$= \sum_{i \backslash \{i^\star, i^{\star\star}\}} \sum_j p_{ij} q_{ji},$$

which comes from the fact that in P' we did not change any row other than $i^\star$ and $i^{\star\star}$, and that for any column $i \neq i^\star, i^{\star\star}$ *in* Q *and in* Q' all elements that are in the position of the $i^\star$th or $i^{\star\star}$th row are 0 anyway. For $i^\star$ we have that

$$\sum_j p'_{i^\star j} q'_{ji^\star} = 1,$$

$$\sum_j p_{i^\star j} q_{ji^\star} = 1.$$

And

$$\sum_j p'_{i^\star j} q_{ji^\star} = 1,$$

$$\sum_j p_{i^\star j} q'_{ji^\star} = p_{i^\star j^\star} \in (0, 1).$$

Hence, if we consider the $i^\star$th row of P (column of Q) in isolation, the mutant (P', Q') loses relative to the resident type (P, Q). But for $i^{\star\star}$ we have that

$$\sum_j p'_{i^{\star\star}j} q'_{ji^{\star\star}} = 1,$$
$$\sum_j p_{i^{\star\star}j} q_{ji^{\star\star}} = 0,$$
$$\sum_j p'_{i^{\star\star}j} q_{ji^{\star\star}} = 0,$$
$$\sum_j p_{i^{\star\star}j} q'_{ji^{\star\star}} = p_{i^{\star\star}j^{\star\star}} \in [0, 1).$$

As a consequence, *in sum* the mutant gains more against itself than it loses against the resident type—compared to the payoff of the resident type against itself. More explicitly, itself. More explicitly,

$$f(L', L') - f(L, L) = \sum_i \sum_j p'_{ij} q'_{ji} - \sum_i \sum_j p_{ij} q_{ji}$$
$$= 1,$$

whereas

$$f(L, L) - f(L', L) = \sum_i \sum_j p_{ij} q_{ji}$$
$$- \left(\frac{1}{2} \sum_i \sum_j p'_{ij} q_{ji} + \frac{1}{2} \sum_i \sum_j p_{ij} q'_{ji} \right)$$
$$= \frac{1 - p_{i^\star j^\star} - p_{i^{\star\star}j^{\star\star}}}{2}$$
$$< 1$$

and therefore,

$$f(L', L') + f(L', L) > 2f(L, L),$$

which means that

$$\rho' > \frac{1}{N}.$$

(ii) If Q has no zero column, then by an analogous argument to what we have seen above for P, it follows that Q has a row with at least two elements strictly between 0 and 1 that are unique maxima of their respective columns in Q. Suppose that these are the $i^{\star\star}$th and $i^{\star\star\star}$th column of Q. These unique column maxima can only appear in some row $j^{\star\star\star} \notin \mathrm{argmax}_j(q_{ji^\star})$—as for all $j \in \mathrm{argmax}_j(q_{ji^\star})$ the full mass of 1 is already at its element in the $i^\star$th position. As P is a best response to Q, both $p_{i^{\star\star}j^{\star\star\star}}$ and $p_{i^{\star\star\star}j^{\star\star\star}}$ are maximal elements of the $j^{\star\star\star}$th column in P. As P has no multiple maximal elements that are not equal to 1 and no zero-column, both $p_{i^{\star\star}j^{\star\star\star}}$ and $p_{i^{\star\star\star}j^{\star\star\star}}$ are equal to 1.

In constructing P' from P, and Q' from Q, we proceed as above for the $i^\star$th and the $i^{\star\star}$th row in P', and the $j^{\star\star}$th row in Q'. Upon that we exchange the $j^{\star\star\star}$th row in Q' such that $q'_{i^{\star\star\star}j^{\star\star\star}} = 1$ and $p'_{j^{\star\star\star}i} = 0$ for all $i \neq i^{\star\star\star}$. Otherwise the entries in P' and Q' stay the same as in P and, respectively, Q. Note that $p_{i^{\star\star\star}j^{\star\star\star}}$ is already equal to 1. We then find that

$$\sum_j p'_{i^\star j} q'_{ji^\star} = 1,$$
$$\sum_j p_{i^\star j} q_{ji^\star} = 1,$$
$$\sum_j p'_{i^\star j} q_{ji^\star} = 1,$$
$$\sum_j, p_{i^\star j} q'_{ji^\star} = p_{i^\star j^\star} \in (0, 1).$$

And

$$\sum_{i^{\star\star}, i^{\star\star\star}} \sum_j p'_{ij} q'_{ji} = 2,$$
$$\sum_{i^{\star\star}, i^{\star\star\star}} \sum_j p_{ij} q_{ji} = 1,$$
$$\sum_{i^{\star\star}, i^{\star\star\star}} \sum_j p'_{ij} q_{ji} = q_{j^{\star\star\star}i^{\star\star\star}} \in (0, 1),$$
$$\sum_{i^{\star\star}, i^{\star\star\star}} \sum_j p_{ij} q'_{ji} = 1.$$

As a consequence,

$$\begin{aligned} f(L',L')-f(L,L) &= \sum_i \sum_j p'_{ij}q'_{ji} - \sum_i \sum_j p_{ij}q_{ji} \\ &= 1, \end{aligned}$$

whereas

$$\begin{aligned} f(L,L)-f(L',L) &= \sum_i \sum_j p_{ij}q_{ji} \\ &\quad - \left(\frac{1}{2}\sum_i \sum_j p'_{ij}q_{ji} + \frac{1}{2}\sum_i \sum_j p_{ij}q'_{ji}\right) \\ &= \frac{1-p_{i^\star j^\star}}{2} + \frac{1-q_{j^{\star\star\star} i^{\star\star\star}}}{2} \\ &< 1. \end{aligned}$$

As before, the mutant wins 1 against itself relative to the payoff that the resident type gets against itself; it loses against the resident type, but this loss is strictly less than 1, and therefore,

$$f(L',L')+f(L',L) > 2f(L,L),$$

which implies that,

$$\rho' > \frac{1}{N}. \qquad \square$$

Illustration of the proof.

(i) Q has a zero column.

$$P = \begin{array}{cccc} & j^\star & j^{\star\star} & \\ i^\star & 1-\alpha & \alpha & 0 \\ i^{\star\star} & 0 & 0 & 1 \\ & 0 & 0 & 1 \end{array}, \quad Q = \begin{array}{cccc} & i^\star & i^{\star\star} & \\ j^\star & 1 & 0 & 0 \\ j^{\star\star} & 1 & 0 & 0 \\ & 0 & 0 & 1 \end{array},$$

$$P' = \begin{array}{cccc} & j^\star & j^{\star\star} & \\ i^\star & 1 & 0 & 0 \\ i^{\star\star} & 0 & 1 & 0 \\ & 0 & 0 & 1 \end{array}, \quad Q' = \begin{array}{cccc} & i^\star & i^{\star\star} & \\ j^\star & 1 & 0 & 0 \\ j^{\star\star} & 0 & 1 & 0 \\ & 0 & 0 & 1 \end{array}.$$

(ii) Q has no zero column.

$$P = \begin{array}{cccc} & j^\star & j^{\star\star} & j^{\star\star\star} \\ i^\star & 1-\alpha & \alpha & 0 \\ i^{\star\star} & 0 & 0 & 1 \\ i^{\star\star\star} & 0 & 0 & 1 \end{array}, \quad Q = \begin{array}{cccc} & i^\star & i^{\star\star} & i^{\star\star\star} \\ j^\star & 1 & 0 & 0 \\ j^{\star\star} & 1 & 0 & 0 \\ j^{\star\star\star} & 0 & \beta & 1-\beta \end{array},$$

$$P' = \begin{array}{cccc} & j^\star & j^{\star\star} & j^{\star\star\star} \\ i^\star & 1 & 0 & 0 \\ i^{\star\star} & 0 & 1 & 0 \\ i^{\star\star\star} & 0 & 0 & 1 \end{array}, \quad Q' = \begin{array}{cccc} & i^\star & i^{\star\star} & i^{\star\star\star} \\ j^\star & 1 & 0 & 0 \\ j^{\star\star} & 0 & 1 & 0 \\ j^{\star\star\star} & 0 & 0 & 1 \end{array}.$$

References

Bomze, I., Weibull, J., 1995. Does neutral stability imply Lyapunov stability? Games Econ. Behav. 11, 173–192.

Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S.L., Wiebe, V., Kitano, T., Monaco, A.P., Pääbo, S., 2002. Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418, 869–872.

Hofbauer, J., Sigmund, K., 1988. The Theory of Evolution and Dynamical Systems. Cambridge University Press, Cambridge.

Hofbauer, J., Sigmund, K., 1998. Evolutionary Games and Population Dynamics. Cambridge University Press, Cambridge.

Hurford, J., 1989. Biological evolution of the Saussurean sign as a component of the language acquisition device. Lingua 77, 187–222.

Imhof, L.A., Nowak, M.A., 2006. Evolutionary game dynamics in a Wright–Fisher process. J. Math. Biol. 52, 667–681.

Komarova, N.L., Nowak, M.A., 2003. Language dynamics in finite populations. J. Theor. Biol. 221, 445–457.

Lessard, S., Ladret, V., 2007. The probability of fixation of a single mutant in an exchangeable selection model. J. Math. Biol. 54, 721–744.

Lewis, D., 1969. Convention: A Philosophical Study. Harvard University Press, Cambridge, MA.

Maynard Smith, J., 1982. Evolution and the Theory of Games. Cambridge University Press, Cambridge.

Maynard Smith, J., 1988. Can a mixed strategy be stable in a finite population? J. Theor. Biol. 130, 247–251.

Niyogi, P., 2006. The Computational Nature of Language Learning and Evolution. MIT Press, Cambridge.

Nowak, M.A., 2006. Evolutionary Dynamics: Exploring the Equations of Life. Belknap Press of Harvard University Press.

Nowak, M.A., Krakauer, D.C., 1999. The evolution of language. Proc. Natl Acad. Sci. USA 96, 8028–8033.

Nowak, M.A., Sigmund, K., 2004. Evolutionary dynamics of biological games. Science 303, 793–799.

Nowak, M.A., Plotkin, J.B., Krakauer, D.C., 1999. The evolutionary language game. J. Theor. Biol. 200, 147–162.

Nowak, M.A., Komarova, N.L., Niyogi, P., 2002. Computational and evolutionary aspects of language. Nature 417, 611–617.

Nowak, M.A., Sasaaki, A., Taylor, C., Fudenberg, D., 2004. Emergence of cooperation and evolutionary stability in finite populations. Nature 428, 246–650.

Pawlowitsch, C., 2007. Why evolution does not always lead to an optimal signaling system. Games Econ. Behav., forthcoming.

Schaffer, M.E., 1988. Evolutionarily stable strategies for a finite population and a variable contest size. J. Theor. Biol. 132, 469–478.

Selten, R., 1980. A note on evolutionarily stable strategies in asymmetric animal contests. J. Theor. Biol. 84, 93–101.

Taylor, P.D., Jonker, L., 1978. Evolutionary stable strategies and game dynamics. Math. Biosci. 40, 145–156.

Thomas, B., 1985. On evolutionarily stable sets. J. Math. Biol. 22, 105–115.

Trapa, P., Nowak, M.A., 2000. Nash equilibria for an evolutionary language game. J. Math. Biol. 41, 172–188.

Traulsen, A., Claussen, J.C., Hauert, C., 2005. Coevolutionary dynamics: from finite to infinite populations. Phys. Rev. Lett. 95, 238701-4.

Traulsen, A., Nowak, M.A., Pacheco, J.M., 2006. Stochastic dynamics of invasion and fixation. Phys. Rev. E 74, 011909-5.

Wärneryd, K., 1993. Cheap talk, coordination and evolutionary stability. Games Econ. Behav. 5, 532–546.