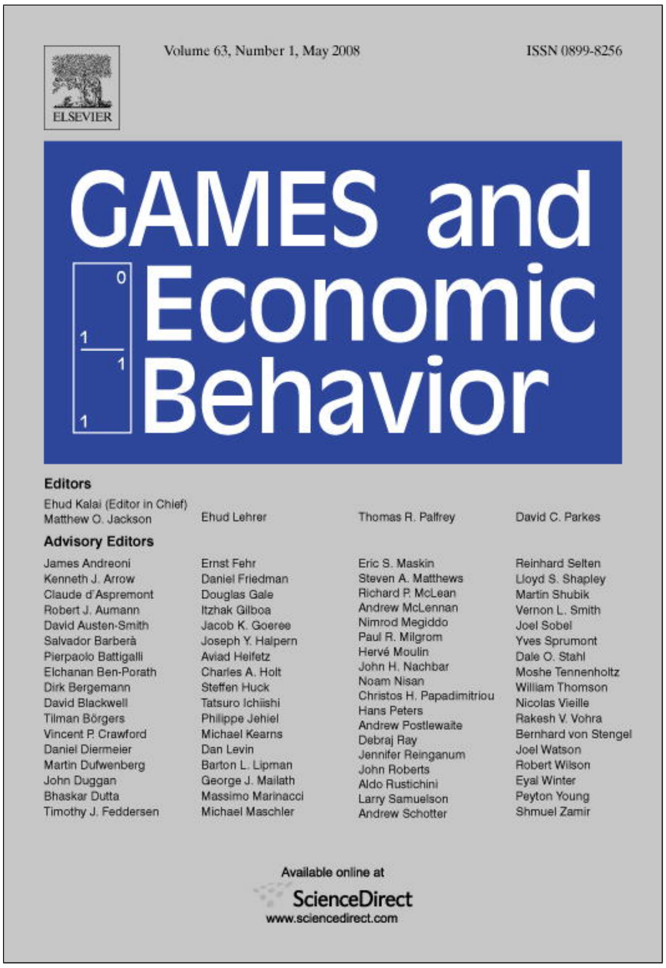

Volume 63, Number 1, May 2008

ISSN 0899-8256

ELSEVIER

GAMES and Economic Behavior

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Games and Economic Behavior 63 (2008) 203–226

GAMES and Economic Behavior

www.elsevier.com/locate/geb

Why evolution does not always lead to an optimal signaling system

Christina Pawlowitsch [*,**]

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

Received 27 July 2006

Available online 23 October 2007

Abstract

This paper gives a complete characterization of *neutrally stable strategies* for sender–receiver games in the style of Lewis, or Nowak and Krakauer [Lewis, D., 1969. Convention: A Philosophical Study. Harvard Univ. Press, Cambridge, MA; Nowak, M., Krakauer, D., 1999. The evolution of language. Proc. Nat. Acad. Sci. USA 96, 8028–8033]. Due to the dynamic implications of neutral stability, the replicator dynamics of this model does not necessarily lead to the rise of an optimal signaling system, where every state of the world is bijectively linked to one signal and vice versa, but it can be trapped in suboptimum situations where two (or more) signals are used for the same event, or two (or more) events are associated with one and the same signal.

JEL classification: C72

Keywords: Language evolution; Sender–receiver game; Suboptimality; Neutral stability; Replicator dynamics; Lyapunov stability

1. Introduction

The idea to use game theoretic methods for analyzing phenomena of natural language dates back relatively early into the history of game theory. In his book “Convention: A philosophical

[*] Fax: +43 1 4277 37495.

[**] Present address: Program for Evolutionary Dynamics, Harvard University, One Brattle Square, Cambridge, MA 02138, USA. Fax: +1 617 496 4629.

E-mail addresses: pawlowit@fas.harvard.edu, christina.pawlowitsch@univie.ac.at.

0899-8256/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.geb.2007.08.009

study" David K. (1969) put forward the concept of Nash equilibrium to explain the conventional character of natural languages.

Lewis' contribution addressed a long-standing debate in Philosophy. W.V. Quine and others had argued that natural languages cannot be like the well–understood cases of central conventions, since we could not have agreed on them without the use of any, even rudimentary, communication device (see, for example, Quine, 1936, 1960, and White, 1956). Lewis, who had gotten into contact with the theory of games through T.C. Schelling, defended the view that languages are conventions; however not in the sense of centrally organized institutions, but in the sense of *self-enforcing agreements*. Provided that everybody is doing his or her part of a convention, no individual agent has an incentive to deviate—in the language of game theory, this is a *Nash equilibrium*. These equilibria, Lewis argues, are supported by agents proceeding by precedent and basing their choices on a system of mutually interacting higher-order expectations. Once established, therefore, such equilibria would persist without the need of any centrally coordinating authority. Interestingly, this led Lewis to one of the first notions of *common knowledge*. Aumann and Heifetz (2002) write: "Lewis (1969) was the first to define common knowledge, which of course is a multi-person, interactive concept; though verbal, his treatment was entirely rigorous."

To make his point of view more precise, Lewis (1969) introduces a simple coordination game that can be used to explain the conventional character of signaling systems. In this game, there are two types of players, senders and receivers. A strategy for a sender is a mapping from events (potential objects of communication) to signals, and a strategy for a receiver is a mapping from signals to events. A combination of strategy choices such that the sender's strategy is a bijective mapping from the set of events into the set of available signals and the receiver's strategy is the inverse of this mapping is a Nash equilibrium of this game—in fact it is even a strict Nash equilibrium. Lewis calls such equilibria *conventional signaling systems*. But this game also admits other, non-strict Nash equilibria, where two (or more) events are mapped to the same signal, or where two (or more) signals are mapped to the same event. Lewis claims that due to its *salient* properties, however, a conventional signaling system eventually will prevail.

What seems to lie behind Lewis' line of reasoning is the idea of some dynamic process that operates in a population of individual agents. But without explicitly formulating such a dynamics, Lewis' argument for the rise of a conventional signaling system hardly goes beyond an *ad hoc* assertion.

Besides his notion of common knowledge, Lewis' important conceptual innovation is to have given an interpretation of the meaning of arbitrary signs as an equilibrium property of a game. Yet, this is only half of a reply to the critique of Quine and others. It does not explain *how* a convention of language can come into being in the first place. Of course, this can only be answered in a dynamic, or evolutionary framework.

Wärneryd (1993) takes up Lewis' model exactly with this perspective. He considers a symmetrized version of this game, where individual agents are assumed to find themselves in the roles of senders and receivers with equal probabilities. Wärneryd shows that conventional signaling systems in the sense of Lewis are not only the only *evolutionarily stable strategies* of this game, but also the only *weakly evolutionarily stable strategies*. In the literature weak evolutionary stability is also referred to as *neutral stability* (see, for example, Maynard Smith, 1982, or Bomze and Weibull, 1995). Evolutionary stability implies asymptotic stability in the replicator dynamics (Taylor and Jonker, 1978), whereas neutral stability implies Lyapunov stability in the replicator dynamics (Thomas, 1985; see also Bomze and Weibull, 1995). Wärneryd suggests these dynamic implications as a foundation for the rise of a conventional signaling system in the sense of Lewis by some trial-and-error process.

However, Wärneryd's characterization of neutrally stable strategies *does not* take into account the mixed strategies of this game. But in a dynamic, population based setting, mixed strategies necessarily arise as the non-monomorphic states of the population, where it is composed of different types of otherwise identical agents who use different pure strategies. It is not difficult to find mixed strategies of this game that are neutrally stable, but not evolutionarily stable. A complete treatment of the evolutionary aspects of this model, therefore, has to be complemented by an analysis of neutral stability that also takes into account the possibility of mixed strategies.

As a model in mixed strategies, Lewis' sender–receiver game is equivalent to the sender–receiver games that have been studied in the context of the origins of language by Hurford (1989), and in a series of articles by M.A. Nowak and others; for example, Nowak and Krakauer (1999), Nowak et al. (1999), or Trapa and Nowak (2000).

From computer simulations with this model in the style of a replicator dynamics Nowak and Krakauer (1999) conjecture that event-to-signal relations where one signal is used for two (or more) events can be stable in an evolutionary context. Trapa and Nowak (2000) show that for the model in mixed strategies it is also true that a strategy is evolutionarily stable if and only if every event is bijectively linked to one signal and vice versa. Hence, a strategy where one signal is used for two (or more) events *cannot* be evolutionarily stable in the strict sense. But this still leaves the question whether we can identify some other selection criterion that can help us to explain why an evolutionary dynamics can be blocked in a suboptimum state where ambiguities in concept-to-sign mappings persist in the population. Neutral stability, as it turns out, provides one possible explanation.

This paper does two things. First, it gives a complete characterization of neutrally stable strategies for a population-based version of the sender–receiver game studied in Lewis (1969) and Wärneryd (1993). Second, it explores the consequences of neutral stability for the long run behavior of the replicator dynamics of this model. Section 2 introduces the model. Section 3 discusses the properties of Nash equilibria and reviews the results on evolutionary stability. Section 4 derives a complete characterization (necessary and sufficient conditions) for a neutrally stable strategy. Section 5 is concerned with the long-run behavior of the replicator dynamics. Section 6 discusses related literature and concludes with some comments on extensions of the model and future work.

2. The model

The basic idea of Lewis' (1969) sender–receiver game is that there is an agent who is informed about the state of the world—but who is not directly called for action—and an agent who is *not* informed about the state of the world, but who has to take an action that determines the payoffs for both agents in a perfectly symmetric way. The informed agent has access to a set of arbitrary signs, one of which she can send to the uninformed agent *before* he takes the payoff relevant action. For each event there is exactly one action that has to be taken, such that every action can be identified with the particular event by which it is called for. If the uninformed agent takes the "correct" action, then this yields a payoff of 1 for both players, and 0 otherwise. Taking this action can be interpreted as understanding the state of the world, and players can be referred to as sender and receiver according to their role in latent communication.

Suppose that there are n events that potentially become the object of communication, and that there are m available signals. A sender's strategy can be represented by a matrix

$$P \in \mathcal{P}_{n\times m} = \bigl\{P \in \mathbb{R}_+^{n\times m}\colon\ \forall i,\ p_{ij} = 1 \text{ for some } j = j'(i) \text{ and } p_{ij} = 0\ \forall j \neq j'(i)\bigr\}, \tag{1}$$

where $p_{ij} = 1$ if event i is mapped to signal j, and zero otherwise. A receiver's strategy can be represented by a matrix

$$Q \in \mathcal{Q}_{m\times n} = \bigl\{Q \in \mathbb{R}_+^{m\times n}\colon\ \forall j,\ q_{ji} = 1 \text{ for some } i = i'(j) \text{ and } q_{ji} = 0\ \forall i \neq i'(j)\bigr\}, \tag{2}$$

where $q_{ji} = 1$ if signal j is associated with event i and 0 otherwise. There are m^n elements in $\mathcal{P}_{n\times m}$, and n^m elements in $\mathcal{Q}_{m\times n}$. Note that P and Q can be seen as special cases of stochastic matrices, since their row elements add up to 1.

The payoff function for both the sender and the receiver can be written as

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

In the sequel, $\operatorname{tr}(PQ)$ is also referred to as the *communicative potential* of a pair (P,Q). Equations (1)–(3) describe an asymmetric game between a sender and a receiver.

When we study language as a social phenomenon, we do not find agents who are either only senders or only receivers. Being a sender or being a receiver are rather *social roles* that agents adopt depending on their relative position in various situations of interaction with other agents. Formally we can capture this by looking at the symmetrized version of the asymmetric game. We assume that each individual agent finds herself or himself in the role of the sender or the receiver with equal probabilities. A strategy of the symmetrized game, then, is *a pair* of a sender and a receiver matrix,

$$(P,Q) \in \mathcal{P}_{n\times m} \times \mathcal{Q}_{m\times n}, \tag{4}$$

and the *payoff function* is given by

$$F\bigl[(P,Q),(P',Q')\bigr] = \frac{1}{2}\operatorname{tr}(PQ') + \frac{1}{2}\operatorname{tr}(P'Q). \tag{5}$$

Note that this payoff function is symmetric,

$$F\bigl[(P,Q),(P',Q')\bigr] = F\bigl[(P',Q'),(P,Q)\bigr],$$

giving rise to a so-called *doubly symmetric* or *partnership game*. A game is called doubly symmetric if all players have the same strategy set and payoff function, and if in addition to that this payoff function is symmetric (for more on partnership games and symmetrized games, see Hofbauer and Sigmund, 1998; or Cressman, 2003). In the sequel,

$$\mathcal{G}_{n,m} = \bigl\{\mathcal{P}_{n\times m} \times \mathcal{Q}_{m\times n}, F\bigl[(P,Q),(P'Q')\bigr]\bigr\} \tag{6}$$

refers to this game. Abstracting from notation, $\mathcal{G}_{n,m}$ is the game studied in Wärneryd (1993).

With each pair $(P_l,Q_l) \in \mathcal{P}_{n\times m} \times \mathcal{Q}_{m\times n}$ we identify a particular *type of agent* who performs this strategy whenever he or she is called for playing the game $\mathcal{G}_{n,m}$. Note that there are $L = m^n \times n^m$ elements in $\mathcal{P}_{n\times m} \times \mathcal{Q}_{m\times n}$. A *state of the population*, then, is a vector of type frequencies,

$$x = (x_1, x_2, \dots, x_L) \in S_L = \left\{x \in \mathbb{R}_+^L\colon\ \sum_{l=1}^{L} x_l = 1\right\}.$$

For every vector of type frequencies $x \in S_L$, the *population's average strategy* is

$$\left(\bar{P}_x, \bar{Q}_x\right) = \sum_1^L x_l (P_l, Q_l).$$

This can be written as

$$\left(\bar{P}_x, \bar{Q}_x\right) = \left[\begin{pmatrix} \bar{p}_{11} & \dots & \bar{p}_{1j} & \dots & \bar{p}_{1m} \\ \vdots & & \vdots & & \\ \bar{p}_{i1} & \dots & \bar{p}_{ij} & \dots & \bar{p}_{im} \\ \vdots & & \vdots & & \\ \bar{p}_{n1} & \dots & \bar{p}_{nj} & \dots & \bar{p}_{nm} \end{pmatrix}, \begin{pmatrix} \bar{q}_{11} & \dots & \bar{q}_{1i} & \dots & \bar{q}_{1n} \\ \vdots & & \vdots & & \\ \bar{q}_{j1} & \dots & \bar{q}_{ji} & \dots & \bar{q}_{jn} \\ \vdots & & \vdots & & \\ \bar{q}_{m1} & \dots & \bar{q}_{mj} & \dots & \bar{q}_{mn} \end{pmatrix}\right],$$

where

$$\bar{p}_{ij} = \sum_{l:p^l_{ij}=1} x_l \quad \text{and} \quad \bar{q}_{ji} = \sum_{l:q^l_{ji}=1} x_l. \tag{7}$$

That is, $\bar{p}_{ij}$ is the sum of all type frequencies whose i, jth entry in P is equal to 1, and analogously for $\bar{q}_{ji}$. Note that $\sum_{j=1}^m \bar{p}_{ij} = 1$, for all i, and $\sum_{i=1}^n \bar{q}_{ji} = 1$, for all j. The average strategy profile, therefore, is a pair of two stochastic matrices,

$$(\bar{P}, \bar{Q}) \in \mathcal{P}^\Delta_{n\times m} \times \mathcal{Q}^\Delta_{m\times n},$$

where

$$\mathcal{P}^\Delta_{n\times m} = \left\{P \in \mathbb{R}^{n\times m}_+ \colon \sum_{j=1}^m p_{ij} = 1,\ \forall i\right\}, \quad \text{and}$$

$$\mathcal{Q}^\Delta_{m\times n} = \left\{Q \in \mathbb{R}^{m\times n}_+ \colon \sum_{i=1}^n q_{ji} = 1,\ \forall j\right\}.$$

The game

$$\Gamma_{n,m} = \left\{\mathcal{P}^\Delta_{n\times m} \times \mathcal{Q}^\Delta_{m\times n}, F\left[(P, Q), (P', Q')\right]\right\} \tag{8}$$

can be considered as the mixed-strategies version of the game $\mathcal{G}_{n,m}$. Formally, $\Gamma_{n,m}$ is equivalent to the sender–receiver game considered in Nowak et al. (1999) and Trapa and Nowak (2000).

3. Nash strategies and evolutionary stability

In the tradition of evolutionary game theory, a Nash equilibrium is interpreted as an equilibrium composition of the population. Following a usual convention, a strategy played in a symmetric Nash equilibrium—that is, a strategy that is a best response to itself—is called a *Nash strategy*. A strategy is called a *strict Nash strategy* if it is a *unique* best response to itself.

Let $B(P)$ and $B(Q)$ be the *set of best responses* to P and respectively Q in the asymmetric game,

$$B(P) = \left\{Q \in \mathcal{Q}^\Delta_{m\times n} \colon \operatorname{tr}(PQ) \geqslant \operatorname{tr}(PQ')\ \forall Q' \in \mathcal{Q}^\Delta_{m\times n}\right\}, \quad \text{and}$$

$$B(Q) = \left\{P \in \mathcal{P}^\Delta_{n\times m} \colon \operatorname{tr}(PQ) \geqslant \operatorname{tr}(P'Q)\ \forall P' \in \mathcal{P}^\Delta_{n\times m}\right\}.$$

Note that both $B(P)$ as well as $B(Q)$ are always non-empty.

Lemma 1. *A pair $(P,Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ is*

(a) *a* Nash strategy *of the game $\Gamma_{n,m}$ if and only if $P \in B(Q)$ and $Q \in B(P)$; and it is*
(b) *a* strict Nash strategy *of $\Gamma_{n,m}$ if and only if P is the unique element in $B(Q)$ and Q is the unique element in $B(P)$.*

Proof. (a) Suppose that $P \notin B(Q)$. Then there is some $P' \in \mathcal{P}^{\Delta}_{n\times m}$, $P' \neq P$, such that $\operatorname{tr}(P'Q) > \operatorname{tr}(PQ)$. Consider (P', Q) as an alternative strategy. Then $\operatorname{tr}(P'Q) + \operatorname{tr}(PQ) > \operatorname{tr}(PQ) + \operatorname{tr}(PQ)$. But this cannot be true if (P,Q) is a Nash strategy. For (b), suppose that $P \in B(Q)$ and $Q \in B(P)$, but that there is some $P' \in B(Q)$ with $P' \neq P$. Then $\operatorname{tr}(P'Q) + \operatorname{tr}(PQ) = \operatorname{tr}(PQ) + \operatorname{tr}(PQ)$, but this cannot be true if (P,Q) is a *strict* Nash strategy. □

Supplementing Lemma 1 with a characterization of best responses in the asymmetric game allows us to characterize the Nash strategies of the symmetrized game $\Gamma_{n,m}$.

Lemma 2 *(Best-response properties). Let $\bar{P} \in \mathcal{P}^{\Delta}_{n\times m}$ and $\bar{Q} \in \mathcal{Q}^{\Delta}_{m\times n}$. If $Q \in B(\bar{P})$, then*

$$\sum_{i \in \operatorname{argmax}_i(\bar{p}_{ij^\star})} q_{j^\star i} = 1 \quad \textit{and} \quad q_{j^\star i} = 0 \quad \forall i \notin \operatorname{argmax}_i(\bar{p}_{ij^\star});$$

if $P \in B(\bar{Q})$, then

$$\sum_{j \in \operatorname{argmax}_j(\bar{q}_{ji^\star})} p_{i^\star j} = 1 \quad \textit{and} \quad p_{i^\star j} = 0 \quad \forall j \notin \operatorname{argmax}_j(\bar{q}_{ji^\star}).$$

Remark 1. For fixed $\bar{P}$,

$$\max_Q\bigl(\operatorname{tr}(\bar{P}Q)\bigr) = \max_{q_{ji}}\Biggl(\sum_j \sum_i \bar{p}_{ij} q_{ji}\Biggr) = \sum_j \max_i(\bar{p}_{ij});$$

analogously for the roles of P and Q reversed.

Proof. The proof is given for fixed $\bar{P}$; perfectly analogous reasoning holds true for fixed $\bar{Q}$. As the elements in Q are row-wise bounded to add up to 1, it is convenient to think of the operator $\operatorname{tr}(\bar{P}Q)$ as multiplying the jth *column* in $\bar{P}$ with the jth *row* of Q, and then summing over all j. Finding a Q that maximizes $\operatorname{tr}(\bar{P}Q)$ then amounts to choosing optimal weights q_{ji} for their corresponding elements $\bar{p}_{ij}$ such that $\sum_i \bar{p}_{ij} q_{ji}$ is maximal for every j (in the sequel I will often refer to p_{ij} as the "corresponding element" of q_{ji}, and vice versa.) Fix, for example, the $j^\star$th column of $\bar{P}$. Suppose first that it contains a unique maximal element, say $\bar{p}_{i^\star j^\star}$. Then in order to maximize $\sum_i \bar{p}_{ij^\star} q_{j^\star i}$ it is the unique optimal solution to put "full weight" to $\bar{p}_{i^\star j^\star}$—that is, to set $q_{j^\star i^\star}$ equal to 1. If, on the other hand, the $j^\star$th column of $\bar{P}$ contains more than one maximal element, then all the vertex solutions, where full weight is put to any of the maximal elements in the $j^\star$th column of $\bar{P}$, as well as any of their convex combinations fulfill the task of maximizing $\sum_i \bar{p}_{ij^\star} q_{j^\star i}$. For any $p_{ij^\star}$ that is not a maximal element in the $j^\star$th column of $\bar{P}$, the corresponding element in Q, $q_{j^\star i}$, has to be equal to zero—which concludes the proof of Lemma 2. But no matter *how* the total mass of 1 is distributed to the elements of the $j^\star$th row of Q, there is no way of doing better than to "extract" from the $j^\star$th column of $\bar{P}$ the value of its maximum. Summing over all j gives the claim of the remark. □

Example 1. Let

$$P_1 = \begin{pmatrix} 1-x & x & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad Q_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1-y & y \end{pmatrix}, \quad x, y \in (0,1).$$

Then the set of best responses to P_1 and Q_1 is given by

$$B(P_1) = \left\{ \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1-y' & y' \end{pmatrix} :\ y' \in [0,1] \right\}$$

and respectively

$$B(Q_1) = \left\{ \begin{pmatrix} 1-x' & x' & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} :\ x' \in [0,1] \right\}.$$

Note that $B(P_1)$ as well as $B(Q_1)$ include the two vertex solutions where y' and respectively x' are either 0 or 1. For any $Q \in B(P_1)$, $\operatorname{tr}(P_1 Q) = 2$, the sum of the column maxima of P_1. And for any $P \in B(Q_1)$, $\operatorname{tr}(P Q_1) = 2$, the sum of the column maxima in Q_1. We see that P_1 is a best response to Q_1, and that Q_1 is a best response to P_1. The pair (P_1, Q_1) therefore is a Nash strategy of the game $\Gamma_{3,3}$. Indeed, it can be checked easily that in this example any pair $[B(Q_1), B(P_1)]$ constitutes a Nash strategy of the game $\Gamma_{3,3}$, including the cases where both the P matrix and Q matrix contain a zero-column, which is shown in the following example.

Example 2.

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

These examples illustrates two important properties of Nash strategies:

- There can be two (or more) events that are linked to the same signal, or two (or more) signals that are associated with the same event.
- And there can be a zero-column in P or in Q, or in both P and Q.

In linguistics, the phenomenon that two or more objects of communication are linked to the same signal is called *homonymy*, whereas *synonymy* refers to a situation where the same object is linked to more than one signal.

Trapa and Nowak (2000) focus on a particular class of Nash strategies, which is defined by the condition that *neither P nor Q* contains any zero-column. The pair (P_1, Q_1) is of that type. Other examples are:

Example 3.

$$(P_3, Q_3) = \left[\begin{pmatrix} 0.3 & 0.7 & 0 \\ 0.3 & 0.7 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0.6 & 0.4 & 0 \\ 0.6 & 0.4 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right].$$

Example 4.

$$(P_4, Q_4) = \left[\begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \end{pmatrix}\right].$$

Lemma 3. *(See Trapa and Nowak, 2000.) Let $(P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ and assume that neither P nor Q contains any column that consists entirely of zeros. Then (P, Q) is a Nash strategy if and only if there exist real numbers $p_1, \ldots, p_n$ and $q_1, \ldots, q_m$ such that*

(a) *for each j, the jth column of P has its entries drawn from $\{0, p_j\}$, and $p_{ij} = p_j$ if and only if $q_{ji} = q_i$; and*
(b) *for each i, the ith column of Q has its entries drawn from $\{0, q_i\}$, and—as a matter of consistency—$q_{ji} = q_i$ if and only if $p_{ij} = p_j$.*

Proof. A possible way to see this is by crosswise exploitation of the best-response properties between P and Q. If p_{ij} is positive, then by the contrapositive of Lemma 2 this means that q_{ji} is a maximal element of its respective column. Since by assumption no column can consist of zero elements only, this means that q_{ji} necessarily has to be positive, but if q_{ji} is positive, this in turn implies that p_{ij} is a maximal element of the respective column in P. So the support of one matrix has to be the same as the support of the transpose of the other matrix, and every non-zero element is equal to the maximum of its respective column. Sufficiency follows from Lemma 2. □

Here, instead, I want to draw attention to the role of zero-columns. The difficulty with zero-columns is that they destroy the property that in a Nash strategy the support of one matrix has to coincide with the support of the transpose of the other matrix (see Example 2). Another consequence of the presence of zero-columns in a Nash strategy is that a non-zero element is *not* automatically a maximal element of its respective column in P or respectively Q.

Example 5.

$$(P_5, Q_5) = \left[\begin{pmatrix} 1-x & x & 0 \\ 1-x-\epsilon & x-\epsilon & 2\epsilon \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right], \quad x \in (0, 1).$$

As long as $\epsilon \geqslant 0$, (P_5, Q_5) is indeed a Nash strategy.

Note that it is generally true that whenever we start with a Nash strategy that has a zero-column, say in Q, then in order to preserve the Nash property, it is irrelevant how we assign the entries in the corresponding row of P *as long as we preserve the column maxima* of P. But even in the presence of zero-columns, the properties of best responses still imply conditions on the entries of P and Q that can be interpreted in the sense of some minimal consistency criteria.

Lemma 4 *(Minimal consistency). Let $P \in \mathcal{P}^{\Delta}_{n\times m}$ and $Q \in \mathcal{Q}^{\Delta}_{m\times n}$.*

(a) *If $Q \in B(P)$, then*:

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

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

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

Proof. If for some $Q \in B(P)$, $q_{j^\star i^\star} \neq 0$, then by the contrapositive of Lemma 2, $p_{i^\star j^\star} \in \text{argmax}_i(p_{ij^*})$. This maximum can be zero. But if it is zero, then, of course, all the elements in that column are zero. Analogous reasoning holds true for the roles of P and Q reversed. □

On the level of the population's average strategy profile, a zero-column in $\bar{P}$ means that there is a signal that remains perfectly idle across the whole population. A zero-column in $\bar{Q}$ means that there is an event that is never possibly inferred by any resident type, which we can interpret in the sense that nobody in this population has a concept of the relevant event.

Lemma 4 then reads as follows: If there is some resident type who infers event $i^\star$ from signal $j^\star$ ($\bar{q}_{j^\star i^\star} \neq 0$), then there is either at least some type who uses signal $j^\star$ to communicate event $i^\star$, and there cannot be more agents who use signal $j^\star$ in order to communicate another event; or, if this is not the case, then signal $j^\star$ remains idle throughout the whole population. Analogously, if there is some resident type who uses signal $j^\star$ in order to communicate event $i^\star$ ($\bar{p}_{i^\star j^\star} \neq 0$), then there is either some type who infers event $i^\star$ from signal $j^\star$, and there cannot be more agents who link signal $j^\star$ to another event; or, if this is not the case, then there is nobody in this population who has a concept of event $i^\star$.

An apparent feature of this model is that there is an abundance of Nash equilibria. The most central refinement concept in evolutionary game theory is that of an *evolutionarily stable strategy* (see Maynard Smith, 1982; Hofbauer and Sigmund, 1998; Cressman, 2003).

Definition 1. A strategy $(P, Q) \in \mathcal{P}^\Delta_{n\times m} \times \mathcal{Q}^\Delta_{m\times n}$ is *evolutionarily stable* if

(i) it is a Nash strategy, and if
(ii) whenever $F[(P,Q),(P,Q)] = F[(P',Q'),(P,Q)]$ for some $(P',Q') \in \mathcal{P}^\Delta_{n\times m} \times \mathcal{Q}^\Delta_{m\times n} \setminus \{(P,Q)\}$, then

$$F\big[(P,Q),(P',Q')\big] > F\big[(P',Q'),(P',Q')\big].$$

A strict Nash strategy necessarily is an evolutionarily stable strategy. For *symmetrized asymmetric games*, evolutionary stability also implies the strict Nash property, so that for this class of games, evolutionary stability and the strict Nash property indeed coincide (this follows from Selten, 1980).

From Lemma 2 it is easily seen that for a *strict* Nash strategy—that is, a pair (P, Q) such that P is the unique best-response to Q and vice versa—there has to be exactly one 1 in each column of P and respectively Q, such that $q_{ji} = 1$ whenever $p_{ij} = 1$. Note that strict Nash strategies do only exist in the case where $m = n$. Consequently, (P, Q) is an *evolutionarily stable strategy* of the game $\Gamma_{n,n}$ if and only if P is a permutation matrix and Q the transpose of P (Trapa and Nowak, 2000). For $m = n$ there are therefore exactly $n!$ evolutionarily stable strategies.

4. Neutrally stable strategies

Evolutionary stability is a rather strict concept. A game might have no evolutionarily stable strategy; for example in our case if $m \neq n$. Maynard Smith (1982) introduces a weaker notion of stability in an evolutionary context, where the strict inequality in Definition 1 is replaced by a weak inequality sign.

Definition 2. A strategy $(P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ is *neutrally stable* if

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

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

In the literature *neutral stability* is sometimes also referred to as *weak evolutionary stability* (for example, Thomas, 1985; for a clarification of terms, see Bomze and Weibull, 1995). Though formally a static concept, evolutionary stability implicitly carries the idea that a strategy can protect itself against the invasion of mutant strategies, in the sense that it can *drive out* other strategies. Neutral stability, instead, is more apt to express the condition that a strategy can protect itself from *being driven out* by other, potentially intruding, strategies.

Due to the symmetry of the payoff function, there is a more convenient way to state neutral stability.

Lemma 5. *A strategy* $(P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ *is a* neutrally stable strategy *if*

(i) *$P \in B(Q)$ and $Q \in B(P)$, and if in addition to that*
(ii) *whenever $P' \in B(Q)$ and $Q' \in B(P)$ this implies that*

$$\operatorname{tr}(PQ) \geqslant \operatorname{tr}(P'Q').$$

Proof. Point (i) is just the condition that the pair (P, Q) be a Nash strategy (Lemma 1). Point (ii): As $F[(P, Q), (P', Q')] = F[(P', Q'), (P, Q)]$, the inequality in the definition of neutral stability can be written as

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

By the supposition this condition this can be written as

$$F\big[(P, Q), (P, Q)\big] \geqslant F\big[(P', Q'), (P', Q')\big],$$

which means that

$$\operatorname{tr}(PQ) \geqslant \operatorname{tr}(P'Q'). \qquad \Box$$

For evolutionary stability, of course, an analogous statement holds true, only with the weak inequality sign replaced by a strict inequality sign.

Hence, if we want to check for neutral stability of a particular Nash strategy $(P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$, all we have to do is to compare its communicative potential $\operatorname{tr}(PQ)$ with the communicative potential of any alternative best reply, $\operatorname{tr}(P'Q')$, where $P' \in B(Q)$ and $Q' \in B(P)$.

In pure strategies, non-strict Nash strategies are always of the form such that each of the two matrices P and Q has at least one zero-column (Example 2 is such a case). Wärneryd (1993) shows that a zero-column in both the sender and the receiver matrix is in general sufficient to destroy neutral stability.

Lemma 6. *(See Wärneryd, 1993.) Let $(P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ be a Nash strategy. If each of the two matrices, P and Q, contains at least one column that consists entirely of zeros, then (P, Q) cannot be a neutrally stable strategy.*

Proof. Just note that if a pair (P, Q) is *not* neutrally stable with respect to all the pure strategies of a game, then it is also not neutrally stable with respect to all the mixed strategies. □

The intuition behind this is straightforward: An invading strategy that does not alter the existing links between events and signals, but that connects the previously idle signal to the event that before has never been potentially understood, is not doing worse against the resident type, but is doing strictly better against itself.

To see this in more detail for (P_2, Q_2) from Example 2, consider as an alternative best reply the pair (P', Q') with

$$P' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad Q' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{9}$$

Clearly, $P' \in B(Q_2)$ and $Q' \in B(P_2)$. But

$$\mathrm{tr}(P'Q') = 3 > 2 = \mathrm{tr}(P_2 Q_2),$$

and therefore (P_2, Q_2) cannot be a neutrally stable strategy.

However, *for the game in mixed strategies, there are neutrally stable strategies that are not evolutionarily stable.* An example of this is the pair (P_1, Q_1) in Example 1. Its communicative potential is $\mathrm{tr}(P_1, Q_1) = 2$. From $B(P_1)$ we can see that for every element $Q_1' \in B(P_1)$ the sum of its column maxima is 2. Therefore, with whatever sender matrix P_1' we multiply any particular $Q_1' \in B(P_1)$, the resulting communicative potential $\mathrm{tr}(P_1' Q_1')$ can never exceed 2. The same is true for the roles of P and Q reversed. There is consequently no alternative best reply (P_1', Q_1') whose communicative potential $\mathrm{tr}(P_1', Q_1')$ exceeds the communicative potential of the original Nash strategy $\mathrm{tr}(P_1 Q_1)$. Thus, the pair (P_1, Q_1) is a neutrally stable strategy. Note, however, that there are alternative best replies to (P_1, Q_1) that have exactly the same communicative potential as (P_1, Q_1)—for example, a pair (P_1', Q_1') that is of the same form as (P_1, Q_1), but with a different $x' \in (0, 1)$, $x' \neq x$, or a different $y' \in (0, 1)$, $y' \neq y$. This is exactly why (P_1, Q_1) is just a neutrally stable strategy, but not an evolutionarily stable strategy.

An obvious characteristic of (P_1, Q_1) as opposed to (P_2, Q_2) is that it contains no zero-column. However, *the absence of zero-columns does not guarantee neutral stability*.

As we have seen above, (P_3, Q_3) and (P_4, Q_4) are Nash strategies. They have no zero-column, but they also fail to be neutrally stable. (This can also be seen by taking the pair (P', Q') from Eq. 9 as an alternative best reply.) What destroys neutral stability in the case of (P_3, Q_3) and (P_4, Q_4) is the fact that P and Q contain columns that have *multiple maximal elements that are strictly between* 0 *and* 1. Other instances of Nash strategies that fail to be neutrally stable are sender–receiver pairs where one of the two matrices, P or Q, contains a column with multiple maximal elements strictly between 0 and 1, and the other matrix contains a zero-column.

Example 6.

$$(P_6, Q_6) = \left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1-\beta & \beta \\ 0 & 1-\beta & \beta \end{pmatrix} \right], \quad \beta \in (0, 1).$$

Interestingly, if one of the two matrices contains a column with multiple maximal elements that are strictly between 0 and 1, by the Nash property alone, the opposite matrix is already bound to contain also a column with multiple maximal elements strictly between 0 and 1, or a zero-column.

Lemma 7. *Let $(P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ be a Nash strategy. If P $[Q]$ contains at least one column that has multiple maximal elements strictly between* 0 *and* 1*, then*

(a) *Q $[P]$ contains at least two columns that have multiple maximal elements strictly between* 0 *and* 1*, or a zero-column*; *and*
(b) *(P, Q) cannot be a neutrally stable strategy.*

A proof of Lemma 7 is given in Appendix A. We then have as a necessary condition for a neutrally stable strategy that at least one of the two matrices P or Q has no zero-column, and that neither P nor Q contains a column with multiple maximal elements that are strictly between 0 and 1. These conditions are, of course, also necessary for evolutionary stability. But for neutral stability they also can be shown to be sufficient.

Lemma 8. *Let $(P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ be a Nash strategy. If P $[Q]$ has no zero-column and no column with multiple maximal elements that are strictly between* 0 *and* 1*, then*

(a) *Q $[P]$ has no column with multiple maximal elements that are strictly between* 0 *and* 1; *and*
(b) *(P, Q) is a neutrally stable strategy.*

A proof is given in Appendix A. In combination with Lemma 6 and Lemma 7, this yields a complete characterization of neutrally stable strategies.

Theorem 1. *Let $(P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ 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.

We can interpret this in the sense that *in a neutrally stable strategy there can be some but not too much ambiguity*. One signal can be linked to two or more events—homonymy, but if this is the case, then these events cannot be communicated by any other signal. One event can be linked to two or more signals—synonymy, but if this is the case, then these signals cannot be used to communicate any other event. In addition to that, there cannot be any idle signal as long as there is an event that is never potentially understood, and vice versa.

For the interpretation of $\Gamma_{n,m}$ as a population game, Theorem 1 has several important implications. Let $\mathcal{U}_\epsilon(x)$ be an open ϵ-neighborhood of x relative to the simplex S_L,

$$\mathcal{U}_\epsilon(x) = S_L \cap \{x' \in \mathbb{R}_L\colon\ |x' - x| < \epsilon\}, \quad \epsilon > 0. \tag{10}$$

Proposition 1.

(a) *If x is a state in the relative interior of S_L, then $(\bar{P}_x, \bar{Q}_x)$ cannot be a neutrally stable strategy.*
(b) *If for some $x \in S_L$, (P_x, Q_x) is a neutrally stable strategy, then there is no $0 < \epsilon < 1$ such that $\mathcal{U}_\epsilon(x)$ contains an $x' \neq x$ for which $(\bar{P}_{x'}, \bar{Q}_{x'})$ is an evolutionarily stable strategy.*

A proof is given in Appendix A. In words, part (b) of Proposition 1 means that evolutionarily stable strategies are always at opposite vertices relative to neutrally stable strategies.

As a additional remark it shall be pointed out here that for the case where $m = n$ neutrally stable strategies *do not* constitute *neutrally stable sets* (see Thomas, 1985; and for a more general treatment, see also Cressman, 2003). This is the case, since the sets of neutrally stable strategies of this model, which lie along particular boundary faces, are not closed at the vertices. Consider, for example, the pair (P_1, Q_1) from Example 1. As long as x *or* y is strictly between 0 and 1, the pair (P_1, Q_1) is a neutrally stable strategy. But if both, x *and* y, are either 0 or 1, (P_1, Q_1) is no longer a neutrally stable strategy—even though it is still a Nash strategy.

5. The replicator dynamics

The replicator dynamics (Taylor and Jonker, 1978) translates the idea that over time different types expand, or contract, according to the difference of their performance with respect to the average performance in the population.

If we assume that individual agents are matched randomly to play the game $\mathcal{G}_{n,m}$, the *average performance of type l* is given by

$$f_l(x) = \sum_{l'=1}^{L} x_{l'} F\big[(P_l, Q_l), (P_{l'}, Q_{l'})\big], \tag{11}$$

and the *average performance in the population* is given by

$$\bar{f}(x) = \sum_{l=1}^{L} x_l f_l(x). \tag{12}$$

In a biological context, $f_l(x)$ is usually interpreted as the *fitness of type l*, and $\bar{f}(x)$ as the *average fitness in the population*. For the model discussed here, the fitness of type l can be written as its payoff from play against the population's average strategy,

$$f_l(x) = F\big[(P_l, Q_l), \big(\bar{P}_x, \bar{Q}_x\big)\big] = \frac{1}{2}\operatorname{tr}\big(P_l \bar{Q}_x\big) + \frac{1}{2}\operatorname{tr}\big(\bar{P}_x Q_l\big), \tag{13}$$

and the average fitness in the population can be written as the payoff of the population's average from play against itself,

$$\bar{f}(x) = F\big[\big(\bar{P}_x, \bar{Q}_x\big), \big(\bar{P}_x, \bar{Q}_x\big)\big] = \operatorname{tr}\big(\bar{P}_x \bar{Q}_x\big). \tag{14}$$

The replicator dynamics then can be written as

$$\frac{\dot{x}_l}{x_l} = f_l(x) - \bar{f}(x), \quad l = 1, \ldots, L, \tag{15}$$

where $\dot{x}_l$ denotes the derivative of x_l with respect to time. A state of this system is a vector of L type frequencies, $x \in S_L$. A state $x^\star \in S_L$ for which $\dot{x}_l^\star = 0$ for all l is called a *rest point* of the dynamics. Of course, if $x^\star$ is a rest point, then the performance of every pure strategy that occurs with some positive frequency has to be equal to the average performance in the population. Every population state that corresponds to a Nash strategy, therefore, is a rest point of the replicator dynamics. A rest point $x^\star \in S_L$ is said to be *locally asymptotically stable* if after a small perturbation in the state variables, which stays within sufficiently small boundaries around the rest point, the dynamics eventually will lead back to this point; it is said to be *Lyapunov stable* if every neighborhood $\mathcal{U}_\epsilon(x^\star)$ contains a neighborhood $\mathcal{U}_{\epsilon'}(x^\star)$ such that for all initial states in $\mathcal{U}_{\epsilon'}(x^\star)$ the dynamics does not leave $\mathcal{U}_\epsilon(x^\star)$ as time proceeds.

The replicator dynamics can be interpreted in terms of both biological as well as cultural transmission of strategies from one generation to the next. Agents who communicate more successfully, it can be argued, are more successful in getting food, escaping dangers, finding mates, etc. Therefore, they have a direct or indirect advantage in reproduction. Assuming that parents directly transmit their P's and Q's to their kids, the P's and Q's of agents who communicate more successfully will therefore reproduce with a higher rate. Or, in terms of cultural evolution, agents who communicate more successfully are more likely to be imitated, and therefore their communicative strategies will spread with a higher rate.

Given its high dimensionality, it is generally not possible to solve explicitly for the replicator dynamics of this model. One way out of this is by computer simulations. Hurford (1989) and Nowak and Krakauer (1999) take that route. Here, instead, we want to see how far we can go with analytical methods.

In language evolution, we are not so much interested in any particular solution path, but in the *qualitative regularity patterns* of a linguistic trait that can be expected to arise in the long run, if any. A possible way to approach this question—and here we follow Wärneryd (1993)—is to exploit the dynamic implications that are dormant in the static refinement concepts of Nash equilibrium. For the replicator dynamics there are quite good results; in particular for games with a symmetric payoff function.

For games with a symmetric payoff matrix, the replicator dynamics constitutes a *gradient system*, which has several important implications for its qualitative behavior (see, for example, Hofbauer and Sigmund, 1998; or Cressman, 2003):

(a) The average fitness/payoff function is a strict Lyapunov function for the replicator dynamics;
(b) every orbit converges to some rest point; and
(c) the locally asymptotically stable rest points coincide with the evolutionarily stable strategies, and are given by the locally strict maxima of the average payoff function.

In combination with the static analysis of evolutionary stability, this directly tells us that for the replicator dynamics of the sender–receiver game at hand there are exactly $n!$ locally asymptotically stable rest points at those vertices of the simplex S_L for which $\bar{P}_x$ is a permutation matrix and $\bar{Q}_x$ is the transpose of $\bar{P}$. No other rest point can be asymptotically stable *as a point*. Yet this does not imply that the replicator dynamics will almost always converge to an evolutionarily stable strategy.

As already pointed out in Wärneryd (1993), *neutrally stable strategies* are *Lyapunov stable* in the replicator dynamics (see Thomas, 1985; and Bomze and Weibull, 1995). We have seen above that evolutionarily stable strategies are always at opposite vertices of the populations space relative to neutrally stable strategies that are not evolutionarily stable. This means that once the replicator dynamics has come sufficiently close to a neutrally stable strategy—which corresponds to a polymorphic population state—it will stay close to this state and it will not converge to an evolutionarily stable strategy. Indeed we can make this more precise by having a closer look at the local properties of neutrally stable strategies.

Lemma 9. *Let $x \in S_L$ be a composition of the population such that $(\bar{P}_x, \bar{Q}_x) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ is a neutrally stable strategy that is not evolutionarily stable. Then there exists some $\epsilon^{\star} > 0$ such that for all x' in $\mathcal{U}_{\epsilon^{\star}}(x)$ the following implication is true: If x' is a rest point of the replicator dynamics, then $(\bar{P}_{x'}, \bar{Q}_{x'})$ is also a neutrally stable strategy, but not an evolutionarily stable strategy.*

A proof is given in Appendix A. We then have the following proposition:

Theorem 2. *Let $x \in S_L$ be a composition of the population such that $(\bar{P}_x, \bar{Q}_x) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ is a neutrally stable strategy that is not evolutionarily stable. Then there exists some $\epsilon^{\star} > 0$ such that for all x' in $\mathcal{U}_{\epsilon^{\star}}(x)$ the replicator dynamics converges to a neutrally stable, but not evolutionarily stable strategy.*

Proof. Let $x \in S_L$ be a state that corresponds to a neutrally stable strategy that is not evolutionarily stable strategy. As x is Lyapunov stable, every neighborhood $\mathcal{U}_{\epsilon}(x)$ contains a neighborhood $\mathcal{U}_{\epsilon'}$ such that for all initial states in $\mathcal{U}_{\epsilon'}$ the replicator dynamics does not leave $\mathcal{U}_{\epsilon}(x)$ as time proceeds. It then suffices to consider $\mathcal{U}_{\epsilon^{\star}}(x)$, where $\epsilon = \epsilon^{\star}$ is chosen according to Lemma 9. As the dynamics does not leave $\mathcal{U}_{\epsilon^{\star}}(x)$ and all rest points within $\mathcal{U}_{\epsilon^{\star}}(x)$ are neutrally stable, but due to the gradient property, every orbit converges to some rest point, it necessarily converges to a neutrally stable strategy. □

Thus, whenever the composition of the population is sufficiently close to a properly neutrally stable strategy—that is, a neutrally stable strategy that is not evolutionarily stable—then the replicator dynamics will indeed converge to such a properly neutrally stable strategy.

Bomze (2002) shows that for games with a *symmetric payoff function* and pairwise interaction, neutral stability does indeed coincide with Lyapunov stability in the replicator dynamics. We can therefore exclude that in addition to neutrally stable strategies there are other rest points that are Lyapunov stable.

6. Interpretation and conclusions

A consequence of Theorem 2 is that the replicator dynamics *will not almost always* converge to an evolutionarily stable strategy, where every event is bijectively linked to one signal and vice versa. Instead, it can be blocked in a suboptimum state where ambiguities in event-to-signal or signal-to-event mappings persist in the population.

This phenomenon is also reflected in the computer simulations reported in Nowak and Krakauer (1999). For 5 events and 5 signals, the emerging signaling system has the property that two events share the use of one signal while another signal remains idle. The authors correctly conclude that such a strategy can be stable in an evolutionary context. Here we have seen that it is not evolutionary stability in its strict sense, but rather evolutionary stability in its weak form, *neutral stability*, that accounts for this type of limiting behavior.

Wärneryd (1993) cannot fully explore the consequences of neutral stability, since he does not take into account the mixed strategies of this game. If we consider pure strategies only, then neutral stability will indeed coincide with evolutionary stability. *But if we want to give a dynamic argument in terms of the replicator dynamics, we are bound to consider mixed strategies as well.*

Komarova and Niyogi (2004) discuss neutrally stable strategies—or in the terminology that they use, *weakly evolutionarily stable strategies*—of a sender–receiver game in the style of Nowak et al. (1999) and Trapa and Nowak (2000). However, they base their characterization of neutrally stable strategies on a characterization of Nash strategies that assumes that neither P nor Q contains any zero-column (Lemma 3; Trapa and Nowak, 2000). Apart from the fact that it is not clear how this restriction translates into a meaningful assumption on the strategies of individual types, this just means to assume away one of the peculiarities of this model—the fact that *there can be a zero-column in a neutrally stable strategy*, either in P or in Q, but not in both P and Q.

The present paper captures both the possibilities of mixed strategies and of zero-columns. We have seen that in a neutrally stable strategy, there can be two (or more) events that are linked to one and the same signal, or two (or more) signals that are associated with one and the same event, but that there *cannot* be two (or more) events that are linked to two (or more) signals in parallel, and that there *cannot* be a signal that remains idle in the presence of an event that is never potentially understood. As a consequence, *the replicator dynamics can, but does not almost always* lead to the rise of an optimally designed signaling system—a *conventional signaling system* in the sense of Lewis (1969). This is not a refutation of Lewis' point that a conventional signaling system will always prevail due to its *salient* properties. What it tells us is that replication alone—though it *can* lead to the rise of a conventional signaling system—is not strong enough to guarantee this outcome for almost all initial conditions. However, the results presented here crucially depend on some implicit assumptions. First and foremost that there is an infinitely large population; second assumptions that are implicitly present in the normal-form description of the base game.

The replicator dynamics as well as the notions of neutral and evolutionary stability presuppose that there is an *infinitely large population,* where the weight of any individual agent vanishes. In this framework, expected payoff directly translates into offspring and there are no effects of random drift.

The specific normal-form description of the base game defines a strategy as a deterministic program that tells individuals which signal to send if they observe a particular event, and what to do if they receive a particular signal, independently of any previous communicative success or failure. This seems to imply a truly evolutionary context. Taking the potential of communication over all possible events as the payoff of the game can be interpreted in the sense that every individual is exposed to all relevant events according to their objective frequencies. As being a sender or being a receiver is not a physical characteristic of any individual, but a social role, it then also makes sense to assume that any individual agent of the given linguistic community encounters all events according to their respective frequencies—once in the role of the sender, and once in the role of the receiver (for treatment of asymmetric contests see, for example, Taylor, 1979; or Schuster et al., 1981). Throughout, the symmetry of the payoff function (which is already present in the asymmetric game) presupposes cooperation or reciprocity.

Besides that there is a further simplifying assumption as reflected in the fact that all events that potentially become the object of communication enter the payoff function with the same weight. Weights of events can be thought of as representing event frequencies together with their relative importance. Formally, the introduction of different weights of events amounts to multiplying every *row in the sender matrix* with the respective weight. Assuming that all weights are indeed different, this destroys the possible sources of multiplicity of best responses in terms of the receiver matrix: Clearly, whenever in the position of the sender there are, say, two events that are linked to the same signal, but if one of these events is more important, then it will always be the unique best response for the receiver to associate this signal with the event that is relatively more important. As a consequence this eliminates the possibility of homonymy in a neutrally stable strategy. Analogously, introducing different "costs" of signals, eliminates the source of synonymy in a neutrally stable strategy. It is therefore more accurate to state our results in the form that evolution can lead to *homonymous use of a signal for objects that are equally important*, and to *synonymous meaning of signals that are of the same costs*.

Tightly related with the concept of an infinitely large population is the assumption that there is no local or any other structure of the population. Agents are matched randomly to interact with other agents. The importance of events and costs of signals could change with the environment. People migrating to different neighborhoods typically leads to so-called bottleneck or founder effects through random drift. In general, drift is an integral feature of evolution in small populations. There is good evidence that in the beginning of language, small populations size and drift played a crucial role (see, for example, Cavalli-Sforza, 1997).

To get a better understanding for the factors that drive language evolution, it will be interesting to look at extensions of the model with small or locally structured populations that allow us to take into account the effects of drift or changing conditions of the environment.

Acknowledgments

I would like to thank Immanuel Bomze and an anonymous referee for most valuable comments and support.

Appendix A. Proofs

Proof of Lemma 7. Let $(P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ be a Nash strategy, and suppose that the $i^\star$th column of Q contains more than one maximum element that is strictly between 0 and 1. As Q is a best response to P, and as $\max_j(q_{ji^\star}) \neq 0$, by Lemma 4, whenever $j^\star \in \operatorname{argmax}_j(q_{ji^\star})$, that is, whenever $q_{j^\star i^\star}$ is a maximal element of the $i^\star$th column of Q, then $p_{i^\star j^\star}$ is *a maximal element* of the $j^\star$th column in P. But since $q_{j^\star i^\star} \neq 1$, any such $p_{i^\star j^\star}$ *cannot be the unique maximal element* of the $j^\star$th column in P (by the contraposition of best-response properties.) Since P also has to be a best response to Q, the sum over all $p_{i^\star j}$ such that $j \in \operatorname{argmax}_j(q_{ji^\star})$ has to be equal to 1—which means in particular that at least some $p_{i^\star j}$ for which $j \in \operatorname{argmax}_j(q_{ji^\star})$ has to be positive. As by assumption, $\operatorname{argmax}_j(q_{ji^\star})$ has at least two elements, this implies that P has at least two columns with multiple maximal elements strictly between 0 and 1, or a zero-column, which proves the first part of the lemma.

For part (b) we have to show that there is some $P' \in B(Q)$ and some $Q' \in B(P)$, such that

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

Without loss of generality, assume that $j^{\star\star} \in \operatorname{argmax}_j(q_{ji^\star})$ such that $p_{i^\star j^{\star\star}} \neq 0$. As Q is a best response to P, it must be true that the sum over all $q_{j^{\star\star}i}$ such that $i \in \operatorname{argmax}_i(p_{ij^{\star\star}})$ equals 1. Of course, $i^\star \in \operatorname{argmax}_i(p_{ij^{\star\star}})$, but since by assumption, $q_{j^{\star\star}i^\star} \in (0, 1)$, there must be some be some other element $p_{i^{\star\star}j^{\star\star}}$ that is a *maximal element* of the $j^{\star\star}$th column of P, with $j^{\star\star} \neq j^\star$, such that $q_{j^{\star\star}i^{\star\star}} \in (0,1)$. Since $p_{i^{\star\star}j^{\star\star}} \neq 0$, and P is a best response to Q, by Lemma 4, $q_{j^{\star\star}i^{\star\star}}$ is a maximal element of the $i^{\star\star}$th column of Q. For later use note that

$$\sum_{j \in \operatorname{argmax}_j(q_{ji^*})} \sum_i p_{ij} q_{ji} = 1, \tag{16}$$

which comes from the fact that all elements in the $i^\star$th row of P for which $j \in \operatorname{argmax}_j(q_{ji^*})$ are maximal elements of their respective columns and that any Q that is a best response to P "extracts" from every column of P exactly the value of its maximum.

Now take as a candidate Q' the original Q but exchange the entries in its $j^{\star\star}$th row such that

$$q'_{j^{\star\star}i^{\star\star}} = 1, \quad \text{and} \quad q'_{j^{\star\star}i} = 0 \ \forall i \neq i^{\star\star},$$

and for some other $j^\star \in \operatorname{argmax}_j(q_{ji^*})$ $j^\star \neq j^{\star\star}$ exchange the entries in its $j^\star$th row such that

$$q'_{j^\star i^\star} = 1 \quad \text{and} \quad q'_{j^\star i} = 0 \ \forall i \neq i^\star.$$

As $p_{i^\star j^\star}$ is a maximal element of the $j^\star$th column of P, and $p_{i^{\star\star}j^{\star\star}}$ is a maximal element of the $j^{\star\star}$th column of P—which we know from above—it is easily checked that Q' is indeed a best response to P.

As an alternative P' take the original P but exchange the entries in its $i^{\star\star}$th and respectively $i^\star$th row such that

$$p'_{i^{\star\star}j^{\star\star}} = 1 \quad \text{and} \quad p'_{i^{\star\star}j} = 0 \ \forall j \neq j^{\star\star},$$
$$p'_{i^\star j^\star} = 1 \quad \text{and} \quad p'_{i^\star j} = 0 \ \forall j \neq j^\star.$$

As $q_{j^\star i^\star}$ is a maximal element of the $i^\star$th column of Q, and $q_{j^{\star\star}i^{\star\star}}$ is a maximal element of the $i^{\star\star}$th column of Q, P' is indeed a best response to Q.

What remains to be done, then, is to compare $\operatorname{tr}(P'Q')$ to $\operatorname{tr}(PQ)$. Since $p'_{i^\star j^\star} q'_{j^\star i^\star} = 1$ and $p'_{i^{\star\star}j^{\star\star}} q'_{j^{\star\star}i^{\star\star}} = 1$, we have that

$$\sum_{j \in \operatorname{argmax}_j(q_{ji^*})} \sum_i p'_{ij} q'_{ji} \geqslant 2 > 1 = \sum_{j \in \operatorname{argmax}_j(q_{ji^*})} \sum_i p_{ij} q_{ji}, \tag{17}$$

where the last inequality comes from Eq. (16). As $p_{i^{\star\star}j^{\star\star}} = p_{i^\star j^{\star\star}} \neq 0$, it is easily checked that

$$\sum_j \sum_i p'_{ij} q'_{ji} = \operatorname{tr}(P'Q') > \operatorname{tr}(PQ) = \sum_j \sum_i p_{ij} q_{ji},$$

which proves the claim of the lemma. □

Proof of Lemma 8. Let $(P,Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ be a Nash strategy, and suppose that P has no zero-column as well as no column with multiple maximal elements that are strictly between 0 and 1. Then for every fixed column of P there are only three possible cases: its maximum is either (i) unique and equal to 1; or (ii) unique, but not equal to 1; or (iii) equal to 1, but not unique. In order to show that (P,Q) is a neutrally stable state, by Lemma 5, we have to show that for any $P' \in B(Q)$ and any $Q' \in B(P)$,

$$\operatorname{tr}(P'Q') \leqslant \operatorname{tr}(PQ).$$

(i) Suppose that $p_{i^\star j^\star} = 1$ is the unique maximal element in the $j^\star$th column of P. Since Q is a best response to P, $q_{j^\star i^\star} = 1$. As Q has no elements that are greater than 1, this automatically implies that $q_{j^\star i^\star}$ is a maximal element of the $i^\star$th column of Q. As the elements in each row of

P add up to 1, $p_{i^\star j} = 0$ whenever $j \neq j^\star$. As there are no zero-columns in P, any $p_{i^\star j} = 0$ can never be a maximal element of its respective column in P. Therefore, as Q is a best response to P, $q_{ji^\star} = 0$ whenever $j \neq j^\star$, which means that $q_{j^\star i^\star} = 1$ is not only a but *the unique maximal element* of the $i^\star$th column of Q.

We now turn to $Q' \in B(P)$ and $P' \in B(Q)$. As $p_{i^\star j^\star} = 1$ is the unique maximal element in the $j^\star$th column of P,

$$q'_{j^\star i^\star} = 1 = q_{j^\star i^\star}, \quad \forall Q' \in B(P).$$

Note that this, of course, means that $q'_{j^\star i} = 0 = q_{j^\star i}$ for all $i \neq i^\star$, for all $Q' \in B(P)$. On the other hand, as $q_{j^\star i^\star} = 1$ is the unique maximal element in the $i^\star$th column of Q, we also have that

$$p'_{i^\star j^\star} = 1 = p_{i^\star j^\star}, \quad \forall P' \in B(Q).$$

Summing over all i, we have that

$$\sum_i p'_{ij^\star} q'_{j^\star i} = 1 = \sum_i p_{ij^\star} q_{j^\star i}, \tag{18}$$

for any $P' \in B(Q)$ and for any $Q' \in B(P)$. Figure 1(1) illustrates this case.

(ii) Synonymy. Suppose that $0 < p_{i^\star j^\star} < 1$ is the unique maximal element in the $j^\star$th column of P. As Q is a best response to P, we have that $q_{j^\star i^\star} = 1$, which means 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 $j \in \operatorname{argmax}_j(q_{ji^\star})$ 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 $j \in \operatorname{argmax}_j(q_{ji^\star})$ 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 $j \in \operatorname{argmax}_j(q_{ji^\star})$, 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 $j \in \operatorname{argmax}_j(q_{ji^\star})$, $0 < p_{i^\star j} < 1$ is indeed the *unique maximal element* of its respective column in P.

As the elements in each row of Q are bound to sum up to 1, for all $j \in \operatorname{argmax}_j(q_{ji^\star})$, $q_{ji} = 0$ whenever $i \neq i^\star$; and as for all $j \notin \operatorname{argmax}_j(q_{ji^\star})$, $p_{i^\star j} = 0$, but P has no zero-column, $q_{ji^\star} = 0$ whenever $j \notin \operatorname{argmax}_j(q_{ji^\star})$.

We now turn to $Q' \in B(P)$ and $P' \in B(Q)$. As for all $j \in \operatorname{argmax}_j(q_{ji^\star})$, $0 < p_{i^\star j} < 1$ is *the unique maximal element* of its respective column in P,

$$j \in \operatorname{argmax}_j(q_{ji^\star}) \Rightarrow q'_{ji^\star} = 1 = q_{ji^\star}, \quad \text{and}$$
$$q'_{ji} = 0 = q_{ji}, \ \forall i \neq i^\star; \quad \forall Q' \in B(P).$$

As the $i^\star$th column of Q has multiple maximal elements, we only have that

$$\sum_{j \in \operatorname{argmax}_j(q_{ji^\star})} p'_{i^\star j} = 1 = \sum_{j \in \operatorname{argmax}_j(q_{ji^\star})} p_{i^\star j}, \quad \forall P' \in B(Q).$$

However, summing over all i and over all $j \in \operatorname{argmax}_j(q_{ji^\star})$ we have that

$$\sum_{j \in \operatorname{argmax}_j(q_{ji^\star})} \sum_i p'_{ij} q'_{ji} = 1 = \sum_{j \in \operatorname{argmax}_j(q_{ji^\star})} \sum_i p_{ij} q_{ji}, \tag{19}$$

for any $P' \in B(Q)$ and for any $Q' \in B(P)$. Figure 1(2) illustrates this case.

$$\begin{tikzcd}
P = \begin{pmatrix} \cdot & \cdot & <1 \\ \cdot & \cdot & <1 \\ 0 & 0 & 1 \end{pmatrix} \arrow[r, leftrightarrow, "Q\in B(P)", "P\in B(P)"'] \arrow[dr, "{Q'\in B(P),\ P'\in B(Q)}" description] & Q = \begin{pmatrix} \cdot & \cdot & 0 \\ \cdot & \cdot & 0 \\ 0 & 0 & 1 \end{pmatrix} \arrow[dl] \\
P' = \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 0 & 0 & 1 \end{pmatrix} & Q' = \begin{pmatrix} \cdot & \cdot & 0 \\ \cdot & \cdot & 0 \\ 0 & 0 & 1 \end{pmatrix}
\end{tikzcd}$$

(1) One event exclusively linked to one signal.

$$\begin{tikzcd}
P = \begin{pmatrix} \cdot & <1-\alpha & <\alpha \\ \cdot & <1-\alpha & <\alpha \\ 0 & 1-\alpha & \alpha \end{pmatrix} \arrow[r, leftrightarrow, "Q\in B(P)", "P\in B(P)"'] \arrow[dr, "{Q'\in B(P),\ P'\in B(Q)}" description] & Q = \begin{pmatrix} \cdot & \cdot & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \arrow[dl] \\
P' = \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 0 & 1-\alpha' & \alpha' \end{pmatrix} & Q' = \begin{pmatrix} \cdot & \cdot & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}
\end{tikzcd}$$

(2) Synonymy: $\alpha \in (0,1)$, $\alpha' \in [0,1]$.

$$\begin{tikzcd}
P = \begin{pmatrix} \cdot & \cdot & <1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \arrow[r, leftrightarrow, "Q\in B(P)", "P\in B(P)"'] \arrow[dr, "{Q'\in B(P),\ P'\in B(Q)}" description] & Q = \begin{pmatrix} \cdot & 0 & 0 \\ \cdot & 0 & 0 \\ 0 & 1-\beta & \beta \end{pmatrix} \arrow[dl] \\
P' = \begin{pmatrix} \cdot & \cdot & \cdot \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} & Q' = \begin{pmatrix} \cdot & 0 & 0 \\ \cdot & 0 & 0 \\ 0 & 1-\beta' & \beta' \end{pmatrix}
\end{tikzcd}$$

(3) Homonymy: $\beta \in [0,1]$, $\beta' \in [0,1]$.

Fig. 1. Proof of Lemma 8.

(iii) Homonymy. Suppose that the $j^\star$th column of P has multiple maximal elements that are equal to 1. As Q is a best response to P, the sum over all $q_{j^\star i}$ such that $i \in \operatorname{argmax}_i(p_{ij^\star})$ has to be equal to 1, and $q_{j^\star i} = 0$ whenever $i \notin \operatorname{argmax}_i(p_{ij^\star})$.

As the elements in each row add up to 1, for all $i \in \operatorname{argmax}_i(p_{ij^\star})$, if $j \neq j^\star$, then $p_{ij} = 0$. Since by assumption P does not contain any zero-column, a zero-element in P can never be a maximal element of its respective column in P. Therefore, as Q is a best response to P, for all $i \in \operatorname{argmax}_i(p_{ij^\star})$, if $j \neq j^\star$, then $q_{ji} = 0$, which implies that for all $i \in \operatorname{argmax}_i(p_{ij^\star})$, $q_{j^\star i}$ is a maximal element of its respective column in Q. As Q may contain zero-columns, the respective element in Q is not necessarily positive, which means that it is not necessarily the unique maximal element of its respective column in Q. However, if for some $i^\star \in \operatorname{argmax}_i(p_{ij^\star})$, $q_{j^\star i^\star} \neq 0$, then it will be the *unique maximal element* of its respective column in Q.

We now turn to $Q' \in B(P)$ and $P' \in B(Q)$. As the $j^\star$th column of P has multiple maximal elements equal to 1, we only have that

$$\sum_{i \in \operatorname{argmax}_i(p_{ij^\star})} q'_{j^\star i} = 1 = \sum_{i \in \operatorname{argmax}_i(p_{ij^\star})} q_{j^\star i}, \quad \text{and}$$

$$q'_{j^\star i} = 0 = q_{j^\star i}, \ \forall i \notin \text{argmax}_i(p_{ij^\star}), \quad \forall Q' \in B(P).$$

On the other hand, whenever $q_{j^\star i^\star} \neq 0$ for some $i^\star \in \text{argmax}_i(p_{ij^\star})$, which, as we have just seen, means that it its a unique maximal element of its respective column in Q, then $p'_{i^\star j^\star} = 1$, for any $P' \in B(Q)$. Summing over all i we have that

$$\sum_i p'_{ij^\star} q'_{j^\star i} \leqslant \sum_i p_{ij^\star} q_{j^\star i} = 1, \tag{20}$$

for any $P' \in B(Q)$ and for any $Q' \in B(P)$. Figure 1(3) illustrates this case.

Taking all three cases together, we see that the maximum of any column in Q is either *unique* or *equal to* 1 *or* 0. Combining Eqs. (18), (19), and (20) and summing over all j, we have that

$$\sum_j \sum_i p'_{ij} q'_{ji} = \text{tr}(P'Q') \leqslant \text{tr}(PQ) = \sum_j \sum_i p_{ij} q_{ji},$$

for any $Q' \in B(P)$ and any $P' \in B(Q)$, which shows that (P, Q) is a neutrally stable strategy. □

Proof of Proposition 1. (a) If *every* pure strategy is present somewhere in the population, then *all elements* in $\bar{P}$ and in $\bar{Q}$ are necessarily positive, and trivially then neither $\bar{P}$ nor $\bar{Q}$ can have any zero-column. By Lemma 3 (Trapa and Nowak, 2000) this implies that there exist real numbers $0 < p_j < 1$, $j = 1, \ldots, m$ and $0 < q_i < 1$, $i = 1, \ldots, n$ such that $p_{ij} = p_j \forall i = 1, \ldots, n$ and $q_{ji} = q_i \forall j = 1, \ldots, m$. From Theorem 1 we directly see that a Nash strategy of that form cannot be neutrally stable.

The claim of (b) is immediate for a neutrally stable strategy that is also evolutionarily stable, as those strategies can only be at vertices of the simplex. Suppose that $(\bar{P}_x, \bar{Q}_x)$ is a neutrally stable strategy that is not evolutionarily stable. From Theorem 1 we know that at least $\bar{P}_x$ or $\bar{Q}_x$ has no zero-column. Without loss of generality, let $\bar{P}_x$ be that matrix. As $(\bar{P}_x, \bar{Q}_x)$ is not evolutionarily stable, $\bar{P}_x$ has at least one column with a unique maximal element that is strictly between 0 and 1, or a column with multiple maximal elements equal to 1—that is, a case of (ii) or respectively (iii) as treated in the proof of Lemma 8. A case of (ii), as can be seen from the proof of Lemma 8, implies that $\bar{Q}_x$ has a column with multiple maximal elements equal to 1. In a case of (iii), $\bar{P}_x$ itself has a column with multiple maximal elements equal to 1. If $\bar{p}_{i^\star j^\star}(x) = 1$, then this means that for all l such that $x_l \neq 0$, $p^l_{i^\star j^\star} = 1$. If there is a column in P_x with multiple maximal elements equal to 1, then this implies that *no type* that is present in x can be of the form where P is a permutation matrix; analogously for $\bar{Q}_x$. There is therefore no $\epsilon \in (0, 1)$ such that $\mathcal{U}_\epsilon(x)$ contains a state x' for which $(\bar{P}_{x'}, \bar{Q}_{x'})$ is a pair of permutation matrices and therefore it also cannot be an evolutionarily stable strategy. □

Proof of Lemma 9. Let $(\bar{P}_x, \bar{Q}_x) \in \mathcal{P}^\Delta \times \mathcal{Q}^\Delta$ be a neutrally stable strategy that is not evolutionarily stable, and assume, without loss of generality, that $\bar{P}_x$ is the matrix that has no zero-column. If $x' \in \mathcal{U}_\epsilon(x)$, then $|\bar{p}_{ij}(x') - \bar{p}_{ij}(x)| < \epsilon$, and an analogous expression holds true for any generic entry q_{ji} in the corresponding receiver matrices $\bar{Q}_x$ and $\bar{Q}_{x'}$. Let $\max_{i \setminus \max}(\bar{p}_{ij})$ be the value of the second highest entry of the jth column in $\bar{P}_x$, and define an analogous expression for $\bar{Q}_x$. To prove the claim of the lemma if suffices to choose $\epsilon^\star$ such that

$$\epsilon^\star \leqslant \big(\max_i(\bar{p}_{ij}) - \epsilon^\star\big) - \big(\max_{i \setminus \max}(\bar{p}_{ij}) + \epsilon^\star\big)$$
$$\Leftrightarrow \epsilon^\star \leqslant \frac{\max_i \bar{(p_{ij})} - \max_{i \setminus \max}(\bar{p}_{ij})}{3}$$

holds true for every column j in $\bar{P}_x$, and such that an analogous condition holds true for $\bar{Q}_x$. That is,

$$\epsilon^\star = \min\left\{\min_j\left[\frac{\max_i(\bar{p}_{ij}) - \max_{i\backslash\max}(\bar{p}_{ij})}{3}\right], \min_i\left[\frac{\max_j(\bar{q}_{ji}) - \max_{j\backslash\max}(\bar{q}_{ji})}{3}\right]\right\}.$$

A state $x' \in S_L$ is a rest point of the replicator dynamics if and only if all types present at x' get the same payoff from communication with the population's average sender–receiver pair $(\bar{P}_{x'}, \bar{Q}_{x'})$, that is,

$$\operatorname{tr}(P_l\bar{Q}_{x'}) + \operatorname{tr}(\bar{P}_{x'}Q_l) = \operatorname{tr}(\bar{P}_{x'}\bar{Q}_{x'}) + \operatorname{tr}(\bar{P}_{x'}\bar{Q}_{x'}) \quad \forall l \text{ s.t. } x'_l \neq 0.$$

The idea of the proof now is the following: Given the restrictions imposed by $\epsilon^\star$, all the types present at x' that have already been present at x, of which there is a fraction of at least $1-\epsilon^\star$, will continue to play a best response to the new population's average sender–receiver pair $(\bar{P}_{x'}, \bar{Q}_{x'})$. For x' to be a rest point, then, all the new types present at x' also have to play a best response to $(\bar{P}_{x'}, \bar{Q}_{x'})$. Starting with x and given $\epsilon^\star$ this implies a particular form of the P and Q matrices used by individual types. It is then just a matter of applying Theorem 1 to see that $(\bar{P}_{x'}, \bar{Q}_{x'})$ is indeed of the form of a neutrally stable strategy.

As in the proof of Lemma 8, since $\bar{P}_x$ has no zero-column, the maximal element of each single column in $\bar{P}_x$ is either (i) unique and equal to 1, (ii) unique, but not equal to 1, or (iii) not unique, but equal to 1.

(i) Let $\bar{p}_{i^\star j^\star}(x) = 1$ be the unique maximal element of the $j^\star$th column in $\bar{P}_x$. From the proof of Lemma 8 we know that in this case $\bar{q}_{j^\star i^\star}(x) = 1$ is the unique maximal element in the $i^\star$th column of $\bar{Q}_x$. Claim (i) now is that for any $(\bar{P}_{x'}, \bar{Q}_{x'})$ where $x' \in \mathcal{U}_{\epsilon^\star}(x)$ is to be a rest point of the replicator dynamics we also have that $\bar{p}_{i^\star j^\star}(x') = 1$ and that $\bar{q}_{j^\star i^\star}(x') = 1$.

By construction of $\epsilon^\star$, $\bar{p}_{i^\star j^\star}(x')$ is also the unique maximal element of the $j^\star$th column in $\bar{P}_{x'}$. As $\bar{q}_{j^\star i^\star}(x) = 1$, all types present at x' that have already been present at x—we will call them the "old types" and index them by "o"—will continue to exploit this maximum. Therefore,

$$\sum_i \bar{p}_{ij^\star}(x')q^o_{j^\star i} \geqslant \sum_i \bar{p}_{ij^\star}(x')\bar{q}_{j^\star i}(x').$$

The expression above holds with equality if and only if $\bar{q}_{j^\star i^\star}(x') = 1$. If this is the case, then

$$\sum_i p^o_{ij^\star}\bar{q}_{j^\star i}(x') \geqslant \sum_i \bar{p}_{ij^\star}(x')\bar{q}_{j^\star i}(x').$$

Suppose for the moment that for evaluating the payoff of types against the new population's average sender–receiver pair we can restrict attention to summation over all i of $j^\star$. We then can conclude that $\bar{p}_{i^\star j^\star} = 1$ as well; otherwise the old types would always extract a strictly higher payoff than all the other types present at x', which cannot be true if x' is to be a rest point. What remains to be done, then, for this case is to show that $\bar{q}_{j^\star i}(x')$ has to be indeed equal to 1.

If $\bar{q}_{j^\star i}(x') \neq 1$, then there is some type present at x' who does not extract from the $j^\star$th column of $\bar{P}_{x'}$ the value of its maximum. What this new type loses in this particular column of $\bar{P}_{x'}$ relative to the old types is *at least* $\epsilon^\star$. On the other hand, if a new type does not set $q_{j^\star i^\star}$ equal to 1, but sets another entry in its $j^\star$th row of Q equal to 1, this will also generate a positive entry in the respective position of $\bar{Q}_{x'}$ that has been 0 in $\bar{Q}_x$. If some of the new types puts a 1 to the corresponding element in its sender matrix this can lead to a situation where

$$\sum_i p^o_{ij^\star}\bar{q}_{j^\star i}(x') \leqslant \sum_i p^n_{ij^\star}\bar{q}_{j^\star i}(x'),$$

where n indicates this new type. However, by construction of $\epsilon^\star$ the difference between the left-hand side and the right-hand side of the above expression is always *strictly smaller* than $\epsilon^\star$. Therefore, if a new type deviates from $q_{j^\star i^\star} = 1$, all else being equal, this will always reduce his payoff relative to the old types.

(ii) If $\bar{p}_{i^\star j^\star}(x)$ is the unique maximal element of the $j^\star$th column of $\bar{P}_x$ that is strictly between 0 and 1, we have seen in the proof of Lemma 8 that in this case the $i^\star$th column of $\bar{Q}_x$ is a column with multiple maximal elements equal to 1, as well as that for any $j \in \operatorname{argmax}[\bar{q}_{ji^\star}(x)]$, $\bar{p}_{i^\star j}(x)$ is strictly between 0 and 1 and that it is the *unique maximal element* of its respective column in $\bar{P}_x$.

Claim (ii) now is that for any rest point $x' \in \mathcal{U}_{\epsilon^\star}(x)$, for all $j \in \operatorname{argmax}[\bar{q}_{ji^\star}(x)]$, $\bar{q}_{ji^\star}(x') = 1$, and that the sum over all $\bar{p}_{i^\star j}(x')$ such that $j \in \operatorname{argmax}[\bar{q}_{ji^\star}(x)]$ is equal to 1. The proof now is analogous to the previous case, only that we have to keep track of summation over all i and over all $j \in \operatorname{argmax}[\bar{q}_{ji^\star}(x)]$.

By construction of $\epsilon^\star$, the relevant elements in the $i^\star$th row of $\bar{P}_{x'}$ are also the *unique maximal elements* of their respective columns in $\bar{P}_{x'}$. All the old types will exploit these maxima. That is,

$$\sum_{j\in\operatorname{argmax}[\bar{q}_{ji^\star}(x)]}\sum_i \bar{p}_{ij^\star}(x')q^o_{j^\star i} \geqslant \sum_{j\in\operatorname{argmax}[\bar{q}_{ji^\star}(x)]}\sum_i \bar{p}_{ij^\star}(x')\bar{q}_{j^\star i}(x').$$

If the expression above holds with equality, that is, if $\bar{q}_{ji^\star} = 1$ for all $j \in \operatorname{argmax}[\bar{q}_{ji^\star}(x)]$, then 1 is the multiple maximum of the $i^\star$th column in $\bar{Q}_{x'}$. Since the sum over all $\bar{p}_{i^\star j}(x)$ such that $j \in \operatorname{argmax}[\bar{q}_{ji^\star}(x)]$ is equal 1,

$$\sum_{j\in\operatorname{argmax}[\bar{q}_{ji^\star}(x)]}\sum_i p^o_{ij^\star}\bar{q}_{j^\star i}(x') \geqslant \sum_{j\in\operatorname{argmax}[\bar{q}_{ji^\star}(x)]}\sum_i \bar{p}_{ij^\star}(x')\bar{q}_{j^\star i}(x').$$

If restrict attention to summation over all i and over all $j \in \operatorname{argmax}[\bar{q}_{ji^\star}(x)]$, we can conclude that the sum over all $\bar{p}_{i^\star j}(x')$ such that $j \in \operatorname{argmax}[\bar{q}_{ji^\star}(x)]$ has to be equal to 1 as well. Otherwise the old types would always extract a strictly higher communicative potential from the relevant sub-matrix in $\bar{Q}_{x'}$ than all the other types present at x'. What remains to be done is to show that for all $j \in \operatorname{argmax}[\bar{q}_{ji^\star}(x)]$, $\bar{q}_{ji^\star}(x')$ is indeed equal to 1.

If for some $j \in \operatorname{argmax}[\bar{q}_{ji^\star}(x)]$ $\bar{q}_{ji^\star}(x') \neq 1$, this means that there is some new type in x' who refrains from exploiting the maximum of the respective column in $\bar{P}_{x'}$. What this deviant type loses from this column in $\bar{P}_{x'}$ relative to the old types is *at least* $\epsilon^\star$. But, by the same argument as in case (i) above, the extra payoff that such a deviation can generate by putting some weight to an element in $\bar{Q}_{x'}$ that has been zero in $\bar{Q}_x$ is always *strictly smaller* than $\epsilon^\star$. Hence, any deviation from $\bar{q}_{ji^\star}(x') = 1$ for all $j \in \operatorname{argmax}[\bar{q}_{ji^\star}(x)]$, all else being equal, would always reduce the payoff of a new type relative to the old types.

(iii) For the case where the $j^\star$th column of P_x has multiple maximal elements equal to 1, we have seen in the proof of Lemma 8 that for any $i \in \operatorname{argmax}[\bar{p}_{ij^\star}(x)]$, if $\bar{q}_{j^\star i} \neq 0$ then $\bar{q}_{j^\star i}$ will be the unique maximal element in its respective column in Q_x.

By a similar line of reasoning as in the cases above one can show that in this case for any $i \in \operatorname{argmax}[\bar{p}_{ij^\star}(x)]$, *if* $\bar{q}_{j^\star i}(x) \neq 0$ *then* $\bar{p}_{ij^\star}(x') = 1$, and that the sum over all $\bar{q}_{j^\star i}(x')$ such that $i \in \operatorname{argmax}[\bar{p}_{ij^\star}(x')]$ has to be equal 1 as well. Otherwise, all else being equal, there would always be a new type that has a lower payoff than the old types.

Taking all 3 cases together, it can be seen that the old types can indeed never lose in terms of their payoff against the new population's average sender–receiver pair $(\bar{P}_{x'}, \bar{Q}_{x'})$ relative to the new types. Given the additive structure of the payoff function, we can therefore indeed treat the three cases separately, and than sum over all j. Then, for x' in $\mathcal{U}_{\epsilon^\star}(x)$ to be rest point, the patterns of the new population's average sender and receiver matrices as claimed in (i) to (iii) have to be true. If these conditions are satisfied, all types present at x' will indeed play a best response to the new population's average sender–receiver pair $\bar{P}_{x'}$, $\bar{Q}_{x'}$, the condition for a Nash strategy. From Theorem 1 it then follows that $(\bar{P}_{x'}, \bar{Q}_{x'})$ is indeed of the form of a neutrally stable strategy. □

References

Aumann, R., Heifetz, A., 2002. Incomplete information. In: Aumann, R., Hart, S. (Eds.), In: Handbook of Game Theory, vol. 3. Elsevier, Amsterdam/New York, pp. 1665–1686.

Bomze, I., 2002. Regularity vs. degeneracy in dynamics, games, and optimization: A unified approach to different aspects. SIAM Rev. 44, 394–414.

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

Cavalli-Sforza, L., 1997. Genes, peoples, and languages. Proc. Nat. Acad. Sci. USA 94, 7719–7724.

Cressman, R., 2003. Evolutionary Dynamics and Extensive Form Games. MIT Press, Cambridge, MA.

Hofbauer, J., Sigmund, K., 1998. Evolutionary Games and Population dynamics. Cambridge Univ. Press, Cambridge, UK.

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

Komarova, N., Niyogi, P., 2004. Optimizing the mutual intelligibility of linguistic agents in a shared world. J. Art. Intell. 154, 1–42.

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

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

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

Nowak, M., Plotkin, J., Krakauer, D., 1999. The evolutionary language game. J. Theoret. Biol. 200, 147–162.

Quine, W., 1936. Truth by convention. In: Lee, O. (Ed.), Philosophical Essays for A. N. Withhead. Longmans, New York.

Quine, W., 1960. Word and Object. MIT Press, Cambridge, MA.

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

Schuster, P., Sigmund, K., Hofbauer, J., Wolff, R., 1981. Self-regulation of behavior in animal societies II. Games between two populations without self-interaction. Biol. Cybern. 40, 9–15.

Taylor, P., 1979. Evolutionarily stable strategies for two types of players. J. Appl. Prob. 16, 76–83.

Taylor, P., 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., 2000. Nash equilibria for an evolutionary language game. J. Math. Biol. 41, 172–188.

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

White, M., 1956. Toward Reunion in Philosophy. Harvard Univ. Press, Cambridge, MA.