WORKING PAPERS

Christina PAWLOWITSCH

Why evolution does not always lead to an optimal proto-language.
An approach based on the replicator dynamics

June 2006

Working Paper No: 0604

DEPARTMENT OF ECONOMICS

UNIVERSITY OF VIENNA

All our working papers are available at: http://mailbox.univie.ac.at/papers.econ

Why evolution does not always lead to an optimal proto–language

Christina Pawlowitsch*

June 2006

Abstract

Sender–receiver models in the style of Lewis (1969), Hurford (1989), or Nowak and Krakauer (1999) can be used to explain *meaning* of signals in situations of cooperative interaction. This paper provides a complete characterization of neutrally stable strategies of this game purely in terms of properties of the lexical matrices that agents use for sending and receiving messages. It is show that in a neutrally stable strategy there can be instances of both homonymy and synonymy as long as the degree of ambiguity is not too high. There be two (or more) events that are linked to the same signal or two (or more) signals that are linked to the same event, but there cannot be two (or more) signals that are linked to two (or more) events in parallel, and there cannot be no signal that remains idle in the presence of an event that is never possibly inferred. This has considerable consequences for the regularity patterns of the signaling system that can be explained to arise from a replicator dynamics in a population of individual agents. Building on a result by Bomze (2002) it can be shown that such an evolutionary dynamics does not necessarily lead to an optimal signaling system, but that it can be trapped in suboptimal situations, where due to ambiguous event–signal relations some of the potential of communication is left unexploited.

Keywords: Origin of language – sender–receiver game – suboptimality – neutral stability – replicator dynamics – Lyapunov stability

*Department of Economics, University of Vienna, Hohenstaufengasse 9, 1010 Vienna, Austria, Phone: +43-4277-37410, Fax: +43-4277-37495, E–mail: christina.pawlowitsch@univie.ac.at

1 Introduction

The adoption of game theoretic methods for analyzing phenomena of natural language dates back relatively early into the history of game theory.

In his 1969 book "Convention: A philosophical study" David K. Lewis put forward the concept of Nash equilibrium to explain the conventional character of natural languages. By that time, the classical position that languages are social conventions had been challenged by the critiques of W.V. Quine and others, who argued that natural languages cannot be like the well–understood cases of central conventions, since we could not possibly have agreed on them without the use of any, even rudimentary, communication device. Against this, Lewis, himself a student of Quine, who had gotten into contact with the theory of coordination games by Thomas C. Schelling, defended the view that languages are conventions—however not in the sense of centrally organized institutions, but in the sense of so–called *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 exactly what is known as *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.[1] Once established, therefore, such equilibria will persist without the need of any centrally coordinating authority.

To make this point of view more precise, Lewis (1969) introduces a simple coordination game that can be used to explain the conventional character of simple signaling systems. In this game, there are two types of players; a player who is informed about the state of the world, and an uninformed player who, depending on the state of the world, is called to take an action that is payoff relevant for both players in a perfectly coinciding way, so that there is no element of conflict between the two parties involved. This game is embedded in a round of pre–play communication, where the strategy set of the informed player is a set of mappings from states of the world to signals, and the strategy set of the uninformed player is a set of mappings form signals to actions. It is assumed that for each state of the world there is exactly one action that has to be taken—taking this action, therefore, can be interpreted as "understanding" the state of the world and actions can be identified with the states of the world by which they are called for. Players can be referred to as senders and receivers according to their respective roles in latent communication, leading to a *pure coordination game*, where senders and receivers have to coordinate on meaning that is attributed to signals.

Clearly, a combination of strategy choices such that the sender's strategy is a bijective mapping form states of the world to signals and the receiver's strategy the inverse of this mapping is a Nash equilibrium of this game. Not only this, it is even a Nash equilibrium that achieves the maximal available payoff. Lewis

[1] Interestingly, this led Lewis to one of the first formulations of *common knowledge*. In their article on incomplete information in the Handbook of Game Theory Aumann and Heifetz (2001) write that "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."

calls such equilibria *conventional signaling system*. *Meaning* in this context can be understood as a property of these equilibria. However, without restricting strategy sets to one–to–one mappings, this game also admits Nash equilibria that do not attain the maximal available payoff. In such situations there are two (or more) events mapped to the same signal, or two (or more) signals mapped to the same action.

Besides its important conceptual innovations, Lewis' contribution therefore still leaves us with two problems. First, the selection of what he calls *conventional signals systems* remains confined to more purely ad hoc considerations. Second, his central argument that languages are conventions in the sense of self–enforcing agreements does not address the corner stone of the critique of Quine and others—it does not explain how these conventions had come into being in the first place. Of course, this is a reflection of the static nature of Nash equilibrium itself. Explaining how conventions of language can come into being in the first place necessarily calls for an evolutionary approach.

1.1 Evolutionary stability

Wärneryd (1993) takes up Lewis' model exactly with this perspective. He shows that conventional signaling systems in the sense of Lewis are the only *evolutionarily stable strategies* and—abstracting from the possibility of mixed strategies—also the only *neutrally stable strategies* of this game. Taylor and Jonker (1978) show that evolutionary stability implies asymptotic stability in the replicator dynamics, whereas Thomas (1985), even more fundamentally, shows that neutral stability implies Lyapunov stability in the replicator dynamics. Wärneryd suggests this as an argument for the rise of a conventional signaling system in the sense of Lewis by a trial–and–error process.

In the presence of mixed strategies, however, it is no longer true that conventional signaling systems are the only neutrally stable strategies. The problem with this, though, is that in a dynamic, population based setting, mixed strategies necessarily arise as the non–monomorphic states of this system, where the population is composed of different types of otherwise identical agents using different pure strategies. A complete treatment of the evolutionary aspects of this model—and this is the first motivation for the present paper—therefore, has to be complemented by an analysis of neutral stability that also takes into account the possibility of mixed strategies.

1.2 Origins of a proto–language

Equivalent versions of this model have been introduced independently by James Hurford in his 1989 article "On the biological evolution of the Saussurean Sign" as well as by Martin A. Nowak and David Krakauer in "The evolution of language" (Hurford, 1989; Nowak and Krakauer, 1999). Both of these articles—having originated in linguistics on the one hand and mathematical biology on the other hand—can be said to have been seminal contributions of the then newly emerging literature on language evolution that uses formal evolutionary

arguments and which recently seems to evolve into a unified interdisciplinary research program[2]. While Hurford (1989) wants to give an evolutionary argument for a particular design feature of concept–signal relations on which human language essentially builds—he wants to explain the rise of the so–called Saussurean sign, that is, the property that these relations are symmetric—Nowak and Krakauer (1999) are more interested in explaining how a common set of concept–signal relations, which they interpret in the sense of a proto–language, can emerge from a previously completely non–linguistic population by a simple replication mechanism that gives positive feedback to successful communication.

In the Hurford–Nowak–Krakauer version of this model, every individual agent has a so–called sender matrix and a receiver matrix, where the sender matrix gives the probabilities with which this individual will produce the various available signals in case a particular event is observed, and the receiver matrix gives the probabilities with which this individual will associate incoming signals with the various events. From this, for a given distribution of event frequencies, one can calculate the rate of successful communication between any pair of a sender and receiver matrix. Payoffs in this framework then are directly taken to be in terms of successful communication. However, what lurks behind this game, clearly is a situation where an informed player can send a signal to an uninformed player who then has to take a payoff relevant action. Formally, the game in its Hurford–Nowak set–up can be seen as a mixed strategy version of the Lewis–Wärneryd model.

Both Hurford (1989) as well as Nowak and Krakauer (1999) approach their respective questions with computer simulations. Whereas Hurford (1989) focuses on comparing the communicative performance of several behavioral rules that differ with respect to the individual's adjustment between sender and receiver matrices, the simulation reported in Nowak and Krakauer (1999) is more in the style of a pure replicator dynamics where individual agents are simple automatons.

Starting from randomly drawn sender and receiver matrices, in every period, every individual agent once communicates with every other individual agent in each of the two roles. Payoffs are calculated, and in the following period every individual strategy, that is, *a pair of a sender and receiver matrix*, replicates itself according to its payoff relative to the population's average payoff. They find that after a certain number of rounds, indeed, specific signals start to correspond with specific events, and finally there seems to be convergence to what can be called a common proto–language.

1.3 Can homonymy be evolutionarily stable?

Interestingly, this common proto–language has the property that some signals remain idle while there are events that share the use of one and the same signal. In linguistics the phenomenon that one signal is used for more then one referent is known as *homonymy*, whereas *synonymy* refers to a situation where one event

[2]See Christiansen and Kirby (2003)

is linked to more than one signal. As a consequence, in such a situation, some of the potential of communication that is in principle available is left unexploited. From this Nowak and Krakauer (1999) conjecture that

> "evolution does not always lead to the optimum solution, but certain suboptimum solutions, in which the same signal is used for two (or more) objects, can be evolutionarily stable."

This, however, needs explanation in view of the formal result shown in Wärneryd (1993) that only those sender–receiver structures that induce a bijective mapping from events to signals can be evolutionarily stable, and which later on independently has been shown by Trapa and Nowak (2000) for the slightly different framework with sender and receiver matrices, in which it reads that only those sender–receiver pairs where the sender matrix is a permutation matrix and the receiver matrix is the transpose of the sender matrix (and, of course, vice versa) can be evolutionarily stable. As a consequence of this result, of course, a proto–language where two (or more) events are linked to the same signal or where two (or more) signals are linked to the same event, cannot be evolutionarily stable—at least not in its strict sense.

The interesting question arising from this—and this is the second motivation for the present paper—is whether we can account for the results of simulation reported in Nowak and Krakauer (1999) from an analytical point of view. That is, I want to ask whether there are, and, if so, what are the properties of Nash strategies that cannot be evolutionary stable (not in the strict sense), but that still can protect themselves from being driven out by the replicator dynamics of this model. As it turns out, neutral stability and its dynamic consequences provide the key to explaining this.

1.4 The plan of the work

The organization of the paper is as follows: Section 2 introduces the model in the conceptual framework of Hurford (1989) and Nowak and Krakauer (1999) and shows how to derive it in this form as the symmetrized, mixed–strategies version of the game in the Lewis–Wärneryd framework. Section 3 elaborates on Nash strategies and the best–response properties in terms of sender and receiver matrices. Section 4 introduces evolutionary stability and neutral stability formally. Section 5 uses the best response properties of sender and receiver matrices to derive a complete characterization of neutrally stable strategies that encompasses the possibility of mixed strategies. Section 6 reformulates the model under study as a population game, where the population is composed of different types of otherwise identical players, each of which links every event to exactly one signal with probability 1, and associates every received signal with exactly one event with probability 1. In this framework, the population's average strategy is a pair of two stochastic matrices, as in the Hurford–Nowak–Krakauer version of this model. In particular, this section discusses the properties of Nash strategies and evolutionarily stable strategies with respect to the cognitive abilities of individual agents in the context of population games. Section 7 is concerned

with the qualitative analysis of the replicator dynamics of this model. This essentially makes use of a result by Bomze (2002), who establishes equivalence of neutral stability and Lyapunov stability in the replicator dynamics for doubly symmetric games with pairwise interaction. Section 8, finally, discusses the results and sheds some light on extensions of the model and future work.

2 The model

In the basic framework as introduced by Hurford (1989) and Nowak and Krakauer (1999) there is a large number of individual agents among whom communication potentially takes place. There are n events that possibly become the object of communication, and there are m signals that can be used to communicate these events. The $n \times m$–matrix

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{i1} & \cdots & p_{ij} & \cdots & p_{im} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nj} & \cdots & p_{nm} \end{pmatrix},$$

denotes an individual's *sender matrix*, where p_{ij} indicates the probability with which signal j will be transmitted if event i is to be communicated, so that

$$P \in \mathcal{P}^{\Delta}_{n\times m} = \{P \in R^{+}_{n\times m} : \sum_{j=1}^{m} p_{ij} = 1, \forall i\}.$$

On the other hand, the $m \times n$–matrix

$$Q = \begin{pmatrix} q_{11} & \cdots & q_{1i} & \cdots & q_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ q_{j1} & \cdots & q_{ji} & \cdots & q_{jn} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ q_{m1} & \cdots & q_{mi} & \cdots & q_{mn} \end{pmatrix},$$

with

$$Q \in \mathcal{Q}^{\Delta}_{m\times n} = \{Q \in R^{+}_{m\times n} : \sum_{i=1}^{n} q_{ji} = 1, \forall i\},$$

denotes an individual's *receiver matrix*, where q_{ji} gives the probability with which event i will be inferred if signal j is received.

If an individual with receiver matrix P wants to communicate event i to an individual with receiver matrix Q, then the probability that he or she is successfully doing so, is

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

Assuming that all events that possibly become the object of communication occur with the same probability and are equally important, as well as that all possibly used signals are of the same costs, the sum over all these column–times–row products,

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

can be taken as a measure for the *communicative potential between an individual with sender matrix P and an individual with receiver matrix Q.*

2.1 The asymmetric game

This set–up can be rephrased as an *asymmetric game* with two types of players, senders and receivers, where senders choose a strategy $P \in \mathcal{P}^{\Delta}_{n\times m}$ and receivers choose a strategy $Q \in \mathcal{Q}^{\Delta}_{m\times n}$. Assuming that communication is mutually beneficial, the communicative potential $\mathrm{tr}(PQ)$ can be taken as the payoff that both parties get out of their interaction, such that $F_P(P,Q)$, the payoff function for senders, as well as $F_Q(P,Q)$, the payoff function for receivers, is given by

$$F_P(P,Q) = \mathrm{tr}(PQ) = F_Q(P,Q). \tag{2}$$

This form of this payoff function can be interpreted in the sense that the very act of communication is beneficial in itself, for example, as a means to establish kin relations that transcend the limits of physical grooming, or it can be interpreted as the reflection of some richer structure, where a player who is informed about the state of the world can send a signal to an uninformed player, who then has to take some payoff relevant action, the gains of which are shared between the two players. As an example of this we can think of two agents coordinating their actions in some common undertaking like the hunting of game or the fishing of fish. Alternatively, if we are willing to interpret the sharing of the payoff in the sense of some element of direct or reciprocal altruism, we can think of one agent informing the other of some event of nature in the face of which the previously uninformed agent has to take a particular action in order to protect his well being or that of fellow members of the same group.

2.2 A symmetrized, doubly symmetric, game

Nowak and Krakauer (1999), which later on has been analyzed in more detail in Trapa and Nowak (2000), consider a symmetrized version of this game[3], where a strategy is *a pair* of a sender and a receiver matrix,

$$(P,Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}, \tag{3}$$

and the *payoff function* is given by

$$F[(P,Q),(P',Q')] \;=\; \frac{1}{2}\mathrm{tr}(PQ') + \frac{1}{2}\mathrm{tr}(P'Q). \tag{4}$$

[3]a game is symmetric if all player have the same strategy sets and payoff functions

The choice of the weights in this payoff function can be interpreted in the sense that there is a homogenous group of individual agents all of whom find themselves in the roles of sender and receiver with equal probabilities.

Note that this payoff function is symmetric,

$$F[(P,Q),(P',Q')] \quad = \quad \frac{1}{2}\mathrm{tr}(P'Q) + \frac{1}{2}\mathrm{tr}(PQ') = F[(P',Q'),(P,Q)],$$

giving rise to a so-called *doubly symmetric game*, that is, a symmetric game with a symmetric payoff function [4].

In the sequel, I shall use

$$\mathcal{G}^{\Delta}_{m,n} = \left\{\mathcal{P}^{\Delta}_{n\times m}, \mathcal{Q}^{\Delta}_{n\times m}, F_P(P,Q) = \mathrm{tr}(PQ), F_Q(P,Q) = \mathrm{tr}(PQ)\right\} \tag{5}$$

to refer to the asymmetric sender–receiver game, and

$$\Gamma^{\Delta}_{m,n} = \left\{\mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{n\times m}, \frac{1}{2}\mathrm{tr}(PQ') + \frac{1}{2}\mathrm{tr}(P'Q)\right\} \tag{6}$$

to refer to the symmetrized, doubly symmetric, version of this game.

2.3 Pure strategies

Formally, if we restrict the strategy sets of senders and receivers to the set of all $n \times m$ and respectively $m \times n$ matrices that have exactly one 1 in every row and zero otherwise, that is, if

$$P \in \mathcal{P}_{n\times m} = \left\{P \in R^{+}_{n\times m} : p_{ij} = 1 \text{ for some } j'(i) \text{ and } p_{ij} = 0 \,\forall\, j \neq j'(i),\ \forall\, i\right\}, \tag{7}$$

and

$$Q \in \mathcal{Q}_{m\times n} = \left\{Q \in R^{+}_{m\times n} : q_{ji} = 1 \text{ for some } i'(j) \text{ and } q_{ji} = 0 \,\forall\, i \neq i'(j),\ \forall\, j\right\}, \tag{8}$$

we get the set of all sender and receiver matrices that link every event deterministically to exactly one signal and every signal to exactly one event. Note that this does not rule out there to be two or more events that are linked to the same signal, or two or more signals that are associated with the same event. There are m^n such deterministic sender matrices in $\mathcal{P}_{n\times m}$ and n^m such deterministic receiver matrices in $\mathcal{Q}_{m\times n}$.

If we can identify a state of the world with the action that it calls for, these deterministic sender and receiver matrices give us equivalent expressions for the strategy sets of informed and respectively uninformed players that Lewis (1969) and Wärneryd (1993) use for their model of a signaling system.

[4] In the literature symmetric games with a symmetric payoff function are sometimes also referred to as *partnership games*.

Due to the linear structure of the payoff function, we can "decompose" the sender–receiver games $\mathcal{G}^{\Delta}_{m,n}$ and respectively $\Gamma^{\Delta}_{m,n}$ into their equivalents in pure strategies, the asymmetric sender–receiver game

$$\mathcal{G}_{m,n} = \{\mathcal{P}_{n\times m}, \mathcal{Q}_{n\times m}, F_P(P,Q) = \mathrm{tr}(PQ), F_Q(P,Q) = \mathrm{tr}(PQ)\}\,, \tag{9}$$

and its symmetrized, doubly symmetric version,

$$\Gamma_{m,n} = \left\{\mathcal{P}_{n\times m} \times \mathcal{Q}_{n\times m}, \frac{1}{2}\mathrm{tr}(PQ') + \frac{1}{2}\mathrm{tr}(P'Q)\right\}, \tag{10}$$

respectively.

The game in its Hurford–Nowak–Krakauer form, therefore, can be seen as the symmetrized, mixed–strategy version of the Lewis–Wärneryd sender–receiver model.

Example 1. Let $m = 2 = n$. In this case the set of all sender and receiver matrices in the asymmetric sender–receiver game in pure strategies is given by

$$\mathcal{P}_{2\times 2} = \left\{P_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, P_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \right.$$
$$\left. P_3 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, P_4 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}\right\},$$

and

$$\mathcal{Q}_{2\times 2} = \left\{Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, Q_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \right.$$
$$\left. Q_3 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, Q_4 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}\right\},$$

respectively. By an appropriate choice of weights, possibly not unique, we can decompose every element P in $\mathcal{P}^{\Delta}_{2\times 2}$ into a convex combination of the 4 pure strategies contained in $\mathcal{P}_{2\times 2}$, and analogously for every Q in $\mathcal{Q}^{\Delta}_{2\times 2}$. Consider, for instance,

$$P = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} \in \mathcal{P}_{2\times 2},$$

which can be written as

$$\begin{aligned} \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} &= \frac{p}{2}\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} + \frac{p}{2}\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \\ &+ \frac{1-p}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{1-p}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \text{for any } p \in [0,1]. \end{aligned}$$

3 Nash strategies

Trapa and Nowak (2000) analyze Nash strategies, strict Nash strategies and evolutionarily stable strategies of the game $\Gamma^{\Delta}_{m,n}$, the sender–receiver game in its symmetrized, mixed–strategies form.

In the tradition of evolutionary game theory, a strategy played in a symmetric Nash equilibrium—that is, a strategy that is a best reply to itself—is called a *Nash strategy*.

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

(a) a Nash strategy *if*

$$F[(P^*, Q^*), (P^*, Q^*)] \geq F[(P, Q), (P^*, Q^*)] \quad \forall\, (P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{n\times m};$$

(b) it is called a strict Nash strategy *if*

$$F[(P^*, Q^*), (P^*, Q^*)] > F[(P, Q), (P^*, Q^*)] \quad \forall\, (P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{n\times m}, \; (P, Q) \neq (P^*, Q^*).$$

Let

$$\begin{aligned} B(P) &= \{Q \in \mathcal{Q}^{\Delta}_{m\times n} \,:\, \mathrm{tr}(PQ) \geq \mathrm{tr}(PQ')\, \forall\, Q' \in \mathcal{Q}^{\Delta}_{m\times n}\} \quad \text{and} \\ B(Q) &= \{P \in \mathcal{P}^{\Delta}_{n\times m} \,:\, \mathrm{tr}(PQ) \geq \mathrm{tr}(P'Q)\, \forall\, P' \in \mathcal{P}^{\Delta}_{n\times m}\} \end{aligned}$$

be the *set of best responses* to P and respectively Q in the asymmetric game. Of course, since for fixed $\bar{P}$ the continuous function $\mathrm{tr}(\bar{P}Q)$ attains a maximum on $\mathcal{Q}^{\Delta}_{n\times m}$, and since for fixed $\bar{Q}$, $\mathrm{tr}(P\bar{Q})$ attains a maximum on $\mathcal{P}^{\Delta}_{m\times n}$, both $B(\bar{P})$ as well as $B(\bar{Q})$ are always non–empty .

Following Selten (1980), as a general property of symmetrized games, for a Nash strategy, the strategy choices in the two roles have to be best responses to each other, and they have to be unique best responses in a strict Nash strategy. That is, $(P^*, Q^*) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{n\times m}$ is

(a) a Nash strategy of the game $\Gamma^{\Delta}_{m,n}$ if and only if $P^* \in B(Q^*)$ and $Q^* \in B(P^*)$; and

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

If we can complement this with a characterization of best–responses in terms of the P and Q matrices we will be enabled to characterize the Nash strategies of the game $\Gamma^{\Delta}_{m,n}$.

3.1 Best–responses in terms of P and Q matrices

Some extra notation helps exposition. Let

$$A(p_{\cdot j^*}) \;=\; \mathrm{argmax}_i(p_{ij^*}) \tag{11}$$

be the set of all row indices i in the j^*–the column of P such that p_{ij^*} is a maximal element of that column, and analogously let

$$A(q_{\cdot i^*}) \quad = \quad \text{argmax}_j(q_{ji^*}), \tag{12}$$

be the set of all row indices j in the i^*–th column of Q such that q_{ji^*} is a maximal element of that column.

Suppose now we are given some fixed $\bar{P}$ and we want to find all receiver matrices Q that maximize $\text{tr}(\bar{P}Q)$. Since the elements in Q are row–wise bounded to add up to 1, for fixed $\bar{P}$, it is convenient to understand the operator $\text{tr}(\bar{P}Q)$ as multiplying the j-th *column* in $\bar{P}$ with the j-th *row* of Q, and then summing over all j,

$$\begin{aligned}
\text{tr}(\bar{P}Q) &= \sum_i \sum_j \bar{p}_{ij} q_{ji} = \sum_j \sum_i \bar{p}_{ij} q_{ji} \\
&= \bar{p}_{11} q_{11} + \bar{p}_{21} q_{12} \cdots + \bar{p}_{n1} q_{1n} \\
&+ \bar{p}_{12} q_{21} + \bar{p}_{22} q_{22} \cdots + \bar{p}_{n2} q_{2n} \\
&\quad \vdots \qquad (13) \\
&+ \bar{p}_{1m} q_{m1} + \bar{p}_{2m} q_{m2} \cdots + \bar{p}_{nm} q_{mn}. \qquad (14)
\end{aligned}$$

Finding a Q that maximizes $\text{tr}(\bar{P}Q)$ then amounts to choosing optimal "weights" q_{ji} to their corresponding elements $\bar{p}_{ij}$ such that $\sum_i \bar{p}_{ij} q_{ji}$ is maximal for every j.[5]

Fix, for example, the $j^\star$–th column of $\bar{P}$ and suppose 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 clearly the optimal choice to put "full weight" to $\bar{p}_{i^\star j^\star}$—that is, to set $q_{j^\star i^\star}$ equal to 1, and all the other elements in the $j^\star$-th row of Q equal to zero. If, on the other hand, the $j^\star$-th column of $\bar{P}$ contains more than one maximal element, then there is more than one optimal appointment of the elements in the $j^\star$-th row of Q. All the corner 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 $\sum_i \bar{p}_{ij^\star} q_{j^\star i}$. But no matter *how* the total mass of 1 is attached to the elements in the $j^\star$-th column of $\bar{P}$, there is no way of doing better than to "extract" from the $j^\star$-th column of $\bar{P}$ the value of its maximum.

If Q is fixed and the entries in P are to be chosen optimally so that $\text{tr}(PQ)$ is maximized, exactly the same logic applies —only with the roles of P and Q reversed. Note that, in this case, one proceeds by columns in Q and rows in P.

The following lemma summarizes these observations.

Lemma 1 (Best response properties). *Let $\bar{P} \in \mathcal{P}^{\Delta}_{n\times m}$ and $\bar{Q} \in \mathcal{Q}^{\Delta}_{n\times m}$.*

[5] Whenever in the sequel I talk about the "corresponding" element (in Q) to some element p_{ij}, I mean it to be the element q_{ji}; analogously I refer to p_{ij} as the "corresponding element" of q_{ji}.

(a) For any $Q \in B(\bar{P})$

$$\sum_{i \in A(p_{\cdot j^\star})} q_{j^\star i} = 1 \quad \textit{and} \quad q_{j^\star i} = 0 \quad \forall\, i \notin A(p_{\cdot j^\star});$$

for any $P \in B(\bar{Q})$

$$\sum_{j \in A(q_{\cdot i^\star})} p_{i^\star j} = 1 \quad \textit{and} \quad p_{i^\star j} = 0 \quad \forall\, j \notin A(q_{\cdot i^\star}).$$

(b) For fixed $\bar{P}$*,*

$$\max_{Q} \left(tr(\bar{P}Q) \right) = \max_{q_{ji}} \left(\sum_j \sum_i \bar{p}_{ij} q_{ji} \right) = \sum_j \max_i \left(\bar{p}_{ij} \right);$$

for fixed $\bar{Q}$*,*

$$\max_{P} \left(tr(P\bar{Q}) \right) = \max_{p_{ij}} \left(\sum_i \sum_j p_{ij} \bar{q}_{ji} \right) = \sum_i \max_j \left(\bar{q}_{ji} \right).$$

If in point (a) $\bar{p}_{i^\star j^\star}$ is the *unique maximal element* in the j^*–th column of $\bar{P}$, then of course for any Q that is a best response to $\bar{P}$, $q_{j^\star i^\star} = 1$. The proposition in (b) directly follows from (a); any Q that is optimal to some fixed $\bar{P}$ never can "extract" from $\bar{P}$ more than the sum of its column maxima. Analogous reasoning applies for the roles of P and Q reversed.

Example 2. Suppose we are given the following sender matrix,

$$P_1 = \begin{pmatrix} 1-x & x & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \text{ with } x \in (0,1).$$

The set of all receiver matrices that are best responses to P_1 is given by

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

Note that $B(P_1)$ also includes the two corner solutions

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where y equals 0 and 1 respectively. It is easily checked that

$$tr(P_1 Q) = 2 = \sum_j \max_i p_{ij} = x + (1-x) + 1$$

for all $Q \in B(P_1)$.

On the other hand, for fixed

$$Q_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1-y & y \end{pmatrix} \text{ with } y \in (0,1),$$

the set of best responses in terms of P is given by

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

With Lemma 1 now it is fairly easy to identify the Nash strategies of this game. All we have to do is to look at each column in P and Q and to check whether the corresponding rows in Q and P, respectively, fulfill the best–response criteria. For example, the pair (P_1, Q_1) from above clearly is a Nash strategy—P_1 is a best response to Q_1, and Q_1 is a best response to P_1.

3.2 Minimal consistency

Lemma 1 together with the condition that in a Nash strategy the sender and the receiver matrices have to be best responses to each other can be interpreted in the sense of some minimal consistency criteria between the entries that are used for sending and receiving messages.

Lemma 2. *The pair $(P,Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{n\times m}$ is a Nash strategy if and only if the following conditions hold true:*

(a) whenever , $q_{j^\star i^\star} \neq 0$, then $p_{i^\star j^\star} \neq 0$ or $p_{ij^\star} = 0 \, \forall i$ and

(b) whenever $p_{i^\star j^\star} \neq 0$, then $q_{j^\star i^\star} \neq 0$ or $q_{ji^\star} = 0 \, \forall j$.

Proof. The proof is just by the contrapositive of Lemma 1. If for some $Q \in B(P)$, $q_{j^\star i^\star} \neq 0$, then from Lemma 1 we know that, $p_{i^\star j^\star} \in A(p_{\cdot j^\star})$. That is, whenever $q_{j^\star i^\star}$ is positive, then the corresponding element $p_{i^\star j^\star}$ in any P to which Q is a best–response only can be a maximal element of that column. This maximum can be everything between zero and one, including zero and one. However, whenever it is zero, then there are indeed no non–zero elements in this column. Analogous implications hold true for the roles of P and Q reversed. □

A zero column in P means that there is a signal that is never used by an individual endowed with this sender matrix. A zero column in Q, on the other hand, means that there is an event that is never possibly inferred, which we can interpret in the sense that an individual with such a receiver matrix has no concept of the respective event. With this in mind the above lemma reads as follows: Whenever in the position of the sender an event $i^\star$ is linked to signal $j^\star$ with some probability, then in the position of the receiver, event $i^\star$ is either also associated with signal $j^\star$ with at least some probability, or, if this is not the case,

then the individual under study has indeed no concept of event $i^\star$. Analogously, whenever in the position of the receiver, event $i^\star$ is associated with signal $j^\star$ with some probability, then in the position of the sender, event $i^\star$ is either also linked to signal $j^\star$ with some probability, or, if this is not the case, then signal $j^\star$ remains idle.

3.3 Features of Nash strategies

There are two features of Nash strategies that are particularly interesting:

(1) 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

(2) there can be zero columns in P or Q, or in both, the P and the Q matrix.

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.

In fact, we have already encountered instantiations of these phenomena in Example 2 above. Sender matrix P_1 links both event 2 as well as event 3 to signal 3—homonymy, whereas receiver matrix Q_1 associates both signal 1 as well as signal 2 with event 1—synonymy.

If we vary Example 2 slightly, we can get examples of Nash strategies that have zero columns in the sender or the receiver matrix or also in both, the sender and the receiver matrix.

Example 3.

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

The problem with zero columns is that they destroy the otherwise fairly uniform patterns of Nash strategies. Suppose we start out with some Nash strategy, (P^*, Q^*). Whenever there is a zero column in one of the two matrices, 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 do not change the column maxima of that matrix*. This following example illustrates this.

Example 4.

$$(P_3, Q_3) = \left[\begin{pmatrix} 1-x & x & 0 \\ 1-x-\epsilon & x-\epsilon & 2\epsilon \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right]$$

Note that the Q matrix is the same as in the previous example. Since the second column in Q consists entirely of zeros, in order for P to be a best response to Q it is irrelevant how we assign the full mass of 1 to the various elements in the second row of P, but we have to make sure that Q still is a best response to P, which, for the present case, is indeed guaranteed as long as p_{21} stays ϵ below p_{11} and p_{22} stays ϵ below p_{12}.

Trapa and Nowak (2000) characterize the class of Nash strategies that is defined by the condition that neither P nor Q contains a zero columns.

Lemma 3 (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*

(i) for each j, the j-th 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

(ii) for each i, the i-th 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 $p_{ij}=p_j$.

The multiplicity of event–signal relations that might occur in a Nash strategy is not bound to occur in only one dimension. In addition to isolated cases of homonymy or synonymy—as we have seen them in Example 2—there also can be combinations of these phenomena, where two (or more) events that are linked to the same signal with some positive probability are also linked to some other signal with some positive probability.

Example 5.

$$(P_4,Q_4) = \left[\begin{pmatrix} 0.3 & 0.7 & 0 \\ 0.3 & 0.7 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0.6 & 0.4 & 0 \\ 0.6 & 0.4 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right]$$

Example 6.

$$(P_5,Q_5) = \left[\begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \end{pmatrix} \right].$$

From Lemma 1 it is easily seen that in order to have a strict Nash strategy—that is, a pair (P^*,Q^*) such that P^* is a unique best–response to $Q^\star$ and vice versa—there has to be exactly one 1 in each column of P^* and respectively Q^*, such that $q^*_{ji}=1$ whenever $p^*_{ij}=1$. That is, in case $n=m$, a pair (P^*,Q^*) is a strict Nash strategy, if and only if P^* is a permutation matrix[6] and Q^* the transpose of P^*, which first has been shown by Trapa and Nowak (2000).

4 Evolutionary stability criteria

The most important stability concept in evolutionary game theory is that of an *evolutionarily stable strategy.*

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

[6] An $n\times n$ matrix A is called a permutation matrix if every row and every column of A contains exactly one 1 and all the other elements are zero.

(i) it is a Nash strategy, and if

(ii) $F[(P^*, Q^*), (P^*, Q^*)] = F[(P, Q), (P^*, Q^*)]$ *implies that* $F[(P^*, Q^*), (P, Q)] > F[(P, Q), (P, Q)] \quad \forall\, (P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{n\times m}$, $(P, Q) \neq (P^*, Q^*)$.

In words, a strategy is evolutionarily stable if it is a Nash strategy, and if, in addition to that, whenever there is an alternative best reply, then the original Nash strategy is a better reply to this alternative best reply than this alternative best reply is to itself. A weaker version of this concept, where the strict inequality in (ii) in the definition above is replaced by a weak inequality sign is known as *neutral stability*.

Definition 3. *A strategy* $(P^*, Q^*) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{n\times m}$ *is called a* neutraly stable strategy *if*

(i) it is a Nash strategy, and if

(ii) $F[(P^*, Q^*), (P^*, Q^*)] = F[(P, Q), (P^*, Q^*)]$ *implies that* $F[(P^*, Q^*), (P, Q)] \geq F[(P, Q), (P, Q)]\ \forall\, (P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{n\times m}$.

Though, formally a static concept, evolutionary stability has been conceived to capture the idea that a strategy can protect itself against the invasion of mutant strategies, or put differently, that it *can drive out* mutant strategies. Neutral stability, instead, expresses the idea that a strategy *cannot be driven out* by other possibly invading strategies, at least not in the absence of neutral drift. On the other hand, this means that neutral stability describes a situation where *there is room for neutral drift*.

From the definitions above we directly see that the following chain of implications holds true:

$$\text{strict Nash} \;\Rightarrow\; \text{evolutionarily stable} \;\Rightarrow\; \text{neutrally stable} \Rightarrow\; \text{Nash}.$$

Selten (1980) shows that for the class of symmetrized games evolutionary stability implies strict Nash, so that for symmetrized games a strategy is evolutionarily stable, if and only if it is a strict Nash strategy. For the symmetrized sender–receiver game $\Gamma_{m,n}$ this means that a pair (P^*, Q^*) is an evolutionarily stable strategy if and only if P^* is a permutation matrix and Q^* is the transpose of P, which first also has been shown by Trapa and Nowak (2000).

In fact, this reflects an earlier result by Wärneryd (1993), who shows that for the sender–receiver game in pure strategies, that is, where a sender's strategy links every event to exactly one signal and receiver's strategy links every signal to exactly one action, an equilibrium is evolutionarily stable if and only if the sender uses a bijective mapping from events to signals and the receiver uses the inverse of this mapping to link signals to actions, where every action can be identified with the event of nature in the light of which it is being called for, such that agents coordinate on a perfectly informative signaling system.

As also has been shown by Wärneryd (1993), for the sender–receiver game in pure strategies, evolutionarily stable strategies are also the only neutrally stable

strategies, or *weakly evolutionarily stable strategies* in the terminology that he uses. Evolutionary stability, as it has been shown by Taylor and Jonker (1978), and Hofbauer, Schuster and Sigmund (1979), implies asymptotic stability in the replicator dynamics, whereas *weak evolutionary stability*, as it has been shown by Thomas (1985), implies *Lyapunov stability* in the replicator dynamics. Based on these results, Wärneryd (1993) argues that a perfectly informative signaling system will arise from a trial–and–error process in the long run.

However, if we want to think of a trial–and–error process that operates in a *population* of individual agents, we are unavoidably faced with mixed strategies as the polymorphic states of this system, where not every individual uses the same (pure) strategy. So if we want to make this model fruitful for evolutionary linguistics we do have to enrich its analysis by a characterization of neutrally stable strategies that encompasses mixed strategies as well.

4.1 Evolutionary stability and neutral stability for doubly symmetric games

Before we move on to characterize neutrally stable strategies, it is helpful to elaborate a little bit on definitions.

Remark 1. *Let* $(P^*, Q^*) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ *be a Nash strategy. Then for any* $Q' \in B(P^\star)$ *and for any* $P' \in B(Q^\star)$*,*

$$tr(P^*Q') = tr(P^*Q^*) = tr(P'Q^*). \tag{15}$$

Proof. Of course, if (P^*, Q^*) is a Nash strategy, then $Q^\star \in B(P^\star)$ and $P^\star \in B(Q^\star)$. Since $Q' \in B(P^*)$, $\mathrm{tr}(P^*Q') = \mathrm{tr}(P^*Q^*)$, and since $P' \in B(Q^\star)$, $\mathrm{tr}(P'Q^*) = \mathrm{tr}(P^*Q^*)$. □

Remark 2. *A strategy* $(P^*, Q^*) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$

(a) is an evolutionarily stable strategy *if and only if*

(i) $P^* \in B(Q^*)$ *and* $Q^* \in B(P^*)$*, and in addition to that*

(ii) $P \in B(Q^*)$ *for some* $P \in \mathcal{P}^{\Delta}_{n\times m}$ *and* $Q \in B(P^*)$ *for some* Q *in* $\mathcal{Q}^{\Delta}_{m\times n}$ *with* $P \neq P^\star$ *or* $Q \neq Q^\star$ *implies that*

$$tr(PQ) < tr(P^*Q^*).$$

(b) (P^*, Q^*) *is a* neutrally stable strategy *if and only if*

(i) $P^* \in B(Q^*)$ *and* $Q^* \in B(P^*)$*, and in addition to that*

(ii) $P \in B(Q^*)$ *for some* $P \in \mathcal{P}^{\Delta}_{n\times m}$ *and* $Q \in B(P^*)$ *for some* Q *in* $\mathcal{Q}^{\Delta}_{m\times n}$ *implies that*

$$tr(PQ) \leq tr(P^*Q^*).$$

Proof. The proof is given for point (a); for (b) it is essentially the same only with the strict inequality replaced by a weak inequality sign. Point (i) is just the condition that the pair (P^*, Q^*) be a Nash strategy. Point (ii) replaces the condition that whenever there is an alternative best reply, that is, a pair (P, Q) with $P \neq P^\star$ or $Q \neq Q^\star$ such that

$$F[(P^*, Q^*), (P^*, Q^*)] = F[(P, Q), (P^*, Q^*)], \tag{16}$$

then it should be the case that

$$F[(P^*, Q^*), (P, Q)] \quad > \quad F[(P, Q), (P, Q)]. \tag{17}$$

By the symmetry of the payoff function, $F[(P^*, Q^*), (P, Q)] = F[(P, Q), (P^*, Q^*)]$. The inequality in (17) therefore is equivalent to

$$F[(P, Q), (P^*, Q^*)] > F[(P, Q), (P, Q)],$$

which by the supposition of the condition (16) implies that

$$F[(P^*, Q^*), (P^*, Q^*)] \quad > \quad F[(P, Q), (P, Q)],$$

which, of course, only can be true if

$$\mathrm{tr}(P^*Q^*) \quad > \quad \mathrm{tr}(PQ).$$

□

Remark 2 says that if we want to check for evolutionary, or respectively neutral stability, of a particular Nash strategy (P^*, Q^*), all we have to do is to compare its communicative potential with itself $\mathrm{tr}(P^*Q^*)$, which we sometimes also shall refer to as the *eigen communicative potential* of a strategy, to the eigen communicative potential of any alternative best reply, that is, $\mathrm{tr}(PQ)$ for which $P \in B(Q^*)$ and $Q \in B(P^*)$.

5 Neutrally stable strategies

Let us first try to find out what we can learn about neutral stability from the examples of the previous section.

Example 1 (continued)

We have seen before that the pair (P_1, Q_1) of Example 2 with $x, y \in (0, 1)$ is a Nash strategy of the game $\mathcal{G}^{\Delta}_{3\times 3}$. Clearly, (P_1, Q_1) cannot be an evolutionarily stable strategy since the two matrices are neither permutation matrices nor is one the transpose of the other. Here we want to see whether despite its failure to satisfy evolutionary stability it still can be a neutrally stable strategy. What we have to do in order to check for neutral stability, by Remark 2, is to see

whether in case there are alternative best replies $Q_1' \in B(P_1)$ and $P_1' \in B(Q_1)$, it is guaranteed that $\mathrm{tr}(Q_1'P_1') \geq \mathrm{tr}(Q_1P_1)$.

The communicative potential of (P_1, Q_1) with respect to itself, that is, $\mathrm{tr}(P_1, Q_1)$, equals 2. In Example 2 we have already determined the set of best responses $B(P_1)$ and $B(Q_1)$. From this we see that for every element $Q_1' \in B(P_1)$ the sum of its column maxima is 2. So with whatever sender matrix P_1' we multiply Q_1', the resulting communicative potential $\mathrm{tr}(P_1'Q_1')$ can never exceed 2. The same is true if we fix any of the sender matrices that are in $B(Q_1)$; the sum of its column maxima is 2 and therefore with whatever receiver matrix we multiply it, the communicative potential we can extract from it is bounded from above by 2. So there can be no alternative best reply (P_1', Q_1') whose eigen communicative potential $\mathrm{tr}(P_1', Q_1')$ is strictly greater than that of (P_1, Q_1). However, there are alternative best replies that have the same eigen communicative potential as the original (P_1, Q_1). As an example for such alternative best replies take a P_1' that is of the same structure as P_1 just with a different $x \in (0, 1)$, and an alternative Q_1' that is of the same structure as Q only with a different $y \in (0, 1)$. Despite the fact that (P_1, Q_1) is not an evolutionarily stable strategy, it therefore still is a neutrally stable strategy.

Example 2 (continued)

Consider instead the Nash strategy (P_2, Q_2) from Example 3 and take as an alternative best reply the pair (P_2', Q_2') with

$$P_2' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \quad Q_2' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Clearly, $P_2' \in B(Q_2)$ and $Q_2' \in B(P_2')$. However,

$$\mathrm{tr}(P_2'Q_2') = 3 > 2 = \mathrm{tr}(P_2Q_2),$$

that is, (P_2', Q_2') attains a strictly higher eigen communicative potential than (P_2, Q_2), and so (P_2, Q_2) cannot be a neutrally stable strategy.

An obvious characteristic of (P_2, Q_2) as opposed to (P_1, Q_1) is that each of its two matrices, P_2 as well as Q_2, contains a column that consists entirely of zero entries. As we have seen before, a zero column in P means that there is a signal that is never used, whereas a zero column in Q means that there is an event that is never possibly understood. In going from (P_2, Q_2) to (P_2', Q_2') nothing has changed but to link the previously idle signal to the event that before has been never understood.

It can be shown that a zero column in both the sender and the receiver matrix is in general sufficient to destroy neutral stability. The idea behind this is exactly the same as in the example above: An invading proto–language that changes nothing about the existing linkages between events and signals, but that in addition to that links the idle signal to the event that is never understood, clearly is not doing worse against the resident language but can do better against itself.

Lemma 4. *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.*

A formal proof of this is given in the Appendix.

However, zero columns in both P and Q is neither the only thing that can happen to prevent a strategy form being neutrally stable, nor is it the case that no zero column in neither P nor Q is sufficient to guarantee neutral stability.

Examples 4 and 5 (continued)

We have seen above that (P_4, Q_4) is a Nash strategy. However, it also fails to satisfy neutral stability, which also can be seen from talking the pair (P_2', Q_2') as an alternative best reply. As before, $Q_2' \in B(P_4)$ and $P_2' \in B(Q_4)$, but the eigen communicative potential of (P_2', Q_2') is higher than that of the original Nash strategy (P_4, Q_4),

$$\mathrm{tr}(P_2'Q_2') = 3 > 2 = \mathrm{tr}(P_4Q_4),$$

and so (P_4, Q_4) cannot be a neutrally stable strategy. Note that this is true even though neither P_4 nor Q_4 contains a column that consists entirely of zeros. What destroys neutral stability in this case is the fact that in both matrices, P_4 and Q_4, there are columns with multiple maximal elements that are positive but not equal to 1. For exactly the same reason strategy (P_5, Q_5), which we considered in Example 6, also cannot be a neutrally stable strategy—which again can be checked by considering (P_2', Q_2') as an alternative best reply.

Another instantiations of Nash strategies that are not neutrally stable are cases 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 7.

$$(P_6, Q_6) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \text{ and } \quad Q_5 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1-\beta & \beta \\ 0 & 1-\beta & \beta \end{pmatrix}, \ \beta \in (0,1).$$

Again by (P_2', Q_2') it can be checked that (P_6, Q_6) also fails to be a neutrally stable strategy.

The crucial thing in the case where one of the two matrices contains a column with multiple maximal elements that are strictly between 0 and 1, is that the Nash property already implies that the other matrix is bound to contain a zero column or a column with multiple maximal elements that are strictly between 0 and 1.

Lemma 5. *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 non-zero multiple maximal elements that are not equal to* 1*, then Q [P] contains*

(i) at least two columns that have non-zero multiple maximal elements that are not equal to 1*, or*

(ii) a zero column,

and (P, Q) *cannot be a neutrally stable strategy.*

A formal proof of this is given in the Appendix.

What lies behind this result is that if the ambiguity created by instances of homonymy or synonymy works into both directions, then this leaves enough room for rearranging the existing linkages between events and signals such that a single mutation can raise its eigen communicative potential over the eigen communicative potential of the original Nash strategy without reducing its communicative potential with the original Nash strategy.

5.1 Necessary and sufficient conditions for neutral stability

Lemmas 4 and 5 together constitute a necessary condition for neutral stability.

Lemma 6. *Let* $(P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ *be a neutrally stable strategy, then*

(i) at least one of the two matrices P *or* Q *has no zero column; and*

(ii) neither P *nor* Q *contains a column with multiple maximal elements that are strictly between* 0 *and* 1*.*

Proof. Combining Lemmas 4 and 5 we have that a Nash strategy *cannot* be a neutrally stable strategy if

(i) P and Q contain a zero column, or if

(ii) P or Q contains a column with multiple maximal elements that are strictly between 0 and 1.

The contrapositive of this statement yields the claim of the proposition. □

Since by definition an evolutionarily stable strategy is necessarily a neutrally stable strategy, these conditions, of course, also have to be met by evolutionarily stable strategies. For neutral stability, however, they also can be shown to be sufficient.

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*] has no column with multiple maximal elements that are not equal to* 1*,*

(i) then Q *[*P*] has no column with non–zero multiple maximal elements that are not equal to* 1*; and*

(ii) (P, Q) *is a neutrally stable strategy.*

The proof of this, which essentially relies on exploiting the best–response properties that must hold true between P and Q' as well as Q and P' of any invading language, again is given in the Appendix.

Combining Lemma 6 and Lemma 7 we finally have a complete characterization of neutrally stable strategies.

Proposition 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 zero and 1.

This means 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 make use of any other signal to get communicated. 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, in a neutrally stable strategy, there cannot be any idle signal as long as there are events that are never possibly understood and vice versa. The reason behind this is that both of these situations leave too much room for shifting around the entries in P and Q, such that there are mutant strategies that hold equal against the original Nash strategy, but that can raise its *eigen* communicative potential over the *eigen* communicative potential of the original Nash strategy.

Next we want to consider what are the evolutionary consequences of this for a particular evolutionary dynamics that formalizes the idea of a trial and error process in a population of individual agents—the so–called replicator dynamics. Before we do that, we have to find a way of specifying the state space on which this dynamics should operate so that we keep the analysis tractable. A possible way of doing so is provided by the theory of population games.

6 Population games

Up until now—somewhat silently—we have considered a pair $(P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ as representing the probabilities with which an individual agent will link events to signals and signals to events. Now we want to look at it as representing the distribution of actions in a *population* of individual agents.

As we have seen above, the symmetrized sender–receiver game $\Gamma^{\Delta}_{m,n}$ can be understood as the mixed–strategies version of a game with finitely many pure strategies, where a pure strategy is *a pair* of a sender and a receiver matrix

$$(P, Q) \in \mathcal{P}_{n\times m} \times \mathcal{Q}_{n\times m}, \tag{18}$$

such that for general n and m there are $L = m^n \times n^m$ pure strategies. Assigning different weights,

$$x = (x_1, x_2, \ldots, x_L) \in S_L,$$

where S_L is the simplex in R^L, that is,

$$\sum_{l=1}^{L} x_l = 1,$$

to these $L = m^n \times n^m$ pure strategies we can span the whole strategy set $\mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{n\times m}$.[7]

In the framework of population games, mixed strategies can be interpreted as the *population's average strategy* resulting from the frequency distribution of *different types of otherwise identical players* each of which uses a particular *pure strategy* whenever he or she is called for playing the game. In this case, x_l is the fraction of individual agents using pure strategy l, and $x = (x_1, x_2, \ldots, x_L)$ is a vector of L type frequencies.

6.1 The average population profile and the fitness in the population

Suppose we are given a vector of type frequencies $x_l = (x_1, x_2, \ldots, x_L) \in S_L$, then the *average strategy profile* is

$$(\bar{P}(x), \bar{Q}(x)) = \sum_{1}^{L} x_l (P_l, Q_l),$$

which also can be written as

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

where $\bar{p}_{ij}$ is the sum of all type frequencies whose i,j–th entry in P is equal to 1, and $\bar{q}_{ji}$ is the sum of all type frequencies whose j,i–th entry in Q is equal to 1, that is,

$$\bar{p}_{ij} = \sum_{l:p^l_{ij}=1} x_l \quad \text{and} \quad \bar{q}_{ji} = \sum_{l:q^l_{ji}=1} x_l, \tag{19}$$

where superscript l indicates the type.

Assuming that agents are matched randomly according to their type frequencies, the *average performance of type l* for a given distribution of type

[7] Of course, if we wish to, from the payoff function (4) we can compute the *payoff matrix* $A_{L\times L}$ for the symmetrized game in pure strategies simply by letting each pure strategy $(P,Q) \in \mathcal{P}_{n\times m} \times \mathcal{Q}_{n\times m}$ play against every other pure strategy that is contained in this set. Due to the symmetry of the payoff function, the payoff matrix $A_{L\times L}$ is of course symmetric.

frequencies is

$$f_l(x) = \sum_{l'=1}^{L} x_{l'} F\left[(P_l, Q_l), (P_{l'}, Q_{l'})\right]; \tag{20}$$

the *average performance in the population* is

$$\bar{f}(x) = \sum_{l=1}^{L} x_l f_l(x). \tag{21}$$

In an evolutionary context $f_l(x)$ can be interpreted as the *fitness of type* l, and $\bar{f}(x)$ as the *average fitness in the population*[8].

For the sender–receiver model discussed here, the fitness of type l can be written as its payoff from play against the population's average strategy,

$$f_l(x) = F\left[(P_l, Q_l), \left(\bar{P}(x), \bar{Q}(x)\right)\right] = \frac{1}{2}\mathrm{tr}\left(P_l \bar{Q}(x)\right) + \frac{1}{2}\mathrm{tr}\left(\bar{P}(x) Q_l\right), \tag{22}$$

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

$$\bar{f}(x) = F\left[\left(\bar{P}(x), \bar{Q}(x)\right), \left(\bar{P}(x), \bar{Q}(x)\right)\right] = \mathrm{tr}\left(\bar{P}(x)\bar{Q}(x)\right). \tag{23}$$

6.2 Nash strategies in population games

With the interpretation of an element $(P, Q) \in \Delta_m^n \times \Delta_n^n$ as the population's average strategy resulting from a specific distribution of types frequencies, we also have to reconsider the interpretations that pertain to the properties of Nash strategies and its refinements in terms of evolutionarily and neutral stability.

As it is generally true for populations games, if the average strategy profile is a Nash strategy, this implies that *given the composition of the population*, $x = (x_1, \ldots, x_l, \ldots, x_L)$, every type l that is present in this population reaches the same fitness, that is,

$$f_l(x) = \bar{f}(x), \; \forall \; l \text{ whenever } x_l \neq 0.$$

For the sender–receiver game discussed here this means that every type l that is present in this population with some non–negligible fraction $x_l > 0$ gets the same payoff from communicating with the population's average,

$$F\left[(P_l(x), Q_l(x)), \left(\bar{P}(x), \bar{Q}(x)\right)\right] = F\left[\left(\bar{P}(x), \bar{Q}(x)\right), \left(\bar{P}(x), \bar{Q}(x)\right)\right],$$

for all l for which $x_l \neq 0$.

[8] In a more standard notation, given the payoff matrix of the underlying stage game, the fitness function of type l can be written as $f_l(x) = (Ax)_l$, indicating the l–th entry of the product of the payoff matrix A times the vector of type frequencies; and the average fitness in the population can be written as $\bar{f}(x) = xAx$.

If in addition to that there is *no type* that yields the same payoff from communicating with the current population's average, but that yields a strictly higher eigen communicative potential as compared to the eigen communicative potential of the population's average, then this population state is *neutrally stable*; and it is *evolutionarily stable* if upon that there is also no type that yields the same payoff from communicating with the current population's average, but that holds equal in terms of its eigen communicative potential as compared to the eigen communicative potential of the population's average.

In the framework of Example 1, where there are 2 events and 2 signals, consider a population where 40 percent use pure strategy

$$(P_3, Q_3) \quad = \quad \left[\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \right],$$

and where 60 percent use pure strategy

$$(P_3, Q_4) \quad = \quad \left[\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \right].$$

In this case, the average population profile is given by

$$[\bar{P}(x), \bar{Q}(x)] \quad = \quad \left[\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{pmatrix} \right].$$

Since $\bar{P}(x)$ is a best response to $\bar{Q}(x)$, and $\bar{Q}(x)$ is a best response to $\bar{P}(x)$, the pair $(\bar{P}(x), \bar{Q}(x))$ definitely is a Nash strategy. The average performance, or fitness, of the type that uses (P_3, Q_3) is

$$\begin{aligned} F[(P_3, Q_3), (\bar{P}(x), \bar{Q}(x))] \quad &= \quad \frac{1}{2}\text{tr}\left[\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{pmatrix} \right] \\ &+ \quad \frac{1}{2}\text{tr}\left[\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \right] = 1, \end{aligned}$$

and the average performance of the type using (P_3, Q_4) is

$$\begin{aligned} F[(P_3, Q_4), (\bar{P}(x), \bar{Q}(x))] \quad &= \quad \frac{1}{2}\text{tr}\left[\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{pmatrix} \right] \\ &+ \quad \frac{1}{2}\text{tr}\left[\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \right] = 1, \end{aligned}$$

which, as it should be if the population's average strategy corresponds to a Nash strategy, are both equal to the average performance in the population,

$$\begin{aligned} F[(\bar{P}(x), \bar{Q}(x)), (\bar{P(x)}, \bar{Q(x)})] \quad &= \quad \text{tr}\left(\bar{P}(x)\bar{Q}(x)\right) \\ &= \quad \text{tr}\left[\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{pmatrix} \right] = 1. \end{aligned}$$

However, the situation is not neutrally stable. If a type that uses pure strategy

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

enters this population, it will get the same payoff from communicating with the population's average as the population's average gets from communicating with itself,

$$F[(P_1, Q_1), (\bar{P}(x), \bar{Q}(x))] = 1 = F\left[\left(\bar{P}(x), \bar{Q}(x)\right), \left(\bar{P}(x), \bar{Q}(x)\right) \right],$$

but it attains a strict higher *eigen* communicative potential than the population's average,

$$\mathrm{tr}(P_1 Q_1) = \mathrm{tr}\left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right] = 2 > 1 = \mathrm{tr}\left(\bar{P}(x)\bar{Q}(x)\right),$$

and so the original composition of the population cannot be neutrally stable. Of course, it then also cannot be evolutionarily stable.

6.3 Individuals' consistency?

We have seen above that the properties of Nash strategies can be interpreted in the sense of some minimal consistency criteria that must be true between the lexical entries that are used for sending and receiving messages (Lemma 2). Here I want to ask what we can learn about *individuals' consistency* if the *population's average strategy* corresponds to a Nash strategy. Some more terminology helps to clarify the discussion.

I shall call a pair $(P, Q) \in \mathcal{P}_{n\times m} \times \mathcal{Q}_{m\times n}$ *perfectly consistent* if it is a strict Nash strategy, that is, if P and Q are permutation matrices and one matrix is the transpose of the other; I shall call it *perfectly inconsistent* if both matrices, P and Q, are permutation matrices, but if $p_{i,j} = 1$ implies that $q_{j,i} = 0$, and if $q_{j,i} = 1$ implies that $p_{i,j} = 0$. So an agent with perfectly inconsistent sender and receiver matrices will link every event unambiguously to one signal that is not used for any other event, but when he or she receives this signal, it will associate with a different event.

In framework of Example 1, that is, in the case of 2 events and 2 signals, perfectly consistent sender–receiver pairs are,

$$(P_1, Q_1) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ and } (P_2, Q_2) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

whereas perfectly inconsistent sender–receiver pairs are

$$(P_1, Q_2) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \text{ and } (P_2, Q_1) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Remarkably, there are compositions of the population that correspond to a Nash strategy, where *all* individual types use perfectly inconsistent sender and receiver matrices.

Suppose, for example, that one half of the population uses pure strategy (P_1, Q_2) and the other half uses pure strategy (P_2, Q_1). In this case, the population's average strategy is

$$\begin{aligned}\left[\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}\right] &= \frac{1}{2}\left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right] \\ &+ \frac{1}{2}\left[\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]. \quad (24)\end{aligned}$$

Of course, the fact that minimal consistency on the level of the population's average does not necessarily extend to the same consistency criterion on the level of individual agents is just a reflection of the more general fact that, if the composition of the population corresponds to a Nash strategy, this does not necessarily imply that every type that is present in this population with some non–negligible fraction uses a Nash strategy.

However, from the characterization of neutrally stable strategies in the previous section, we see easily that a population's average as in (24) cannot be neutrally stable, since there are multiple column maxima strictly between zero and 1 in both the average sender and the average receiver matrix. Indeed, this population is highly vulnerable to be invaded by some mutant strategy—either (P_1, Q_1) or (P_2, Q_2).

More generally, it can be shown that in a neutrally stable population there can be no type that uses perfectly inconsistent sender and receiver matrices.

Lemma 8. *Let $(\bar{P}, \bar{Q}) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ be a population's average strategy that is neutrally stable. Then no type l playing one of the pure strategies $(P^l, Q^l) \in \mathcal{P}_{n\times m} \times \mathcal{Q}_{m\times n}$ that is present in this population with some non–negligible fraction $x_l > 0$ can have perfectly inconsistent sender and receiver matrices.*

Proof. The proof is an application of Lemma 5 together with the best–response properties of Nash strategies. Suppose there is a fraction $\alpha > 0$ of agents who use permutation matrix P^1 for sending messages and permutation matrix Q^2 for receiving messages, where P^1 and Q^2 are perfectly inconsistent, that is, if $p^1_{i^\star j^\star} = 1$, then $q^2_{j^\star i^\star} = 0$ but $q^2_{j^\star i'} = 1$ for some $i' \neq i^\star$. So we know that for the average strategy profile of the population, $\bar{p}_{i^\star j^\star} \geq \alpha$ as well as $\bar{q}_{j^\star i'} \geq \alpha$ for some $i' \neq i^\star$, for all $j^\star$. By the supposition of the proposition the average strategy profile $(\bar{P}, \bar{Q})$ is a Nash strategy. Since $\bar{q}_{j^\star i'} \neq 0$, by the best–response properties that must hold true in a Nash strategy, we know that $\bar{p}_{i' j^\star}$ must be a maximal element of the $j^\star$–th column in $\bar{P}$. Since $\bar{p}_{i^\star j^\star} \geq \alpha$, we must have that $\bar{p}_{i' j^\star} \geq \alpha$ as well. However, since P^1 is a permutation matrix, we know that there is some $j' \neq j^\star$ such that $p^1_{i' j'} = 1$, which implies that $\bar{p}_{i' j^\star}$ can never be 1. Therefore, $0 < \alpha \leq \bar{p}_{i' j^\star} < 1$, that is, $\bar{p}_{i' j^\star}$ is strictly between zero and 1, but from Lemma 5 we know that this cannot be true in a neutrally stable strategy. Consequently, there cannot be a positive fraction of agents using a perfectly inconsistent pair of a sender and receiver matrix in a population whose average strategy profile is a neutrally stable strategy. □

Unfortunately, *neutral stability is not strong enough to sort out partly inconsistent individual behavior.*

Example 8. As an example for this, consider a population where the composition of the average sender matrix is given by

$$\begin{aligned} & \alpha \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \beta \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ + \quad & \gamma \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} + 1 - \alpha - \beta - \gamma \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \\ = \quad & \begin{pmatrix} \alpha + \beta & 1 - \alpha - \beta & 0 \\ \alpha + \gamma & \beta & 1 - \alpha - \beta - \gamma \\ 0 & 0 & 1 \end{pmatrix} = \bar{P}, \end{aligned} \tag{25}$$

with α, β and γ strictly between zero and 1, and $\beta > \gamma$ as well as $1 - \alpha - \beta > \beta$, so that $\bar{p}_{11} = \alpha + \beta$ is the unique maximal element in the first column of $\bar{P}$, $\bar{p}_{12} = 1 - \alpha - \beta$ is the unique maximal element in the second column of $\bar{P}$, and $\bar{p}_{33} = 1$ is the unique maximal element in the third column of $\bar{P}$. The receiver matrix used by all types be

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \bar{Q}. \tag{26}$$

Clearly, the pair $(\bar{Q}, \bar{P})$ as given by (25) and (26) is a Nash strategy; it even is a neutrally stable strategy, as it can be checked by Proposition 1. Still there is a β–fraction of agents who uses pure strategy

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

which does not satisfy minimal consistency: P is a best response to Q, but Q is not a best response to P.

If, on the other hand, the population's average strategy satisfies evolutionary stability—that is, if $\bar{P}$ is a permutation matrix and $\bar{Q}$ is the transpose of $\bar{P}$—then, of course, we are in a monomorphic population state where all individual agents use the same pair of a sender and a receiver matrix that mimics $(\bar{P}, \bar{Q})$, and all individual agents will have *perfectly consistent* sender and receiver matrices.

6.4 No zero columns and interior states

An important consequence of neutral stability in the framework of population games arises with respect to zero columns, or better their absence.

In the framework of population games, a zero column in $\bar{P}$ simply means that there is a particular signal that is never used by anybody in this population, whereas a zero column in $\bar{Q}$ means that there is an event that is never understood, which we can interpret in the sense that nobody in this population has acquired an abstract concept of this event yet. Intuitively, then, it is irrelevant what one would understand if a signal that nobody uses was received, or what one will send in the face of an event for which nobody has a concept. Formally, this is reflected by the possibility of zero columns in a Nash strategy.

On the other hand, if every pure strategy is present somewhere in the population—that is, if the composition of the population corresponds to an interior point of the simplex S_L—then all elements in $\bar{P}$ and $\bar{Q}$ are positive, and trivially then neither $\bar{P}$ nor $\bar{Q}$ will have any zero column.

As a consequence, every Nash strategy in the interior of S_L has to be of the Trapa–Nowak type—more precisely, it has to be of a particular subclass of this type.

Lemma 9. *If $(\bar{P}, \bar{Q})$ is a Nash strategy in the interior of S_L, then there exist real numbers $0 < p_j < 1$, $j = 1, \ldots, m$ and $0 < q_i < 1$, $i = 1, \ldots, n$ such that*

$$\begin{aligned} p_{ij} &= p_j \,\forall\, i = 1, \ldots, n, \text{ and} \\ q_{ji} &= q_i \,\forall\, j = 1, \ldots, m. \end{aligned}$$

Proof. The proof is immediate from the result of Trapa and Nowak (2000) in combination with the fact that at an interior state all entries $\bar{p}_{ij}$ and $\bar{q}_{ji}$ are necessarily strictly between zero and one for all i and for all j. □

Example 9. Let $m = 2 = n$. Then the set of all Nash strategies that are in the interior of the simplex is given by

$$\mathcal{N}_{interior} = \left\{ \left[\begin{pmatrix} \alpha & 1-\alpha \\ \alpha & 1-\alpha \end{pmatrix}, \begin{pmatrix} \beta & 1-\beta \\ \beta & 1-\beta \end{pmatrix} \right]_{\alpha,\beta \in (0,1)} \right\}.$$

From Proposition 1 we directly see that a Nash strategy of that form cannot be neutrally stable.

Lemma 10. *If $(\bar{P}, \bar{Q})$ is a Nash strategy in the interior of S_L, then it cannot be a neutrally stable strategy.*

Proof. Lemma 9 together with Proposition 1. □

Neutrally stable strategies, therefore, can only be at the boundary of the simplex. As we will see, this has important consequences for the replicator dynamics of this model.

7 The replicator dynamics

The replicator dynamics, whose relations to game theoretical solution concepts have been studied thoroughly by evolutionary game theorists[9] translates the idea that over time strategies corresponding to different types expand, or contract, according to the difference of their performance with respect to the average performance in the population.

In continuous time, this can be written as a system of differential equations,

$$\begin{aligned} \frac{\dot{x}_1}{x_1} &= f_1(x) - \bar{f}(x) \\ &\vdots \\ \frac{\dot{x}_l}{x_l} &= f_l(x) - \bar{f}(x) \\ &\vdots \\ \frac{\dot{x}_L}{x_L} &= f_L(x) - \bar{f}(x), \end{aligned} \tag{27}$$

where $\dot{x}_l$ is the derivative of x_l with respect to time, that is, $\dot{x}_l/x_l$ is the growth rate of x_l, $f_l(x)$ is the fitness of strategy l and $\bar{f}(x)$ the average fitness in the population as defined by equations (22) and (23), respectively. A state of this system is a vector of L type frequencies, $x \in S_L$.

As such this dynamics is not more than a black–box mechanism that gives positive feedback to relatively more successful strategies. For the model discussed here this can be interpreted in terms of biological as well as cultural transmission of strategies from one generation to the next. Agents who communicate more successfully are more successful in getting good food, escaping dangers, etc. Therefore, it can be argued, they have a direct advantage in reproduction, and, assuming that parents directly transmit their P's and Q's to their offspring, either biologically or culturally, the P's and Q's of agents who communicate better will reproduce with a higher rate. Or, what is a slightly different approach, the replicator dynamics also can be interpreted in the sense of some imitation mechanism. Agents who are more successful in escaping dangers, getting good food etc. are more likely to be imitated and therefore also their communicative strategies will reproduce more successfully.

A state $x^\star$ for which $\dot{x}_l^* = 0$ for all l—that is, a state for which all movement comes to an end—is a *rest point* of this dynamics. Note that at such a state the performance of every pure strategy that occurs with some positive frequency has to be equal to the average performance in the population, which means that a population state whose average strategy corresponds to a Nash strategy necessarily is a rest point of this dynamics.

As usually, a rest point $x^\star$ is said to be *locally asymptotically stable* if after a small perturbation in the state variables that stays within sufficiently small boundaries around the rest point the dynamics eventually will lead back to this

[9] See in particular Hofbauer and Sigmund (1988)

point; and it is said to be *Lyapunov stable* if every neighborhood $B(x^\star)$ contains a neighborhood $B'(x^\star)$ such that for all initial states in $B'(x^\star)$ the dynamics does not leave $B(x^\star)$ as time proceeds.

Given its high dimensionality—for 2 events and 2 signals there are already 16 pure types—it is generally not possible to solve explicitly for this dynamics. However, in the context discussed here we are not so much interested in any particular solution path, but more in the qualitative regularity patterns of the proto–language, if any, that can be explained to occur under this dynamics in the long run.

One of the advantages of formulating the model under study as a game is that we can make use of the rich body of results established by evolutionary game theorists that link the static equilibrium analysis to the qualitative properties of specific evolutionary dynamics. One of the most important results—due to Taylor and Jonker (1978), Hofbauer, Schuster and Sigmund (1979), as well as Zeeman (1979)—is that every *evolutionarily stable* strategy is a *locally asymptotically stable* rest point of the replicator dynamics. Thomas (1985), which later on also has been shown in a more general way by Bomze and Weibull (1995), shows that an analog implication holds true between the weaker versions of these static and dynamic stability concepts: every *neutrally stable* strategy is *Lyapunov stable* in the replicator dynamics. Importantly, none of the converse of these results is true in general. However, for the model discussed here, help comes from the double symmetry of the game.

7.1 The replicator dynamics for doubly symmetric games

For games with a symmetric payoff matrix, the replicator dynamics constitutes a *gradient system*, which induces a strictly monotonic increase in the average fitness along every non–stationary solution path; that is, the average fitness $\bar{f}(x) = \mathrm{tr}(\bar{P}(x)\bar{Q}(x))$ is a strict Lyapunov function for this dynamics[10]. In fact this can be seen as an instance of what in biology is known as *Fisher's Fundamental Theorem of Natural Selection*, that the average fitness in a population increases in the process of evolution. For the long–run behavior on this system this has several implications:

(a) In this case—as established by Akin and Hofbauer (1982) and Losert and Akin (1983)—every orbit, indeed, *converges to some rest point*; and

(b) as shown by Hofbauer and Sigmund (1988), evolutionarily stable strategies *coincide* with the locally asymptotically stable rest points of the replicator dynamics, and are given by the locally strict maxima of the average payoff function.

In combination with the static analysis of evolutionary stability for the sender–receiver game discussed here (Trapa and Nowak, 2000) this directly yields the following:

[10]For a more detailed account see Hofbauer and Sigmund or Weibull

Proposition 2. *For the replicator dynamics of the symmetrized sender–receiver game* $\Gamma_{n,n}$*:*

(a) Every orbit converges to some rest point; and

(b) there are exactly $n!$ *locally asymptotically stable rest points located at those vertices of the simplex* S_L *where* $\bar{P}$ *is a permutation matrix and* $\bar{Q}$ *is the transpose of* $\bar{P}$*; all the other rest points cannot be locally asymptotically stable.*

Note that if $\bar{P}$ is a permutation matrix and $\bar{Q}$ is the transpose of $\bar{P}$, the the average payoff attains its maximum value and the full potential of communication is realized in this population.

This, however it is not the complete picture. From the fact that only the strict Nash strategies can be locally asymptotically stable, we cannot conclude that the system will almost always lead to such a state. Withstanding the fact that each single rest point that corresponds to a non–strict Nash strategy cannot be locally asymptotically stable, there can be sets of non–strict Nash strategies, which, as a set, do attract some positive measured set of states. Neutral stability, as it turns out, is key to explaining this.

Bomze (2002) shows that for *doubly symmetric games* and pairwise interaction *Lyapunov stability* in the replicator dynamics *implies neutral stability*, so that for this case the two concepts coincide. For the model discussed here this implies that the replicator dynamics almost always flows to the boundary of simplex.

Proposition 3. *For the replicator dynamics of the symmetrized sender–receiver game* $\Gamma_{m=n}$*, a rest point* $x^* \in S_L$

(a) is Lyapunov stable if and only if the corresponding population's average strategy $(\bar{P}(x^*), \bar{Q}(x^*))$ *satisfies the condition that (i) at least* $\bar{P}(x^*)$ *or* $\bar{Q}(x^*)$ *has no zero column, and (ii) neither* $\bar{P}(x^*)$ *nor* $\bar{Q}(x^*)$ *has any column with multiple maximal elements that are strictly between zero and* 1*;*

(b) a rest point in the interior of the simplex S_L *never can be Lyapunov stable.*

Proof. The proof of (a) is a direct application of the equivalence result in Bomze (2002) together with Proposition 1, the static characterization of neutrally stable strategies. Point (b) follows from (a) together with Lemma 10. Note that for (b) we really need that Lyapunov stability implies neutral stability; otherwise we could not conclude that a strategy that is not neutrally stable also cannot be Lypunov stable. □

Taking Propositions 2 and 3 together, this means that for *almost all* initial conditions, the replicator dynamics of the symmetrized sender–receiver game will always flow to the boundary of the simplex, to a neutrally stable strategy, but not necessarily to an evolutionarily stable strategy.

8 Interpretation and conclusions

So combining the static analysis of neutrally stable strategies for the symmetrized sender–receiver game $\Gamma^{\Delta}_{n\times m}$ with the qualitative properties of the replicator dynamics for doubly symmetric games, we see that this dynamics generically will lead to a proto–language that satisfies the following regularity patterns on the level of the population's average sender and receiver matrices: There cannot be two (or more) events that are linked to (two) or more signals in parallel, and there can be no signal that remains idle in the presence of events that are never possibly understood; however there can be isolated cases of homonymy or synonymy where two (or more) events share the use of one and the same signal, or where two (or more) signals are associated with one and the same event.

This is exactly what is reflected in the simulation reported in Nowak and Krakauer (1999), where for the same number of signals and events the replicator dynamics that they model seems to converge to a common proto–language where two events share the use of one signal while another signal remains idle. The analytical result presented here, therefore, supports the conjecture that Nowak and Krakauer (1999) draw from their simulations: *Evolution does not necessarily lead to an optimal proto–language*, but it can be blocked in a state where due to ambiguities in event–signal relations some of the potential of communication will be left unexploited. However, what hinders the dynamics to converge to an optimum solution is not the fact that ambiguities in event–signal relations can be evolutionarily stable in the strict sense, but that they can be *neutrally stable.*

On the individuals' level this means that *not every type* will necessarily end up with a *perfectly consistent* proto–language. There are polymorphic populations states that are neutrally stable, and therefore also Lyapunov stable, where not every type satisfies minimal consistency in the sense that the individual's receiver matrix is a best–response to the individual's sender matrix and vice versa. However, as we can see from Proposition 8, in a neutrally stable population *no type* will use *perfectly inconsistent* lexical matrices for sending and receiving messages.

8.1 Future work

The results presented here depend on the assumption that all events that possibly become the object of communication enter the payoff function with the same weight as well as that all available signals also enter the payoff function with the same weights.

Formally, the introduction of different weights of events amounts to multiplying every row element in the sender matrix with the respective weight, whereas the introduction of weights of signals amounts to multiplying every row element in the receiver matrix with the respective weight. Assuming that all weights are indeed different, this destroys the possible sources of homonymy and respectively synonymy in a neutrally stable strategy. Whenever in the position of the

sender there are two (or more) events linked to the same signal with probability 1, but where one of these events is relatively more important than the others, then, of course, the best response to this in terms of the receiver matrix is to associate this signal with the event that is relatively more important. Therefore this cannot be a Nash strategy. Analogously, the introduction of different weights to signals destroys the multiplicity of best–responses in terms of the sender matrix, and therefore synonymy as a phenomenon in a neutrally stable strategy. This means that the results presented here allow us to account for *homonymy of events that are equally important* and *synonymy of signals that are of the same costs*.

Another assumption that we made throughout, almost silently, is that there is a large population where agents are matched randomly according to the frequencies of types in the overall population. In particular we did not take into account any spatial structure that is imposed on the population. Differences in the frequency or importance of events as well as costs of signals typically vary with the environment. On the other hand, language differentiation is one of the general linguistic facts that needs explanation. Interesting extensions of this model therefore point into the direction of considering the combined effects of locally structured population and differences in the frequencies of events or costs of signals in different environments.

References

1. Aumann, R.J. and A. Heifetz (2002). Incomplete Information. In R. Aumann and S. Hart (Ed.), *Handbook of Game Theory with Economic Applications*, Vol. 3 (pp. 1665-1686). Amsterdam: North–Holland.

2. Bomze, I. (2002). Regularity vs. degeneracy in dynamics, games, and optimization: a unified approach to different aspects. *SIAM Review* 44, 394-414.

3. Bomze, I. and J. Weibull (1995). Does neutral stability imply Lyapunov stability? *Games and Economic Behavior* 11, 173–192.

4. Akin, E. and J. Hofbauer (1982). Recurrence of the Unfit. *Mathematical Biosciences* 61, 51–62.

5. Hofbauer, J. and K. Sigmund (1988). *The Theory of Evolution and Dynamical Systems*. Cambridge, UK: Cambridge University Press.

6. Hofbauer, J. and K. Sigmund (1998). *Evolutionary Games and Population Dynamics*. Cambridge, UK: Cambridge University Press.

7. Hofbauer, J., P. Schuster, and K. Sigmund (1979). A note on evolutionarily stable strategies and game dynamics. *Journal of Theoretical Biology* 81, 609–612.

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

9. Lewis, D. (1969). *Convention: A philosophical study.* Cambridge, MA: Cambridge University Press.

10. Losert, V. and E. Akin (1983). Dynamics of games and genes: Discrete versus continuous time. *Journal of Mathematical Biology*, 17, 241–251.

11. Nowak, M. and D. Krakauer (1999). The evolution of language. *Proceedings of the National Academy of Science USA* 96, 8028–8033.

12. Selten, R. (1980). A note on evolutionarily stable strategies in asymmetric animal contests. *Journal of Theoretical Biology*, 84, 93–101.

13. Taylor, P. and L. Jonker (1978). Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences* 40, 145–156.

14. Thomas, B. (1985). On evolutionarily stable sets. *Journal of Mathematical Biology* 22, 105–115.

15. Trapa, P. and M. Nowak (2000). Nash equilibria for an evolutionary language game. *Journal of Mathematical Biology*, 41, 172–188.

16. Weibull, J. (1995). *Evolutionary game theory.* Cambridge, MA: MIT Press, .

17. Wärneryd, K. (1993). Cheap talk, coordination and evolutionary stability. *Games and Economic Behavior*, 5, 532–546.

Appendix

Proof of Lemma 4.

Proof. Suppose that $(P,Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ is a Nash strategy, and assume that the $j^\star$–th column of P as well as the $i^\star$–th column of Q consist entirely of zero elements.

In order to show that (P,Q) is not neutrally stable, by Remark 2, we have to show that there is some $P' \in B(Q)$ and some $Q' \in B(P)$ with $P' \neq P$ or $Q' \neq Q$ such that

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

Now, take as a candidate P' the original P but with the entries in its $i^\star$–th *row* substituted by the vector

$$p'_{i^\star j} = \begin{cases} 1 & \text{for } j = j^\star \\ 0 & \text{otherwise} \end{cases} ;$$

and take as a candidate Q' the original Q but with the entries in its $j^\star$-th *row* substituted by the vector

$$q'_{j^\star i} = \begin{cases} 1 & \text{for } i = i^\star \\ 0 & \text{otherwise} \end{cases} .$$

We first check that P' is indeed a best response to Q. Since the elements of the $i^\star$–th *column* of Q are all zero, the product of the $i^\star$–th *column* of Q with the $i^\star$–th *row* of any sender matrix will be zero. So, whatever the elements in the $i^\star$–th *row* of P might have been, we "loose" nothing by setting $p'_{i^\star j^\star}$ equal to 1. Since in constructing P' from P we did not change the elements of any other row, $\mathrm{tr}(P'Q) = \mathrm{tr}(PQ)$. By an analogous argument with the roles of sender and receiver matrices reversed, we have that $\mathrm{tr}(PQ') = \mathrm{tr}(PQ)$, which means that $Q' \in B(P)$. What remains to be done, then, is to show that $\mathrm{tr}(P'Q') > \mathrm{tr}(PQ)$.

Note that

$$p'_{ij}q'_{ji} = p_{ij}q_{ji} \quad \text{whenever} \quad i \neq i^\star \text{ or } j \neq j^\star.$$

On the other hand,

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

whereas

$$p_{i^\star j^\star}q_{j^\star i^\star} = 0.$$

Summing over all i and j we therefore have that

$$\sum_i \sum_j p'_{ij}q'_{ji} = \sum_i \sum_j p_{ij}q_{ji} + 1,$$

which means that,

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

and (P, Q) cannot be neutrally stable. □

Proof of Lemma 5.

Proof. Suppose that $(P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$ is a Nash strategy. As for Lemma 4, we have to show that there is some $P' \in B(Q)$ and some $Q' \in B(P)$, such that

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

We give the proof for the case where the condition of the proposition applies to Q. Analogous reasoning holds true for the case where it applies to P.

Suppose that the $i^\star$–th column of Q contains more than one maximum element that is positive but not equal to 1. Since Q is a best response to P, for all $j \in A(q_{\cdot i^\star})$, $p_{i^\star j}$ is a maximal, but not the unique maximal element of its respective column in P: If $p_{i^\star j}$ with $j \in A(q_{\cdot i^\star})$ was not a maximal element of

the j–th column of P, then $q_{ji^\star}$ could not be positive. If, on the other hand, $p_{i^\star j}$ with $j \in A(q_{\cdot i^\star})$ was the unique maximal element of the j–th column of P, then $q_{ji^\star}$ would have to be exactly equal to 1. Note that this does not exclude the possibility that for some $j \in A(q_{\cdot i^\star})$ the j–th column of P is a zero column.

On the other hand, P is a best response to Q, which implies that

$$\sum_{j \in A(q_{\cdot i^\star})} p_{i^\star j} = 1,$$

and that $p_{i^\star j} = 0$ whenever $j \notin A(q_{\cdot i^\star})$. This means that even though some of the $p_{i^\star j}$ with $j \in A(q_{\cdot i^\star})$ might be zero, not all of them can be zero. At least one of them has to be positive—and if it is really the only one, it has to be exactly equal to 1. Since, by assumption, $A(q_{\cdot i^\star})$ has at least two elements, this implies that P has at least (i) two columns with multiple maximal elements strictly between 0 and 1, or (ii) a zero column, which proves the first part of the lemma.

Suppose that for $j^{\star\star}$ with $j^{\star\star} \in A(q_{\cdot i^\star})$, $p_{i^\star j^{\star\star}} \neq 0$. Note that we do not rule out the possibility that $p_{i^\star j^{\star\star}}$ is equal to 1. Since Q is a best response to P, it must be true that $\sum_{i \in A(p_{\cdot j^{\star\star}})} q_{j^{\star\star} i} = 1$. Remember, we know from above that $i^\star \in A(p_{\cdot j^{\star\star}})$. But since $0 < q_{j^{\star\star} i^\star} < 1$ (and since the $j^{\star\star}$–th column of P is not a zero column) there must be some $i^{\star\star} \neq i^\star$ with $i^{\star\star} \in A(p_{\cdot j^{\star\star}})$ such that $q_{j^{\star\star} i^{\star\star}} \neq 0$. Of course, since P is a best response to Q, $q_{j^{\star\star} i^{\star\star}}$ is a maximal element of the $i^{\star\star}$–th column of Q. In the case where $\max_i(p_{ij^{\star\star}}) = 1$, $q_{j^{\star\star} i^{\star\star}}$ might well be the unique maximal element of this column.

For later use note that

$$\sum_{j \in A(q_{\cdot i^\star})} \sum_i p_{ij} q_{ji} = 1.$$

To see why this is so, consider the following argument: We know from above that $\sum_{j \in A(q_{\cdot i^\star})} p_{i^\star j} = 1$ and that all these elements in the $i^\star$–th row of P for which $j \in A(q_{\cdot i^\star})$ are maximal elements of their respective columns; therefore

$$\sum_{j \in A(q_{\cdot i^\star})} \max_i p_{ij} = 1.$$

However, any Q that is a best response to P "extracts" from P exactly the sum of its column maxima, which gives the claim of the statement.

Now, we try to create an alternative Q' that is doing as well against P, as Q is doing against P; and an alternative P' that is doing as well against Q, as P is doing against Q.

Take as a candidate Q' the original Q but exchange the entries in its $j^{\star\star}$–th row by the vector

$$q'_{j^{\star\star} i} = \begin{cases} 1 & \text{for } i = i^{\star\star} \\ 0 & \text{otherwise} \end{cases},$$

and exchange its $j^\star$–th row by the vector

$$q'_{j^\star i} = \begin{cases} 1 & \text{for } i = i^\star \\ 0 & \text{otherwise} \end{cases},$$

where $j^\star$ is some $j \in A(q_{\cdot i^\star})$ with $j^\star \neq j^{\star\star}$. Note that, since $q_{j^\star i^\star} \neq 0$ and Q is a best response to P, this also implies that $p_{i^\star j^\star}$ is a maximal element of the $i^\star$–th column of P.

Since the original Q was a best response to P, and since in constructing Q' form Q we did not change any rows other than the $j^{\star\star}$–th and $j^\star$–th row of Q, all we have to do in order to show that $\mathrm{tr}(PQ') = \mathrm{tr}(PQ)$, is to show that the $j^{\star\star}$–th and the $j^\star$–th row of Q' successfully extract the maximum value of $j^{\star\star}$–th and the $j^\star$–th column of P, respectively. We know from above that $p_{i^{\star\star}j^{\star\star}}$ is a maximal element of the $j^{\star\star}$–th column of P, and that $p_{i^\star j^\star}$ is a maximal element of the $j^\star$–th column of P. So setting their corresponding elements in Q', $q_{j^{\star\star}i^{\star\star}}$ and $q_{j^\star i^\star}$, equal to 1, clearly is optimal in order to maximize $\mathrm{tr}(PQ)$, and so $\mathrm{tr}(PQ') = \mathrm{tr}(PQ)$.

As an alternative P' take the original P but exchange the entries in its $i^{\star\star}$–th row by the vector

$$p'_{i^{\star\star}j} = \begin{cases} 1 & \text{for } j = j^{\star\star} \\ 0 & \text{otherwise} \end{cases},$$

and the entries in its $i^\star$–th row by the vector

$$p'_{i^\star j} = \begin{cases} 1 & \text{for } j = j^\star \\ 0 & \text{otherwise} \end{cases}$$

We also know from above that $q_{j^\star i^\star}$ is a maximal element of the $i^\star$–th column of Q (by assumption definitely not its unique maximal element) and that $q_{j^{\star\star}i^{\star\star}}$ is a maximal element of the $i^{\star\star}$–th column of Q. As pointed out above, it even might be the unique maximal element of this column. In any case, setting $p_{i^\star j^\star}$ and $p_{i^{\star\star}j^{\star\star}}$ equal to 1 clearly is an optimal choice in order to maximize $\mathrm{tr}(PQ)$, and so $\mathrm{tr}(P'Q) = \mathrm{tr}(PQ)$.

What remains to be done, is to compare $\mathrm{tr}(P'Q')$ to $\mathrm{tr}(PQ)$. 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

$$\sum_{j \in A(q_{\cdot i^\star})} \sum_i p'_{ij} q'_{ji} \geq 2, \tag{28}$$

whereas, as we have noted above,

$$\sum_{j \in A(q_{\cdot i^\star})} \sum_i p_{ij} q_{ji} = 1.$$

We distinguish two cases now: Suppose first that for the $j^{\star\star} \in A(q_{\cdot i^\star})$ that we have chosen above, $p_{i^\star j^{\star\star}} = 1$. Then, of course, also $p_{i^{\star\star}j^{\star\star}} = 1$. But this implies that $p_{i^{\star\star}j} = 0$ for all $j \notin A(q_{\cdot i^\star})$, and so

$$\sum_{j \notin A(q_{\cdot i^\star})} \sum_i p'_{ij} q'_{ji} = \sum_{j \notin A(q_{\cdot i^\star})} \sum_i p_{ij} q_{ji}. \tag{29}$$

Summing over all j, we have that

$$\sum_j \sum_i p'_{ij} q'_{ji} \geq \sum_j \sum_i p_{ij} q_{ji} + 1,$$

and we are done.

The case where $0 < p_{i^\star j^{\star\star}} < 1$ is a little bit more complicated. Equation (28) still holds, but equation (29) is no longer necessarily true. It might well be that there are some $j \notin A(q_{\cdot i^\star})$ for which $p_{i^{\star\star} j} \neq 0$. Hence, in constructing P' form P, when we replace the $i^{\star\star}$ row in P by the vector $(p'_{i^{\star\star} j})$ that is 1 for $j = j^{\star\star}$ and 0 otherwise, it might well be the case that we nullify some positive entries $p_{i^{\star\star} j}$ for which $j \notin A(q_{\cdot i^\star})$ *that might be attributed some positive weight to by* Q! So when we multiply the elements $p'_{i^{\star\star} j}$ for $j \notin A(q_{\cdot i^\star})$ with their corresponding elements in Q'—which definitely are unchanged for $j \notin A(q_{\cdot i^\star})$—it might well be that we "lose" something as compared to the same expressions in $\mathrm{tr}(PQ)$. Nevertheless, since $p_{i^\star j^{\star\star}} \neq 0$, what we "lose" this way cannot be greater than 1, and so overall we still have that

$$\sum_j \sum_i p'_{ij} q'_{ji} > \sum_j \sum_i p_{ij} q_{ji},$$

that is,

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

and therefore the original Nash strategy (P, Q) cannot be a neutrally stable strategy. □

Proof of Lemma 7

Proof. Suppose that $(P, Q) \in \mathcal{P}^{\Delta}_{n \times m} \times \mathcal{Q}^{\Delta}_{m \times n}$ is a Nash strategy. We give the proof for the case where the condition of the proposition is satisfied for P; analogous reasoning holds true for the case where it is satisfied for Q.

In order to show that (P, Q) is a neutrally stable state, by Remark 2, we have to show that for *any* $P' \in B(Q)$ and $Q' \in B(P)$,

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

If P has no column with multiple maximal elements that are not equal to 1, then for every fixed column of P there are only three possible cases. Its maximum is either

(1) unique and equal to 1, or

(2) unique, but not equal to 1, or

(3) equal to 1, but not unique.

Note that, in particular, there is no zero column in P.

Starting with the assumptions on the columns of P—for each of these three cases in turn—we will first try to exploit all the information we can get about

the corresponding rows in Q and the other columns in P that derive from the fact that P and Q are best responses to each other. In the following step we will consider the consequences for the corresponding rows of *all possible* receiver matrices that are best responses to P, that is, $Q' \in B(P)$, and the corresponding columns of *all possible* sender matrices that are best responses to Q, that is, $P' \in B(Q)$. Multiplying columns with their corresponding rows, we will see that, for each of these three cases, these column–times–row products for P' and Q' are always smaller than or equal to their corresponding expressions for the original P and Q. Summing over all these products finally yields the result.

Case 1. (One event exclusively linked to one signal.) 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, $q_{j^\star i^\star} = 1$, and $q_{j^\star i} = 0$ whenever $i \neq i^\star$. Note that, since there are no entries greater than 1, this immediately implies that $q_{j^\star i^\star}$ is *a* maximal element of the $i^\star$–th column of Q.

Since the elements in each row of P add up at most to 1, we also have that $p_{i^\star j} = 0$ whenever $j \neq j^\star$. Since by assumption there are no columns in P that consist entirely of zeros, all these elements in the $i^\star$-th *row* of P that are equal to 0, cannot be maximal elements of their respective *columns* $j \neq j^\star$. Since Q is a best response to P, $q_{ji^\star} = 0$ whenever $j \neq j^\star$. So,

$$q_{ji^\star} = \begin{cases} 1 & \text{for } j = j^\star \\ 0 & \text{otherwise} \end{cases} . \tag{30}$$

This means that $q_{j^\star i^\star} = 1$ is not only *a* but *the unique* maximal element in the $i^\star$–th *column* of Q.

Now, we turn to $Q' \in B(P)$ and $P' \in B(Q)$. Since by assumption, $p_{i^\star j^\star}$ is the unique maximal element in the $j^\star$–th column of P, we have that

$$q'_{j^\star i^\star} = 1 = q_{j^\star i^\star} \text{ and } q'_{j^\star i} = 0 = q_{j^\star i} \, \forall \, i \neq i^\star, \tag{31}$$

for all $Q' \in B(P)$.

Since in this case, we also have that $q_{j^\star i^\star} = 1$ is the unique maximal element in the $i^\star$–th column of Q, Lemma 1 also tells us that

$$p'_{i^\star j^\star} = 1 = p_{i^\star j^\star}, \tag{32}$$

for all $P' \in B(Q)$.

Taking (31) and (32) together, we have that

$$\sum_i p'_{ij^\star} q'_{j^\star i} = 1 = \sum_i p_{ij^\star} q_{j^\star i}, \tag{33}$$

for all (P', Q') such that $P' \in B(Q)$ and $Q' \in B(P)$.

Figure 1.1 illustrates this case.

Case 2. (Synonymy.) Suppose that $0 < p_{i^\star j^\star} < 1$ is the unique maximal element in the $j^\star$–th column of P.

As in the previous case, from Q being a best response to P, we have that $q_{j^\star i^\star} = 1$, and that $q_{j^\star i} = 0$ for all $i \neq i^\star$. Since there are no elements greater than 1, this again implies that $q_{j^\star i^\star}$ is *a* maximal element of the $i^\star$–th *column* of Q. But now, since $p_{i^\star j^\star} \neq 1$, by Lemma 1, the fact that P is a best response to Q, implies that $q_{j^\star i^\star}$ cannot be the *unique* maximal element in the $i^\star$–th column of Q, and, of course, since the elements in each row may not add up to something greater than 1, we have that *for all* $j \in A(q_{\cdot i^\star})$, $q_{ji} = 0$ whenever $i \neq i^\star$. But since the maximum of the $i^\star$–th column of Q is not equal to zero, by Lemma 1 we also have that $\sum_{j \in A(q_{\cdot i^\star})} p_{i^\star j} = 1$, and that $p_{i^\star j} = 0$ for all $j \notin A(q_{\cdot i^\star})$. Together with the assumption that $0 < p_{i^\star j^\star} < 1$, this in turn implies that for all $j \in A(q_{\cdot i^\star})$, $p_{ji^\star} \neq 1$. On the other hand, since for all $j \in A(q_{\cdot i^\star})$, $q_{ji^\star} = 1 \neq 0$ and Q is a best response to P, by Lemma 1, we also know that for all $j \in A(q_{\cdot i^\star})$, $p_{ji^\star}$ is *a* maximal element of its respective column in P. Together with the assumption that P has no zero column and no column with multiple maximal elements strictly between 0 and 1, this implies that *for all* $j \in A(q_{\cdot i^\star})$, $0 < p_{i^\star j} < 1$ is *the unique* maximal element of its respective column in P.

In perfect analogy to the previous case, the fact that $p_{i^\star j} = 0$ for all $j \notin A(q_{\cdot i^\star})$ together with the assumption that P does not contain any zero column implies that $q_{ji^\star} = 0$ whenever $j \notin A(q_{\cdot i^\star})$. So,

$$q_{ji^\star} = \begin{cases} 1 & \text{for } j \in A(q_{\cdot i^\star}) \\ 0 & \text{otherwise} \end{cases} . \tag{34}$$

We now turn again to Q' and P'. Since *for all* $j \in A(q_{\cdot i^\star})$ it is true that $0 < p_{i^\star j} < 1$ is *the unique* maximal element of its respective column, by Lemma 1 we have that that *for all* $j \in A(q_{\cdot i^\star})$

$$q'_{ji^\star} = 1 = q_{ji^\star}, \text{ and } q'_{ji} = 0 = q_{ji} \; \forall \, i \neq i^\star, \tag{35}$$

for all $Q' \in B(P)$. On the other hand, the fact that the $i^\star$–th column of Q has multiple maximal elements implies that

$$\sum_{j \in A(q_{\cdot i^\star})} p'_{i^\star j} = 1 = \sum_{j \in A(q_{\cdot i^\star})} p_{i^\star j}, \tag{36}$$

for all $P' \in B(Q)$.

Putting (35) and (36) together we have that

$$\sum_{j \in A(q_{\cdot i^\star})} \sum_i p'_{ij} q'_{ji} = \sum_{j \in A(q_{\cdot i^\star})} \sum_i p_{ij} q_{ji} = 1, \tag{37}$$

for all (P', Q') such that $P' \in B(Q)$ and $Q' \in B(P)$.

This case is illustrated by Figure 1.2.

Case 3. (Homonymy.) Suppose that $p_{i^\star j^\star}$ is equal to 1, but not the unique maximal element in the $j^\star$-th column of P. So, $i^\star \in A(p_{\cdot j^\star})$, but there is at least one $i^{\star\star} \neq i^\star$ such that $i \in A(p_{\cdot j^\star})$.

In this case, from Q being a best response to P, by Lemma 1 we only have that $\sum_{i \in A(p_{\cdot j^\star})} q_{j^\star i} = 1$, and that $q_{j^\star i} = 0$ whenever $i \notin A(p_{\cdot j^\star})$. In particular, this does not imply that $q_{j^\star i} \neq 0$ for all $i \in A(p_{\cdot j^\star})$.

On the other hand, from the constraint that the elements in each row cannot add up to something greater than 1, we also have that *for all* $i \in A(p_{\cdot j^\star})$, $p_{ij} = 0$ whenever $j \neq j^\star$. Since by assumption P does not contain any column that consist entirely of zeros, any zero element in P can never be a maximal element of its respective column, and since Q is a best response to P, we have that *for all* $i \in A(p_{\cdot j^\star})$, $q_{ji} = 0$ for all $j \neq j^\star$. This implies that all the $q_{j^\star i}$ such $i \in A(p_{\cdot j^\star})$ are *maximal* elements of their respective *columns*. But since not all them are necessarily non–zero, this means that they are not necessarily the unique maximal element of this column, and there might be a zero column in Q. But, if $q_{j^\star i}$ with $i \in A(p_{\cdot j^\star})$ is equal to some positive value, then it definitely will be *the unique* maximal element of its column in P. So,

$$q_{ji} = \begin{cases} \max_j(q_{ji}) & \text{for } j = j^\star \\ 0 & \text{otherwise} \end{cases} . \tag{38}$$

We now turn to Q' and P'. By Lemma 1 we have that

$$\sum_{i \in A(p_{\cdot j^\star})} q'_{j^\star i} = 1 = \sum_{i \in A(p_{\cdot j^\star})} q_{j^\star i}, \text{ and } q'_{j^\star i} = 0 = q_{j^\star i} \; \forall \, i \notin A(p_{\cdot j^\star}), \tag{39}$$

for all $Q' \in B(P)$.

The case of $P' \in B(Q)$ is a little bit more complicated. As we have seen above, whenever $q_{j^\star i} \neq 0$ for some $i \in A(p_{\cdot j^\star})$, then it definitely will be the unique maximal element of its respective column in Q. By Lemma 1, its corresponding element in P' then *has to be* equal to 1. But if for some $i \in A(p_{\cdot j^\star})$, $q_{j^\star i} = 0$, then $p'_{ij^\star}$ *does not have to be* equal to 1—even though it can be equal to 1 or to some other positive value. So, if for some $i \in A(p_{ij^\star})$,

$$p'_{ij^\star} \neq 0 \Rightarrow p_{ij^\star} = 1. \tag{40}$$

Taking (39) and (40) together, we have that

$$\sum_i p'_{ij^\star} q'_{j^\star i} \leq \sum_i p_{ij^\star} q_{j^\star i} = 1, \tag{41}$$

for all (P', Q') such that $P' \in B(Q)$ and $Q' \in B(P)$.

Figure 1.3 illustrates this case.

So, whatever the cases out of these three possible cases that might be captured by any sender matrix P that is part of a Nash strategy $(P, Q) \in \mathcal{P}^{\Delta}_{n\times m} \times \mathcal{Q}^{\Delta}_{m\times n}$, we have that

(i) the maximum of each column in Q is either *unique* or *equal to* 1; and

(ii) summing *over all* j and over all i, we see that

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

which means that (P, Q) is a neutrally stable strategy.

Just note that (i) comes from (30), (38) and (34), and (ii) from (33), (41) and (37) together. □

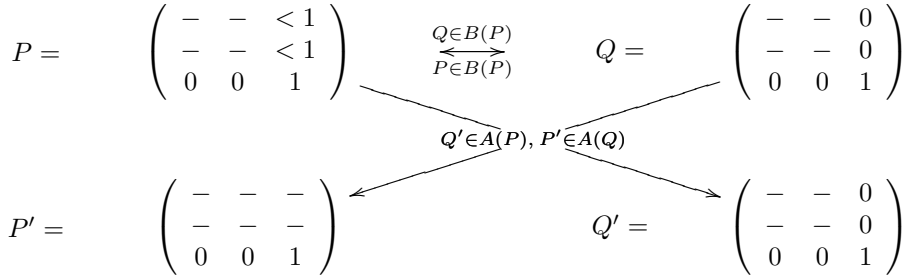

Figure 1.1. One event exclusively linked to one signal.

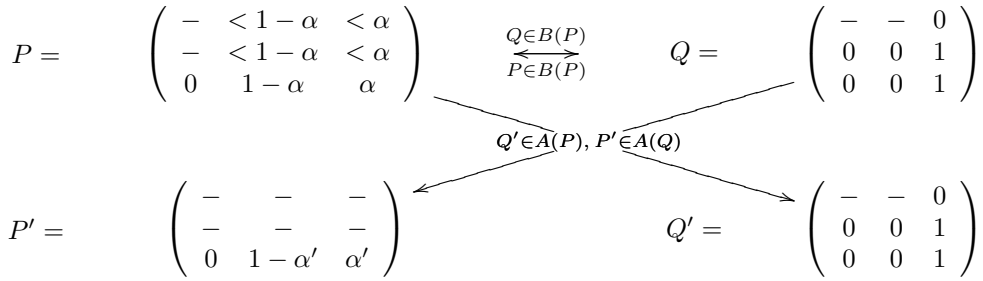

Figure 1.2. Synonymy.

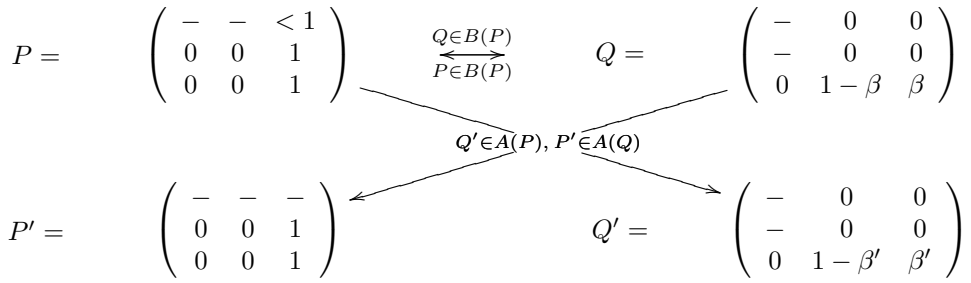

Figure 1.3. Homonymy.

Figure 1. Proof of Proposition 7. For all three cases in turn, we always apply the following basic line of reasoning. First, exploit the fact that P and Q are best responses to each other. Second, consider the properties of *any* P' and Q' that are best responses to Q and respectively P. And, finally, show that the trace of any $P' \in B(Q)$ times any $Q' \in B(P)$ can never be strictly greater than the trace of the original P times the original Q.

WORKING PAPERS

Christina PAWLOWITSCH

Why evolution does not always lead to an optimal proto-language.
An approach based on the replicator dynamics

June 2006

Working Paper No: 0604

DEPARTMENT OF ECONOMICS

UNIVERSITY OF VIENNA

All our working papers are available at: http://mailbox.univie.ac.at/papers.econ