Neutral stability, drift, and the diversification of languages

Christina Pawlowitsch*

Paris School of Economics, France, and Department of Economics, University of Vienna, Austria

Panayotis Mertikopoulos

Department of Economics, École Polytechnique, France

Nikolaus Ritt

Department of English, University of Vienna, Austria

Abstract

The diversification of languages is one of the most interesting facts about language that seek explanation from an evolutionary point of view. Conceptually the question is related to explaining mechanisms of speciation. An argument that prominently figures in evolutionary accounts of language diversification is that it serves the formation of group markers which help to enhance in-group cooperation. In this paper we use the theory of evolutionary games to show that language diversification on the level of the meaning of lexical items can come about in a perfectly cooperative world solely as a result of the effects of *frequency-dependent selection*. Importantly, our argument does not rely on some stipulated function of language diversification in some coevolutionary process, but comes about as an endogenous feature of the model. The model that we propose is an evolutionary language game in the style of Nowak et al. [1999, The evolutionary language game. J. Theor. Biol. 200, 147–162], which has been used to explain the rise of a signaling system or protolanguage from a prelinguistic environment. Our analysis focuses on the existence of *neutrally stable polymorphisms* in this model, where, on the level of the population, a signal can be used for more than one concept or a concept can be inferred by more than one signal. Specifically, such states *cannot* be invaded by a mutation for bidirectionality, that is, a mutation that tries to resolve the existing ambiguity by linking each concept to exactly one signal in a bijective way. However, such states are not resistant against *drift* between the selectively neutral variants that are present in such a state. Neutral drift can be a pathway for a mutation for bidirectionality that was blocked before but that finally will take over the population. Different directions of neutral drift open the door for a mutation for bidirectionality to appear on different resident types. This mechanism—which can be seen as a form of *shifting balance*—can explain why a word can acquire a different meaning in two languages that go back to the same common ancestral language, thereby contributing to the splitting of these two languages. Examples from currently spoken languages, for instance, English *clean* and its German cognate *klein* with the meaning of "small," are provided.

Keywords: Language evolution, evolutionary language game, neutrally stable polymorphisms, shifting balance.

1. Introduction

Language is our legacy, language is what makes us uniquely human. And yet we can communicate effectively only with those of our conspecifics who have grown up in the same linguistic community, typically the same geographical region. There are at present about 7,000 languages spoken in the world (Lewis, 2009). Languages differ at all levels of linguistic expression: the lexicon, morphology, phonology, syntax, semantics. From an evolutionary point of view, the differentiation and diversification of languages is one of the most interesting facts about language that seek explanation (see, for example, Hurford, 2003 or the recent target article and debate in Evans and Levinson, 2009). Conceptually, the question is related to explaining mechanisms of *speciation* (for an overview of models of speciation see, for example, Gavrilets, 2004).

Evolutionary accounts of the origin of language typically evoke the communicative *function* of language—language helps us to exchange information about the world, enhances cooperation, and thereby increases fitness. It has been argued that functionalist-adaptionist approaches to language make it difficult to account for the fact that natural human languages have always tended to diversify into dialects and eventually split into

---
*Corresponding author.
Email address: christina.pawlowitsch@ens.fr (Christina Pawlowitsch)

separate and mutually unintelligible languages.[1]

An argument that has been advanced in an evolutionary context to explain language diversification is that it serves the formation of group markers which can be exploited to enhance in-group cooperation (for example, Dunbar, 1998). Ultimately this argument boils down to postulating that there is a function, or if one wishes, a preference for diversity.

In this paper we use the theory of evolutionary games to show that language diversification on the level of the meaning of lexical items can come about in a perfectly cooperative world—a world where everybody wants to cooperate with everybody—solely as a result of the effects of *frequency-dependent selection*. Importantly, our argument does not rely on some stipulated function of language diversification, but comes about as an endogenous feature of the model.

1.1. Language change

Language change, like biological evolution, can be understood as a process of descent with modification. While it affects all levels of linguistic description (sound, grammar, lexicon, meaning), a good way of appreciating its effects is to look at word pairs such as Modern English *clean* and Modern German *klein*: systematic correspondences between their sound structures prove that they derive from a common ancestor in what once must have been a single language (in this case West Germanic). Yet English *clean* means "clean" while German *klein* means "small." Thus, either one or both words must have adopted a new meaning and thereby contributed to splitting West Germanic into English and German. Table 1.1 contains some more well documented examples of cognates in Modern English and Modern German that exhibit a shift in meaning.

While there is a huge discussion among linguists about what it is that makes a language, and what it is that all languages have in common (see, for example, Evans and Levinson 2009), most linguistic theories, if not explicitly so tacitly, postulate form-meaning correspondences (lexical mappings in the narrow sense but also form-meaning correspondences that encode a grammatical feature like markers of tense, mood, etc.) as a basic building block of language. A major reason why linguistic change is notoriously difficult to account for is that it affects both *changes of form* (for example, changes in the sound shape of lexical items) as well as *changes in meaning* (which form is mapped to which concept). In this paper, we will focus on change in the meaning of lexical items while abstracting

English	German
dish	*Tisch* ("table")
knave	*Knabe* ("boy")
knight	*Knecht* ("servant")
tide	*Zeit* ("time")
town	*Zaun* ("fence")
to starve	*sterben* ("to die")
to worry	*würgen* ("to retch")
to reckon	*rechnen* ("to calculate")
clean	*klein* ("small")
silly	*selig* ("blessed")
true	*treu* ("faithful")

Table 1.1: Cognates in Modern English and Modern German exhibiting a shift in meaning

from changes in the form of lexical items. For the example given above this means that we look at *clean* and *klein* as two instances of the same form, and wish to account for the fact that they acquire a different meaning in two languages that go back to a common ancestral language.

1.2. Language games

The model that we present is an *evolutionary language game* in the style of Nowak et al. (1999a), which has been proposed as a model for the evolution of a signaling system or protolanguage, that is, a collection of form-meaning correspondences (see also, Nowak and Krakauer, 1999; Trapa and Nowak, 2000; Komarova and Nowak, 2001; Nowak et al., 2002; Komarova and Niyogi, 2004). Evolutionary game theory (Maynard Smith and Price, 1973; Maynard Smith, 1982; Hofbauer and Sigmund, 1988, 1998; Weibull, 1995; Cressman, 2003; Nowak, 2006; Sandholm, 2011) provides a formal framework for studying frequency-dependent selection. Language is a typical case where the performance or fitness of a type depends on the frequencies of the other types present in the population; it therefore naturally lends itself to an analysis in terms of evolutionary games.

In the Nowak et al. language game the evolving entities—*strategies*—are lexical mappings. More precisely, a strategy is a pair of two mappings: a mapping from the set of concepts to the set of available signals (a strategy in the role of the sender), and a mapping from signals to concepts (a strategy in the role of the receiver). Signals are of no cost and the concepts to be potentially communicated are of no differential weight. There is a homogeneous population of individuals with perfectly coinciding interests, and whenever two individuals correctly communicate a concept, this will give both of them a positive payoff which translates into an incremental fitness advantage. Similar formulations of this model can be found in Lewis (1969)—see also Skyrms (1996, 2002)—and Hurford (1989); extensions have been studied in Nowak et al. (1999b), Donaldson et al. (2007), Jäger (2008), and Hofbauer and Huttegger (2008).

---

[1]See, for example, Piattelli-Palmarini (2000), who expresses this critique most explicitly: "Different human communities speak different and, for the most part, not mutually understandable languages. This fact is a mighty challenge to all naive functionalist and adaptionist explanations of the origins and structure of language. Had language been the result of the need to communicate, then linguistic diversity should not have been possible."

Provided that there is the same number of signals as there are concepts to be potentially communicated, an *optimum signaling system* or *optimum protolanguage* is a pair of mappings such that each concept is bijectively linked to one signal and vice versa, and an optimum in the population will be attained if one such signaling system has become fixed in the population.

In Lewis (1969), one can find the idea that some kind of trial-and-error process that operates in a population of agents will lead to the emergence of such an optimum signaling system. Lewis—who writes just before the advent of evolutionary game theory—motivates this by the "salient" character of these strategies. Later, when Lewis' model has been taken up under the use of methods which in the meantime had been introduced by evolutionary game theory, it has been shown that there is indeed a formal foundation for the selection of optimum signaling systems: optimum signaling systems are the only *evolutionarily stable strategies* in this game (Wärneryd, 1993; see also Trapa and Nowak, 2000). However, computer experiments with this model have given rise to the conjecture that some form of suboptimality—as expressed by one signal being used for more than one concept or one concept being inferred by more than one signal—can have some form of evolutionary stability (see, for example, Nowak and Krakauer, 1999). More recently it has been shown analytically that for this game some well-defined evolutionary dynamics, most importantly the *replicator dynamics* (Taylor and Jonker, 1978), will indeed *not* almost always converge to an optimum signaling system, but instead can lead to suboptimum states where on the level of the population's average strategy—the idealized "language" of the population—two or more concepts are linked to the same signal, or where two or more signals are linked to the same concept (Huttegger, 2007; Pawlowitsch, 2008). While such states are not *evolutionarily stable* in the strict sense as defined by Maynard Smith and Price (1973), they do satisfy a weaker version of this notion known as *neutral stability* or *weak evolutionary stability* (Maynard Smith, 1982; Thomas, 1985). Neutrally stable states are Lyapunov stable in the replicator dynamics (Bomze and Weibull, 1995), which is why the replicator dynamics can be blocked in these suboptimum states.

Ensuing research on the Lewis-Hurford-Nowak language game has to a good part focused on the question whether some other dynamic processes, or perturbations of the replicator dynamics, will or will not lead to the rise of an optimal signaling system (for an overview of this literature, see, for example, Huttegger and Zollman 2011). What has received much less attention so far—but which, in our mind, leaves a number of questions to be investigated from a linguistic point of view—is the fact that neutral stability supports *polymorphic states* where different types resolve the ambiguity in concept-to-signal or signal-to-concept mappings that appears on the level of the population in different ways. We consider this an interesting property of the model since language change,

like biological evolution, essentially thrives on *variation in the population*.

In this paper, we take a closer look at the specific form of variation that can persist in a neutrally stable state, and we will show that the variation sustained by neutral stability is rich enough to account for the branching of languages. We will consider *neutral drift*—that is, a random shift in the relative type frequencies—among the variants that can coexist in a neutrally stable state, and we will see that this can be a pathway for mutations that so far have been blocked. But different directions of neutral drift may open the door to different mutations, which will eventually lead to different long-run outcomes. It is in tracing these different evolutionary paths that we will encounter the phenomenon that the meaning of a signal may shift, or switch, between two populations that go back to the same common ancestral population. Formally, the mechanism that we describe can be seen as a particular case of *shifting balance* (Wright, 1931) where different fitness peaks can be reached by drift along (locally stable) ridges of high, but not globally maximal fitness (see, for example, Gavrilets and Hastings, 1996).

2. The model

There are n concepts that potentially become the object of communication, and there are m signals (words or morphemes that encode a grammatical feature) that are available to individual agents. We assume that by its very nature no signal is any more or less "fit" to represent a particular concept. In other words, signals are of no differential costs, which we will express formally by assuming that signals are of no cost at all. In particular, this implies that the cost of a signal does not depend on the state of the world, so that observation of a particular signal would not reveal any information about the state of the world. In this sense, signals are "arbitrary."

We aim at modeling certain aspects of natural language. In doing so we make a very broad assumption about the cooperative nature of language: We assume that there is a homogeneous population, where (*i*) over their lifetimes, individuals randomly and repeatedly engage in potential communication over all possible concepts with everybody else in the population, (*ii*) the sender and the receiver benefit from successful communication in equal terms, and (*iii*) individuals appear in the role of the sender or the receiver with equal probabilities.

A (pure) *strategy* for an individual in the role of the sender is a mapping from potential objects of communication to available signals. We represent this by a matrix

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

where p_{ij} is either 0 or 1, and there is exactly one 1 in each row of P—the interpretation being that if $p_{ij} = 1$, then concept i is mapped to signal j. That is, if this individual wants to communicate concept i, he or she will use signal j. Formally,

$$P \in \mathcal{P}_{n \times m} = \big\{ P \in \mathbb{R}_+^{n \times m} : \forall i, \, p_{ij} = 1 \text{ for some } j \equiv j(i)$$
$$\text{and } p_{ij} = 0 \text{ if } j \neq j(i) \big\}. \quad (2)$$

Similarly, a strategy for an individual in the role of the receiver is a mapping from potentially received signals to objects, which we represent by a matrix

$$Q = \begin{pmatrix} q_{11} & \cdots & q_{1i} & \cdots & q_{1n} \\ \vdots & & \vdots & & \\ q_{j1} & \cdots & q_{ji} & \cdots & q_{jn} \\ \vdots & & \vdots & & \\ q_{m1} & \cdots & q_{mi} & \cdots & q_{mn} \end{pmatrix}, \quad (3)$$

where q_{ji} is either 0 or 1, and there is exactly one 1 in every row of Q—the interpretation being that if $q_{ji} = 1$, then signal j is mapped to concept i. That is, if this individual receives signal j, then he or she will link it mentally to concept i. Formally,

$$Q \in \mathcal{Q}_{m \times n} = \big\{ Q \in \mathbb{R}_+^{m \times n} : \forall j, \, q_{ji} = 1 \text{ for some } i \equiv i(j)$$
$$\text{and } q_{ji} = 0 \text{ if } i \neq i(j) \big\}. \quad (4)$$

For given m and n, there are m^n P-matrices and n^m Q-matrices. Note that the restrictions on P and Q do not preclude that, in the role of the sender, there can be a signal that is used for more than one concept, or in the role of the receiver, a concept that is associated with more than one signal: there can be more than one 1 in a column of P, or respectively Q.

If a sender who uses strategy P interacts with a receiver who uses strategy Q, then a specific concept, say $i^\star$, will be correctly communicated between these two if there is a signal $j^\star$ such that $p_{i^\star j^\star} = 1 = q_{j^\star i^\star}$. We take the sum of all correctly communicated concepts between a sender P and a receiver Q as a measure for the *communicative potential* between P and Q (we adopt this terminology from Hurford 1989). In the notation that we use here, the communicative potential between P and Q can be written as

$$\pi(P, Q) = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} q_{ji} = \mathrm{tr}(PQ) \quad (5)$$

We identify the communicative potential with the *payoff* that both the sender and the receiver get out of their interaction. The payoff functions $\pi_1(P, Q) = \pi(P, Q)$ and $\pi_2(P, Q) = \pi(P, Q)$ together with the strategy sets $\mathcal{P}_{n \times m}$ and $\mathcal{Q}_{m \times n}$ define an asymmetric game (with common interests).

We look at the symmetrization of this game, where an individual adopts the role of a sender or of a receiver with equal probabilities. A strategy for an individual then is a

$$\pi(P, Q) = p_{11} q_{11} + \cdots + p_{1j} q_{j1} + \cdots + p_{1m} q_{m1}$$
$$\cdots$$
$$+ p_{i1} q_{1i} + \cdots + p_{ij} q_{ji} + \cdots + p_{im} q_{mi}$$
$$\cdots$$
$$+ p_{n1} q_{1n} + \cdots + p_{nj} q_{jn} + \cdots + p_{nm} q_{mn}$$
$$= \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} q_{ji} = \mathrm{tr}(PQ)$$

Figure 1: Sender matrix P and receiver matrix Q (c_i stands for concept i, s_j for signal j, etc.), and the communicative potential $\pi(P, Q)$.

pair of a sender and a receiver matrix $(P, Q) \in \mathcal{P}_{n \times m} \times \mathcal{Q}_{m \times n}$, and the *payoff* of strategy (P_k, Q_k) from interaction with (P_l, Q_l) is given by

$$f[(P_k, Q_k), (P_l, Q_l)] = \frac{1}{2} \big[\pi(P_k, Q_l) + \pi(P_l, Q_k) \big]. \quad (6)$$

Note that for fixed n and m, there are $N = m^n \times n^m$ such "pure strategies." Note also that $f[(P_k, Q_k), (P_l, Q_l)] = f[(P_l, Q_l), (P_k, Q_k)]$, that is, the payoff function is symmetric; in other words, the payoff that (P_k, Q_k) gets out of interaction with (P_l, Q_l) is the same as the payoff that (P_l, Q_l) gets out of interaction with (P_k, Q_k). Symmetric games with a symmetric payoff function are sometimes called *doubly symmetric games*. In our case, this property is, of course, a consequence of the identity of payoffs in the underlying asymmetric game and the symmetry of weights for the two roles (for more on symmetrized asymmetric games, in particular on their dynamic properties, see, Cressman, 2003).

2.1. The classical case of an infinitely large population

We take the symmetrized game in pure strategies as the base game of a *population game* that is played in an infinitely large population (the basic model in evolutionary game theory; see, for example, Hofbauer and Sigmund, 1998; Weibull, 1995; or Cressman, 2003). With every strategy $(P, Q) \in \mathcal{P}_{n \times m} \times \mathcal{Q}_{m \times n}$ we identify a particular *type of player* and we represent a *state of the population* by a vector

$$x = (x_1, \ldots, x_l, \ldots, x_N), \quad \sum_{l=1}^{N} x_l = 1, \quad (7)$$

where x_l is the relative frequency of type (P_l, Q_l). To every vector of type frequencies x we can assign the population's average strategy (P_x, Q_x), where $P_x = \sum_l x_l P_l$ is the population's average sender matrix, and $Q_x = \sum_l x_l Q_l$

the population's average receiver matrix. P_x will then be a row-stochastic matrix of dimensions $n \times m$, that is,

$$P_x \in \mathcal{M}_{n \times m} = \left\{ M \in \mathbb{R}_+^{n \times m} : \sum_j m_{ij} = 1, \forall\, i. \right\}, \quad (8)$$

and Q_x a row-stochastic matrix of dimensions $m \times n$,

$$Q_x \in \mathcal{M}_{m \times n} = \left\{ M \in \mathbb{R}_+^{m \times n} : \sum_i m_{ji} = 1, \forall\, j. \right\}. \quad (9)$$

Note that $\mathcal{M}_{n \times m}$ is indeed spanned by $\mathcal{P}_{n \times m}$, that is, every element in $\mathcal{M}_{n \times m}$ can be represented by a convex combination of elements in $\mathcal{P}_{n \times m}$ (possibly not unique); and $\mathcal{M}_{m \times n}$ is spanned by $\mathcal{Q}_{m \times n}$.[2]

The *fitness of type l* is the average payoff that a type l individual gets from interaction with all other types present in the population proportional to their type frequencies, $f_l(x) = \sum_k x_k f[(P_l, Q_l), (P_k, Q_k)]$. This can be written as the payoff of type l from play against the population's average strategy,

$$f_l(x) = f\left[(P_l, Q_l), (P_x, Q_x) \right]$$
$$= \frac{1}{2} \left[\pi\left(P_l, Q_x\right) + \pi\left(P_x, Q_l\right) \right]. \quad (10)$$

The *average fitness in the population*, $\bar{f} = \sum_l x_l f_l(x)$, can be written as the payoff of the population's average strategy from play against itself,

$$\bar{f}(x) = f\left[(P_x, Q_x), (P_x, Q_x) \right] = \pi\left(P_x, Q_x\right). \quad (11)$$

We call $\pi\left(P_x, Q_x\right)$ the *eigen communicative potential* of a "language" (P_x, Q_x).

Example 1. Let $n = m = 3$ and suppose that there are only two types present in the population,

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

and

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

and let the corresponding type frequencies be $x_1 = 0.75$ and $x_2 = 0.25$. Then the population's average strategy is

$$(P_x, Q_x) = \left[\begin{pmatrix} .75 & .25 & 0 \\ .25 & .75 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} .75 & .25 & 0 \\ .25 & .75 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right].$$

The fitness of type 1 is $f_1(x) = 2.5$, and the fitness of type 2 is $f_2(x) = 1.5$. The fitness of a type clearly depends on the relative frequencies of all types. Consider a different vector of type frequencies, for example, x', where $x_1' = 0.5$ and $x_2' = 0.5$. Then the population's average strategy is

$$(P_{x'}, Q_{x'}) = \left[\begin{pmatrix} .5 & .5 & 0 \\ .5 & .5 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} .5 & .5 & 0 \\ .5 & .5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right];$$

$f_1(x') = 2$, and the fitness of type 2, $f_2(x')$, is equally 2.

The infinite-population scenario is conceptually intimately linked to the *replicator dynamics*, a simple feedback process where the frequency of a type grows proportionally to its fitness difference relative to the average fitness in the population (Taylor and Jonker 1978, Hofbauer et al. 1979). In our model,

$$\dot{x}_l = x_l[f[(P_l, Q_l), (P_x, Q_x)] - f[(P_x, Q_x), (P_x, Q_x)]]. \quad (12)$$

Usually this dynamics is interpreted in terms of biological evolution, but it can also be interpreted in terms of cultural evolution or learning—for example, it can be derived from a process where individuals imitate strategies that do better than their current strategy (Schlag, 1998; see also Traulsen et al., 2005 and Sandholm, 2011). A rest point of the replicator dynamics is a state where all resident types attain the same fitness. Such a state is called a *population equilibrium*. In the Example 1 above, the state where $x_1' = 0.5$ and $x_2' = 0.5$ is a population equilibrium.

2.2. Evolutionary stability

A characteristic of the present model is that it has many equilibria; in fact infinitely many. However, not all of these satisfy the same stability properties. A strategy $(P_x, Q_x) \in \mathcal{M}_{n \times m} \times \mathcal{M}_{m \times n}$ is an *evolutionarily stable strategy* (ESS) in the sense of Maynard Smith and Price (1973) if

(i) $f[(P_x, Q_x), (P_x, Q_x)] \geq f[(P, Q), (P_x, Q_x)]$ for all $(P, Q) \in \mathcal{M}_{n \times m} \times \mathcal{M}_{m \times n}$; and

(ii) whenever (i) holds with equality for some $(P, Q) \in \mathcal{M}_{n \times m} \times \mathcal{M}_{m \times n}$ with $P \neq P_x$ or $Q \neq Q_x$, then

$$f[(P_x, Q_x), (P, Q)] > f[(P, Q), (P, Q)]. \quad (13)$$

The first condition states that (P_x, Q_x) has to be a best response to itself—the condition for a symmetric *Nash equilibrium*. The second condition states that whenever there is an alternative best response (P, Q) to the original Nash-equilibrium strategy (P_x, Q_x), then this alternative best response has to yield a strictly lower payoff against itself than the original Nash-equilibrium strategy yields against the alternative best response.

For the game discussed here, these conditions can be simplified. Condition (i) holds if and only if:

$$\pi(P_x, Q_x) \geq \pi(P_x, Q), \quad \text{for all } P \in \mathcal{M}_{n \times m} \quad (14a)$$

[2]In the papers by Nowak et al. the game is defined right away on $\mathcal{M}_{n \times m} \times \mathcal{M}_{m \times n}$ as the strategy space. Here we build the game explicitly from a model with a finite number of types, in order to have a proper framework for defining the standard replicator dynamics on this game and connecting the evolutionary-stability analysis to the analysis of this dynamics.

and

$$\pi(P_x, Q_x) \geq \pi(P, Q_x), \quad \text{for all } Q \in \mathcal{M}_{m \times n}. \qquad (14b)$$

That is, Q_x has to be a best response to P_x, and P_x has to be a best response to Q_x.[3] The following remark characterizes best responses in terms of properties of the P and Q matrices.

Remark 1 (Best-response properties of P and Q). (1) Consider a fixed $\bar{P} \in \mathcal{M}_{n \times m}$. Then

(1.a) $\bar{Q} \in \mathcal{M}_{m \times n}$ is a *best response to $\bar{P}$ in $\mathcal{M}_{m \times n}$* (that is, an argument Q that maximizes $\pi(\bar{P}, Q)$ in $\mathcal{M}_{m \times n}$) if and only if for all j, $\sum_{i'} \bar{q}_{ji'} = 1$, where $i' \in \text{argmax}_i(\bar{p}_{ij})$; and

(1.b) $\max_Q (\pi(\bar{P}, Q)) = \sum_j \max_i(\bar{p}_{ij})$.

(2) Similarly, for fixed $\bar{Q} \in \mathcal{M}_{m \times n}$,

(2.a) $\bar{P} \in \mathcal{M}_{n \times m}$ is a *best response to $\bar{Q}$ in $\mathcal{M}_{n \times m}$* (that is, an argument P that maximizes $\pi(P, \bar{Q})$ in $\mathcal{M}_{n \times m}$) if and only if for all i, $\sum_{j'} \bar{p}_{ij'} = 1$, where $j' \in \text{argmax}_j(\bar{q}_{ji})$; and

(2.b) $\max_P (\pi(P, \bar{Q})) = \sum_i \max_j (\bar{q}_{ji})$.

(1.a) tells us that a receiver who responds optimally to a given sender matrix $\bar{P}$ will infer concept i from signal j if and only if this is one of the concepts that have maximal probability of being meant by signal j. (1.b) tells us that for given $\bar{P} \in \mathcal{M}_{n \times m}$ the communicative potential $\pi(\bar{P}, Q)$ is bounded by the sum of the column maxima in $\bar{P}$. Similarly for (2.a) and (2.b). A more detailed discussion of these conditions, including a proof, can be found in Pawlowitsch (2008).

Condition (ii), in (13) above, is equivalent to requiring that *if* there is a $P \in \mathcal{M}_{n \times m}$ that is an alternative best response to Q_x and a $Q \in \mathcal{M}_{m \times n}$ that is an alternative best response to P_x, with $P \neq P_x$ or $Q \neq Q_x$, then

$$\pi(P_x, Q_x) > \pi(P, Q). \qquad (15)$$

That is, the eigen communicative potential of the original Nash-equilibrium strategy (P_x, Q_x) has to be higher than the communicative potential of any pair of alternative best responses.[4]

Example 1 continued. With conditions (14) and (15) at hand, together with the best-response properties of the

P and Q matrices (Remark 1), it is easy to see that in Example 1 above, the state $(x_1', x_2') = (0.5, 0.5)$ corresponds to a Nash-equilibrium strategy, but is not evolutionarily stable: (i) $P_{x'}$ is a best response to $Q_{x'}$ and $Q_{x'}$ is a best response to $P_{x'}$; (ii), as it should be true for a Nash equilibrium in mixed strategies, P_1 is an alternative best response to $Q_{x'}$ and Q_1 is an alternative best response to $P_{x'}$, but $\pi(P_1, Q_1) = 3$ while $\pi(P_{x'}, Q_{x'}) = 2$. Similarly, P_2 is an alternative best response to $Q_{x'}$ and Q_2 is an alternative best response to $P_{x'}$, but $\pi(P_2, Q_2) = 3$. However, compare this now to the state where the entire population is of type (P_1, Q_1), $(x_1'', x_2'') = (1, 0)$, in which case $(P_{x''} Q_{x''}) = (P_1, Q_1)$. From the best-response properties of the P and Q matrices we can easily see that $P_{x''} = P_1$ is not only *a*, but the *unique best response* to $Q_{x''} = Q_1$, and that $Q_{x''} = Q_1$ is not only *a*, but the *unique best response* to $P_{x''} = P_1$. In other words, $(P_{x''}, Q_{x''})$ is a strict Nash-equilibrium strategy (there are no alternative best responses), and hence it is evolutionarily stable. Similarly, P_2 is the unique best response to Q_2 and Q_2 is the unique best response to P_2, and hence, the state where the entire population is of type (P_2, Q_2) will be evolutionarily stable. Note that both (P_1, Q_1) and (P_2, Q_2) establish a bijection between concepts and signals.

Evolutionary stability captures the idea that a state is resistant against the invasion of mutant strategies. For a variety of selection dynamics, most importantly the replicator dynamics, this can be given a precise formulation in terms of dynamic stability properties: If a strategy (P_x, Q_x) is evolutionarily stable, then the corresponding state x will be an *asymptotically stable* rest point of the replicator dynamics (Taylor and Jonker, 1978).[5] That is, if the system starts close enough to such a rest point, then it will always remain close to it and will eventually converge to it.

It can be shown that an evolutionarily stable strategy of this game will exist if and only if $m = n$, that is, there is the same number of signals as there are concepts to be communicated, and that $(P_x, Q_x) \in \mathcal{M}_{n \times n} \times \mathcal{M}_{n \times n}$ will be an evolutionarily stable strategy if and only if both P_x and Q_x have the form of a permutation matrix (a matrix that has exactly one 1 in every row and in every column) and one matrix is the transpose of the other (Trapa and Nowak, 2000). That is, an evolutionarily stable strategy can only be a "language" that bijectively links every concept to exactly one signal such that the mapping used in the role of the receiver is the inverse of the mapping used in the role of the sender; in other words, a language that is unambiguous.[6] (P_1, Q_1) and (P_2, Q_2), which we have seen above in Example 1, are of this form. If such a strat-

---

[3]This is a general property of symmetrized asymmetric games: Suppose that there is a $Q \in \mathcal{M}_{n \times m}$ such that $\pi(P_x, Q) > \pi(P_x, Q_x)$, and consider the pair (P_x, Q). Then $f[(P_x, Q), (P_x, Q_x)] = \frac{1}{2}[\pi(P_x, Q_x) + \pi(P_x, Q)] > \frac{1}{2}[\pi(P_x, Q_x) + \pi(P_x, Q_x)] = f[(P_x, Q_x), (P_x, Q_x)]$, yielding a contradiction to condition (i). Similarly for the roles of P and Q reversed.

[4]This comes from the symmetry of the payoff function: $f[(P_x, Q_x), (P_x, Q_x)] = f[(P, Q), (P_x, Q_x)] = f[(P_x, Q_x), (P, Q)] > f[(P, Q), (P, Q)]$, and hence $\pi(P_x, Q_x) > \pi(P, Q)$.

[5]Note that the converse is not true in general; an example of an asymptotically stable rest point that is not an evolutionarily stable state can be found in Taylor and Jonker (1978).

[6]Restricting attention to pure strategies, or what in our model corresponds to states where the entire population is of the same type, this first has been shown by Wärneryd (1993). In view of Selten's 1980 general

egy is adopted by the entire population, the maximum communicative potential will be attained.

2.3. The Lyapunov function—fitness landscapes

Due to the symmetry of the payoff function, the replicator dynamics of this model has a special property: the average fitness function constitutes a Lyapunov function for the dynamics, that is, a function that is increasing along every trajectory.[7] In other words, the dynamical system satisfies Fisher's fundamental theorem of natural selection (Fisher, 1930)—the average fitness increases along every evolutionary path.[8] This can be represented by a fitness landscape, where evolution is represented by moving along its uphill directions. The strict local maximum points of the Lyapunov function coincide with the evolutionarily stable states, and as a consequence, the evolutionarily stable states *coincide* with the asymptotically stable rest points of the replicator dynamics (Hofbauer and Sigmund, 1988, 1998).

2.4. Monomorphic ESS—no language change

An important aspect of the results above is that an evolutionarily stable state, and fitness peak, can only be attained in a *monomorphic population state* where a language that bijectively links every concept to exactly one signal has become fixed in the population. But languages, like biological organisms, change on the basis of existing or newly occurring variation. Once evolution has settled down to such an evolutionarily stable state—with all variation having been driven out and resistance to any possible mutant strategy—all descendant populations will be of exactly the same type with the same (P, Q) being fixed throughout. And, importantly, this will be the case *even if populations get isolated and evolve separately*: for, no matter how we draw subgroups of the original population, since there is no variation, the same (P, Q) will resurface in all descendant populations, and this identically inherited (P, Q) will be a form that is resistant against any possible mutant. Hence, the same perfectly bijective (P, Q)

will be fixed in all descendant populations, and there will be no change in the meaning of signals. However—and this is the first step in our argument—the replicator dynamics does not necessarily converge to an evolutionarily stable state.

3. Neutral stability—stable polymorphisms

There are equilibrium states in this model that are not evolutionarily stable but that satisfy a weaker condition known as *neutral stability* (Maynard Smith, 1982) or *weak evolutionary stability* (Thomas, 1985) and that do allow for variation in the population. Formally, a strategy $(P_x, Q_x) \in \mathcal{M}_{n \times m} \times \mathcal{M}_{m \times n}$ is *neutrally stable* if

(i) $f[(P_x, Q_x), (P_x, Q_x)] \geq f[(P, Q), (P_x, Q_x)] \, \forall \, (P, Q) \in \mathcal{M}_{n \times m} \times \mathcal{M}_{m \times n}$; and

(ii) whenever (i) holds with equality for some $(P, Q) \in \mathcal{M}_{n \times m} \times \mathcal{M}_{m \times n}$, then

$$f[(P_x, Q_x), (P, Q)] \geq f[(P, Q), (P, Q)]. \quad (16)$$

This condition is similar to the notion of evolutionary stability (13), only that the strict inequality in the second condition is replaced by a weak inequality. We have already seen above in (14) that the first condition simplifies to requiring that P_x and Q_x be best responses to each another. By the symmetry of the payoff function, the second condition simplifies analogously to what we have seen in (15): If there is a $P \in \mathcal{M}_{n \times m}$ that is a best response to Q_x and a $Q \in \mathcal{M}_{m \times n}$ that is a best response to P_x, then it should be true that

$$\pi(P_x, Q_x) \geq \pi(P, Q). \quad (17)$$

Note that a state that is evolutionarily stable will also be neutrally stable. We call a state that is neutrally stable but not evolutionarily stable, *properly neutrally stable*. Example 2 discusses a typical properly neutrally stable state.

Example 2. Suppose there are 4 resident types who all have the same sender matrix P_0, but different receiver matrices,

$$(P_0, Q_1) = \left[\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right],$$

$$(P_0, Q_2) = \left[\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right],$$

$$(P_0, Q_3) = \left[\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \right],$$

$$(P_0, Q_4) = \left[\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \right],$$

---

result that for asymmetric games—and as a consequence also for symmetrized games—evolutionarily stable strategies can only be strict Nash equilibria, and hence in pure strategies, the two results are equivalent. From the best-response properties of the P and Q matrices (Remark 1) it is not difficult to see that for a pair (P, Q) to be a strict Nash equilibrium strategy (that is, a pair (P, Q) such that P is the unique best responses to Q and Q the unique best response to P), both P and Q have to be permutation matrices and one has to be the transpose of the other. The result then is immediate.

[7]In fact, for this game—and doubly symmetric games in general—the average fitness is not only a Lyapunov function, but a potential function and the replicator dynamics constitutes a gradient system with respect to the Shashahani metric; for more on this see Hofbauer and Sigmund (1998) and Huttegger (2007).

[8]This is a rather special property for a model of frequency-dependent selection. For a number of games that prominently have been studied in an evolutionary context this is not true, and there are even games, for example, the Prisoner's Dilemma, where the average fitness *decreases* along any evolutionary path.

and let the corresponding type frequencies be $(x_1, x_2, x_3, x_4) = (0.3, 0.3, 0.2, 0.2)$. Then the population's average strategy is

$$(P_x, Q_x) = \left[\begin{pmatrix} \mathbf{1} & 0 & 0 \\ \mathbf{1} & 0 & 0 \\ 0 & 0 & \mathbf{1} \end{pmatrix}, \begin{pmatrix} \mathbf{.5} & \mathbf{.5} & 0 \\ .3 & .3 & .4 \\ 0 & 0 & \mathbf{1} \end{pmatrix} \right].$$

With the help of the characterization of best responses (Remark 1), it is straightforward to check that $P_x = P_0$ is a best response to Q_x, and that Q_x is a best response to P_x (in order to have an immediate glance at the column maxima, we have highlighted them in boldface). In fact, $P_x = P_0$ is not only *a* best response, but the *unique* best response to Q_x. Hence, for any pair of alternative best responses $(P', Q') \in \mathcal{M}_{n \times m} \times \mathcal{M}_{m \times n}$ to (P_x, Q_x) we will have that $P' = P_x$. And this is in fact sufficient to see that (P_x, Q_x) is neutrally stable, since for *any* $Q \in \mathcal{M}_{m \times n}$ (irrespective of whether it will be a best response to P_x or not) we will have

$$\pi(P_x, Q) \leq 2 = \pi(P_x, Q_x).$$

Hence, the communicative potential of any pair of alternative best responses to the original Nash equilibrium strategy $\pi(P', Q') = \pi(P_x, Q')$ will always be bounded by the eigen communicative potential of the original Nash equilibrium strategy $\pi(P_x, Q_x)$. Note in particular that (P_x, Q_x) *cannot* be invaded by a mutant who switches to

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

or

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

However, while (P_x, Q_x) is neutrally stable, it fails to be evolutionarily stable, since there are alternative best responses $Q' \in \mathcal{M}_{m \times n}$ to P_x such that $\pi(P_x, Q') = \pi(P_x, Q_x)$. Of course, every pure-strategy $Q_l, l = 1, 2, 3, 4$, that is in the support of Q_x is already a best response to P_x, but, more generally, every $Q' \in \mathcal{M}_{m \times n}$ that is of the form

$$Q' = \begin{pmatrix} q_{11} & q_{12} & 0 \\ q_{21} & q_{22} & q_{23} \\ 0 & 0 & 1 \end{pmatrix},$$

will be a best response to P_x, and for any such Q' we will have $\pi(P_x, Q') = 2 = \pi(P_x, Q_x)$.

While evolutionary stability translates the idea that a strategy can protect itself against the invasion of mutant strategies—in the strict sense that it can drive out mutant strategies—neutral stability is more apt to capture the idea that the currently resident types *cannot be driven out* by other, potentially intruding, strategies. Instead,

there can be *coexistence of types*. Again, for the replicator dynamics this can be given a precise formulation in terms of dynamic stability properties: If a strategy (P_x, Q_x) is neutrally stable, then the corresponding state x will be a *Lyapunov stable* rest point of the replicator dynamics (Bomze and Weibull, 1995). That is, if the system starts close enough to such a rest point, then it will always remain close to it, but need not converge to it. For doubly symmetric games, the converse is true as well (Bomze, 2002), and hence for the game discussed here, a state is neutrally stable if and only if it is Lyapunov stable in the replicator dynamics. And, again, this comes from the fact that the average fitness is a Lyapunov function for the dynamics: the local maxima of this function coincide with the neutrally stable states.

3.1. Patterns in the P and Q matrices

It can be shown that a Nash-equilibrium strategy $(P_x, Q_x) \in \mathcal{M}_{n \times m} \times \mathcal{M}_{m \times n}$ is *neutrally stable* if and only if the following condition holds:

(i) at least one of the matrices P_x or Q_x (or both) has no zero column, and

(ii) none of the two matrices, neither P_x nor Q_x, has a column with multiple maximal elements that are strictly between 0 and 1 (Pawlowitsch, 2008).

Necessity of these conditions is quite intuitive: (i) a zero column in P means that there is a signal that is never used, and a zero column in Q means that there is a concept that is never possibly inferred. It is straightforward to show then that a mutant who links this empty signal to this unknown concept can do as well against the population but can do better against itself. (ii) By the use of the best-response properties of the P and Q matrices, it is easy to show that if, in equilibrium, there is one column with multiple maximal elements strictly between 0 and 1, then there will always be another column with multiple maximal elements strictly between 0 and 1. In terms of the language model this means that there are two (or more) signals that are simultaneously linked to two (or more) concepts. It is straightforward to show then that a mutant who resolves this ambiguity by linking one of these concepts bijectively to one of these signals, and the other concept to the other signal, can do as well against the population as all the resident types but can do strictly better against itself. Example 1 illustrates this case. Sufficiency follows from a generalization of the argument that we have seen in Example 2. A complete proof can be found in (Pawlowitsch, 2008).

A couple of observations follow:

(a) *A monomorphic population can never be properly neutrally stable.* From the best-response properties of the P and Q matrices it is easy to see that in a monomorphic Nash-equilibrium state there will be necessarily a zero column in both P and Q. By condition (i), then, (P, Q) cannot not be neutrally stable.

8

(b) *A population state at the interior of the state space can never be neutrally stable.* This follows from condition (ii): In a population state at the interior of the state space, the population's average P_x, and respectively Q_x, will always be of the form that each element is strictly between 0 and 1. Best-response properties of the P and Q matrices then imply that the elements in each column of P_x, and respectively Q_x, have to be identical, and hence there will be multiple maximal column elements strictly between 0 and 1.

(c) *Minimal consistency.* The characterization of neutrally stable strategies above can be interpreted in the sense of some *minimal consistency* criteria between the sender and the receiver matrix. Condition (i) has a straightforward interpretation: it tells us that there can be no signal that remains idle (a zero column in P_x) as long as there is a concept that is never possibly inferred (a zero column in Q_x). Condition (ii), together with the best-response properties of the P_x and Q_x matrices (Remark 1), implies the following: There can be different resident types who use different signals to communicate a particular concept (multiple entries between 0 and 1 in a row of P_x), but if this is the case, then *all resident types* will infer this particular concept from any of the signals that *some resident type* uses to communicate this concept (a column with multiple 1s in Q_x), or this concept is never inferred by any resident type (a zero column in Q_x). And, similarly for the roles of P_x and Q_x reversed: there can be different resident types who infer different concepts from the same signal (multiple entries between 0 and 1 in a row of Q_x; in Example 2, the first and the second row of Q_x), but if this is the case, then *all resident types* will use this particular signal to communicate all the concepts that some resident type infers from this signal (a column with multiple 1s in P_x; in Example 2, the first column in P_x), or this signal is never used by any resident type (a zero column in P_x; in Example 2, the second column of P_x).

(d) *Stabilizing variation and uninvadability by perfectly bijective strategies.* While a monomorphic population that is not evolutionarily stable can always be invaded by another strategy, it is the very *coexistence of types* that stabilizes multiplicities in concept-to-signal (or signal-to-concept) mappings in a neutrally stable state. Example 2 illustrates this property.

Example 2 continued. Since there are some types who, in the role of the receiver, will map signal 1 to concept 1 and some types who will map signal 1 to concept 2 (and these concepts are no more likely to be inferred by any other signal), the unique optimal response to this aggregate receiver behavior, in the role of the sender, is to link both concept 1 and concept 2 to signal 1. In a Nash equilibrium all resident types will do so, as reflected in the two 1s in the first column of
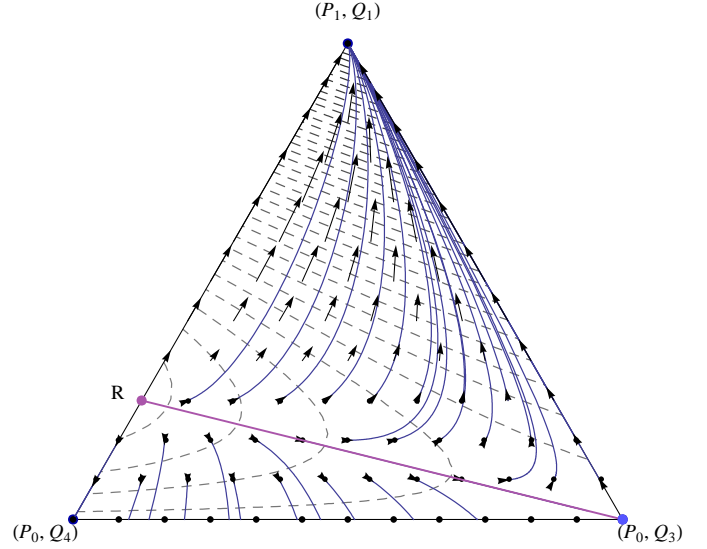

Figure 2: The phase portrait of the replicator dynamics when the population consists only of the three types (P_1, Q_1), (P_0, Q_3), and (P_0, Q_4) mentioned in Example 2. The light purple stationary point $R = 1/4(P_1, Q_1) + 3/4(P_0, Q_4)$ is unstable and, in fact, it corresponds to the global minimum of the average population fitness $\bar{f}(x)$ with respect to these three types; the dashed gray contours represent the level sets of $\bar{f}$, which, for this model, is a Lyapunov function for the dynamics—it is increasing along every trajectory. Moreover, we see that the line which joins R to the semi-stable type (P_0, Q_3) (light blue) is actually a *separatrix* of the system: it is invariant under the replicator dynamics and separates the state space into two regions that are themselves invariant as well. Every point on the face spanned by (P_0, Q_3) and (P_0, Q_4), except for the vertex (P_0, Q_3), is neutrally stable. Hence, even though the type (P_1, Q_1) corresponds to the global maximum of the average population fitness $\bar{f}$, we see that there is a positive measure of initial conditions which *do not* converge to it. (Note that (P_0, Q_4) is neutrally stable for the truncation of the game to the subset of strategies considered here, but is not neutrally stable in the complete strategy space $\mathcal{P}_{3\times3} \times \mathcal{Q}_{3\times3}$; it can be invaded by (P_2, Q_2), see also Figure 5.)

the population's sender matrix $P_x = P_0$, but it is only this property of the population's sender matrix that enables variation in the role of the receiver. Though variation in Q is not imposed by Nash-equilibrium conditions (for Q_x to be a best response to P_x, it is not necessary that both q_{11} and q_{12} are strictly positive), once it is there (once q_{11} and q_{12} have taken values strictly between 0 and 1), any best response to Q_x will have to set both p_{11} and p_{21} equal to 1, thereby freezing the multiplicity in the population's average sender matrix and blocking off mutations who try to resolve the existing ambiguity by linking each concept to exactly one signal. As we have seen above, (P_x, Q_x) cannot be invaded by any of the perfectly bijective strategies (P_1, Q_1) or (P_2, Q_2). In Appendix Appendix A we discuss this phenomenon in the context of the so-called "bidirectional Saussurean sign."

3.2. Convergence to neutrally stable states

An important consequence of the average fitness being a Lyapunov function for the dynamics is that every tra-

jectory will converge to a rest point (Akin and Hofbauer, 1982). For the game discussed here it can be shown that for every properly neutrally stable state there is a neighborhood in which every rest point of the dynamics is a neutrally stable state (Pawlowitsch, 2008). From this, together with the fact that neutrally stable states are Lyapunov stable, one can see that there are in fact components of properly neutrally stable strategies that have a basin of attraction of non-zero measure. More precisely, properly neutrally stable states occur in connected (but not closed) sets of Nash equilibria at the boundary of the state space. Of course, the replicator dynamics can converge to an evolutionarily stable state, but it will not do so "almost always" (Huttegger, 2007; Pawlowitsch, 2008). Figures 2 and 3 illustrate this for a truncation of the game.

In a neutrally stable state the forces of selection and mutation exert no further pressure on the coexistence of types: selection has come to an end since all resident types get the same payoff, and it is the very coexistence of these types that protects the population from being taken over by mutants "from outside." However, as we shall see in the next section, this form of neutrally stable coexistence of types can be destabilized by a redistribution of the *relative frequencies* of the types who are present in such a state.

## 4. A third evolutionary force: neutral drift

Evolutionary and neutral stability, or more precisely their dynamic counterparts in the form of asymptotic stability and Lyapunov stability, test locally around an equilibrium against small perturbations in the state of the population. These concepts do not test against a scenario where a larger fraction of the population *simultaneously* switches to a new strategy, or where a major shift in the relative type frequencies of the types already present in the population occurs. The first type of change is in fact hard to argue in an evolutionary setting, or a setting where individual strategies are updated in a decentralized way. The second type of change, however, a redistribution of the types *already present in the population*, does not seem artificial for a scenario of language change. Such a shift can be brought about by a pronounced reduction in population size, so-called bottlenecks, or it can be the result of a subset of the population migrating to a different neighborhood. Archeological, genetic, and linguistic evidence suggests that such events have dramatically shaped human evolution and the geographic distribution of languages (see, for example, Cavalli-Sforza, 1997; Cavalli-Sforza and Feldman, 1981).

4.1. Neutral drift as a pathway for a mutation for bidirectionality

If the population has reached a state where all agents are of the same type, in whichever way we draw subsets of the original population, all descendant populations will be of exactly the same type. However, if the
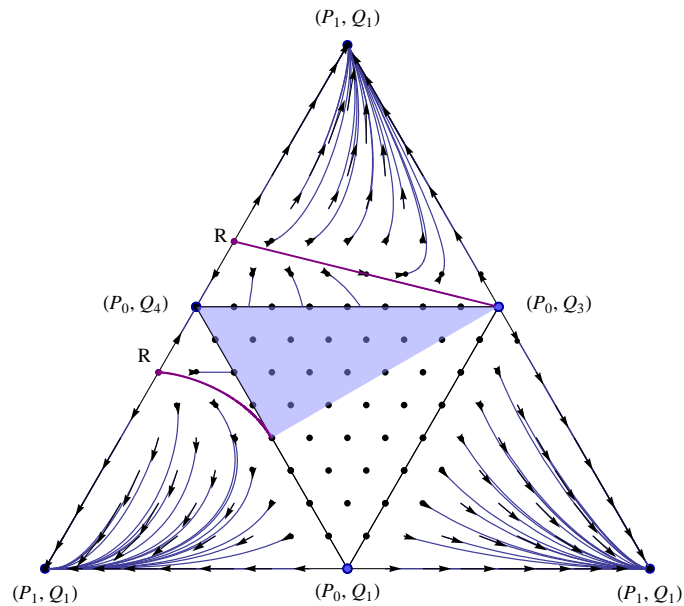
(a) 2-dimensional foldout of the faces of the replicator dynamics for four types
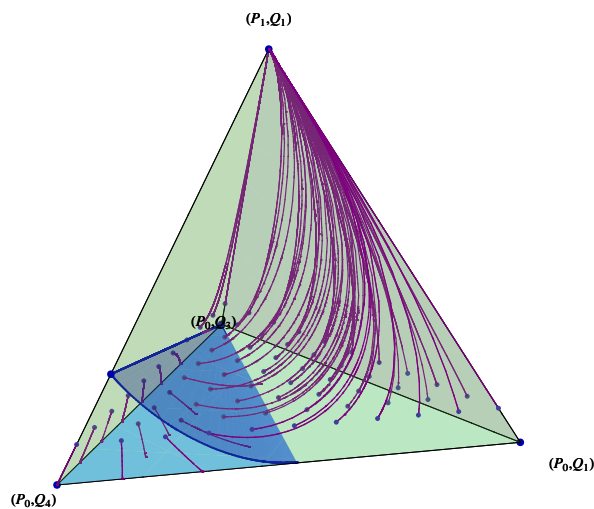
(b) Phase portrait of the replicator dynamics for four types

Figure 3: 3(a) A 2-dimensional foldout of the faces of the phase portrait of the replicator dynamics when the 3-type population of Figure 2 is augmented by the fourth type (P_0, Q_1). Since the types containing P_0 all yield the same payoff when paired against each other, the corresponding face (center) consists entirely of fixed points; however, not all of them are neutrally stable. The light blue triangle shows the set of neutrally stable states. This set is convex, but not closed; it is open all along the boundary that connects the vertex (P_0, Q_3) to the midpoint on the face $(P_0, Q_4) - (P_0, Q_1)$. (The vertex (P_0, Q_4) is neutrally stable for the truncation of the game considered here, but is not neutrally stable in the complete strategy space $\mathcal{P}_{3\times3} \times \mathcal{Q}_{3\times3}$; it can be invaded by (P_2, Q_2).) The light purple lines represent *separatrices* of the system and show that the type (P_1, Q_1), which maximizes population fitness, is not globally attracting. In Fig. 3(b) we present the full 3-dimensional phase portrait of the replicator dynamics for this subset of strategies. The separtrices that we can see in 3(a) delineate the boundary of a higher-dimensional separatrix (semi-transparent plane) that breaks up the state space in two distinct invariant sets. Both sets have positive measure. So, even though the type (P_1, Q_1) is evolutionarily stable, there is a positive measure of initial conditions which do not converge to it, but which, instead, converge to neutrally stable states in the face spanned by the other three types.
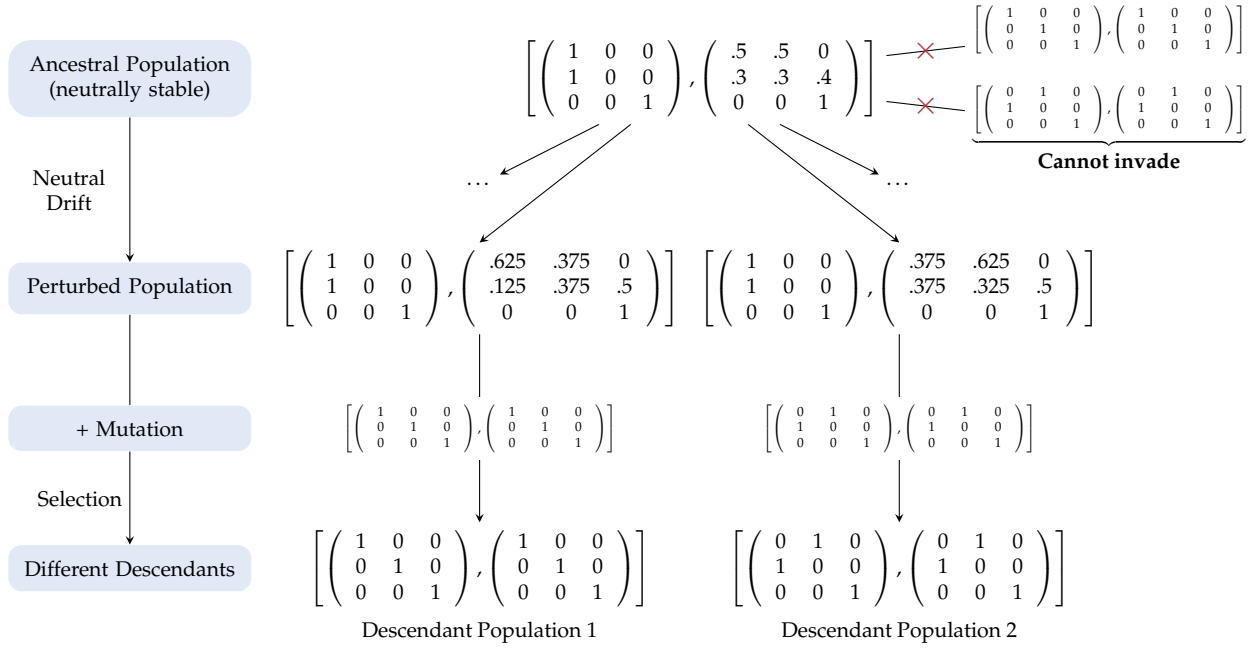
10

Figure 4: The ancestral population is neutrally stable; in particular, it cannot be invaded by mutants who try to resolve the existing ambiguity by establishing a bijection between concepts and signals. However, a shift in the relative type frequencies (neutral drift) can overcome neutral stability and open the door for such mutations. Different directions of neutral drift are pathways for different mutations, which finally lead to the fixation of different languages. As a result we can observe a switch in the meaning of lexical items in two languages that go back to the same common ancestor.

population is composed of different types, we cannot expect that the type frequencies in the descendant population will be an exact image of the ancestral population.

Suppose that we are in a neutrally stable state as we have seen it in Example 2 with the type frequencies given by $(x_1, x_2, x_3, x_4) = (0.3, 0.3, 0.2, 0.2)$ so that

$$(P_x, Q_x) = \left[\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} .5 & .5 & 0 \\ .3 & .3 & .4 \\ 0 & 0 & 1 \end{pmatrix} \right],$$

but that now there is an exogenous random event that brings about a shift in the relative frequencies of the resident types such that after this shift, for example, $(x'_1, x'_2, x'_3, x'_4) = (0.375, 0.125, 0.25, 0.25)$. The average sender-receiver pair then is

$$(P_{x'}, Q_{x'}) = \left[\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} .625 & .375 & 0 \\ .125 & .375 & .5 \\ 0 & 0 & 1 \end{pmatrix} \right].$$

Now, $P_{x'}$ and $Q_{x'}$ are still best responses to each other, and hence $(P_{x'}, Q_{x'})$ still is a Nash-equilibrium strategy. But, as we can readily see from the multiple maximal elements in the second column of $Q_{x'}$, it is no longer neutrally stable. If a small fraction of the population now switches to (P_1, Q_1)—as it could come about, for example, by a mutation for bidirectionality, as we discuss it in Appendix Appendix A, that appears on type 1 (a mutation that was blocked before)—then this mutant strategy will do as well against $(P_{x'}, Q_{x'})$ as any of the resident types,

but will do strictly better against itself, and hence under a monotonic selection dynamics will eventually become fixed in the population. If—all else being equal—an even more pronounced reduction of type 2 comes about, then the resulting population state will still be a rest point of the replicator dynamics (since all resident types gain the same payoff against each other), but it will no longer be a Nash equilibrium. (P_1, Q_1) will then be a *better* response to $(P_{x'}, Q_{x'})$ than $(P_{x'}, Q_{x'})$ is to itself, and hence a mutant who switches to (P_1, Q_1) will be immediately on its way to fixation.

A similar scenario will obtain if type 4 goes extinct, bringing about a shift in the relative type frequencies to $(x''_1, x''_2, x''_3, x''_4) = (0.375, 0.375, 0.25, 0)$. In this case,

$$(P_{x''}, Q_{x''}) = \left[\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} .625 & .375 & 0 \\ .375 & .375 & .5 \\ 0 & 0 & 1 \end{pmatrix} \right].$$

A mutation to (P_1, Q_1) will now also be able to invade, and will finally take over the population.

4.2. Different histories of change

A different outcome, however, obtains if a shift in the relative frequency of the resident types to the detriment of type 1 or type 3 occurs. Consider, for example a shift that results in $(x'''_1, x'''_2, x'''_3, x'''_4) = (0.125, 0.375, 0.25, 0.25)$.

Then,

$$(P_{x'''}, Q_{x'''}] = \left[\begin{pmatrix} \mathbf{1} & 0 & 0 \\ \mathbf{1} & 0 & 0 \\ 0 & 0 & \mathbf{1} \end{pmatrix}, \begin{pmatrix} \mathbf{.375} & \mathbf{.625} & 0 \\ \mathbf{.375} & .125 & .5 \\ 0 & 0 & \mathbf{1} \end{pmatrix} \right].$$

This produces also a change in the column maxima of Q_x, but this time in a different column. If a small fraction of mutants now switches to (P_2, Q_2)—for example a mutation for bidirectionality that appears on type 2 —then this mutant strategy will do as well against (P_x, Q_x) as all the resident types, but will do strictly better against itself and hence will finally take over the population. The same mutation will be enabled if type 3 goes extinct. So, depending on the direction of drift between selectively neutral variants, the same ancestral population can give rise to two different, perfectly bidirectional, languages. Figure 4 illustrates these different evolutionary outcomes. Figure 5 provides a graphical representation of this phenomenon for the case that the ancestral population consists of the two types (P_0, Q_3) and (P_0, Q_4) only: The face connecting (P_0, Q_3) and (P_0, Q_4) is a ridge of selectively neutral variants. As long as both (P_0, Q_3) and (P_0, Q_4) are present in the population, the population is in a neutrally stable state. In this case, the population's average sender-receiver pair (P_x, Q_x) is of the form

$$\left[\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} x_3 & x_4 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \right],$$

and neither (P_1, Q_1) nor (P_2, Q_2) can invade. But if (P_0, Q_4) goes extinct—which means that the population drifts to the boundary of this ridge of selectively neutral variants—the population can be invaded by (P_1, Q_1), and from thereon selection will take it towards the point where everybody will be of type (P_1, Q_1). If, on the other hand, (P_0, Q_3) goes extinct, the population will drift to the opposite boundary of this neutral ridge, and will finally be invaded by (P_2, Q_2). Note that along any of these possible paths, the average fitness stays constant or is increasing, but is never decreasing.

4.3. Shifting balance

The mechanism that we describe can be seen as a form of *shifting balance* (Wright, 1931) where for different species (here "languages") represented by fitness peaks to be reached, populations do not have to cross fitness valleys, but where different species ("languages") can form by drift along ridges of high, locally maximal, fitness. Drift along ridges of selectively neutral variants naturally comes about if populations are subject to exogenous random shocks like bottlenecks or migration leading to the founders' effect (Mayr, 1963). Gavrilets and Hastings (1996) have analyzed scenarios of shifting-balance–founder-effect speciation in a classical population-genetics model. In economics, Binmore and Samuelson (1999) have studied the role of drift along
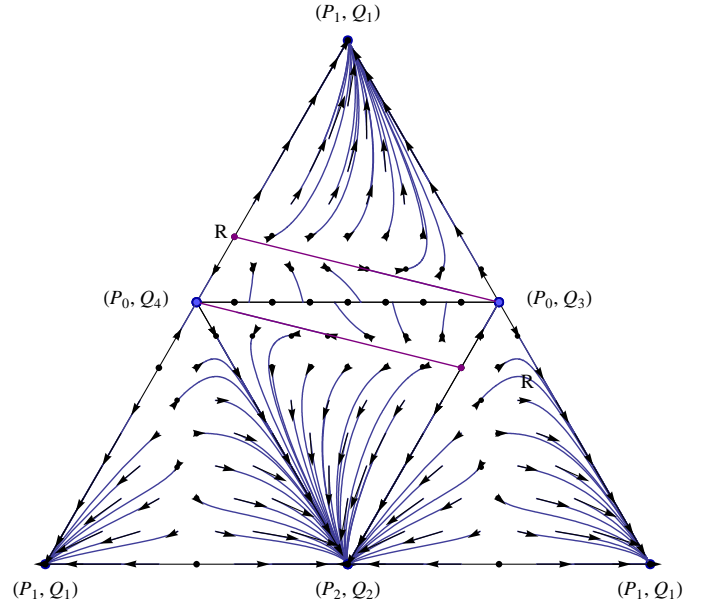

Figure 5: A 2-dimensional foldout of the faces of the phase portrait of the replicator dynamics when the population consists of the four types $(P_0, Q_3), (P_0, Q_4), (P_1, Q_1)$, and (P_2, Q_2) mentioned in Example 2. The vertices (P_1, Q_1) and (P_2, Q_2) are evolutionarily stable states. The face connecting (P_0, Q_3) and (P_0, Q_4) is a ridge of selectively neutral variants. As long as both (P_0, Q_3) and (P_0, Q_3) are present in the population, the population is in a neutrally stable state, and neither (P_1, Q_1) nor (P_2, Q_2) can invade. But if (P_0, Q_4) goes extinct, the population can be invaded by (P_1, Q_1); and if (P_0, Q_3) goes extinct, the population can be invaded by (P_2, Q_2).

ridges of neutral evolution in its relation to classical game theoretic—strategic—approaches to equilibrum selection. Importantly, the shape of the fitness landscape that gives rise to this phenomenon is not an ad hoc assumption,[9] but comes about as an endogenous feature of the model—in our case the underlying language game. The proposed mechanism is in line with empirical findings which support a serial founder effect model of language expansion (for a recent contribution focusing on a phonological trait see, Atkinson, 2011).

5. Conclusions

In linguistics the question whether all human languages can be traced back to one unique common ancestor remains the subject of a heated debate (see, for example, Evans and Levinson, 2009). A standard argument in evolutionary accounts for the diversification of language is that it serves as an in-group marker to enhance cooperation in small groups (see, for example, Dunbar, 1998). From a conceptual point of view, postulating a function of language differentiation for enhancing in-group cooperation (which will result in increased material payoffs)

---

[9]A critique that sometimes has been addressed to Wright's shifting balance theory is that the fitness landscapes that he used to illustrate the theory were drawn in an ad hoc manner.

amounts to postulating a preference for diversification in the language game, that is, making it an assumption of the model, and therefore, in our mind, cannot be considered something explained by the model.

Another mechanism for language diversification that is sometimes proposed is *adaptation to different environments*. In the context of a game-theoretic model, this is to assume different payoff functions in different environments, which, conceptually, also boils down to writing part of what has to be explained into the payoff function.

The argument that we give here shows that language diversification can come about in a perfectly cooperative world (a world where everybody wants to cooperate with everybody) solely by the effects of frequency-dependent selection. This does not mean that language differentiation and diversification cannot have other sources (for example, adaption to different environments) or cannot be adapted to serve some function in a co-evolutionary process (possibly as an in-group marker to enhance cooperation), but it shows that *we do not need these quite specific assumptions to give an evolutionary account for the differentiation of languages*. Rather it can come about already under weaker assumptions.

In historical linguistics, language change is often described as two, or more, variant forms coexisting for some time and then one giving way to another (see, for example, Schendl 2001). From an evolutionary point of view, this begs the question how variation in the population is sustained in the first place. The argument that we give here does not only show why languages can differentiate and branch on the basis of actual variation, but also *why* variation in the population, and the resulting ambiguities, can be a locally stable phenomenon in the first place even though ex ante there is no incentive for differentiation, and globally it would always be the best if everybody used the same language that bijectively links every concept to exactly one signal.

The suboptimality that we observe in polymorphic Nash equilibria reflects a problem that has long been known to social philosophers and philosophers of language and to which game theorists have given a precise formulation: the problem of being stuck in a bad equilibrium. We would all be better off if we could simultaneously jump out of the bad equilibrium, and right into another, but as long as there is no central institution that makes us jump simultaneously, we are stuck in the bad equilibrium, since unilaterally nobody has an incentive to deviate from the old one, and in fact would lose if he were the only one to deviate. Language abounds with this type of inefficiencies; the existence of centralized institutions of language regulation seems to testify to this fact. Big coordinated jumps are difficult to argue in an evolutionary setting. What is not difficult to argue in an evolutionary setting is a redistribution of the frequencies of the types that are already present in the population. This form of drift can open the door for mutations that can take us out of these inefficient states—but thereby also opening the door for the diversification of languages.

Appendix A. Excursion: Evolution of the bidirectional Saussurean sign?

Linguists call the property that if a concept *A* is linked to a signal *s*, and if *s*, when received, evokes the image of *A*, *bidirectionality*, and a form-meaning pair that satisfies this property, the *bidirectional Saussurean Sign*. Most linguistic theories postulate such form-meaning pairs as the underlying basic building blocks of language and the ability to grasp and operate with these objects as genetically implemented as part of the human language acquisition device. It is in the search of an evolutionary foundation for bidirectionality that Hurford (1989) has introduced a version of the model that we investigate here in the linguistics literature.

Hurford's approach is to study this model by agent-based computer simulations and to compare the performance of different behavioral types. Specifically he is interested in the question whether types who align their *Q* with their *P* in a bidirectional way will outperform other behavioral types, where the formal criterion for bidirectionality that he uses is that the individual's *Q* has to be a best response to the individuals's *P*. The two other behavioral types that he considers are basically agents who adopt a *P* that is randomly sampled from the population and a *Q* that is *independently* randomly sampled from the population. For some specific initial conditions, Hurford can show that bidirectional types will indeed do better than the two other behavioral types, but altogether the results that he gets are inconclusive. Hurford's 1989 paper—which has been written at a time when evolutionary game theory was very little known outside a small group of mathematicians and biologists—mentions the work of Maynard Smith, but does not evoke the term *evolutionary game theory*, nor does it include an evolutionary stability analysis. So it is also of some methodological interest to ask what we can say about the question of bidirectionality from the point of view of the game-theoretic analysis of the model that is available to us now.

An immediate answer that we can give on the basis of the replicator dynamics, if we take as a formal definition of bidirectionality that the *P* and the *Q* have to be best responses to each other, is that *replication operating on*

the (P, Q) *pairs* can account for the rise of bidirectionality *on the level of the population's average strategy* (P_x, Q_x)—the "language" of the population—but that it is not sufficient to guarantee that each individual type who is present in this population will be perfectly bidirectional with itself. Example 2 illustrates this: All resident types use a Q_l that is a best response to the fixed sender matrix $P_x = P_0$. But while the fixed sender matrix $P_0 = P_x$ is a best response to Q_x, it is *not* a best response to each individual Q_l that is present in the population: it is a best response to Q_3 and Q_4, but it is not a best response to Q_1 or Q_2. That is, type 1 and 2, when they appear in the role of the sender, do not respond optimally to themselves in the role of the receiver. Note that what prevents them from doing so—in equilibrium—is the need to respond optimally to the population's average Q. Hence, the aggregate property that in a Nash equilibrium (P_x, Q_x), P_x and Q_x have to be best responses to each other does not necessarily carry over to the (P, Q) of each individual type that is present in such a state. From a game-theoretic point of view this is not surprising; it simply reflects the fact that in a mixed Nash equilibrium not every pure strategy that is in its support has to constitute a Nash equilibrium in itself.

What is interesting about Example 2 in the context of bidirectionality as Hurford asks the question—namely whether a *behavioral program* for bidirectionality will outperform other behavioral types—is that there can be states where a mutation for bidirectionality *cannot* break through: Suppose we are in a neutrally stable state as we have seen it above with $(x_1, x_2, x_3, x_4) = (0.3, 0.3, 0.2, 0.2)$ and that there is a mutation that appears on type 1 that makes this type want to be consistent with himself and adopt a P that is a best response to his individual Q, that is, a mutation of (P_0, Q_1) to (P_1, Q_1): Under a monotonic selection dynamics this mutation has no chance of invading the population since it will attain a *strictly lower* payoff against the current population's average strategy than any of the resident types. Likewise, a mutation for bidirectionality that appears on type 2—a mutation of (P_0, Q_2) to (P_2, Q_2)—will also be blocked. However, as we discuss in Section 4, drift between the selectively neutral variants that are present in such a state can be a pathway for a mutation for bidirectionality. Different directions of drift may enable mutations for bidirectionality on different resident types and thereby can give rise to different long-run outcomes, where different proto-languages become fixed in the population.

References

Akin, E., Hofbauer, J., 1982. Recurrence of the unfit. Math. Biosci. 61, 51–62.

Atkinson, Q. D., 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science 332, 346–349.

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

Bomze, I. M., Weibull, J. W., 1995. Does neutral stability imply Lyapunov stability? Games Econ. Behav. 11, 173–92.

Binmore, K., Samuelson, L., 1999. Evolutionary drfit and equilibrium selection. Rev. Econ. Studies 66, 363–393.

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

Cavalli-Sforza, L. L., Feldman, M. W., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton, NJ.

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

Donaldson, M. C., Lachmann, M., Bergstrom C. T., 2007. The evolution of functionally referential meaning in a structured world. J. Theor. Biol. 246, 225–233.

Dunbar, R., 1998. Grooming, Gossip, and the Evolution of Language. Harvard University Press, Cambridge, MA.

Evans, N., Levinson, S. C., 2009. The myth of language universals: Language diversity and its importance for cognitive science. Behav. Brain Sci. 32, 429–492.

Fisher, R. A. The Genetical Theory of Natural Selection. Clarendon Press, London.

Gavrilets, S. 2004. Fitness Landscapes and the Origin of Species. Princeton University Press, Princeton.

Gavrilets, S., Hastings, A., 1996. Founder effect speciation: a theoretical reassassment. Am. Naturalist 147, 466–491.

Hofbauer, J., Huttegger, S., Feasibility of communication in binary signaling games. J. Theor. Biol. 254, 843–849.

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

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

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

Hurford, J., 2003. The language mosaic and its evolution. In: Christiansen, M. H., Kirby, S. (Eds.), Language Evolution: the States of the Art. Oxford University Press, Oxford, UK.

Huttegger, S., 2007. Evolution and the explanation of meaning. Philos. Sci. 74, 1–27.

Huttegger, S., Zollman, K., 2011. Signaling games: dynamics of evolution and learning. In: Benz, A., Ebert, C., Jäger, G., van Rooij, R. (Eds.), Language, Games, and Evolution. Trends in Current Research on Language and Game Theory. Springer.

Jäger, G., 2008. Evolutionary stability conditions for signaling games with costly signals. J. Theor. Biol. 253, 131–141.

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

Komarova, N. L., Nowak, M. A., 2001. Evolutionary dynamics of the lexical matrix. Bull. Math. Biol. 63, 451–485.

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

Lewis, M. P. (Ed.), 2009. Ethnologue: languages of the world, 16th Edition. SIL international, Dallas, TX.

Mayr, E., 1963. Animal Species and Evolution. Harvard University Press, Cambridge, MA.

Maynard Smith, J., 1982. Evolution and the theory of games. Cambridge University Press.

Maynard Smith, J., Price, G. R., 1973. Logic of animal conflict. Nature 246, 15–18.

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

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

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

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

Nowak M. A., Krakauer, D. C., Dress, A., 1999. An error limit for the evolution of language, Proc. R. Soc. B 266, 2131–2136.

Piattelli-Palmarini, M., 2000. Tower of Babel. Trends in Ecology and Evolution 15, 173–174.

Pawlowitsch, C., 2008. Why evolution does not always lead to an optimal signaling system. Games Econ. Behav. 63, 203–226.

Sandholm, W. H., 2011. Population Games and Evolutionary Dynamics.

MIT Press, Cambridge, MA.

Schendl, H., 2001. Historical Linguistics. Oxford Introductions to Language Study. Oxford University Press.

Schlag, K. H., 1998. Why imitate, and if so how. a boundedly rational approach to multi-armed bandits. J. Econ. Theory 78, 130–156.

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

Skyrms, B., 1996. Evolution of the Social Contract. Cambridge University Press, Cambridge, UK.

Skyrms, B., 2002. Signals, evolution and the explanatory power of transient information. Philos. Sci. 69, 407–428.

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

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

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

Traulsen, A., Claussen, J. C., Hauert, C. 2005. Coevolutionary dynamics: from finite to infinite populations. Phys. Rev. Lett. 95, 238701–238704.

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

Weibull, J. W., 1995. Evolutionary Game Theory. MIT Press, Cambridge, MA.

Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159.