ECOLE POLYTECHNIQUE

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

NEUTRAL STABILITY, DRIFT, AND THE DIFFERENTIATION OF LANGUAGES

Christina PAWLOWITSCH
Panayotis MERTIKOPOULOS
Nikolaus RITT

March 2011

Cahier n° 2011-08

DEPARTEMENT D'ECONOMIE
Route de Saclay
91128 PALAISEAU CEDEX
(33) 1 69333033
http://www.enseignement.polytechnique.fr/economie/
mailto:chantal.poujouly@polytechnique.edu

Neutral stability, drift, and the differentiation of languages

Christina Pawlowitsch

Paris School of Economics

Panayotis Mertikopoulos

École Polytechnique

Nikolaus Ritt

University of Vienna

Abstract

The differentiation of languages is one of the most interesting facts about language that seek explanation from an evolutionary point of view. An argument that prominently figures in evolutionary accounts of language differentiation is its supposed function in the formation of in-group markers that serve to enhance cooperation in small groups. In this paper we use the theory of evolutionary games to show that language differentiation 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 differentiation in some co-evolutionary 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 is 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 become fixed in the population. Different directions of neutral drift open the door for a mutation for bidirectionality to appear on different resident types. This mechanism can explain why a word can acquire a different meaning in two languages that go back to the same common ancestral language. 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, drift.

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 on all levels of linguistic expression: the lexicon, morphology, phonology, syntax, and 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). In this paper we deal with language differentiation on the level of the lexicon in the form of semantic change, that is, change in the meaning of lexical items. We present a formal model of language evolution in terms of an evolutionary game that can explain why words that have a common ancestor in some predecessor language can acquire different meanings in two descendant languages—for example, English *clean* and German *klein* with the meaning of "small."

1.1. Language change

Language change, like biological evolution, can be understood as a process of descent with modification. On the level of the lexicon, we can distinguish two forms of

Email addresses: christina.pawlowitsch@ens.fr (Christina Pawlowitsch), panayotis.mertikopoulos@polytechnique.edu (Panayotis Mertikopoulos), nikolaus.ritt@univie.ac.at (Nikolaus Ritt)

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 with different meaning.

change: (i) phonological change, that is, change in the sound shapes of lexical items, and (ii) semantic change, that is, change in the *meaning* of lexical items. Here we focus on semantic change, while abstracting from erosions in the sound shape or outer form of lexical items.

In linguistics, words that share a common ancestor are called *cognates*, from Latin *cognatus*—"born together". Most cognates are used with the same meaning. For example, Italian *notte*, Spanish *noche*, French *nuit*, German *Nacht*, Schwedisch *natt*, and English *night*. But there are numerous examples of cognates that do exhibit a shift in meaning. Table 1.1 contains some well documented examples from English–German.

A standard model in historical linguistics is to relate languages by a system of shared ancestry represented by a tree, where a node of the tree is a common ancestor to all nodes at the end of its branches. Constructing family trees of languages is a backward-looking process, where on the basis of linguistic data from present-day languages inferences are drawn about the shared history of these languages (see, for example, Warnow 1997). The building blocks of these methods are overlaps in characters (for example, a semantic slot for the meaning of "hand") together with the assumption that if a character in two languages is not the same, then an event of replacement– represented by a branch of the tree—has taken place. Statistical inference is then used to find the tree that best fits the data. But, these methods do not explain *how* the differentiation and branching of languages comes about. Rather, lexical replacement is taken as an exogenous phenomenon; it is part of the assumptions, and not what is explained by the model. Similarly for computational methods that estimate the rate of linguistic divergence (see, for example, Pagel et al., 2007). In these models too, word replacement comes into the model as an exogenously determined process, but the mechanism that brings about this change is left unexplained. In this paper, we address a complementary issue. We provide an account of a mechanism that brings about lexical replacement—resulting in the branching of languages—as an *endogenous* feature of a model of language change.

Conceptually the question is related to the question of speciation in biology (see, for example, Gavrilets and Gravner 1997 and Gavrilets 2004).

1.2. Language games

The model that we present is an *evolutionary language game* in the style of Nowak et al. (1999), which has been proposed as a model for the evolution of a signaling system, a lexicon, or protolanguage, that is, a collection of form-meaning pairings (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 provides a formal framework for studying frequency-dependent selection (Maynard Smith and Price, 1973; Maynard Smith, 1982; Hofbauer and Sigmund, 1988, 1998; Weibull, 1995; Nowak, 2006; Sandholm, 2011). 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, so it 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). There is a homogenous 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) and Hurford (1989); see also Skyrms (1996, 2002).

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 the 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 shown a different picture and 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, the simulations reported in 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, 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.

1.3. Variation in the population—the basis for language change

Language change, like biological evolution, 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. We will see that what stabilizes these equilibria—that is, what keeps mutants from taking over the population—is the need to respond optimally to multiplicities in concept-to-signal and signal-to-concept mappings that appear on the level of the population: If there are two resident types who, in the role of the sender, map the same concept to different signals (and on the level of the population these signals are not used in higher frequency for another concept), then a type who, in the role of the receiver, will link both signals to this particular concept will do better than a type who is not doing so—and in particular will do better than a mutant, or an individual experimenting, who tries to play a "bidirectional" strategy and resolve the existing ambiguity by linking only one of these signals to the concept in question and using the other, now "free," signal for a concept which so far could never be successfully inferred. In equilibrium *all* resident types will link both incoming signals to the very concept that has triggered the production of these signals, and it is this property of the population's average receiver strategy that opens up the possibility for the coexistence of types in the role of the sender in the first place.

We will consider *neutral drift*—that is, a random shift in the relative type frequencies—among the variants that coexist in a neutrally stable state, and we will see that this may open the door for a mutation that so far has been blocked. Different directions of drift may be pathways for different mutations, and hence will 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.

2. Methods

There are n concepts that potentially become the object of communication, and there are m signals (words) 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 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 an homogenous 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 a sender or receiver with equal probabilities.

A *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

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

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. 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

$$Q = \begin{pmatrix} q_{11} & \cdots & q_{1i} & \cdots & q_{1n} \\ \vdots & & \vdots & & \\ q_{j1} & \cdots & q_{ji} & \cdots & q_{jn} \\ \vdots & & \vdots & & \\ q_{m1} & \cdots & q_{mj} & \cdots & q_{mn} \end{pmatrix}, \tag{2}$$

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. For given m and n, there are m^n such P matrices—the set of which we denote by $\mathcal{P}_{n\times m}$; and there are n^m such Q matrices—the set of which we denote by $\mathcal{Q} \equiv \mathcal{Q}_{m\times n}$. 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 expressed as

$$\begin{aligned} \pi(P,Q) &= \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}q_{ji} \\ &= p_{11}q_{11} + p_{12}q_{21} + \cdots + p_{1m}q_{m1} \\ &+ p_{21}q_{12} + p_{22}q_{22} + \cdots + p_{2m}q_{m2} \\ &\quad \ldots \\ &+ p_{n1}q_{1n} + p_{n2}q_{2n} + \cdots + p_{nm}q_{mn} \end{aligned} \tag{3}$$

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

$$f[(P_k,Q_k),(P_l,Q_l)] = \frac{1}{2}\left[\pi(P_k,Q_l) + \pi(P_l,Q_k)\right]. \tag{4}$$

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 1989, 1998, Weibull 1995, or Cressman 2003). With every strategy $(P,Q) \in \mathcal{P}\times\mathcal{Q}$ we identify a particular *type of player* and we represent a *state of the population* by a vector

$$x = (x_1,\ldots,x_l,\ldots,x_N), \quad \sum_{l=1}^{N} x_l = 1, \tag{5}$$

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 be a row-stochastic matrix of dimensions $n\times m$, and Q_x a row-stochastic matrix of dimensions ~~$m\times n$~~, which we will denote by $P_x \in \mathcal{M}_{n\times m}$ and $Q \in \mathcal{M}_{m\times n}$, respectively. Note that all pure strategies $P \in \mathcal{P}$ will indeed span $\mathcal{M}_{n\times m}$, and all pure strategies $Q \in \mathcal{Q}$ will span $\mathcal{M}_{m\times n}$.[1]

The *fitness of type* l is the average payoff that a type l individual gets from interaction with all other types 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,

$$\begin{aligned} f_l(x) &= f\left[(P_l,Q_l),(P_x,Q_x)\right] \\ &= \frac{1}{2}\left[\pi(P_l,Q_x) + \pi(P_x,Q_l)\right]. \end{aligned} \tag{6}$$

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,

$$\bar{f}(x) = f\left[(P_x,Q_x),(P_x,Q_x)\right] = \pi(P_x,Q_x). \tag{7}$$

We call $\pi(P_x,Q_x)$ the *eigen communicative potential* of (P_x,Q_x).

[1] 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.

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 types. Consider a different vector of type frequencies x', where, for example, $x'_1 = 0.5$ 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.

2.2. *Evolutionary stability*

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,

$$\dot{x}_l = x_l[f[(P_l, Q_l), (P_x, Q_x)] - f[(P_x, Q_x), (P_x, Q_x)]]. \quad (8)$$

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 then their current strategy (Schlag, 1998; see also 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 an *equilibrium* in the population.

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 *evolutionarily stable* 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

$$f[(P_x, Q_x), (P, Q)] > f[(P, Q), (P, Q)]. \quad (9)$$

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, condition (i) holds if and only if:

$$\pi(P_x, Q_x) \geq \pi(P_x, Q), \quad \text{for all } P \in \mathcal{M}_{n\times m} \quad (10a)$$

and

$$\pi(P_x, Q_x) \geq \pi(P, Q_x), \quad \text{for all } Q \in \mathcal{M}_{m\times n}. \quad (10b)$$

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.[2] A characterization of best responses in terms of the P and Q matrices can be found in Appendix A. By the symmetry of $f[(P_x, Q_x), (P, Q)]$, condition (ii) 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

$$\pi(P_x, Q_x) > \pi(P, Q), \quad (11)$$

that is, the eigen communicative potential of (P_x, Q_x) has to be higher than the communicative potential of any pair of alternative best responses to Q_x and P_x.

Example 1 continued. With conditions (10) and (11) at hand, together with the best-response properties of the P and Q matrices (see Appendix A), it is easy to see that in Example 1 above, the state where $(x'_1, x'_2) = (0.5, 0.5)$ corresponds to a Nash-equilibrium strategy, but is not evolutionarily stable. As it should be true for a Nash equilibrium in mixed strategies, each of the pure strategies that are in its support, in this case, (P_1, Q_1) and (P_2, Q_2), is an alternative best response to the mixed strategy $(P_{x'}, Q_{x'})$: P_1 is an alternative best response to $Q_{x'}$ and Q_1 is an alternative best response to $P_{x'}$, as well as P_2 is an alternative best response to $Q_{x'}$ and Q_2 is an alternative best response to $P_{x'}$. But $\pi(P_1, Q_1) = 3$ as well as $\pi(P_2, Q_2) = 3$, while $\pi(P_{x'}, Q_{x'}) = 2$. But compare this now to the situation where the entire population is of type (P_1, Q_1),

[2] 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.

$(x_1'', x_2'') = (1, 0)$. In this 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_1 is not only *a* best response to Q_1, but the *unique best response* to Q_1, and that Q_1 is not only *a* best response to P_1, but the *unique best response* to P_1. In other words, (P_1, Q_1) is a strict Nash-equilibrium strategy, and hence, in the absence of alternative best responses, 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.

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), which means that 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.

For the class of doubly symmetric games, the replicator dynamics has a special property: The average fitness function constitutes a strict Lyapunov function for the dynamics, that is, a function that is increasing along every trajectory. In other words, the system can be represented by a fitness landscape that satisfies Fisher's fundamental theorem of natural selection: the average fitness increases along every evolutionary path The strict local maxima of this function coincide with the evolutionarily stable states, and as a consequence, the asymptotically stable rest points of the replicator dynamics coincide with the evolutionarily stable states (Hofbauer and Sigmund, 1988, 1998).

It can be shown that an evolutionarily stable strategy of this game will exist only if $m = n$, that is, if and only if 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 are permutation matrices (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).[3] That is, an evolutionarily stable strategy can only be a "language" that bijectively links every concept to exactly one signal such that sender and receiver strategies are perfectly aligned, that is, one mapping is the inverse of the other. If such a strategy is adopted by the entire population, the replicator dynamics will have attained a fitness peak and the maximum communicative potential will be exploited in the population.

An important aspect of this result is that an evolutionarily stable state, and fitness peak, can only be attained in a *monomorphic population state*, that is, a state where one type has become fixed throughout the entire population. Languages, like biological organisms, change on the basis of existing or newly occurring variation. Once evolution has settled into such an evolutionarily stable state, with all variation being driven out and no mutant strategy possibly entering the population, all descendant populations will be of exactly the same type *even if populations get isolated and evolve separately.* There will be no change in the population's (P, Q), and hence no change in the meaning of signals. However, the replicator dynamics does not necessarily converge to such a state.

2.3. 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

$$f[(P_x, Q_x), (P, Q)] \geq f[(P, Q), (P, Q)]. \tag{12}$$

This condition is analogous to the notion of evolutionary stability, only that the strict inequality in the second condition is replaced by a weak inequality. We have already seen in (10) above 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 (11): 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

$$\pi(P_x, Q_x) \geq \pi(P, Q). \tag{13}$$

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 ma-

[3] 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 result that for asymmetric games—and as a consequence also for symmetrized asymmetric games—evolutionarily stable strategies can only be ~~in~~ strict Nash equilibria, and hence in pure strategies, the two results are equivalent. From the best-response properties of the P and Q matrices (Appendix A) 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.

trices,

$$
\begin{aligned}
(P_0, Q_1) &= \left[\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right], \\
(P_0, Q_2) &= \left[\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right], \\
(P_0, Q_3) &= \left[\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}\right], \\
(P_0, Q_4) &= \left[\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}\right],
\end{aligned}
$$

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

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

In order to have an immediate glance at the best-response properties (see Appendix A), we have highlighted the column maxima in P_x and Q_x. 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 fact, $P_x = P_0$ is not only *a* 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 need to have that $P' = P_x = P_0$. 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 that

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

and hence the communicative potential of any pair of alternative best responses to the original Nash equilibrium strategy (P_x, Q_x) is always bounded by the eigen communicative potential of the original Nash equilibrium strategy. Note in particular that (P_x, Q_x) cannot be invaded by a mutant who switches to either of the evolutionarily stable strategies below:

$$
(P_1, Q_1) = \left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right] \tag{14}
$$

or

$$
(P_2, Q_2) = \left[\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right]. \tag{15}
$$

However, (P_x, Q_x) 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 that $\pi(P_x, Q') = 2 = \pi(P_x, Q_x)$.

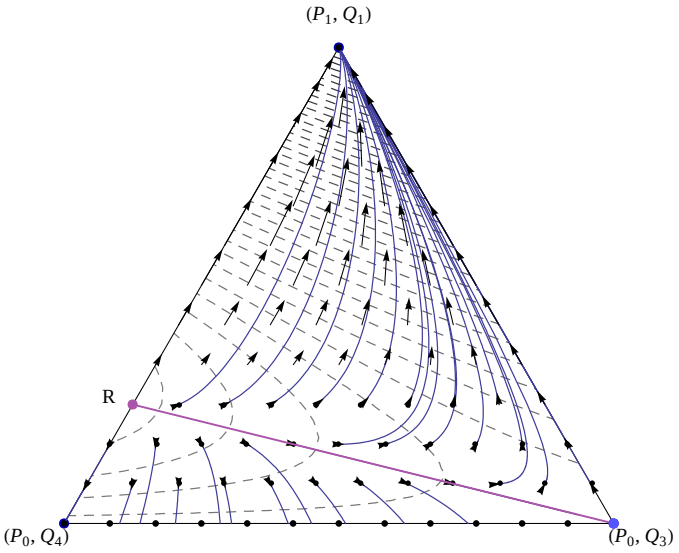


Figure 1: The phase portrait of the replicator dynamics when the population consists only of the three types (P_1, Q_1), (P_0, Q_3), and (P_0, Q_4) mentioned in Example 2. The light purple stationary point $R = 1/4(P_1, Q_1) + 3/4(P_0, Q_4)$ is unstable and, in fact, it corresponds to the global minimum of the average population fitness $\overline{f}(x)$ with respect to these three types; the dashed gray contours represent the level sets of $\overline{f}$, 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 $\overline{f}$, 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 four strategies considered here, but is not neutrally stable in the complete strategy space $\mathcal{P}_{3\times 3} \times \mathcal{Q}_{3\times 3}$; it can be invaded by (P_2, Q_2), see also Figure 5.)

While evolutionary stability is motivated by the idea that a strategy can protect itself against the invasion of mutant strategies—in the strict sense that it can keep mutants from entering the population—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*. Similarly to the implications that we have seen for evolutionary stability, 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), which means that if the system starts close enough to such a rest point, then it will always remain close to

it (but need not converge). 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. Again, this comes from the fact that the average fitness is a strict Lyapunov function for the dynamics; the local maxima of this function correspond to the neutrally stable states.

2.4. *Convergence results for the replicator dynamics*

An important property of the replicator dynamics for doubly symmetric games—also a consequence of the average fitness being a strict Lyapunov function for the dynamics—is that for almost all initial conditions, more precisely for all initial conditions in the interior of the state space, the dynamics will converge to a Nash equilibrium (Akin and Hofbauer, 1982). For the particular 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 one can see that there are components of properly neutrally stable strategies that have a basin of attraction of non-zero measure, which means that if the initial condition comes to lie within this subset of the state space—and there is a non-zero chance that it will do so—the dynamics will converge to a properly neutrally stable state, and not to an evolutionarily stable state. More precisely, properly neutrally stable states occur in connected (but not closed) components 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 1, 2 and 3 illustrate this for a truncation of the game.

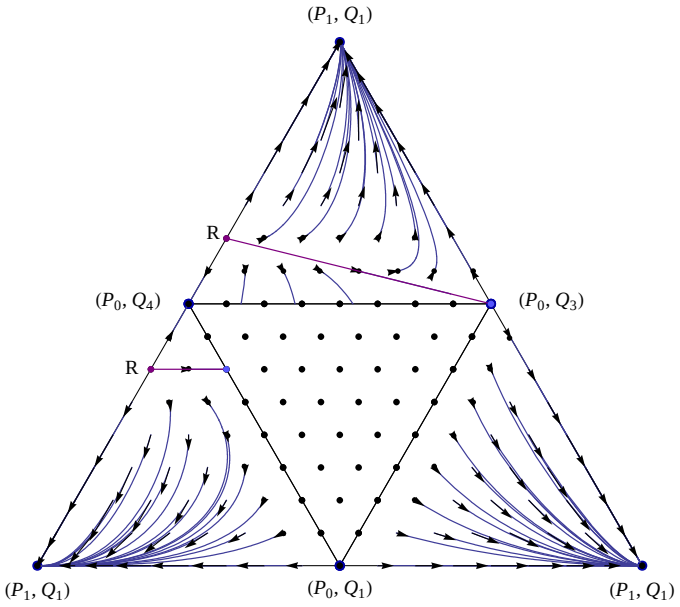


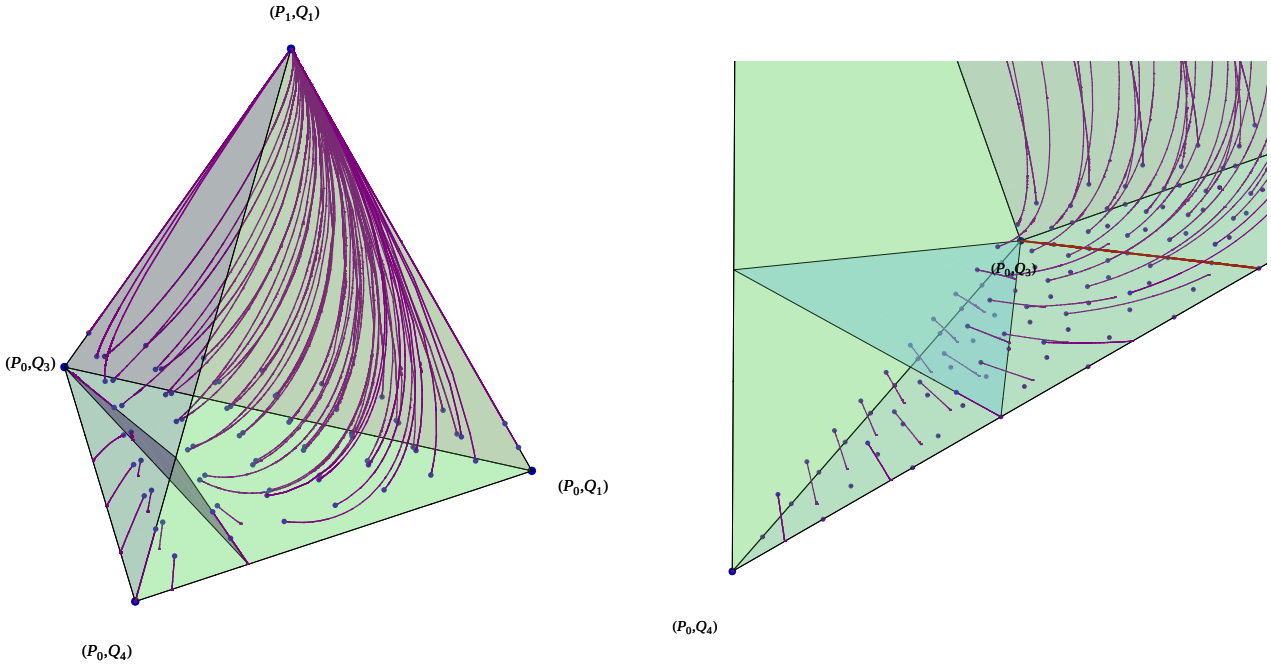
Figure 2: A 2-dimensional foldout of the faces of the phase portrait of the replicator dynamics when the 3-type population of Figure 1 is augmented by the fourth type (P_0, Q_1). Since the types containing P_0 all yield the same payoff when paired against each other, the corresponding face (center) consists entirely of fixed points – however, as we shall see in Figure 3(b) not all of them are neutrally stable. As in Figure 1, the light purple lines represent the *separatrices* of the system and show that the type (P_1, Q_1), which maximizes population fitness, is not globally attracting.

3. Emerging patterns in the P and Q matrices

When we consider this model in the context of language evolution, we are ultimately interested in understanding emerging patterns in the P and Q matrices. We know that the system converges generically to a neutrally stable state—to an evolutionarily stable state or to a properly neutrally stable state. We know the pattern of the P and Q matrices in an evolutionarily stable state—it is a pair of permutation matrices such that one is the transpose of the other (Trapa and Nowak, 2000). It is interesting to ask, then, whether we can identify a pattern in the P and Q matrices that necessarily has to prevail in a neutrally stable state.

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 (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). It is easy to verify that the (P_x, Q_x) that we have seen in Example 2 satisfies this condition: none of the two matrices has a column with multiple maximal elements that are strictly between 0 and 1, and only one of the two matrices, in this case P_x, has a zero column. On the other hand, the Nash-equilibrium strategy $(P_{x'}, Q_{x'})$, with $(x''_1, x''_2) = (0.5, 0.5)$, that we have seen in Example 1 fails to satisfy these conditions: both matrices have columns with multiple maximal elements strictly between 0 and 1.

We can interpret this result 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), and vice versa. Condition (ii), together with the best-response properties of the P_x and Q_x matrices (see Appendix A), implies the following: there can be a signal that will evoke two or more concepts with certain probabilities (multiple entries between 0 and 1 in a row of Q_x, in our example, the first and respectively second row of Q_x), but if this is the case, then either all these concepts have to be mapped to this one particular signal (multiple 1 entries in the corresponding column of P_x, in our example, the first column in P_x), or this signal is indeed never used for any concept (a zero column in P_x, in our example, the second column in P_x). And similarly, for the roles of P_x and Q_x reversed: there can be a concept that is linked to two or more signals with certain probabilities (multiple entries between 0 and 1 in a particular row of P_x), but if this is the case, then this



(a) Phase portrait of the replicator dynamics for four types

(b) A closeup of the phase portrait near neutrally stable states

Figure 3: The full 3-dimensional phase portrait of the replicator dynamics for the four types of Figure 2. As one might expect, the separtrices of Figure 2 delineate the boundary of a higher-dimensional separatrix (semi-transparent plane) which breaks up the state space in two distinct invariant sets, Fig. 3(a). 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 – instead, they converge to neutrally stable states in the face spanned by the other three types. In Fig. 3(b), we present a closeup of the full phase portrait, showing the barrier (red line) between neutrally stable states and unstable ones. As we can see, the set of neutrally stable states (to the west of the red line) attracts a positive measure of initial conditions, whereas unstable states do not.

particular concept will either be inferred from any of the signals that are used to communicate this concept with some positive probability (a column with multiple 1 entries in Q_x), or—and this now can only be true if P_x has no zero column—this concept will indeed never inferred from any signal (a zero column in Q_x). In other words, neutral stability imposes some bounds on the form of ambiguity that can persist on the level of the population's average strategy (P_x, Q_x).

3.1. *Aggreagate vs. individual patterens*

When we talk about regularity patterns of the P and Q matrices, we have to distinguish patterns on two levels: patterns that emerge on the level of the population, and patterns on the level of individual types. Evolutionary game theory provides us with tools that allow us to identify some regularities that arise on the level of the population's average (P_x, Q_x). Here we want to ask what they allow us to say about regularities—and possibly the form of ambiguity—that might be expressed on the level of individual types. In particular we are interested in the interplay between the aggregate and the individual level.

For the specific population-based version of the model that we consider here, ambiguity on the level of the population's average strategy (P_x, Q_x) is a consequence of *variation in the population*. In this perspective, condition (ii) above reads the following way: in a neutrally stable state there can be different resident types who associate a particular signal with different concepts (multiple entries between 0 and 1 in a row of Q_x), but if this is the case, then indeed *all resident types* will use this particular signal to communicate all these concepts (a column with multiple 1 entries in P_x), or this signal is never used by any of the resident types (a zero column in P_x). And similarly, for the roles of P_x and Q_x reversed: different resident types may use different signals to communicate a particular concept (multiple entries between 0 and 1 in a particular row of P_x), but if this is the case, then in fact *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 1 entries in Q_x), or—and this now can only be true if there in no signal that is never used—this concept is never inferred by any of the resident types (a zero column in Q_x). That is, the bounds on the form of ambiguity that is imposed by neutral stability translate into bounds on the form of variation that can be sustained in a neutrally stable state.

Note, on the other hand, that it is the coexistence of types that imposes—and in fact stabilizes—multiplicities in concept-to-signal mappings in a neutrally stable state. Example 2 illustrates this: 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 not inferred by other signals with

higher frequencies), the unique optimal response to this aggregate receiver behavior, in the role of the sender, is to link both concepts, concept 1 and concept 2, to signal 1. In equilibrium all resident types will link both concepts to signal 1, as manifested in the fixed sender matrix $P_0 = P_x$, and it is this property of the population's sender matrix that opens the door for variation in the role of the receiver. Variation is not imposed by best-response conditions, but *once it is there*, it will indeed block mutations that try to resolve the existing ambiguity. In the example, the first column of P_x has two 1 entries, $p_{11} = 1$ and $p_{21} = 1$, but 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. Instead, a Q_x with $q_{11} = 1$ and $q_{12} = 0$, or $q_{11} = 0$ and $q_{12} = 1$ will also be compatible with being a best response to P_x. But once q_{11} and q_{12} have taken values strictly between 0 and 1, any best response to Q_x will have to set p_{11} and p_{21} equal to 1, therby blocking off mutants who try to resolve the existing ambiguity and carburizing the multiplicity in the population's concept-to-signal mappings. In Section 4 we will see that 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, therby giving way to different evolutionary paths that finally might lead to the fixation of different strategies.

Excursion: Evolution of the bidirectional Saussurean sign?

Example 2 has another interesting property. While the fixed sender matrix $P_0 = P_x$ uses signal 1 for two concepts, another signal, signal 2, remains idle. All resident types $l = 1, \ldots, 4$ use a Q_l that is a best response to $P_x = P_0$. But there is no unique way of best-responding to the ambiguity that is manifested in P_0. In particular there is no unique way of best responding to the "empty" signal 2, and this opens the door for the following phenomenon: 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. Hence, 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. We can see from this that the aggregate property that in a Nash equilibrium 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 who is present in such a state.

This is not so much surprising from a game-theoretic point of view—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 by itself—as it is of potential interest from a linguistic point of view. What is interesting about type 1 and 2 in a linguistic context is that even though in the role of the receiver they link each potentially received signal bijectively to exactly one concept, they do not—or better in equilibrium cannot—use the inverse of this mapping in the role of the sender. Or, if one wishes, they are not "consistent" with themselves. This is to pose the question of *bidirectionality*, which stands in fact the beginning of this model in the linguistics literature (Hurford, 1989).

Linguists call the property that if a concept A is linked to a signal σ, and σ 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 a theoretical foundation for the rise of bidirectionality that Hurford (1989) introduces a version of the sender-receiver model that we investigate here in the linguistics literature. His 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 matrix in a bidirectional way will outperform other behavioral types. The formal criterion for bidirectionality that he uses is that the Q matrix has to be a best response to the individual's P matrix, and he calls the type who follows this rule the *Saussurean strategists*. The two other behavioral types that he considers are *imitators*—agents who adopt a randomly drawn P from the population and an independently randomly drawn Q, and *calculators*—agents who, in the role of the sender, adopt a P that is a best response to a randomly drawn Q from the population, and, in the role of the receiver, adopt a Q that is a best response to an independently randomly drawn P. For some initial conditions Hurford can show that Saussurean strategists do better than the two other behavioral types, but altogether the results that he gets do not allow to draw a general conclusion. Hurford's 1989 paper—which is written at a time when evolutionary game theory was very little known outside a small group of biologists and mathematicians and the uniform body of results that we have now in fact only in the making—mentions the work of Maynard Smith, 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 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 is not sufficient to guarantee that each individual type who is present in this population will be bidirectional with itself.

What is particularly interesting about Example 2 in the context of bidirectionality as Hurford asks the ques-

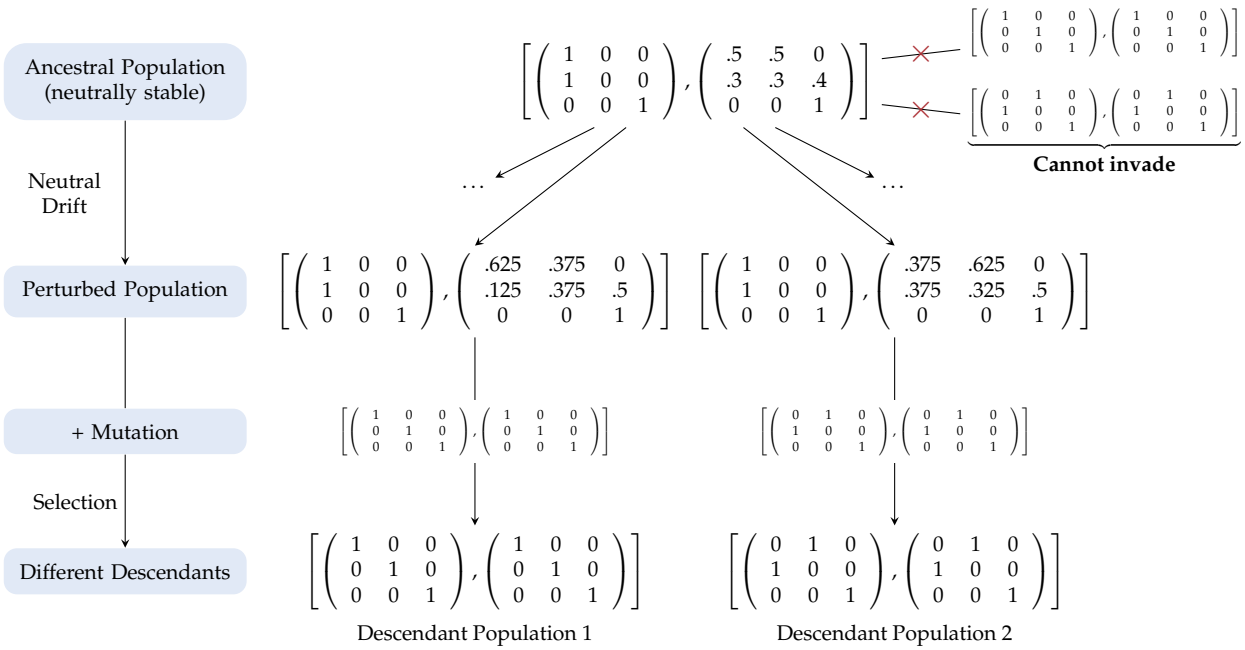


Figure 4: The ancestral population is neutrally stable; in particular, it cannot be invaded by mutants who try to resolve the existing ambiguity by establishing a bijection between concepts and signals. However, a shift in the relative type frequencies (neutral drift) can overcome neutral stability and 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.

tion—namely whether a *behavioral program* for bidirectionality will outperform other behavioral types—is the following: 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 now there is a mutation that appears on type 1 which 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) as we have considered it in (14). Under a monotone 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, that is, a mutation of (P_0, Q_2) to (P_2, Q_2) as we have seen it in (15) will also be blocked. So, even if there was a tendency for agents to adjust their behavior in the roles of the sender and the receiver to make them bidirectional with themselves, there are equilibrium states where such mutations cannot break through, and where as a consequence variation in the population plus the resulting ambiguities will persist. In such a state, selection exercises no further pressure since all resident types attain exactly the same fitness. No mutant strategy can invade, so the forces of selection and mutation have come to an end. But such a state is not immune against *drift* between the selectively neutral resident types.

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 shift 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 shift, 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 (see, for example, Cavalli-Sforza 1997).

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

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 in Example 2 with type frequencies given by $(x_1, x_2, x_3, x_4) = (0.3, 0.3, 0.2, 0.2)$, but that now there is an exogenous random event that brings about a shift in the frequencies of the types already present in the population such that after this shift, $(x_1', x_2', x_3', x_4') = (0.375, 0.125, 0.25, 0.25)$. The average sender-receiver pair then is

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

In this case, $P_{x'}$ 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 that appears on type 1 as we discuss it in the excursion above—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 better against itself, and hence under a monotone selection dynamics will eventually become fixed in the population.

If an even more pronounced diminishment of type 2 comes about, all else being equal, the resulting population state will still be a rest point of the replicator dynamics (since all types gain the same payoff against each other), but it will no longer be a Nash equilibrium. For example, if $(x_1', x_2', x_3', x_4') = (0.39,\ 0.09, 0.26, 0.26)$, the population's average strategy will be

$$(P_{x'}, Q_{x'}) = \left[\begin{pmatrix} \mathbf{1} & 0 & 0 \\ \mathbf{1} & 0 & 0 \\ 0 & 0 & \mathbf{1} \end{pmatrix}, \begin{pmatrix} \mathbf{.65} & \mathbf{.35} & 0 \\ .09 & \mathbf{.39} & .52 \\ 0 & 0 & \mathbf{1} \end{pmatrix} \right].$$

In this case, $P_{x'}$ is no longer a best response to $Q_{x'}$, and hence the pair $(P_{x'}, Q_{x'})$ is no longer a Nash-equilibrium strategy. Since q_{22} is now the maximal element of the second column of $Q_{x'}$, any P that is a best response to $Q_{x'}$ must have $p_{22} = 1$). In fact, (P_1, Q_1) is now a better response to $(P_{x'}, Q_{x'})$ than $(P_{x'}, Q_{x'})$ is against itself, and hence if a mutant to (P_1, Q_1) occurs, it will immediately be on its way to take over the population.

A similar scenario obtains if type 4 goes completely 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,

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

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

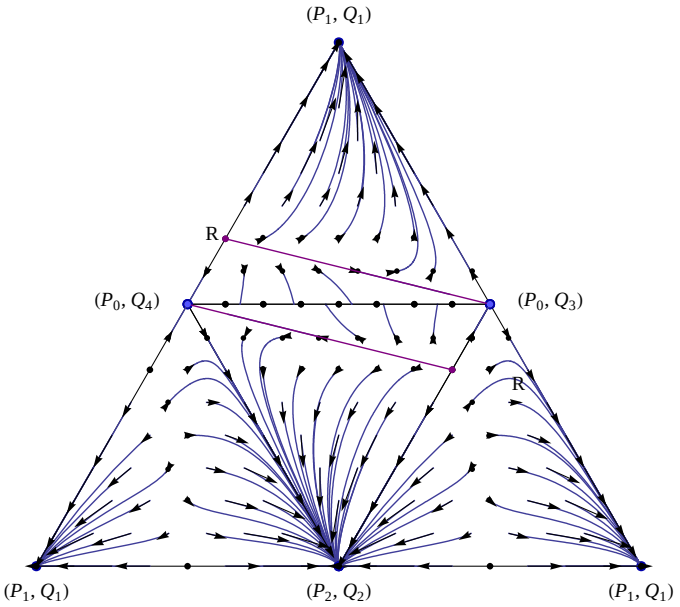


Figure 5: A 2-dimensional foldout of the faces of the phase portrait of the replicator dynamics when the population consists of the four types (P_0, Q_3), (P_0, Q_4), (P_1, Q_1) and (P_2, Q_2) mentioned in Example 2. Every point on the face spanned by (P_0, Q_3) and (P_0, Q_4), except for the two vertices, is neutrally stable.

4.2. *Different histories of change*

A different outcome, however, will obtain if a shift in the relative type frequency 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 finally take over the population. The same mutation will be enabled if type 3 goes completely extinct. Figure 4 illustrates these different evolutionary outcomes. Figure 5 can be read as a graphical representation of this phenomenon for the case that the ancestral population consist only of the two types (P_0, Q_3) and (P_0, Q_4). As long as both (P_0, Q_3) and (P_0, Q_3) are present in the population, 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); if (P_0, Q_3) goes extinct, the population can be invaded by (P_2, Q_2).

So the same ancestral population can give rise to two different perfectly bidirectional protolanguages where the same signal (word) is used for a different meaning.

5. Comments

In linguistics the question whether all human languages can be traced back to one unique common ancestor is the subject of a heated debate. A standard argument in evolutionary accounts for the diversification of language is that is serves as an in-group marker to enhance cooperation in small groups (see, for example, Dunbar, 1998). 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 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*, but that this can come about already under weaker assumptions. From a methodological point of view, postulating a function of language differentiation for enhancing in-group cooperation (which will result in increased material payoffs) amounts to postulating a preference for diversification in the language game, that is, making it an assumption of the model, and therefore cannot be considered something explained by the model.

Language change, like biological evolution, thrives on variation in the population. 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). 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 there is *no* exante incentive for differentiation (as it would be the case, for example, in a model that formalizes the idea that language differentiation serves as an in-group marker), 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 well known to game theorists—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 seems to be full of these type of inefficiencies; the existence of centralized language regulation seems to testify to this effect. Big coordinated jumps are difficult to argue in an evolutionary setting. Neutral drift can open the door for mutations that can take us out of these inefficiencies—but thereby also opening the door for the differentiation of languages.

Acknowledgements: C.P. would like to thank the European Research Council for financial support under the contract "Game Theory and Applications in the Presence of Cognitive Limitation." This work has been supported by Région Ile-de-France.

Appendix A. Best responses in terms of P and Q and the maximum communicative potential

Given a $Q \in \mathcal{M}_{m\times n}$ what is an optimum response in terms of P? For example, let

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \text{ and } Q = \begin{pmatrix} 1-\alpha & \alpha & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

with $\alpha \in (0,1)$. It is easy to see that in this example, P is in fact an optimum response to Q. Due to the quadratic form of the payoff operator $\pi(P,Q) = \sum_i \sum_j p_{ij} q_{ji}$, we can proceed separately for each row i in P, each element of which gets multiplied with the corresponding element in the i-th column in Q. Since the elements in each row i of P are bound to add up to 1, the optimization problem boils down to attributing "optimal weights" to the elements in the i-th column of Q. Now, look at the first column in Q. Since $q_{11} = 1-\alpha$ is the unique maximal element in the first column of Q, and the elements in P are row-wise bound to ad up to 1, in order to maximize $\sum_j p_{1j} q_{j1}$, we have to set $p_{11} = 1$. Similarly, since $q_{12} = \alpha$ is the unique maximum in the second column of Q, in order to maximize $\sum_j p_{2j} q_{j2}$, we have to set $p_{21} = 1$. The third column of Q has two maximal elements, $q_{23} = 1$ and $q_{33} = 1$. Hence, any distribution of weights to the elements in the third row of P that satisfies $p_{32} + p_{33} = 1$ will maximize $\sum_{j=1}^{m} p_{3j} q_{j3}$—and $p_{33} = 1$ is a possible realization of this. So in sum, P is an optimum respnse to Q. If we reverse the problem and fix P and ask what is an optimal response in terms of Q, then, since the elements in Q are row-wise bound to ad up to 1, it is convenient to look at the operator $\pi(P,Q)$ as $\sum_j \sum_i p_{ij} q_{ji}$, that is, first multiplying each column j in P with its corresponding row j in Q, and then summing over all j. It can be easily verified that in the example above, Q is also an optimal response to P: For the first two columns in P, the situation is similar to situations that we have already seen with the columns in Q. Only the case of the third column in P is different: all elements are equal to 0. But since all elements are equal to 0, all elements are maxima of this column, and therefore any distribution of weights to the elements in the third row of Q—and $q_{33} = 1$ is a possible realization of this—is an optimal response to this column.

In general, if we fix $Q \in \mathcal{M}_{m\times n}$, then for any $P \in \mathcal{M}_{n\times m}$ that maximizes $\pi(P,Q)$, we have that:

(a) $\sum_{i^\star} p_{i^\star j} = 1$, where $i^\star \in \text{argmax}_i\{q_{ji}\}$, and
if $p_{i^\star j} \neq 0$, then $i^\star \in \text{argmax}_i\{q_{ji}\}$, for all j; and

(b) $\max_P(\pi(P,Q)) = \sum_j \max_i q_{ji}$.

Similarly, if we fix $P \in \mathcal{P} \in \mathcal{M}_{n\times m}$, then for any $Q \in \mathcal{M}_{m\times n}$ that maximizes $\pi(P,Q)$, we have that:

(a) $\sum_{j^\star} q_{j^\star i} = 1$, where $j^\star \in \text{argmax}_j\{p_{ij}\}$, and
if $q_{j^\star i} \neq 0$, then $p_{ij} = \text{argmax}_{j'}\{p_{ij'}\}$, for all i; and

(b) $\max_Q\{\pi(P,Q)\} = \sum \text{argmax}_{j'}\{p_{ij'}\}$.

That is, for fixed Q, the communicative potential $\pi(P,Q)$, for any $P \in \mathcal{M}_{n\times m}$, is bound by the sum of the column maxima in Q; and for fixed P, the communicative potential $\pi(P,Q)$, for any $Q \in \mathcal{M}_{m\times n}$, is bound by the sum of the column maxima in P.

References

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

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.

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

Dunbar, R., 1998. Grooming, gossip, and the evolution of language. Harvard University Press, Cambridge, MA.

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

Gavrilets, S., Gravner, J., 1997. Percolation on the Fitness hypercube and the evolution of reproductive isolation. J. Theor. Biol. 184, 51–64.

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.

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., 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.

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., 2006. Evolutionary dynamics: exploring the equations of life. Belknap press of Harvard University Press, Cambridge, MA.

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

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

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

Pagel, M., Atkinson, Q., Meade, A., 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature 449, 717–720.

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

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

Warnow, T., 1997. Mathematical approaches to comparative linguistics. Proc. Natl. Acad. Sci. USA 94 (6585-6590).

Weibull, J. W., 1995. Evolutionary game theory. MIT Press, Cambridge, MA.