What is a universal? On the explanatory potential of evolutionary game theory in linguistics*

Gerhard Jäger

University of Tübingen, Institute of Linguistics,
Wilhelmstr. 19, 72072 Tübingen, Germany
gerhard.jaeger@uni-tuebingen.de

Abstract. Natural languages are shaped by evolutionary processes, both in the sense of biological evolution of our species, and, on a much shorter time scale, by a form of cultural evolution. There are long research traditions in theoretical biology and economics (a) to model communication by means of game theory, and (b) to use game theory to study biological and cultural evolution. Drawing mostly on work by Huttegger (2007) and Pawlowitsch (2008), the paper argues that results and methods from game theory are apt to formalize the intuitive notion of linguistic universals as emergent properties of communication.

1 Introduction: Language and evolutionary game theory

One of the central issues in modern linguistics is the search for **universals**, that is properties that are shared by all languages. Empirical research over the past few decades has unearthed a solid amount of universals or quasi-universals (properties that are shared by almost all natural languages), some of them quite contingent.

There is an ongoing intense and at times ideological discussion within the linguistic community what a satisfactory explanation for universals should look like. One school of thought (most prominently defended by Noam Chomsky; see for instance Chomsky 1957, 1995) takes it that universals are reflexes of our innate linguistic competence that is shared basically by all humans and ultimately genetically determined. We will call this the *nativist* position. The other creed (see for instance Bybee 2001, Barlow and Kemmer 2000, or Haspelmath 1999) holds the view that languages are adapted to its functions in communication and cognition. Therefore linguistic universals should be explained in terms of language function. This school of thought is usually dubbed *functionalist*.

Both approaches face severe epistemological problems. It is not much of an explanation to claim that a certain universal is based on innate, genetically determined, properties of the human brain. Without a independent justification

* Thanks to Simon Huttegger, Christina Pawlowitsch and two anonymous reviewers for important comments on an earlier version of the article.

for that assumed innate feature of the brain, the innateness hypothesis amounts to little more than a restatement of the facts. Functionalism, by itself, is no more explanatory either. To show that a certain property of human languages facilitates their usage has to be complemented by a plausible reconstruction of a causality that leads to this kind of adaptation. Additionally, functionalist explanations are frequently *post hoc*, i.e. functionalist research frequently starts out with some empirically established universal and tries to find a function that the universal is an adaptation for, rather than truly predicting universals from function.

However, both approaches seem to be fundamentally true, if properly conceived. Nobody would deny that there are innate constraints on what kind of linguistic items an infant is able to acquire and to put to use. To take a simple example: Many languages employ the *relative pitch* of speech sounds to encode grammatical distinctions. For instance, in English a rising intonation at the end of a sentence can be used to mark the sentence as a question. In other languages, as for instance in Chinese, the same sequence of sounds can acquire a completely different meaning, depending on whether it is pronounced with a rising or a falling intonation. There are no languages, however, that would employ *absolute pitch* to mark linguistic distinctions. A low pitch of a female speaker might still be higher, in terms of absolute frequency, than the high pitch of a male speaker. What matters for natural languages though is the position of a certain pitch within the range of pitches that a particular speaker is able to produce. For all we know, the reliance on relative pitch is a good candidate for a linguistic universal that is based on innate properties of the human brain.

On the other hand, some linguistic universals cannot be reduced to the cognitive state of a single language user. They are irreducibly social in nature. A well-known candidate is Zipf's law (Zipf 1935, 1949). It states that in a sufficiently large corpus of naturally occurring speech, the frequency of occurrence of the words of the language are distributed according to a power law. The most frequent word is used about twice as many times as the second most frequent, about ten times as often as the tenth most frequent word etc. More generally, if you list the words of a language according to their frequency of usage, the rank of a word in this list is approximately inversely proportional to its frequency of occurrence. This law has been validated many times for many languages. It is a good candidate for a universal property of language usage. However, this regularity is certainly not cognitively represented in the minds of language users. Rather, it is an emergent property of the usage of language in social interaction. The two approaches should thus be considered as complementary rather than as mutually exclusive. Evolutionary game theory, I would like to argue, provides a formal framework that model questions of language evolution that allows us to integrate the two approaches, innateness and the social function of language.

2 Modeling language evolution as a game

Game theory is a formal language that allows us to study the *interaction of individual agents*, and the joint outcome of this interaction. The evolutionary branch of game theory explicitly takes into account the *path dependence* of this interaction under some limited form of rationality.

An evolutionary game consists of two parts. First, the so–called *stage game*, that is, the basic situation of interaction that is encountered repeatedly among a group of individuals, where it is defined what are possible action choices of individuals, so–called *strategies*, and how individuals' payoffs are calculated as a function of the strategy profile in the whole population. Second, the *game dynamics*, which described according to which rule individuals' strategies are updated from one period to the next. This rule can be either exogenously given, representing something like a "law of nature", that is, a feedback mechanism of the environment; or it can be derived from some optimizing behavior of individual agents. This is usually seen as the part of the model setup where different degrees of rationality enter the description of the problem. In an economic model it is often plausible to assume that agents are perfectly rational, in the sense that they can calculate their optimal choice of action given their state of knowledge and their beliefs about other agents actions. On the other hand, in most ecological or biological models it rarely makes sense to assume that agents consciously interact rather they perform some predefined program whenever they get the stimulus to do so. Taking these as the two extremes of one scale, I would argue that most linguistic applications lie somewhere in between, depending on the particular problem at hand. If we are interested in an aspect of the origins of language in an anthropological sense we are most probably closer to a biological set up, whereas if we are interested in some aspects of the pragmatics of language use, it definitely makes sense to assume that agents are conscious about their interaction, that they have beliefs about other players actions, that they employ some kind of optimizing behavior given all these considerations etc.

Any optimizing behavior, or even imitation, in the last event, implicitly assumes some innate abilities or properties that are not further explained in the model. Assumptions on innate abilities of individual agents, however, do not only enter the construction of the game dynamics, they also may be part of the description of possible actions choices of individual agents. For example, if a strategy of an agent represents a program to perform a particular utterance whenever he or she observes a particular event of nature, this definitely involves the assumption that this agent has an abstract notion of this event and that he or she has the ability to link this to an arbitrary sign.

What is specific about applications of evolutionary game theory to linguistic questions is that strategies in the stage game are very often not just verbal descriptions like "cooperate" or "do not cooperate", but take the form of a *mathematically complex object*, similar to a quantitative trait in biology. This is a direct consequence of the fact that linguistic interaction is sequential in nature, and thus most naturally modeled as an extensive game. Normalizing an extensive game leads to strategies that are functions of a particular kind. As such these

strategies can display properties that we can describe in some formal language. For the kind of game I am going to discuss below, strategies can be described by stochastic matrices, using the language of linear algebra. Eventually, we aim at learning something about the *regularity patterns* of these strategies that are used in an equilibrium outcome of the model.

Thus, innateness assumptions typically enter the description of a game, whereas the functionalist point of view is reflected in the solution concept applied. This is not to say that innateness enters the model *only* at the level of assumptions. An evolutionary game can also serve the purpose of explaining some property of language that is considered to be innate—but what sustains this property are the equilibrium conditions of individuals interaction. It is in this sense I consider the explanatory value of a game theoretic model as *functionalist*—even though it may rely on assumptions that concern innate abilities of individual agents.

I will discuss this and related issues in more detail for a specific class of games that are widely used in game theoretic approaches to language.

3 Signaling games

Signaling games in the style of Lewis (1969) or Nowak and Krakauer (1999) have received particular attention in the newly arising literature that uses evolutionary game theory to study questions about the evolution of language.

In these games there is a finite number of events that potentially become the object of communication and a finite number of arbitrary signs. In each round of the game, nature presents the sender with on of the events. The sender in turn emits a signal that is visible to the receiver. Finally, the receiver guesses an event, possibly using the signal received as a clue. If the guess of the receiver is correct, both players score a point, otherwise neither receives a payoff.

A strategy in the role of the sender is thus a map from the set of events to the set of signals, and a strategy in the role of the receiver is a map from received signals to events.

As long as domain and range of these functions are finite, it is convenient to use the language of linear algebra and to represent functions as matrices.

A strategy in the role of the sender can be represented by an $n \times m$ matrix P (n being the number of events, and m the number of signals), such that each row contains exactly one cell with the entry 1, while all other cells have the entry 0. The intended interpretation is that $p_{ij} = 1$ if event i is mapped to signal j. Likewise a strategy in the role of the receiver can be represented by an $m \times n$ matrix Q, where here the interpretation is that $q_{ji} = 1$ if signal j is associated event i and 0 otherwise.

For the sake of simplicity, in the context of this paper I confine attention to cases where $n = m$, i.e. there are exactly as many signals as events. Also, I assume that all events occur with the same probability, and that sending or receiving signals does not incur any costs. With these assumptions, the payoff

function is identical for both players, and it can be defined by

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

(The utility that was informally described in the text can be obtained if $\pi(P,Q)$ is divided by n, because each event occurs with probability $\frac{1}{n}$. However, multiplying all payoffs by a constant factor does not alter the structure of a game, and we can thus as well drop this factor.)

In the language of linear algebra, this can conveniently be written as

$$\pi(P,Q) = \mathrm{tr}(PQ),$$

where $\mathrm{tr}(PQ)$ denotes the trace of (PQ).

It is assumed that all individuals of a linguistic community finds themselves in the roles of sender and receiver with equal probabilities. A strategy for the *symmetrized game*, then, is a pair of two matrices, (P,Q), and the payoff function is given by

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

Note that this payoff function is symmetric,

$$F[(P,Q),(P',Q')] = F[(P',Q'),(P,Q)],$$

giving rise to a so–called *doubly symmetric* or *partnership game*, that is, a symmetric game with a symmetric payoff function.

3.1 Mixed strategies and population games

Rephrasing the model at hand in a population based framework, every pair (P,Q) can be identified with a particular *type* of agent. A *state of the population*, then, is a vector of type frequencies,

$$x = (x_1, x_2, \ldots, x_L) \text{ s.t.} \sum_{l=1}^{L} x_l = 1,$$

where x_l is the fraction of agents using pure strategy l. Formally such a vector of type frequencies is equivalent to a *mixed strategy*.

The average payoff of a type is interpreted as its *fitness*,

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

the *average fitness in the population* is denoted by

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

To every vector of type frequencies $x = (x_1, x_2, \ldots, x_L)$ we can assign the *population's average strategy profile* in terms of the P and Q matrices,

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

which can be written as

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

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

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

In general the average strategy profile is a pair of two stochastic matrices, which are denoted by $(\bar{P}, \bar{Q})$. The average payoff or *fitness* of a type $f_l(x)$ then can be written as the payoff of this strategy from play against the population's average strategy,

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

and the *average fitness in the population* is the payoff of the population's average strategy from play against itself,

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

Hurford (1989) introduces essentially the same model, though he does not use the language of game theory. He uses this model to study the evolutionary emergence of a particular linguistic universal known as *bidirectionality*, that is, the property that whenever an individual links a particular sign to a particular concept (object of communication), then this individual will also link this concept to that sign.[1] In a nutshell, this means that each adult speaker of a language uses, in the role of the speaker, a code that she is able understand in the role of the hearer. This feature is so deeply entrenched in our conception of "language" that it seems to be almost too obvious to mention. However, there are plenty of signaling systems that do not have this property. An obvious example are intermediate stages in the acquisition of a natural language (both

[1] Sometimes this is also referred to as the notion of the Saussurean sign.

in first language acquisition by infants and in second language acquisition by adults). New linguistic items are much faster acquired passively than actively. This means that a language learner can *interpret* some linguistic items correctly without being able to *produce* them.

In some signaling games there are strategies that display perfect bidirectionality. For 2 events and 2 signals, for instance, these are

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

But there are also strategies that are not bidirectional at all; rather they display what one would call "perfectly inconsistent behavior" in the role of the sender and the receiver, namely

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

Most linguistic theories tacitly assume that bidirectionality is an innate property, which is ultimately genetically determined. If we look at isolated sender–receiver interaction, bidirectionality is not required for successful communication. But obviously bidirectionality seems to be fixed in humans. Shifting the question to an evolutionary framework, Hurford asks whether bidirectionality could get fixed due to some evolutionary advantage in a population based setting. In order to test this hypothesis, Hurford ran a series of computer experiments, where he lets "Saussurean strategists", that is, individuals who update their Q according to their received P in a best–response fashion according to the asymmetric game, compete against types with other behavioral rules. Individuals who communicate better leave relatively more offspring and parents transmit their type to their kids. This simulation is run for different initial conditions. There seems to be good evidence that for most initial conditions Saussurean strategists indeed do better than other behavioral types. However, this is not true for all initial conditions.

3.2 Equilibrium selection

The kind of evolutionary dynamics that Hurford, Nowak and their coworkers assume in their simulations assigns fitness to communicating agents which is proportional to the average communicative success of that agent within the given population. This success rate, in turn, depends on the relative frequency of communicative strategies within this population. We are thus dealing with frequency dependent selection, which can be modeled by means of evolutionary game theory.

In the tradition of evolutionary game theory, a Nash equilibrium in mixed strategies is interpreted as an equilibrium composition of the population. It generally holds that in a Nash equilibrium in mixed strategies, all strategies that are played with some positive probability must yield the same payoff given the

specific probability mix of all the other players. In the population interpretation of mixed strategies in mind, this translates into the condition that in a Nash–equilibrium state of the population every type l that is present with some positive frequency must yields the same average payoff as all the other resident types—given the actual composition of the population.

For a Nash–equilibrium strategy of a symmetrized game, the strategies choices in the two roles have to be best responses to each other, that is, P is a best response to Q and Q is a best response to P. The condition that Q is a best response to P is the criterion that Hurford (1989) uses to characterize bidirectionality. Adopting this concept of bidirectionality, we may say that in a Nash–equilibrium composition of the population, bidirectionality is satisfied on the level of the population's average sender and receiver matrices. Interestingly, this property does not extend from the population's average to the individuals' level.

Let us consider an example. Suppose $n = m = 2$, i.e. we have two signals and two event. Then the set of all sender matrices is given by

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

and the set of all receiver matrices is given by

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

Now suppose one half of the population uses pure strategy (P_1, Q_2), and the other half uses (P_2, Q_1). Then population's average strategy is

$$(\bar{P}, \bar{Q}) = \left[\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} \right].$$

In this case $\bar{P}$ is a best response to $\bar{Q}$ and vice versa, but still every individual agent is perfectly inconsistent between his or her sender and receiver strategy.

As for most models used in language evolution, the signaling game considered has an abundance of equilibria. In the case of 2 events and 2 signals, there are already 6 symmetric Nash equilibria in pure strategies, (P_1, Q_1), (P_2, Q_2), (P_3, Q_3), (P_3, Q_4), (P_4, Q_3), (P_4, Q_4), and there are whole continua of equilibria in mixed strategies. For these equilibria the average strategy profile is of the form

$$\left[\begin{pmatrix} 1-\alpha & \alpha \\ 1-\alpha & \alpha \end{pmatrix}, \begin{pmatrix} 1-\beta & \beta \\ 1-\beta & \beta \end{pmatrix} \right],$$

where α and β are between 0 and 1—for example, if one half of the population uses pure strategy (P_1, Q_1), and the other half uses (P_2, Q_2). The population's average strategy profile then is

$$\left[\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}\right].$$

$$\left[\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}\right].$$

The problem of equilibrium selection is therefore of major importance.

An important part of the program in evolutionary game theory has focused on establishing links between the static analysis of Nash equilibria, and its refinement concepts, and the stability properties of game dynamics. A great deal of effort has been devoted to the so–called *replicator dynamics*.

In continuous time, the replicator dynamics is given by a system of L differential equations,

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

where $\dot{x}_l$ denotes the derivative of x_l with respect to time. So the growth rate of type l is given by the difference of its average payoff minus the average payoff in the population.

This dynamics can be interpreted in the sense of both biological as well as cultural transmission of strategies from one generation to the next (roughly corresponding to the nativist and the functionalist perspective in linguistics that were mentioned in the beginning). Agents who communicate more successfully are more successful in finding mates, acquiring sufficient sources of food, escaping dangers, and so on, which yields them either a direct or an indirect advantage in reproduction, or increases the chances of their offspring to reach reproduction age. Or, in the context of cultural evolution, agents who communicate more successfully are more likely to be imitated by other agents and therefore the strategies that they use will reproduce with a higher rate.

A Nash–equilibrium strategy is necessarily a rest point of the replicator dynamics. The reverse is not generally true. For example, every monomorphic state, that is, a state where the whole population consists of only one type, trivially is a rest point of the replicator dynamics, but it is not necessarily a Nash–equilibrium strategy.

In language evolution we are not so much interested in any one particular equilibrium; rather we want to understand where a particular dynamics typically will lead us to and what this implies for the qualitative regularity patterns of the communicative strategies that are used by agents.

The most commonly applied refinement concept in an evolutionary context is that of an *evolutionarily stable strategy*. A strategy played in a symmetric Nash equilibrium is evolutionarily stable if either (i) it has no alternative best

replies (that is, if it is a strict equilibrium), or if (ii) in case that there is an alternative best reply, this alternative best reply yields a strictly lower payoff against itself than the original Nash strategy yields against this alternative best reply. As a lighter version of this, if in case (ii) the alternative best reply yields only a lower or equal payoff against the original Nash strategy, then the original Nash strategy is called a *neutrally stable* or *weakly evolutionarily stable strategy*.

Though their name seems to hint at some dynamic story, both evolutionary and neutral stability are as such static refinement criteria for symmetric Nash equilibria. However, as it has been shown by Taylor and Jonker (1978) for the continuous case, and by Hofbauer, Schuster and Sigmund (1979) for the discrete case, every *evolutionarily stable* strategy is a *locally asymptotically stable* rest point of the replicator dynamics. This means that a system that has reached such a state will return there if it is disturbed by a small perturbation. In analogy to this, Thomas (1985), and in a more general context Bomze and Weibull (1995), show that every *neutrally stable* strategy is *Lyapunov stable* in the replicator dynamics. If a system is in a Lyapunov stable state, it will remain within the local environment of this state if a small perturbation occurs.

Unfortunately, none of the converses of these results is true in general. This means that whenever we have found an evolutionarily stable strategy, we know that once a population has attained such a strategy, it is locally asymptotically stable, but this does not rule out there to be other asymptotically stable rest points that do not correspond to an evolutionarily stable strategy; analogously for neutral stability and Lyapunov stability. However, in the case of signaling games, help comes from their symmetry properties.

Akin and Hofbauer (1982) show that for doubly symmetric games each orbit of the replicator dynamics, indeed, *converges to some rest point*. This is related to the property that for such games, the replicator dynamics induces a strictly monotonic increase in the average payoff along every non–stationary solution path, which in biology this is known as *Fisher's fundamental theorem of natural selection*. Hofbauer and Sigmund (1988) show that in this case a locally asymptotically stable rest point is evolutionarily stable, so that for doubly symmetric games, evolutionary stability indeed *coincides* with asymptotic stability. Bomze (2002) shows that an analogous result holds true between neutral stability and Lyapunov stability so that for doubly symmetric games with pairwise interaction a rest point of the replicator dynamics is Lyapunov stable if and only if it corresponds to a neutrally stable Nash equilibrium. So we can rule out that there are other rest points that are Lyapunov stable, and the dynamics typically will lead to a Nash equilibrium that satisfies neutral stability. Once we are able to identify general patterns in the strategies used in population states that satisfy the essentially static criterion of neutral stability, then this will tell us something about the regularity patterns of the communicative strategies that can be expected to arise in the long run.

For the signaling game considered here evolutionarily stable strategies are all pairs of permutation matrices (P, Q) such that one matrix is the transpose of the other (Wärneryd 1993, Trapa and Nowak 2000). For 2 events and 2 signals,

these are the two pairs

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

For 2 events and 2 signals evolutionarily stable strategies are also the only neutrally stable strategies. However, if the number of events or signals is greater or equal to 3, this is no longer true.

For example, if $n = m = 3$ a possible neutrally stable strategy looks like

$$(P, Q) = \left[\begin{pmatrix} 1-\alpha & \alpha & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1-\beta & \beta \end{pmatrix} \right], \ \alpha, \beta \in (0, 1).$$

There is no strategy $(P', Q') \in \mathcal{P}^{\Delta} \times \mathcal{Q}^{\Delta}$ such that $F[(P, Q), (P', Q')] = F[(P, Q), (P, Q)]$ and such that $F[(P, Q), (P', Q')] > F[(P', Q'), (P', Q')]$. The same is true if in the example above *either* α *or* β is 0 or 1, which means that one of the two matrices contains a column that consists entirely of zeros.

This example shows that signaling games may possess equilibrium states that are neutrally but not evolutionarily stable. In fact, Huttegger (2007) and Pawlowitsch (2008) show that games with more than two events and signals have infinitely many such states. These are generally states that involve a certain amount of synonymy and homonomy on the population level; they represent sub-optimal communication strategies.

The mentioned authors furthermore show that this set of neutrally but not evolutionarily stable states always has a basin of attraction that has a Lebesgue measure larger than zero. So if an initial state is chosen at random and each initial state has a positive probability density, there is a positive probability that the population will converge to such a suboptimal state. I performed a numerical approximation which revealed that the joint basin of attraction of the set of neutrally but not evolutionarily stable strategies in a 3×3 game comprises about 1.2% of the state space. (See the Appendix for details.)

4 Evolution of signaling games and linguistic universals

Let us return to the main methodological point of this paper, the potential of evolutionary game theory for illuminating the concept of a linguistic universal. Intuitively, a universal of a certain game under a certain dynamics is a set of population states U such that with probability one, a population will eventually enter U and never leave it afterwards. To make this precise, one has to assume a probability distribution over initial states. Since I assume that the stage game by definition excludes all "impossible" states, I take it that each strategy, pure or mixed, has a positive probability density. Since under the replicator dynamics, each trajectory in a doubly symmetric game converges to some rest point (as shown in Akin and Hofbauer 1982), a set U is a universal if and only if *almost all initial states converge to some point* $x^* \in U$.

Given the considerations in the previous section, we can conclude that any set containing the set of Nash equilibria is a universal. This follows from the facts that (a) all orbits converge to a rest point, and (b) if an interior point converges to a single point, this point is a Nash equilibrium (as shown for instance in Hofbauer and Sigmund 1998:69/70). Since the boundary of a simplex is a null set, almost all points converge to some Nash equilibrium.

The significance of Huttegger's and Pawlowitsch's result is that each universal has to include all neutrally stable states, including those that are not evolutionarily stable. In the model chosen, perfect bidirectionality (in the sense of a one-one map between forms and meanings) is thus not a universal.

Is it possible to give a more precise characterization of the set of universals of this game? Huttegger (2007) proves that almost all points of the state space converge to a Nash equilibrium at the boundary of the state space. Translated into our terminology, this means that it is a universal that not all possible grammars are represented in the population.

It is important to stress that these universals are social in nature, rather than cognitive. Being in a Nash equilibrium state at the boundary of the state space is a property of a population, not of an individual member of such a population.

To draw conclusions about universal properties of the state of individual agents, it would be necessary to narrow down the class of population level universals further. This problem proves to be surprisingly difficult though, and I have to close this section with some conjectures and suggestions for further research.

It might perhaps seem plausible to assume that almost all orbits converge to a Lyapunov stable—and thus neutrally stable—point. This would amount to the claim that any set containing all neutrally stable states is a universal. However, it is possible to come up with games where this is not the case. In the Appendix I discuss a doubly symmetric game which has a non-neutrally stable Nash equilibrium that attracts a set with a positive measure.

Of course, this game is not a signaling game. It is thus possible—and in fact, it seems highly likely—that in signaling games, almost all points converge to a neutrally stable strategy. At the present point, I have to leave this open as an issue for further research.

The notion of a universal depends on the underlying dynamics, and modifying this dynamics may thus lead to different universals. An empirically well-motivated choice would be the kind of stochastic dynamics of finite populations that is studied in (Kandori, Mailath and Rob, 1993) and (Young, 1993). Strictly speaking, these models predict that there are no universals at all except for the trivial one comprising the entire state space. However, linguists frequently operate with the notion of a *statistical universal*, which is a property that is shared by almost all languages. This concept could be formalized as a set with the property that an orbit with a randomly chosen initial state, observed at a randomly chosen time, will be within this set with a sufficiently high probability. Under stochastic evolution, this would correspond to a set containing all stochastically stable states and their environments. In fact, van Rooij (2004) and Jäger (2007)

employ the notion of stochastic stability to derive certain empirically attested statistical universals from a game theoretic model.

5 An example: the evolution of case marking

Let me finally illustrate the theoretical notions developed above by an example that is linguistically somewhat more informed than the highly schematic signaling games considered so far. In Jäger (2007), I give a formalization of the typology of case marking systems in terms of a game that is a slight generalization of a signaling game. The presentation of this model here is necessarily rather dense; the interested reader is referred to the mentioned article.

Let us assume that there are two options to morphologically mark the agent argument of a transitive verb, ergative case or nominative) case, i.e. the morphological form of subjects of intransitive verbs. Likewise, we assume to options for marking the patient argument, accusative or nominative (sometimes called absolutive case in the context of ergative languages). We furthermore assume that nominative case is unmarked, i.e. less costly than both ergative and accusative marking. So both syntactic core roles in a transitive clause may be marked in a unambigous but costly or in an ambiguous but cheap way.

We are interested in split systems, i.e. systems where the assignment of overt ergative/accusative marking is conditioned by semantic properties of the NP in question. For simplicity's sake, only those strategies are considered where pronouns may follow a different case marking paradigm than full NPs. Of course non-split strategies are also taken into account. A sender strategy has to specify

1. whether pronominal agents are realized in ergative or nominative,
2. whether non-pronominal agentes are realized in ergative or nominative,
3. whether pronominal patients are realized in ergative or nominative, and
4. whether non-pronominal patients are realized in ergative or nominative.

Hence a sender strategy can be represented as a binary 4-tuple, where 1 means "unambiguous marking" (ergative or accusative, depending on the role of the NP), and 0 means "nominative marking". There are 16 such strategies.

When interpreting a transitive clause, the receiver observes whether the two NPs in question are pronominal or not, and their case marking. Based on this information, he has to decide which of the two NPs is agent and which is patient. As a further abstraction, we assume that the receiver has to make this decision solely on the basis of this information, i.e. factors such as word order, semantic plausibility etc. are disregarded. If at least one of the two NPs is in a non-nominative case, the case morphology completely disambiguates the role assignment. If both NPs are in nominative and both have the same status with regard to pronominality, the receiver has no choice but to make a random guess. The only scenario where he can make a strategic choice arises if one NP is a pronoun and the other one a full NP, and both are in nominative case. So essentially there are only two receiver strategies. In the scenario just described, the possible choice are:

1. the pronoun is agent and the full NP is patient (abbreviated as pA), or
2. the pronoun is patient and the full NP is agent (abbreviated as pP).

The utilities of sender and receiver as assumed to be identical. If the receiver opts for the correct role assignment, both players score a point. Additionally, each occurrence of a non-nominative case marking in a clause incurs a cost for both players.[2]

In Jäger (2007), the probabilities of the four different combinations of syntactic roles and pronominality status are estimated using a corpus study. Furthermore, a range of various differential costs for case marking are considered. In the present context, I will only discuss one configuration, where each case exponent incurs a cost of 0.1. The normalized asymmetric utility matrix then comes out as in table 1.

sender strategies	*receiver strategies* pA	pP
1111	0.80	0.80
1110	0.88	0.88
1101	0.82	0.82
1100	**0.90**	**0.90**
1011	0.81	0.81
1010	0.85	0.85
1001	0.81	0.83
1000	0.86	0.87
0111	0.89	0.89
0110	**0.97**	**0.26**
0101	0.81	0.81
0110	0.89	0.18
0011	**0.90**	**0.90**
0010	0.94	0.23
0001	0.81	0.82
0000	0.85	0.15

Table 1. Asymmetric normalized utilities for the case marking game

Of the 16 sender strategies, only 3 are *strictly undominated* (shown in bold face). These are

- 1100: all agents are marked in ergative and all patients in nominative, i.e. an unconditional ergative system,
- 0110: ergative marking only occurs with full NPs and accusative marking only with pronouns, i.e. a typical double split system, and

[2] Arguably these costs apply to the sender but not to the receiver. However, it can be shown that the stability properties of the resulting game remain unchanged if signaling costs are assigned to both players.

– 0011: all agents are in nominative and all patients in accusative, i.e. an unconditional accusative system.

It follows directly from the definition of the replicator dynamics that all trajectories converge to a state where only strictly undominated strategies have a positive probability.[3] Therefore we can restrict attention to the sub-game that only comprises undominated strategies. Also, since 1100 and 0011 have the same utility profile, we can collapse them into a single strategy for the purpose of analysis. I will continue to use the name 0011, with the intended meaning that this covers any mixture of 0011 and 1100. The utility matrix is given in table 2. This

sender strategies	*receiver strategies* pA	pP
1100	0.90	**0.90**
0110	**0.97**	0.26

Table 2. Reduced game

game has two pure Nash equilibria (shown in bold). Additionally, it has infinitely many mixed strategy equilibria. There is a threshold $\theta \approx 0.90$ such that each mixed strategy with $p_s(0110) = 0$ and $p_r(pP) > \theta$ is also a Nash equilibrium.

This structure carries over the the symmetrized version of this game, shown in table 3: This game has two pure symmetric Nash equilibria as well (shown

	$1100/pA$	$0110/pA$	$1100/pP$	$0110/pP$
$1100/pA$	0.90	0.93	0.90	0.93
$0110/pA$	0.93	**0.97**	0.58	0.61
$1100/pP$	0.90	0.58	**0.90**	0.58
$0110/pP$	0.93	0.61	0.58	0.26

Table 3. Symmetrized reduced game

in bold). Additionally, each mixed strategy with $p(1100/pA) + p(1100/pP) = 1$ and $p(1100/pP) \geq \theta$ is a symmetric Nash equilibrium.

As was pointed out in Jäger (2007), the equilibrium $0110/pA$ is the only evolutionarily stable strategy in this game (as well as in the larger original game). This does not entail though that a population will evolve towards this strategy from each initial point. The set of strategies with $p(1100/pA) + p(0011/pP) = 1$

[3] This carries to the symmetrized version of the game, i.e. in the symmetrized game only those agents survive that play an undominated strategy in the sender role.

and $p(1100/pP) > \theta$ (note the strict inequality!) consists exclusively of neutrally stable strategies. With an argument along the lines of Pawlowitsch (2007) it can straightforwardly be shown that this continuum of neutrally stable strategies attracts a positive measure of the state space.

As every orbit in the interior of the state space converges to some Nash equilibrium, we can conclude that the set of Nash equilibria of this game (and therefore also of the origninal game comprising all 16 sender strategies) is the smallest universal here. The fact that the notion of a universal is more inclusive than the set of evolutionarily stable strategies is empirically well-justified here. The neutrally stable strategies of this game represent case marking systems without splits. These are empirically well-attested, for instance with a pure accusative system in Hungarian and a pure ergative system in the past tense paradigm of Burushaski.

6 Conclusion

To wrap up, I tried to argue that the notion of a linguistic universal can be illuminated by considering a (cultural) evolutionary dynamics, because this approach promises to give a causal explanation for seemingly functionally motivated features of natural languages. Evolutionary game theory is a suitable mathematical framework to carry out this programme because the dynamics of language use arguably involves frequency dependent selection, and because linguistic interaction can naturally be formalized by game theoretic means. I have to close with a note of caution though: to actually derive universals analytically from a game theoretic model, it is not sufficient to study static equilibrium properties. Rather, a careful exploration of the dynamic properties of the model is inevitable.

References

Akin, E., Hofbauer, J.: Recurrence of the unfit. Mathematical Biosciences **61** (1982) 51–62

Barlow, M., Kemmer, S., eds.: Usage-based models of language. CSLI Publications, Stanford (2000)

Bomze, I.: Regularity versus degeneracy in dynamics, games, and optimization: A unified approach to different aspects. SIAM Review **44**(3) (2002) 394–441

Bomze, I.M., Weibull, J.W.: Does neutral stability imply Lyapunov stability? Games and Economic Behavior **11**(2) (1995) 173–192

Bybee, J.: Phonology and language use. Cambridge University Press, Cambridge, UK (2001)

Chomsky, N.: Syntactic Structures. Mouton, The Hague (1957)

Chomsky, N.: The Minimalist Program. MIT Press, Cambridge, MA (1995)

Haspelmath, M.: Optimality and diachronic adaptation. Zeitschrift für Sprachwissenschaft **18**(2) (1999) 180–205

Hofbauer, J., Schuster, P., Sigmund, K.: A note on evolutionarily stable strategies and game dynamics. Journal of Theoretical Biology **81**(3) (1979) 609–612

Hofbauer, J., Sigmund, K.: The Theory of Evolution and Dynamical Systems. Cambridge University Press (1988)
Hofbauer, J., Sigmund, K.: Evolutionary Games and Population Dynamics. Cambridge University Press, Cambridge, UK (1998)
Hurford, J.R.: Biological evolution of the Saussurean sign as a component of the language acquisition device. Lingua **77** (1989) 187–222
Huttegger, S.H.: Evolution and the explanation of meaning. Philosophy of Science **74** (2007) 1–27
Jäger, G.: Evolutionary Game Theory and typology: a case study. Language **83**(1) (2007) 74–109
Kandori, M., Mailath, G., Rob, R.: Learning, mutation, and long-run equilibria in games. Econometrica **61** (1993) 29–56
Lewis, D.: Convention. Harvard University Press, Cambridge, Mass. (1969)
Nowak, M.A., Krakauer, D.C.: The evolution of language. Proceedings of the National Academy of Sciences **96**(14) (1999) 8028–8033
Pawlowitsch, C.: Why evolution does not always lead to an optimal signaling system. Games and Economic Behavior **63** (2008) 203–226
van Rooij, R.: Signalling games select Horn strategies. Linguistics and Philosophy **27** (2004) 493–527
Taylor, P., Jonker, L.: Evolutionarily stable strategies and game dynamics. Mathematical Biosciences **40** (1978) 145–156
Thomas, B.: On evolutionarily stable sets. Journal of Mathematical Biology **22** (1985) 105–115
Trapa, P., Nowak, M.: Nash equilibria for an evolutionary language game. Journal of Mathematical Biology **41** (2000) 172–188
Wärneryd, K.: Cheap talk, coordination and evolutionary stability. Games and Economic Behavior **5** (1993) 532–546
Young, H.P.: The evolution of conventions. Econometrica **61** (1993) 57–84
Zipf, G.: The Psycho-Biology of Language. MIT Press, Cambridge, Massachusetts (1935)
Zipf, G.K.: Human Behavior and the Principle of Least Effort. Addison Wesley, Cambridge (1949)

Appendix

Numerical estimation of the size of the basin of attraction As mentioned in the text, the joint basin of attraction of the set of neutrally but not evolutionarily stable strategies is about 1.2% of the entire strategy space. The result was obtained by using a Monte-Carlo method. A random initial state s was picked out by a random generator according to the uniform distribution over the 729-dimensional simplex, the replicator dynamics was solved numerically for the initial condition s, and the asymptotic behavior was analyzed at $t = 1000$ (by which time all time series have converged towards a rest point, within the limits of precision imposed by the numerical algorithm used). This procedure was repeated 5,000 times. It turned out that in 4,941 cases the solution converged towards an evolutionarily stable state, and in 59 cases to a neutrally but not evolutionarily stable state. This means that with a confidence of 95%, the true value lies between 0.9% and 1.5%, with a maximum likelihood estimation of 1.2%.

Basins of attraction The utility matrix of the game in question is:

$$A = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 3 & -1 \\ 2 & -1 & 0 \end{pmatrix}$$

This game is doubly symmetric, and it has two Nash equilibria: e_1 and e_2. Consider the linear manifold $E = \{x \in \text{int}S_3 : x_3 = 4x_2\}$. For points $x \in E$, we have

$$\begin{aligned} \dot{x}_2 &= x_2(2x_1 + 3x_2 - x_3 - x \cdot Ax) \\ &= x_2(2x_1 - x_2 - x \cdot Ax) \\ \dot{x}_3 &= x_3(2x_1 - x_2 - x \cdot Ax) \\ &= 4x_2(2x_1 - x_2 - x \cdot Ax), \end{aligned}$$

and thus $\dot{x}_3 = 4\dot{x}_2$. So $\dot{x}$ is always tangential to E, and E is thus an invariant set. As e_1 is the only Nash equilibrium within the closure of this set, all points in E converge towards e_1. Now consider the set $F = \{x \in \text{int}S_3 : x_3 > 4x_2\}$. Since e_2 is not in the closure of this set, a trajectory starting in F and converging towards e_2 would have to leave F. This is impossible though because all interior points remain within the interior under the replicator dynamics—so no orbit can touch a boundary face—and a trajectory cannot cross E due to uniqueness of solutions of autonomous differential equations. We thus conclude that all points in F converge towards e_1. Since F has a positive Lebesgue measure, e_1 has a basin of attraction that is not a null set. Nevertheless e_1 is not neutrally stable, because

$$\begin{aligned} e_2 \cdot Ae_1 &= e_1 \cdot Ae_1 \\ e_2 \cdot Ae_2 &> e_1 \cdot Ae_2 \end{aligned}$$

In figure 1 the phase portrait of the game is sketched. The bold line indicates

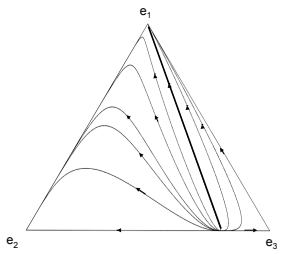

Fig. 1. Phase portrait

the boundary between the basins of attraction of e_1 and e_2. It is easy to see

that e_1 is not Lyapunov stable (because every open environment has a non-empty intersection with the basin of attraction of e_2), but nevertheless attracts a non-null set.