CONVEX MEANINGS AND EVOLUTIONARY STABILITY

GERHARD JAGER

Department of Language and Literature, University of Bielefeld,
PF 10 01 31, D-33501 Bielefeld, Germany
Gerhard.Jaeger@uni-bielefeld.de

Gärdenfors (2000) argues that natural denotations of natural language predicates are convex regions in a conceptual space. Using techniques from evolutionary game theory, the paper shows that this convexity criterion is a consequence of the evolutionary dynamics of language use.

Evolutionary game theory (EGT) is a mathematical framework to model the consequences of interaction between individuals for the evolutionary dynamics of a population. Suppose a certain type of interaction between members of a population has an impact on their fitness. Furthermore, the outcome of the interaction depends on heritable traits ("strategies") of the individuals involved. Under these conditions, the interaction can be modeled as a strategic game, where utility can be identified with fitness.

One of the appealing features of this model is the fact that the attractor points of the ensuing dynamics (the "evolutionarily stable strategies") can be characterized purely in terms of the utility matrix. This is particularly straightforward in the case of asymmetric games. There, exactly the **strict Nash equilibria** are evolutionarily stable (as proved by Reinhard Selten in 1980). A pair of strategies is a strict Nash equilibrium if each strategy is the unique best response to the other strategies.

This model is applicable both to biological and to cultural evolution. Applied to the cultural evolution of language, the individuals involved are speakers/hearers-in-an-utterance, strategies are linguistic constructions, and fitness is the likelihood of a construction to be imitated. We expect to find natural languages in evolutionarily stable states most of the time. The notion of evolutionary stability is thus a way to reduce linguistic universals to the evolutionary dynamics of language use.

Gärdenfors (2000) argues the semantic domains that natural language deals with have a geometrical structure. He gives evidence that simple natural language adjective usually denote natural properties, where a natural property is a **convex** region of such a "conceptual space". In this paper I will argue that under very

natural assumptions about the utility function, convexity of meanings falls out as a consequence of evolutionary stability.

Imagine a simple communication game. The game leader, "Nature", shows one player (the sender S) a point in some continuous Euclidean space. S can send one out of a finite set of signals to the receiver R. R in turn has to guess the point that Nature showed to S.

The problem has the shape of a signaling game. Nature chooses some point in the meaning space, according to some fixed probability function p_i. S and R have the joint goal to maximize the similarity between Nature's choice and R's choice.

Formally, we say that M is the set of meanings (points in the meaning space), S is a function from M into some finite set F of forms, and R is a function from F to M. The utility of the communicators can be defined as

$$u(S, H) = \sum_{m \in M} p_i(m) \cdot \mathit{sim}(m, R(S(m))) \tag{1}$$

where *sim* is a function that measures the similarity between two points. Similarity between two points is a monotonically decreasing function of their distance in the conceptual space. By convention, similarity is always positive, and every point has a similarity of 1 to itself. The interests of S and R are completely identical. Also, the signals themselves come with no costs.

Suppose S knows R, the interpretation function of the receiver. What would be the best coding strategy then? For each possible signal f, S knows R's interpretation, namely $R(f)$. So for a given choice m of Nature, S should choose the signal f that maximizes $\mathit{sim}(m, R(f))$. In other words, each form f corresponds to a uniqe point $R(f)$, and each input meaning m is assigned to the point which is most similar to m, or, in other words, which minimizes the distance to m. This kind of partitioning of the input space is called a **Voronoi Tesselation**. It is easy to show that the Voronoi tessellation based on a Euclidean metric always results in a partioning of the space into convex regions.

Since the meaning space is continuous, there may be pairs of functions from M to F that, though different, differ with a probability 0. (For instance if they only differ with respect to a single point.) If we identify sender strategies with equivalence classes of such functions rather than with such functions itself, there is actually a unique best response to each receiver strategy—the equivalence class of the Voronoi tessellation that is induced by that receiver strategy.

The best response of R to a given coding function S is given by.

$$R(f) = \arg\max_m \int_{S^{-1}(f)} p_i(m')\mathit{sim}(m, m')dm' \tag{2}$$

The precise nature of such an optimal sender's response depends on the details of Nature's probability function p_i, the similarity function, and the geometry of the

underlying space. Suffice it to say that for a closed space and continuous p_i and *sim*, the existence of such a best response is always guaranteed.

From these considerations it follows that in each game of the described class (with a finite Euclidean meaning space and continuous similarity functions and probability distributions), there are evolutionarily stable states, and **in each evolutionarily stable state, the meaning of each form is a continuous region of the meaning space.**

For the purpose of illustration, I did a few computer simulations of the dynamics described above (using a discrete approximation of a continuous space). The meaning space was a set of squares inside a circle. The similarity between two squares is inversely related to its Euclidean distance. All meanings were assumed to be equally likely. The experiments confirmed the evolutionary stability of Voronoi tesselations. The graphics in Figure 1 show stable states for different numbers of forms. The shadings of a square indicates the form that it is mapped to by the dominant sender strategy. Black squares indicate the interpretation of a form under the dominant receiver strategy.

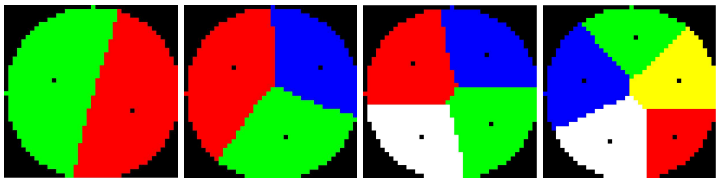

Figure 1. Evolutionarily stable states of the signaling game with a uniform probability distribution over meanings

In the previous simulations, I assumed a uniform probability distribution over the meaning space. This leads to an infinity of evolutionarily stable states. In the remainder of the paper I will illustrate with an example that a skewed probability distribution may reduce the number of equilibria quite drastically.

Let us assume that the meaning space contains finitely many, in fact very few, small regions that are highly frequent, while all other meanings are so rare that their impact on the average utility is negligible. For the sake of concreteness, let us suppose that the meaning space forms a circle, and that there are just four meanings that are frequent. Let us call them Red, Green, Blue and Yellow. Inspired by the middle plane of the color space (supressing the brightness dimension), I assume that all four prominent meanings are close to the periphery, and they are arranged clockwise in the order Red, Yellow, Green, and Blue. Their positions are not entirely symmetric. Rather, they are arranged as indicated in Figure 2.

Since the similarity between two points is inversely related to their distance, it follows that Blue and Green are more similar to each other than Red and Yellow, which are in turn more similar than the pair Green/Yellow and the pair Red/Blue.

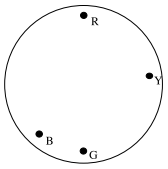

Figure 2. Schematic arrangement of the four prominent meanings in the example

The pairs Blue/Yellow and Red/Green are most dissimilar.

We finally assume that the probabilities of all four prominent meanings are close to 25%, but not completely equal. For the sake of concreteness, let us assume that

$$p_i(\text{Red}) > p_i(\text{Green}) > p_i(\text{Blue}) > p_i(\text{Yellow}) \tag{3}$$

Now suppose the sender has just two forms at her disposal. What are the strict Nash equilibria of this game?

A Voronoi tesselation induces a partition of the set {Red, Yellow, Green, Blue}. The partition {Red, Green}, {Blue,Yellow} is excluded because it is not convex. This leaves seven possible partitions:

1. {**Red, Yellow, Green, Blue**}/{} This is a weak Nash equilibrium; i.e., one of the strategies involved is not the *unique* best response to the other player's strategy. It is thus not evolutionarily stable.

2. {**Red**}/{**Yellow, Green, Blue**} If the sender strategy partitions the meaning space in this way, the best response of the receiver is to map the first signal to Red and the second one to the point that maximizes the average similarity to the elements of the second partition. If the probabilities of Yellow, Green and Blue are almost equal and all other points have a probability close to 0, the average similarity to Yellow, Green and Blue is a function with three local maxima, that are located close to Yellow, Green and Blue respectively. So if the sender uses this partition, the best response of the receiver is to map the first form to Red and the second to the point with the highest average similarity to Yellow, Green and Blue. This is one of the mentioned three local maxima. Since, by assumption, Green is more probable than Yellow and Blue, the maximum close to Green is the global maximum. But this entails that the sender strategy is not the best response to the receiver, and thus this partition does not lead to evolutionary stability.

3. {**Yellow**}/{**Red, Green, Blue**} For similar reasons as in the previous case, this partition thus does not correspond to a Nash equilibrium either.

4. {**Green**}/{**Red, Yellow, Blue**} Since Blue is closer to Green than to Red, this partition does not correspond to an equilibrium for analogous reasons.

5. {**Blue**}/{**Red, Yellow, Green**} This case is analogous because Green is closer to Blue than to Red.

6. {**Red, Yellow**}/{**Green, Blue**} The best response of the receiver here is to map the first form to Red and the second to Green. The best response of the sender to this strategy in turn is to map Red and Yellow to the first form, and Green and Blue to the second. So this partition creates a strict Nash equilibrium.

7. {**Red, Blue**}/{**Yellow, Green**} The best response of the receiver here is to map the first form to Red and the second to Green. The best response of the sender in turn would be to to map Red and Yellow to the first form, and Green and Blue to the second. Hence this partition does not induce a Nash equilibrium.

So it turns out that with two forms, only the bipartition {Red, Yellow}/{Green, Blue} is evolutionarily stable.

Let us turn to the analogous game with three forms. Each sender strategy in this game creates a tripartition of the meaning space. We only have to consider convex tripartitions. All convex bipartitions are trivially also tripartitions, with an empty third cell. It is also immediately obvious that such a partially trivial partition cannot give rise to a strict Nash equilibrium. Besides, there are four more, non-trivial convex tripartitions:

1. {**Red**}/{**Yellow**}/{**Green, Blue**} The best response of the receiver is to map the first signal to Red, the second to Yellow, and the third to Green. The best response of the sender to this strategy is to use the above-mentioned partition, so this leads to a strict Nash equilibrium.

2. {**Yellow**}/{**Green**}/{**Blue, Red**} This does not correspond to a Nash equilibrium because the best response of the receiver is to map the third form to red, and since Blue is closer to Green than to Red, the best response of the sender would be to switch to the previous partition.

3. {**Green**}/{**Blue**}/{**Red, Yellow**} The best response of the receiver is to map the three forms to Green, Blue, and Red respectively, and the best response of the sender in turn is to use the Voronoi tessellation that is induced by these three points. This is exactly the partition in question, so it does lead to a strict Nash equilibrium.

4. {**Red**}/{**Blue**}/{**Yellow, Green**} Since Yellow is closer to Red than to Green, this does not lead to a Nash equilibrium either.

So in this game, we have two partitions that are evolutionarily stable, namely {Red}/ {Yellow}/{Green, Blue} and {Green}/{Blue}/{Red, Yellow}. There is an asymmetry between the two equilibria though. Recall that the evolutionary model assumes that the strategy choice of the players is not fully deterministic but subject to some perturbation. Suppose the system is in one of the two evolutionarily stable

states. Unlikely though it may be, it is possible that very many mutations occur at once, and all those mutations favor the other equilibrium. This may have the effect of pushing the entire system into the other equilibrium. Such an event is very unlikely in either direction. However, it may be that such a switch from the first to the second equilibrium may by more likely than in the reverse direction. This would have the long term effect that in the long run, the system spends more time in the second than in the first equilibrium. Such an asymmetry grows larger as the mutation rate gets smaller. In the limit, the long term probability of the first equilibrium converges to 0 then, and the probability of the second equilibrium to one. Equilibria which have a non-zero probability for any mutation rate in this sense are called *stochastically stable*.

Computer simulations indicate that for the game in question, the only stochastically stable states are those that are based on the partition {Red}/{Yellow}/{Green, Blue}. In the simulation, the system underwent 20,000 update cycles, starting from a random state. Of these 20,000 "generations", the system spent 18,847 in a {Red}/{Yellow}/{Green, Blue} state, against 1,054 in a {Green}/{Blue}/{Red, Yellow} state. A switch from the first into the second kind of equilibrium did not occur.

Figure 3 visualizes the stable states for the game with two, three and four different forms. As in Figure 1, the shade of a point indicates the form to which the sender maps this point, while the black squares indicate the preferred interpretation of the forms according to the dominant receiver strategy. The circles indicate the location of the four focal meanings Red, Yellow, Green and Blue.

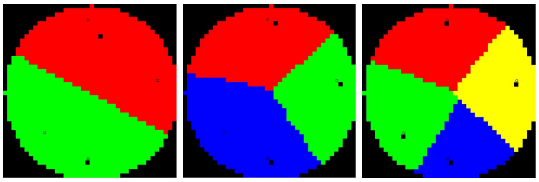

Figure 3. Evolutionarily stable states of the signaling game with focal points

The last example sketched a hypothetical scenario which induces the kind of implicative universals that are observed in natural languages with regard to color universals. My choice of parameters was largely stipulative. However, it is hoped that psycholinguistic research can supply empirically justified values for them. Evolutionary consideration could then establish a link between psycholinguistics and typology.

References

Gärdenfors, P. (2000). *Conceptual spaces.* Cambridge, Mass.: The MIT Press.