EVOLUTIONARY GAMES AND SEMANTIC UNIVERSALS

ROBERT VAN ROOIJ

ILLC, University of Amsterdam, Nieuwe Doelenstraat 15,
Amsterdam, 1012 CP, the Netherlands
R.a.m.vanRooij@uva.nl

An evolutionary perspective on signaling games is adopted to explain some semantic universals concerning truth-conditional connectives; property denoting expressions, and generalized quantifiers. The question to be addressed is: of the many meanings of a particular type that *can* be expressed, why are only some of them expressed in natural languages by 'simple' expressions?

Most work on the evolution of language concentrates on the evolution of syntactic and phonetic rules and/or principles. This is reasonable, because in the generative tradition these disciplines acquired a central place in linguistics. In another sense, however, the under-representation in evolutionary linguistics of work that concentrates on semantics is surprising: how many of us would be interesting in language if it was not the main vehicle used to transmit meanings? Moreover, semantics and pragmatics are by now well-established disciplines within linguistics that study how, across languages, meanings are transmitted by language. In this paper I will concentrate on giving evolutionary motivations for some semantic features shared by all or most languages of the world.

There are in fact many semantic features shared by all languages of the world. For instance, it seems that of all the speech acts that we can express in natural language, only three of them are normally grammaticalized, and distinguished, in mood (i.e., declarative, imperative, and interrogative). In this paper, we will be most interested in similar kinds of universals that make claims about what kinds of meanings are expressed by short and simple terms (e.g. with one word) in natural languages. One of them concerns *indexicals*, short expressions corresponding to the English *I*, *you*, *this*, *that*, *here*, etc., the denotations of which are essentially context-dependent. It seems that all languages have short words that express such meanings (cf. Goddard, 2001), and this fact makes evolutionary sense: it is a useful feature of a language if it can refer to nearby individuals, objects, and places, and we can do so by using short expressions because their denotations can normally be inferred from the shared context between speaker and hearer. In this paper I will be concerned with similar universals involving mainly the connectives, property denoting expressions, and generalized quantifiers.

Signaling games and Connectives In *signaling games* as introduced by David Lewis (1969), signals have an underspecified meaning, and the actual interpretation the signals receive depends on the equilibria of sender and receiver strategy combinations of such games. Recently, these games have been looked upon from an *evolutionary* point of view to study the evolution of language. According to it, a signaling convention can arise in which signal s denotes t if and only if in the *evolutionary stable strategy* (ESS) signal s is only used when the speaker is in situation t. Thinking of meanings as situations, one can show that if there exists a 1-1 mapping between situations and the best actions to be performed there, and there are enough messages, the ESSs, or resulting communication systems, of signaling games always give rise to 1-1 mappings between signals and meanings. It is obvious that in this simple communication system there can be no role for connectives: the existence of a disjunctive or conjunctive message would destroy the 1-1 correspondence between (types of) situations and signals. That gives rise to the question, however, under which circumstances messages with such more complex meanings could arise. In this paper I concentrate only on one particular truth-conditional connective: *disjunction*.

Taking t_i and t_j to be (types of) situations, under which circumstances can a language evolve in which we have a message that means 't_i', one that means 't_j', and yet another with the disjunctive meaning 't_i or t_j'? As indicated above, if there exists a 1-1 function from situations to (optimal) actions to be performed in those situations, a language can evolve with a 1-1 correspondence between signals and situations. The existence of this 1-1 function won't be enough, however, to 'explain' the emergence of messages with a disjunctive meaning. What is required, instead, is a 1-1 function from *sets* of situations to (optimal) actions. We can understand such a function in terms of a payoff table like the following:

	a_1	a_2	a_3	a_4	a_5	a_6	a_7
t_1	4	0	0	3	3	0	2.3
t_2	0	4	0	3	o	3	2.3
t_3	0	0	4	0	3	3	2.3

Notice that according to this payoff table, for each $i \in \{1, 2, 3\}$ action a_i is the unique optimal action to be performed in situation t_i. This table, however, contains more information. Suppose that the speaker (and/or hearer) knows that the actual situation is either t_1 or t_2, and that both situations are equally likely. In that case the best action to perform is neither a_1 nor a_2 – they only have an expected utility of 2 –, but rather a_4, because this action now has the highest expected utility, i.e., 3. Something similar holds for information 't_1 or t_3' and action a_5, and for 't_2 or t_3' and action a_6. Finally, in case of no information, which corresponds with information 't_1 or t_2 or t_3', the unique optimal action to perform is a_7. Thus for all (non-empty) subsets of $\{t_1, t_2, t_3\}$ there exists now a unique best action to be performed. Notice that each such subset may be thought of as an *information*

state, the (complete or incomplete) information an agent might have about the actual situation. Suppose now that we lift the sender-strategy from a function that assigns to each *situation* a unique message to be sent, to one that assigns to each *information state* a unique message to be sent. Now it can be shown that we will end up (after evolution) with a communication system (an ESS) in which there exists a 1-1-1 correspondence between information states (or sets of situations), messages, and actions to be performed.[a] Thus, there will now be messages which have a disjunctive meaning. This by itself doesn't mean yet that we have a separate message that denotes disjunction, but only that we have separate messages with disjunctive meanings in addition to messages with simple meanings. However, as convincingly shown by Kirby and others, a learning bottleneck is a strong force for languages to become compositional. It is reasonable to assume that under such a pressure a complex message will evolve which means $\{t_i, t_j\}$ that consists of three separate signals: one signal denoting $\{t_i\}$, one signal denoting $\{t_j\}$, and one signal that turns these two meanings into the new meaning $\{t_i, t_j\}$ by (set theoretic) *union*. The latter signal might then be called 'disjunction'.

In principle, once we take information states into account, we cannot only state under which circumstances disjunctive messages will evolve, but also when negative and conjunctive messages will evolve.[b] The main difference is that we have to assume more structure of the set of information states. An interesting feature of our evolutionary description of the connectives is that it might answer the question why only humans have communication systems involving (truth-conditional) connectives. In contrast to the signaling games discussed by Lewis, and used to explain the alarm calls of, e,g. vervet monkeys, it was crucial for connectives to evolve to take *information states*, or *belief states* into account, i.e., sender strategies must take *sets* of situations as arguments, and not just situations themselves, and this must be recognized by receivers as well. Perhaps, the existence of such more complicated sender strategies is what that sets us apart from those monkeys.

Why not more connectives? Once we assume that each (declarative) sentence is either true or false, there are *four* potential unary connectives, and as much as *sixteen* potential binary connectives. Although all these potential connectives *can* be expressed in natural language, the question is why only one unary (*negation* and only two (or perhaps three) binary truth-functional connectives (*disjunction* and *conjunction*) are expressed by means of simple words in all (or most) natural languages? That is, can we give natural reasons for why languages don't have the truth-functional connectives that are mathematically possible? For unary connectives this problem is easy to solve. Look at the four possible unary truth-

[a] This is a general result, and not restricted to the particular example discussed above.

[b] More interesting things can be said about why, and of the conditions under which, messages with negative and conjunctive meanings could evolve, but space doesn't allow me to go into this here.

conditional connectives, $c_1, ..., c_4$:

p	$c_1\ p$	$c_2\ p$	$c_3\ p$	$c_4\ p$
1	0	1	0	1
0	1	0	0	1

Connective c_1 is, of course, standard negation. Why we don't see the others in natural language(s) is obvious: they just don't make sense! $c_2\ p$ just has the same truth-value as p itself, and, thus, c_2 is *superfluous*, while the truth values of $c_3\ p$ and $c_4\ p$ are *independent* of the truth value of its argument p, which leaves it unclear why c_3 and c_4 require arguments at all.

For binary connectives the problem is more difficult, but Gazdar & Pullum (1976) show that when we require that all lexicalized binary connectives must be *commutative* and obey the principles of *strict compositionality* and *confessionality*, all potential binary connectives are ruled out except for the following three: conjunction, standard (inclusive) disjunction, and what is known as *exclusive* disjunction. This is an appealing result, because (i) strict compositionality makes perfect sense, (ii) the principle of confessionality – which forbids (binary) connectives which yields the value true when all its arguments are false – can be explained by the psychologically well-established fact that negation is difficult to process, while (iii) the constraint of commutativity is motivated by the not unnatrual idea that the underlying structures of the connected sentences are linearly unordered. The non-existence of a lexicalized exclusive disjunction can be explained, finally, by the standard conversational implicature from *A or B* to *not A and B*, which makes such a connective superfluous.

Properties In extensional terms, any subset of a set of individuals, or objects, can be thought of as a property. Thinking of properties in this way, however, leaves us with many more properties that *can* be expressed, than that there are simple expressions that denote properties in any natural language. This gives rise to the following questions: (i) can we characterize the properties that are denoted by simple expressions in natural language(s), and, if so, (ii) can we give a pragmatic and/or evolutionary explanation of this characterization?

The first idea that comes to mind to limit the use of all possible properties, is that only those properties will be expressed a lot in natural language that are *useful* for sender and receiver. Using our signaling game framework, it is easy enough to show how usefulness can influence the existence of property denoting terms when we either have less messages, or less actions than we have situations.[c] To

[c]These abstract formulations might be used to model other 'real-world' phenomena as well, such as noise in the communication channel which doesn't allow receivers to discriminate enough signals; a limitation of the objects speakers are acquainted with, perhaps due to ever changing contexts; and maybe also non-aligned preferences between sender and receiver.

illustrate the first case, consider a game involving three situations, three actions, but only two messages. Taking the sender and receiver strategies to be functions from situations to messages and messages to situations, respectively, we predict that in equilibrium only two actions will be performed. Which of those actions that will be depends on the utilities and probabilities involved. Consider the following utility tables:

	a_1	a_2	a_3
t_1	8	0	0
t_2	0	4	1
t_3	0	0	2

	a_1	a_2	a_3
t_1	1	0	0
t_2	0	1	0
t_3	0	0	1

In both cases there exists a 1-1 correspondence between situations and messages. If there are three messages, in each situation the sender will send a different message, and the receiver will react appropriately. When there are only two messages, however, expected utility will play a role. In the left-hand table above it is more useful to distinguish t_1 from t_2 and t_3, then to distinguish t_2 from t_3. As a consequence, in equilibrium t_2 and t_3 will not be distinguished from each other and in both situations the same message will be sent. We have implicitly assumed here that the probability of the three situations was equal. Consider now the table on the right-hand side, and suppose that t_1 is much more likely to occur than t_2, which, in turn, is much more likely than t_3. Again, it will be more useful to distinguish t_1 from t_2 and t_3, then to distinguish t_2 from t_3. Thus, also here we find that in equilibrium t_3 will not be distinguished separately, and meshed together with t_2.

A common complaint of Chomskyan linguists (e.g. Bickerton, Jackendoff) against explanations like the one above is that usefulness can't be the only constraint: there are many useful properties, or distinctions 'out there' that are still not really named, or distinguished, in simple natural language terms. Bickerton (1990) mentions *contiguity* (or *convexity*) as an extra constraint, and hypothesizes that the preference for convex properties is an innate property of our brains. Unfortunately, if we think of properties as in standard semantics just as subsets of the universe of discourse, such a constraint cannot even be formulated. For reasons like this, Gärdenfors – following philosophers like van Fraassen and Stalnaker – proposed to use a meaning space to represent meanings in which the notion of convexity makes sense. This meaning space is essentially an n-ary vector space where any subset of this space is (or represents) a property. However, because each point in space can now be characterized in terms of the values of its coordinates, Gärdenfors can make a distinction between 'natural' and 'unnatural' properties: only those subsets can be thought of as natural properties that form *convex regions* of the space.[d] Because only a small minority of all subsets of any struc-

[d] For a set of objects to be a convex region, it has to be closed in the following sense: if x and y are

tured meaning space form convex regions, the hypothesis that (most or all) simple natural language property denoting expressions denote such convex regions is, potentially, a very strong one. Gärdenfors' proposal is quite successful for some categories of property denoting expressions, like colors, and this gives rise to the question what makes convex regions so natural. This question is addressed in Jäger & van Rooij (to appear). It is shown that only those communication systems will be evolutionary stable in a signaling game where the sender strategy is just a function from points in the meaning space to messages, and where the receiver has to guess this point, in case the set of points in the space in which the same signal is sent forms a convex region in this shared meaning space with a prototype.

Gärdenfors (2000) mentions a number of examples (of property-, but also of relation denoting expressions and prepositions) where convexity seems like a natural constraint, and might give rise to semantic universals. We won't go into these examples here, but instead (i) will discuss some examples not discussed by Gärdenfors where convexity can explain some well established semantic universals, and (ii) will speculate a bit on the difference between communication systems of (some) animals, young children and adults humans making use of the above mentioned evolutionary motivation for convexity. I start with the latter.

Basic level properties It is a basic observation that many property denoting expressions used by adults, (e.g. *tool*, *furniture*) denote objects that are not similar to each other, neither with respect to appearance, nor w.r.t. (basic) function. The psychologist Rosch (1978) made a distinction between *basic level* categories/properties (*chair*, *dog*), and sub- and superordinate ones (*armchair*, *furniture*), and proposed that only for the first ones the notion of similarity plays an important role. She also observed that it are the first ones that are learned earlier and easier by children, and – we might speculate –animals never come any further than making basic level category-like distinctions. Now, notice that in terms of meaning spaces, convex sets are defined in terms of a distance measure, where the 'closeness' of two objects to each other depends on their (mutual) resemblance. This gives rise to the hypothesis that in contrast to animals and young children, only 'adult' humans can make use of expressions in their communication systems that denote non-convex properties. Interestingly enough – and in parallel with our above 'explanation' of why only humans make use of connectives –, this contrast might be understood from the complexity of the sender strategies used in signaling games that generate (non-convex) properties. Remember that to explain the emergence of property denoting expressions we assumed that sender strategies were just very simple functions from situations to messages. When we assume that objects exist in structured meaning spaces, all properties that will be expressed in equilibrium form convex regions with obvious prototypes. But this means that

elements of the set, all objects 'between' x and y must also be members of this set.

to explain the existence of those properties that do not denote convex sets (i.e., by hypothesis the sub- and superordinate ones) and/or do not have prototypes, we need either more involved sender strategies (cf. the case of connectives), or utility functions not defined in terms of a very simple measure of similarity. Again, this might explain why only adult humans can make use of non-basic level property denoting expressions. What our analysis also explains is why *conjunction* seems easier to understand and process than *disjunction* and *negation*. Notice that these connectives make sense for properties as well. Now one can show that in contrast to the other connectives the conjunction of two convex properties is guaranteed to be convex as well (this is not true for the connectives of 'quantum logic', though).

Quantifiers and determiners Most work on universals in model-theoretic semantics is concentrated on quantifiers and determiners. This is also very natural, given that the discrepancy between the number of meanings that are predicted to be expressible, and the terms to do so is here much larger than for properties and relations. To get a glimpse of this, in a simple extensional model with only 4 individuals, standard model theoretic semantics predict that there are not less than 2^{2^4} = 65.636 quantifiers that can be expressed, and even the immense number of 2^{4^4} many determiners! Obviously, constraints are in order to limit the meanings that can be expressed by (simple) noun phrases and determiners.

Because a determiner denotes a relation between properties, or, equivalently, a function from properties to quantifiers, any constraint on quantifiers gives rise to a constraint on determiners as well. So we can safely limit ourselves to constraints on determiners. A simple, and very intuitive constraint is *variety*. A determiner shows variety iff it gives rise to a *contingent* meaning: the sentence of type 'Det Noun VP' in which it occurs is neither always true nor always false. More formally, determiner D is said to show variety iff in every model in which the determiner is defined there are A, B such that $D(A, B)$ is true, and A', B' such that $D(A', B')$ is false. It is clear that we can form complex determiners which do not show variety (like *some or no*), but it is generally assumed that all 'simple' determiners satisfy this constraint. An explanation of this fact is easy to imagine: why would a language end up with a simple determiner the use of which doesn't express an informative, and thus useful, proposition?

In this paper we will only explain one semantic universal, stated in essence already in Barwise & Cooper (1981), which says that all 'simple' determiners satisfy the following *continuity* constraint:

For all A, B, B', B'': if $D(A, B')$, $D(A, B'')$ and $B' \subseteq B \subseteq B''$, then $D(A, B)$.

I claim that the notion of convexity can be used to motivate this universal, at least if we assume that the meaning of natural language determiners are *context-independent* and *conservative*. Assume that E and E' are domains of discourse,

and π a permutation function on E'. The *context-independence* constraint then says that if $A, B \subseteq E \subseteq E'$, then $D(\pi(A), \pi(B))$ is true with respect to E iff $D(A, B)$ is true with respect to E'. Intuitively, this means that the meaning of a sentence of the form $D(A, B)$, where, as before, D is the determiner meaning, A is the noun-denotation, while B is the denotation of the VP, doesn't depend on the domain of discourse, and only on the number of individuals in A, B, and $A \cap B$. The further constraint of *conservativity* then says that the meaning of such a sentence depends only on the number of individuals in $A \cap B$ and $A - B$. Intuitively, a determiner is said to satisfy conservativity iff the truth or falsity of a simple sentence of the form NP VP depends only on the denotation of the noun of the NP. An important observation due to van Benthem (1986) is that all quantifiers that satisfy *context-independence* and *conservativity* can be represented geometrically in the so-called 'tree of numbers'. This tree can be thought of as a binary meaning space with as coordinates the numbers of individuals in $A \cap B$ and $A - B$. Each quantifier satisfying the above two constraints can now be represented as a subset of this meaning space, and only some of these subsets form convex regions. One can now show that the continuous quantifiers all give rise to such convex sets. Thus, if the tree of numbers is a natural representation format of generalized quantifiers, our signaling game analysis can help to motivate one very important semantic universal.

The tree of numbers itself can be argued to be a natural geometrical representation format of (most) generalized quantifiers, by motivating the constraints of *context-independence* and *conservativity*. *Conservativity* will be explained, for instance, by the evolutionary preference of languages to follow a topic-comment structure (as for instance already motivated by linguists with as diverse backgrounds as Givon and Bickerton).

References

Barwise J. & R. Cooper (1981), 'Generalized quantifiers in natural language', *Linguistics and Philosophy*, **4**: 159-219.

Benthem, J. van (1986), *Essays in Logical Semantics*, Kluwer, Boston.

Bickerton, D. (1990), *Language and Species*, Univ. of Chicago Press, Chicago.

Gärdenfors, P. (2000), *Conceptual Spaces*, MIT Press, Cambridge, MA.

Gazdar, G. & G.K. Pullum (1976), 'Truth-functional connectives in natural language', *Chicago Linguistic Society*, pp. 220-234.

Goddard, C. (2001), 'Lexico-semantic universals', *Linguistic Typology*, **5**: 1-65.

Jäger, G. and R. van Rooij (to appear), 'Language structure', *Synthese*.

Lewis, D. (1969), *Convention*, Harvard University Press, Cambridge, MA.

Rosch, E. (1978), 'Principles of categorization', In E. Rosch & B. Lloyd (eds.), *Cognition and categorization*, Hilsdale, NJ: Erlbaum.