Prototypes as extreme exemplars: a game theoretical derivation.

Robert van Rooij, ILLC, Amsterdam

Abstract

1 Introduction

Standard linguistic theories in the tradition of formal semantics predicts that (for a simple fragment) any function from individuals to truth values is a property that can be denoted by a property denoting expression. It is obvious, however, that in any language only a tiny fragment of all these functions are, in fact, denoted by simple words or constructions. 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?

To answer question (i), Bickerton (1981) hypothesizes that 'simple' expressions can only denote *connected*, or *convex*, regions of cognitive space, and hypothesizes that the preference for convex properties is an innate property of our brains. Unfortunately, when we think of properties as in standard denotational semantics, it is impossible to distinguish properties that have, from properties that don't have such 'structural features'. Partly for reasons like this Gärdenfors (2000) proposed an alternative, and *richer*, framework to represent meanings: meaning spaces should have some additional structure, i.e., an *a priori* given coordinate structure. 'Natural' properties can now be characterized as those subsets of the meaning space which form *convex regions* of this meaning space. A partition of the meaning space into convex regions with centre points for each of the regions is known as a *Voronoi Tesselation*. Gärdenfors thinks of the region itself as the *descriptive meaning* of the property denoting expression, while the centre point is thought of as its *prototype*.

As for the second question, it is shown in Jäger & van Rooij (2005) that Voronoi Tesselations can be derived as equilibria of signaling games using an Eucledian meaning space with a utility function based on a

notion of similarity. On flat distributions of the points in the meaning spaces, these prototopes will always be in the center of their descriptive meanings. Although this captures an important intuition, it also misses something, and it might even be wrong for simple predicates like color and polar adjectives: intuitively, the prototype (or stereotype, if you want) of being white is being *very* white, but that is not what is coming out. It is also predicted that adjectives like "tall" and prepositional phrases like "above the table" give rise not only to convex meanings, but also to prototypes. Although the former result is appropriate (cf. Zwarts, 1997), the latter result is in conflict with what is standardly assumed in linguistics (e.g. Kamp & Partee, 1995).

In this paper we give an alternative evolutionary game theoretical motivation for the prominence of convex meanings, but one that (i) doesn't make use of utility functions based on a notion of similarity, and (ii) gives rise to prototypes (or stereotypes) that are *extreme* rather than central points in the descriptive meaning of the property denoting expression (adjective).[1]

2 Signaling games with comparison classes

Consider the following signaling game between a sender S and a receiver R. It is known between S and R that there are $n+1$ individuals at the central station. These individuals all have different heights, and there are λ different heights. S wants R to pick out a certain individual, X, but only the former knows the height of this individual, R does not. Receiver R, however, can see the relative heights of all $n+1$ individuals at the central station, and we assume that all λ individuals have a different height. S can influence the choice of R's action by sending her a message in $M = \{m_1, m_2\}$. Although we might think of m_1 as being 'Small', while m_2 as being the message 'Large', these messages don't have any a priori meaning. The utility functions of S and R are identical and give value 1 in case R picks the person S had in mind, and 0 otherwise. What are the equilibria of this type of game?

To get some intuitions, let us fix the numbers n and λ. Let us assume that $\lambda = 4$ and $n = 1$, meaning that individuals can have 4 (relevantly) different heights and that there are 2 persons at the place between which R has to choose. But this means that there are 6 possible situations: $S = \{\langle 1,2\rangle, \langle 1,3\rangle, \langle 1,4\rangle, \langle 2,3\rangle, \langle 2,4\rangle, \langle 3,4\rangle\}$, where $\langle i,j\rangle$ denotes the sit-

[1] The *result* is somewhat similar to the outcome of the game set up by Nowak & Krakauer (1999) to provide a game theoretical motivation for the fact that the vowels of languages are perceptually maximally distinct. The *method* of deriving this 'maximal distinctiveness' result, however, is very different, and it also doesn't require the (extremely) strong *stochastic* evolutionary stability concept implicitly used by the above authors.

uation where the two individuals at the central station are the ones with height i and with height j (with $j > i$). The sender S knows the height of the person he has in mind, and we can associate with each height the set of situations in which the person with this height occurs: $H = \{h_1, h_2, h_3, h_4\}$, where:

$h_1 = \{\langle 1,2\rangle, \langle 1,3\rangle, \langle 1,4\rangle\}$, $h_2 = \{\langle 1,2\rangle, \langle 2,3\rangle, \langle 2,4\rangle\}$, and
$h_3 = \{\langle 1.3\rangle, \langle 2,3\rangle, \langle 3,4\rangle\}$, and $h_4 = \{\langle 1,4\rangle, \langle 2,4\rangle, \langle 3,4\rangle\}$.

His strategy is to send in each of those situations a particular message, thus a function in $H = \{h_1, h_2, h_3, h_4\} \rightarrow M = \{m_1, m_2\}$. The receiver strategy is a function from H to pick either the smallest individual in a situation, or the tallest one, thus a function in $[M \rightarrow \{sm, t\}]$. We say that $\langle \sigma, \rho\rangle$ is a solution (a perfect Bayesian equilibrium)[2] of this game iff

(i) $\forall h \in H : \sigma(t) \in argmax_{m\in M} U^*_{\mathrm{S}}(h, \rho(m))$, and
(ii) $\forall m \in M : \rho(m) \in argmax_{a\in\{sm,t\}} \sum_{h\in H} P(h|\sigma^{-1}(m)) \times U^*_{\mathrm{S}}(h, a)$.

The only thing we have to determine now is the values of $U^*(h_i, sm)$ and $U^*(h_i, t)$. We calculate the value of $U^*(h_i, sm)$ as the expected utility that i is the individual with the smallest height in the situations in h_i. Because each h_i contains three situations, we assume that for each $h_i \in H$ and $s \in h_i$: $P(s|h_i) = \frac{1}{3}$. Thus (where $sm(\langle i, j\rangle) = i$):

$$U^*(h_i, sm) = \sum_{s\in S} P(s|h_i) \times [1, \text{ if } i = sm(s), \ 0 \text{ otherwise}].$$

Similarly, we define $U^*(h_i, t)$ as follows:

$$U^*(h_i, t) = \sum_{s\in S} P(s|h_i) \times [1, \text{ if } i = t(s), \ 0 \text{ otherwise}].$$

Now we can easily see that $\langle \sigma, \rho\rangle$ is a perfect Bayesian equilibrium of this game where S sends m_1 in h_1 and h_2, and m_2 in h_3 and h_4 — $\sigma = \{\langle h_1, m_1\rangle, \langle h_2, m_1\rangle, \langle h_3, m_2\rangle, \langle h_4, m_2\rangle\}$ —, and R responds with sm when he receives m_1, and with t when he receives m_2. To see this, notice that if we fix ρ as above,

$U^*(h_1, \rho(m_1)) = U^*(h_1, sm) = 1 > 0 = U^*(h_1, t) = U^*(h_1, \rho(m_2))$, and
$U^*(h_2, \rho(m_1)) = U^*(h_2, sm) = \frac{2}{3} > \frac{1}{3} = U^*(h_2, t) = U^*(h_2, \rho(m_2))$, and
$U^*(h_3, \rho(m_1)) = U^*(h_3, sm) = \frac{1}{3} < \frac{2}{3} = U^*(h_3, t) = U^*(h_3, \rho(m_2))$, and

[2]This is the standard solution concept for signaling games.

$U^*(h_4, \rho(m_1)) = U^*(h_4, sm) = 0 < 1 = U^*(h_4, t) = U^*(h_4, \rho(m_2))$.

Thus, σ is indeed a best, and indeed *the* best, reply to ρ. But ρ is also the best reply to σ. To see this, we show that for each $f \in \{m_1, m_2\}$, $\rho(m) \in argmax_{a \in \{sm,l\}} \sum_{h \in H} P(h|\sigma^{-1}(m)) \times U^*_{\mathrm{S}}(h, a)$. Abbreviating $\sum_{h \in H} P(h|\sigma^{-1}(m)) \times U^*_{\mathrm{S}}(h, a)$ by $EU_{\mathrm{R}}(a, \sigma, m)$, it suffices to show that $EU_{\mathrm{R}}(sm, \sigma, m_1) > EU_{\mathrm{R}}(l, \sigma, m_1)$ and $EU_{\mathrm{R}}(l, \sigma, m_2) > EU_{\mathrm{R}}(sm, \sigma, m_2)$. We do this by just calculating all these values:

$$
\begin{aligned}
EU(sm, \sigma, m_1) &= [P(h_1|\{h_1, h_2\}) \times U^*(h_1, sm)] + [P(h_2|\{h_1, h_2\}) \times U^*(h_2, sm)] = \\
&\quad [\tfrac{1}{2} \times 1] + [\tfrac{1}{2} \times \tfrac{2}{3}] = \tfrac{5}{6} \\
EU(t, \sigma, m_1) &= [P(h_1|\{h_1, h_2\}) \times U^*(h_1, t)] + [P(h_2|\{h_1, h_2\}) \times U^*(h_2, t)] = \\
&\quad [\tfrac{1}{2} \times 0] + [\tfrac{1}{2} \times \tfrac{1}{3}] = \tfrac{1}{6} \\
EU(sm, \sigma, m_2) &= [P(h_3|\{h_3, h_4\}) \times U^*(h_3, sm)] + [P(h_4|\{h_3, h_4\}) \times U^*(h_4, sm)] = \\
&\quad [\tfrac{1}{2} \times \tfrac{1}{3}] + [\tfrac{1}{2} \times 0] = \tfrac{1}{6} \\
EU(t, \sigma, m_2) &= [P(h_3|\{h_3, h_4\}) \times U^*(h_3, t)] + [P(h_4|\{h_3, h_4\}) \times U^*(h_4, t)] = \\
&\quad [\tfrac{1}{2} \times \tfrac{2}{3}] + [\tfrac{1}{2} \times 1] = \tfrac{5}{6}
\end{aligned}
$$

Thus, the (unique) best response to σ is to pick the smallest individual at the central station if m_1 is sent, and to pick the tallest individual as a response to m_2, which is exactly what ρ does.

Now we might think with Lewis (1969) of $\sigma^{-1}(m)$ as the *descriptive* meaning of m, and of $\rho(m)$ as the *imperative* meaning of m (given equilibrium $\langle \sigma, \rho \rangle$). Notice that $\sigma^{-1}(m_1) = \{h_1, h_2\}$ while $\rho(m_1) = sm$, and $\sigma^{-1}(m_2) = \{h_3, h_4\}$ while $\rho(m_2) = t$.

3 General results

Our example was extremely simple, with only 4 possible heights and 2 individuals at the central station. However, the result holds in general. Also if there are, for instance, 100 possible heights and 40 individuals at the central station, such that R's actions consists of the functions 'pick the $k+1$th tallest individual' ($0 \leq k \leq 39$), there will be only one (relevantly different) non-pooling equilibrium in pure strategies[3] according to which in all types in $\{h_1, \ldots, h_{50}\}$ sender S will send m_1, and R will pick the smallest individual at the central station if he receives m_1, and the same for $\{h_{51}, \ldots, h_{100}\}$: S will send m_2, while R will respond by picking out the tallest individual he sees at the central station. Notice that on a uniform probabability distribution, the imperative meanings of m_1 and m_2 will be always as far apart as possible.

[3]And thus only one relevantly different evolutionary stable strategy.

The above results were all based on the assumption that the probabilities are flatly distributed over the meaning space. The results change dramatically if the meaning space is still finite but the probabilities are not flatly distributed. However, one can prove an important and very surprising fact/theorem if one assumes that the meaning space is *continuous*, like $[0, 1]$: in that case, the descriptive meanings will be convex and equally large, i.e., $[0, 0.5]$ and $(0, 5, 1]$, while the prototype meanings are always as far away as possible. The surprising result (due to Bart Lipman) is that this holds *irrespective of the probability distribution.*

Let there be $n+1$ people at the central station, and let Rs strategy be to choose the $(k+1)^{th}$ tallest person in response to one word and the $(k+d+1)^{th}$ tallest in response to the other, where $d > 0$. It is not difficult to show that given k and d, there is a unique c with the property that S's payoff to sending the first word exceeds his payoff to sending the second iff $h \in [0, c]$. Hence given any pure strategy by R, S's best reply will always be partition $[0, 1]$ into two intervals. Therefore, we know that the language will have S send one message if $h \in [0, c]$ and the other if $h \in (c, 1]$ for some c.

Consider R's best reply if he receives the message corresponding to $(c, 1]$. His payoff to choosing the $(k+1)^{th}$ tallest person is proportional to

$$Pr[h \in (c,1] \text{ and } k \text{ people taller}] = \left(\begin{array}{c} n \\ k \end{array} \right) \int_c^1 [1-F(h)]^k [F(h)]^{n-k} f(h) dh.$$

Let this expression be $\phi(k)$. For $k \geq 1$, we can integrate by parts to obtain

$$\phi(k) = \left(\begin{array}{c} n \\ k \end{array} \right) \tfrac{1}{n-k+1} [1-F(h)]^k [F(h)^{n-k+1}|_c^1 + \left(\begin{array}{c} n \\ k \end{array} \right) \int_c^1 \tfrac{k}{n-k+1} [1-F(h)]^{k-1} [F(h)]^{n-k+1} f(h) dh.$$

But $F(1) = 1$ and $\left(\begin{array}{c} n \\ k \end{array} \right) \frac{k}{n-k+1} = \left(\begin{array}{c} n \\ k-1 \end{array} \right)$, so

$$\phi(k) = \phi(k-1) - \left(\begin{array}{c} n \\ k \end{array} \right) \frac{1}{n-k+1} [1-F(c)]^k [F(c)]^{n-k+1}$$

Because the term being subtracted on the right-hand side must be positive, we see that $\phi(k)$ is decreasing in k. Hence the optimal choice for R is $k = 0$ – that is, to choose the tallest person. An analogous argument shows that when R conditions on $[0, c]$, the optimal choice is the shortest person.

Given these options, it is easy to see that S prefers R to pick the tallest person iff $[F(h)]^n > [1-F(h)]^n$ or $F(h) > \frac{1}{2}$. Thus we must have

c equal to the median height. Because this is the only pure strategy equilibrium, the optimal language is such that S says m_1 when X's height is below the population median and m_2 otherwise. R's strategy is to try the shortest person in the first case and the tallest in the second.

4 Use of the game

The above model can be applied immediately to polar adjectives, with a one-dimentional meaning space. For other property denoting expressions, the game has to be extended a bit. For color adjectives, for instance, we are working with a 3-dimensional meaning space (the dimensions being *hue, saturation*, and *intensity*) and with more than two messages. Such an extension is not a problem: we still get the result that their prototype meanings will be as far away as possible. Combining this with contrast classes seems to give the right results:

> [...] the same color term clearly has a different reference in each domain [...] color terms aren't used so much to refer to particular colors as to maintain the color *contrast* between different referents. Every 'domain' is thus a contrast class, to which we apply color terms of maximal distinctiveness. (Bromström, 1994).

References

[7] Bickerton, D. (1981), *Roots of Language*, Karoma Publishers.

[7] Broström, I (1994), *The Role of Methaphor in Cognitive Semantics*, Lund, Lund University Cognitive Studies, 31.

[7] Gärdenfors, P. (2000), *Conceptual Spaces. The Geometry of Thought*, MIT Press, Cambridge, MA.

[7] Jäger, G. and R. van Rooij (2005), 'Language structure: psychological and social constraints', to appear in *Synthese*.

[7] Kamp, H. and B. Partee (1995), 'Protoytpe theory and compositionality', *Cognition*, **57**: 129-191.

[7] Nowak, M. and D. Krakauer (1999), 'The evolution of language', *Proc. Natl. Acad. Sci. USA*, **96**: 8028-8033.

[7] Zwarts, J. (1997), 'Vectors as relative positions: A compositional semantics of modified PPs, *Journal of Semantics*, **14**: 57-86.