

Metaphor, Cognitive Models, and Language

Comprehensive Module 3
(Special Topics 2A)

Supervisors:
Dr. Jerome Feldman
Dr. George Lakoff
University of California at Berkeley

Steve R. Howell
McMaster University
May 26th, 2000

Metaphor is often considered as something that belongs in poetry, that is more concerned with novel or interesting uses of words than with accepted, everyday practice. Some linguists and other language researchers have a different perspective, however. As has been advanced most extensively by George Lakoff and Mark Johnson (1980, 1999; but see also O’Keefe, 1990), metaphor may, in fact, be far more central to human language, indeed to our very thought. Lakoff and Johnson show how metaphor is pervasive in everyday life, and how it is more than just a matter of language; it may structure our entire conceptual system. As such, Lakoff also provides a new theory of mental concept formation, both linguistic and pre-linguistic: Idealized cognitive models (ICMs). The theory of cognitive models provides a possible mechanism for the operation of metaphor in language, but more than that it can account for our entire understanding of the world, from concrete physical concepts to the most abstract scientific concepts and language (Lakoff, 1987). Furthermore, this theory also provides an important new set of criteria for any language modeling enterprise to address, especially parallel distributed processing neural network (PDP) models.

Metaphor, of course, is the beginning of the theory of ICM’s. Metaphor may indeed be conceptual and hence pre-linguistic, but it is examined most readily when it is expressed in language. For example, take the metaphorical concept “argument is war” (Lakoff & Johnson, 1980). As expressed in our language, this concept results in utterances like:

Your claims are *indefensible*.

His criticisms are *right on target*.

I’ve never *won* an argument with him

He *shot down* all of my arguments.

The military nature of the language used in all of these sentences is quite consistent with the overall concept, which is typical of the example metaphors that Lakoff and Johnson examine throughout their work. It is this very internal consistency, this systematicity of metaphor, that elevates it from the level of an “interesting use of words” to a mental concept. The example of “argument is war” is a metaphor with a limited domain, however. But other metaphorical concepts are much more central to our experience. For example, consider the concept “Time is money”. As Lakoff and Johnson explain, this metaphor is central to western culture (but by no means to all cultures!), is internally consistent, and even offers several other

metaphors which are subcategories of this concept. These subcategories are referred to by Lakoff & Johnson as 'entailment relationships', where the primary metaphor "Time is money" entails that "time is a limited resource", which entails that "Time is a valuable commodity", each at an increasing level of generality. This cluster of metaphors accounts for a great many possible sentences about time, for example:

How do you *spend* your time these days? (Time is money)

Is that *worth* your while? (Time is a limited resource.)

I *lost* a lot of time when I was sick. (Time is a valuable commodity.)

Accounting for the exact nature of the clustering process that leads to these metaphorical entailments seems to be the motivations for Lakoff's (1987) work on cognitive models and categorization, as we shall see below.

The above example metaphors are mappings from one abstract domain to another, first from argument to war, then from time to money. In general, however, the most powerful aspect of Lakoff and Johnson's theory materializes when we examine mappings from more abstract domains to more concrete domains. When the source domain is suitably basic, such as when it deals with human kinesthetic experience or knowledge of the properties of physical objects, then we are no longer just talking about metaphor, but rather about a system for the embodiment of human cognition. Embodiment is sometimes also referred to as semantic or symbol grounding, by which is meant a process for assigning meaning to an arbitrary symbol. Computer systems, for example, excel at symbol manipulation, but rarely are able to explain what the symbols *mean*. This is what we typically mean when we say that computers cannot think, that they cannot have knowledge. Their symbol manipulation is ungrounded. Even if we succeed in making a computer generate text (e.g. Howell, in press), perhaps through statistical or grammatical rules, and even if the resulting text makes some sense to a human reader, the computer that generated it cannot be said to have any knowledge of the text unless we have addressed this issue of embodiment or grounding.

This then, is the most powerful aspect of the theory of ICMs; they ground abstract concepts in meaning indirectly through more concrete concepts. Lakoff identifies a rough hierarchy of ICM from the most concrete to the most abstract. These include:

Image-Schematic

Propositional

Metaphoric

Metonymic

Symbolic

All ICMs can be roughly defined as mental spaces and models that structure those spaces (Lakoff, 1987).

The differences lie in what those spaces relate to and the manner in which they are structured.

Image-schematic ICMs, of course, are the most basic or concrete of all. They consist of basic-level kinesthetic image schemas (Johnson, 1987), the kinds of sensorimotor experiences that begin at the earliest age, and involve the most central objects and actions in our lives. ‘Basic-level’ is meant in the tradition of Berlin, Rosch, etc., as that level of interaction with the external environment at which people function most effectively and accurately. This basic level is characterized by gestalt perception (the whole is more than its parts), mental imagery, and motor movements and our proprioceptive perception of those movements. As Lakoff points out, the studies of basic-level categorization suggest that our human experience is preconceptually structured at that level, and that we are equipped with general capacities for dealing with part-whole structure in real-world objects via gestalt perception, motor movement, and the formation of rich mental images (Lakoff, 1987). With our earliest experience being structured in this way, it is no surprise that successively more complex experience and conceptualization should be built upon it.

Mark Johnson (1987) has identified some of the most basic of the kinesthetic image schemas, ones arguably most central to human experience. These include the **container** schema (a boundary distinguishing an interior from an exterior), the **part-whole** schema (the part-whole structure of bodies and objects), the **link** schema (which secures the location of one thing relative to another e.g. a rope), the **center/periphery** schema (center = identity or importance), and the **source-path-goal** schema (all actions involve a starting point, a trajectory, and an endpoint). Other central schemas include the up-down schema, the front-back schema, and the linear order schema, for example, although the full range and centrality of all of these is under continuous investigation to date.

From a developmental perspective, the means of learning these schemas could easily be quite basic, such as reinforcement learning, or stimulus-response learning. After all, children have a long period of time (at least a year), in which they are mostly helpless and immobile, to observe and interact with their

bodies and their immediate environments and build up these schemas slowly over experience. This may seem simple, but if the straightforward learn-by-experience of kinesthetic image schemas can in turn serve as a foundation in that most complex task, the acquisition of language, then that acquisition can itself be simplified, as we shall see. Moreover, not just language (which corresponds to the symbolic level of the ICM hierarchy above) but all intervening levels are based upon these kinesthetic image schemas. This concept itself should not surprise developmental researchers, of course, but Lakoff and Johnson have finally put the concept into a complete theory.

So using kinesthetic image-schematic models as a foundation, what do we gain? Lakoff offers the ‘Spatialization of Form’ hypothesis, by which he means a metaphorical mapping from the structure of physical space into that of a conceptual, mental space. Here the importance of the above-mentioned primary schemas is noticeable, as they have particular effects. Container schemas provide a method for understanding categorization, with the category as a container and exemplars being in or out of it (or in to some degree, depending). Hierarchical structure can be understood in terms of up-down and part-whole schemas acting together. Relational structure is of course defined by the link schema. Radial structure, as we will discuss later, is grounded by center-periphery schemas. Foreground-background structure, whether in visual scenes or in discourse, can be understood in terms of front-back schemas. And linear quantity scales are understood in terms of up-down and linear order schemas, as in the basic “up is more” metaphor. Image schemas thus both provide for direct understanding of their own structure, and are used metaphorically to structure other complex concepts.(Lakoff, 1987).

Incidentally, it is interesting to note that the central emphasis on early sensorimotor experience that the image-schematic ICM introduces is not new to psychology. Classic developmental work by Piaget (1952) focused on the importance of proper early sensorimotor experience in children’s development. Piaget theorized that such early experience was necessary prior to the child being able to proceed to more mature, ‘propositional’ knowledge such as quantity, number, etc. Of course, Piaget’s work, while influential, is often ignored by less developmentally and more cognitively-oriented researchers. Thus the renewed focus that Lakoff and Johnson have brought to the image-schematic level in cognitive science is invaluable. It would seem that higher thought is indeed predicated upon physical sensory and motor processes.

The next more complex type of ICM is the propositional, which may itself be of many types. The most basic are ‘simple’ and ‘scenario’. Others arise based on what sort of categories the concepts involved lend themselves to, for example ‘feature-bundle’, ‘taxonomy’, or ‘radial category’. The simple propositional ICM, first, is merely a structured arrangement of parts into a whole, perhaps including various semantic relationships between the parts (e.g. agent, experiencer, etc). The scenario ICM is more interesting. It is structured by the source-path-goal schema, but in the *time* domain, where the initial state is the source, the final state is the destination, and the various events of the scenario are locations on the path. These scenarios provide a contextual background against which to experience the people and objects within them, and may in fact alter our perception of those people and things depending on which scenario is currently active for us. As a simplistic example, one’s supervisor may also be a member of one’s softball team. The way that person is perceived will vary depending on which scenario one finds oneself in. These concepts of ‘boss’ and ‘teammate’ are of course themselves conceptual categories, as many such concepts will be, since we typically have many exemplars of both in our experience. The nature of the category structure of any given concept will determine what sort of category effects are applicable. For example, a concept that is defined by a logical rule is a classical category, with membership being either true or false. A feature-bundle with yes-no features thus represents a classical category. A feature-bundle structured by up-down and part-whole schemas becomes a hierarchical taxonomy, and can represent classical either-or inheritance taxonomies, such as scientific naming conventions. If those feature-bundles are graded instead of yes/no (e.g. size instead of “has fur”, then a graded category will result, with a fuzzy boundary. Finally, a complex category with subcategories, structured by the center-periphery schema, is a radial category, and is most important when considering strong, evolving metaphors in everyday usage that are subject to being extended to account for novel events (e.g. what adoption does to the category mother – subcategory “birth-mother”). This interesting concept of ‘radial category’ is well-documented by Lakoff (1987) and is one of the challenges to current language modeling efforts, as is discussed below.

Metaphoric or metonymic ICMs, the next layer, are quite powerful. They involve a partial mapping of structure from a source domain (typically image-schematic) to a target domain. While the mapping may be only partial, the parts of the source domain that are mapped will be internally consistent; elements that are inconsistent or opposed with the metaphor will be consistently excluded. With

metaphoric mapping, the source domain is typically structured by a prepositional or image-schematic ICM. With a metonymic model, the source domain and the target domain are the same, structured by an ICM, and a part **A** of the whole concept **B** is used to stand for **B**. It is interesting to note that this is similar to the concept of ‘abstracting’ that was advanced much earlier by philosopher A. N. Whitehead (1933) to allow a more integrated view of meaning. Instead of meaning being composed of atomist component parts, Whitehead argued, it is instead an integrated system where the whole is greater than the sum of its parts (a gestalt, in Lakoff’s terms). To avoid the false criticism that it would then be impossible to consider any one thing or concept independently, Whitehead proposed that one simply mentally ‘abstracts away’ those relationships in which the object participates that are static or not relevant at that time, without the implication that those relationships are somehow invalidated. In a similar fashion, if the other aspects of a concept are static or irrelevant in this domain, then we can safely treat a thing by whatever distinguishing part is left, hence metonymy.

This brings us to the final part of the theory of ICMs, the symbolic level. This is where we find language and linguistic phenomena. Of course, as might have been obvious, much of the work of explaining the meaning of language has been done already in the discussion of the ICM; language simply builds upon existing conceptual structure by assigning linguistic terms to them. As Lakoff points out (Lakoff, 1987) the traditional grammatical category *noun* turns out to be mostly accurate, since noun is a radial category. That is, the central subcategory of nouns consists of names for physical things (the classic “person, place or thing”) and these are the category’s prototypes. Other nouns are members of other subcategories, radially linked to the central subcategory (e.g. abstract nouns, unusual or out-of-date nouns). As Adele Goldberg suggests (1990;1995), verbs and verb constructions are also a radial category, with the ‘light’ verbs (move, put, give, do) being more prototypical, earlier learned by children, and ideally balanced between maximally informative and maximally general. Even the constructions (e.g. caused motion) that verbs are used in are radially categorized as a result, and light verbs’ constructions are prototypical of these categories, Goldberg argues, and are learned earliest.

The evidence indicates then, that ICMs are indeed how we develop mental concepts, that those concepts become categories over time (with experience of many exemplars), and that some of the most important categories evolve the most complex structure, especially radial structure. Although the concept

of radial categorization is interesting and powerful by itself, explaining as it does how metaphor and hence concepts and language evolve, a yet more interesting question remains. Idealized Cognitive Models seem to explain a great deal, and even hold predictive power (such as on what metaphoric extensions are possible in language), but does the brain actually think that way? The evidence offered by Lakoff from language use is compelling, so if we accept that the brain does operate this way, then the next question is “how”? How are these categories structured in the brain? How is categorization represented neurally?

Well, evidence does exist indicating that metaphorical relationships are learned using the same circuits in the hippocampus that are otherwise used for learning spatial relationships and spatial maps. Indeed, in a manner remarkably similar to Lakoff’s, but from an evolutionary psychology rather than a linguistic perspective, John O’Keefe argues that metaphor is central to human experience, and that the human hippocampus is responsible for it (O’Keefe, 1990). He suggests that while in non-human animals both the left and right hippocampus receive sensory inputs about objects and stimuli from the neocortex, in humans the left hippocampus has specialized to receive a new set of inputs from the language centers of the neocortex. These language inputs would consist of the names of objects and features and not of their sensory attributes.

This seems to indicate the possibility that it is within the hippocampus that Lakoff’s ICMs connect, with the symbolic ICMs coming through the left hippocampus, being structured spatially, and perhaps being paired with the spatio-temporal ICMs that would still be processed through the right hippocampus. Certainly the evidence showing the importance of the hippocampus in spatial maps is uncontroversial; if the left-right specialization exists as O’Keefe argues, then it is only a small step to accepting the integration of ICMs through the two sides of the hippocampus. Further investigation seems definitely warranted!

Unfortunately, with such issues as yet only suggestive and not proven, and neuroscience in general not having much definitive to say on higher-level cognitive processes like language, we must turn to alternative methods such as computational neuroscience and neural modeling or neural networks. What can, for example, neural network modeling work on brain processes tell us about possible neural methods for categorization? Actually, it is quite possible to find evidence of categorization in neural nets. Neural network models with simplified brain-like neurons and connections can be shown to instantiate several different aspects of categorization phenomena. For example, a standard feedforward neural network can

handle feature-bundle or propositional categorization including displaying semantic priming effects (e.g. Cree, McRae, & McNorman, 1999), by developing hidden representations that encapsulate the common relationships among many input examples. If the feature-bundles are allowed to have graded representations, then the network can handle graded categories. Hierarchical taxonomies should also be possible, provided that representations for superordinate and subordinate layers are included in their input representations.

It may even be possible to account for the phenomenon of radial categorization in neural network terms. Radial categories are groupings of subcategories that stem from a central subcategory and extend into peripheral categories in all directions. Following the lead of Lakoff with his spatialization of form hypothesis, we might best imagine radial categories as a topology, a surface in multidimensional space. If we first consider the more easily visualized three-dimensional space, we can imagine the central subcategory of a cluster as a hole or depression in the landscape. Then, within this larger depression, smaller subcategories may lie as secondary depressions or basins. Indeed, if a given subcategory extends quite far into the periphery of the main concept, then it may well overlap the lip of the main basin entirely, extending in part beyond the main concept, and overlapping with the other concepts that it is linked to metaphorically. If we imagine a ball dropped on this three-dimensional surface, we can see that on the average it will tend to fall to the bottom of the main depression, but if it rolls in from a given angle or direction (i.e. from a neighboring semantic space), then it may instead tend to get caught in a local basin, whether entirely within the sides of the main basin, or overlapping between it and another concept.

Of course, one of the drawbacks of this three-dimensional picture is that in three dimensions, there is a limited amount of space on the surface that is adjacent to the conceptual basin, and any neighboring concepts that might overlap or allow for metaphorical extensions would need to lie closely around it. However, it is doubtful that the complexity of semantic relationships could be captured with such a limited range of relationships. Luckily, however, we can extend this basin-and-surface analogy to higher dimensional spaces such as those typically encompassed by the vector co-ordinates of neural network attractor models. In these models, the depressions or basins are termed attractors, since they 'attract' a trajectory (like the imaginary ball mentioned earlier) into their domain. In multi-dimensional space, the dimensions in which attractors neighbor each other can be very numerous (100-200 dimensions are often

used) allowing for complex simultaneous semantic neighborhood structures. A representation like this of Lakoff's radial categories has the advantage over the purely linguistic description that it may actually have predictive power. That is, if we can map and model the existing state of a language's rich radial categories, or at least of a subset of them, then we can determine the positions of the attractors that result.

Metaphorical extensions in a radial category are theorized to operate primarily on the central subcategory, not the peripheral. Thus we should be able to determine from the positioning of a central subcategory in relation to its neighbors where the likely future metaphorical extensions could happen: anywhere that positions are close but an overlapping basin does not yet exist.

Of course, this attractor account of radial categorization is speculative at present. However, it does seem worthy of empirical investigation. Perhaps an attractor network could be trained on concepts from a well-documented test domain, and experiments run to determine if the metaphors that exist map to the closest relationships in the artificial semantic space. Alternatively, some variant on the ever-popular priming experiments using attractor networks could be used to test for the metaphorical relationships between concepts. Radial categorization is a new concept to the PDP modeling community, so perhaps this sort of investigation will open up new directions in semantic and language modeling.

Shifting from the most complicated of categorization effects to the most basic, we must consider how image-schematic concepts could be represented in neural networks. Basic-level effects, central to the image-schematic level of concept formation and categorization, have been claimed to be demonstrated by Rogers and McClelland (1999, in preparation). Their models *do* seem to fit the data on a wide variety of the supposed effects of the basic level, including expertise effects, familiarity effects, etc. However, upon a close examination it seems they are not necessarily demonstrating basic-level categorization. Basic-level effects, as we saw above, are characterized by gestalt perception (the whole is more than its parts), mental imagery, and motor movements and our sensory and proprioceptive perception of those movements. The networks that Rogers and McClelland use are solely at the feature-bundle ('has legs', 'size') level of representation, or the simple propositional ('is a', 'has a'). The effects that they are demonstrating could have arisen solely due to a balancing act between frequency (generality) and distinctiveness.

That is, if a representation is too general, then it is very applicable, and hence frequent. But by corollary, it is not very distinctive; for example, consider 'plant' compared to 'flower' compared to 'rose'.

'Rose' is maximally informative and distinctive, but not very general and hence not very frequent; not many plants are roses. 'Plant' is very general (true of all plants), and hence the proposition 'isa' plant is true of all plants, but is not very informative or distinguishing. 'Flower' is more general than 'rose', more distinctive than 'plant', but strikes the right balance between the two. The hidden layers in a PDP net pick up on these relationships in the process of performing their recodings of the inputs.

However, whether this inverse relationship between frequency and distinctiveness is really the same thing as the conceptual 'basic-level' that emerges from our human experience is another question. The phenomena are similar, true. But as Lakoff has argued (Lakoff, 1999, personal communication) Rogers and McClelland's models don't address the aspects of basic-level phenomena that are arguably most central to our human experience: sensorimotor (image-schematic) experience. Rogers and McClelland's models provide no account of any conceptual level other than the propositional (feature bundles and simple propositions), and thus do not capture any of the conceptual experience of the image-schematic level. Their propositional representations *assume* some sort of preprocessing to create these feature-bundle representations, representations that would otherwise be built up out of image-schematic components. They do not, however, account for this lack. The fact that this is common practice in the psychological and connectionist work with feature-bundle representations is not an excuse, not in light of the above-cited and fairly widespread evidence for the necessity of the image-schematic level. The issue of how to instantiate the gestalt effects of the image-schematic level, for example, poses a possible stumbling block for PDP modelling and deserves much further investigation. One promising approach might be drawn from the similar recent work on figure-ground separation in neural networks using phase-synchronous oscillatory nodes (e.g. Wang, 1996).

However, there might be another way to integrate the two approaches. Despite the criticisms, Rogers and McClelland are able to demonstrate a model that fit some of the human data on the basic level effect. Perhaps the reason that they are able to do so is in *spite* of the fact that they use such pre-processed, non-sensorimotor feature-bundle representations. It is possible that the inverse curve that seems to underlie their results is in fact an emergent phenomenon that occurs in any categorization domain, even including the image-schematic level. Perhaps if their models had included some sort of image-schematic input representations, capturing both sensory and motor relationships, then the same effect would be observed for

semantics in that domain. It is conceivable that sensory and motor experience might, in effect, act to determine the points on the generality/distinctiveness curve that are of significant enough value to us as humans to serve as the locus of basic level effects.

Evidence for this is provided by the expertise effect. In the plant-flower-rose example above, we can imagine the point of maximum utility on the generality-distinctiveness curve changing as our external environment changes. Thus when I take a job as a rose gardener, for example, I might expect to encounter many more varieties of roses than of other flowers, or even of other plants, including various subterms for different types of rose. My sensorimotor experience with roses might lead the word 'rose' to become more basic to me than 'flower'.

Of course, other than the reinterpretation of the expert-level effect just discussed, I can as yet offer no evidence for this integration of neural network models and categorization phenomenon. Still, I believe the hypothesis is plausible, and lends itself to a variety of empirical testing, such as by experimentally altering subjects' external environment as in the rose example and continually re-evaluating their basic-level effects. The frequency by distinctiveness curve itself is emergent from the structure of feedforward neural networks, and as PDP modelers would like to believe, also (in some form) in the brain. Of course, it is also true that metaphor-level cognitive models mapping from abstract domains to concrete domains offer great promise for extending language modeling from the concrete to the abstract, and also deserve further examination. However, ICM levels are somewhat hierarchical. We must learn to instantiate image-schematic cognitive models before we can concern ourselves with metaphor and metonymy.

Thus, the incorporation of more image-schematic, and ultimately perhaps, metaphoric aspects into PDP models can only improve the state of the field, allowing us to move from toy problems that address only a given issue under study to more general models that can deal with the process of language embodiment and begin to truly acquire language. Thus it seems the study of metaphor may have led to several valuable insights; Lakoff's theory of cognitive models, Johnson's kinesthetic image schemas, a possible neural location for their operation (the hippocampus), and perhaps to a means for embodying meaning in neural networks. Indeed, if the theory of cognitive models can be instantiated in neural networks, it may breathe new life into PDP models of language processing and language acquisition. Metaphor is certainly more than just poetry!

References

- Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming., *Cognitive Science*, 23(3), pp. 371-414.
- Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago and London: University of Chicago Press.
- Goldberg, A. E. (1999). The emergence of the semantics of argument structure constructions. In MacWhinney, B. (Ed.) (1999). *The Emergence of Language*. New Jersey: Lawrence Erlbaum Associates.
- Howell, S. R. (in press). Modeling Language at Multiple Temporal Scales, *Proceedings of the Cognitive Science Society 2000*.
- Lakoff, G. and Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York: Basic Books.
- Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind*. Chicago and London: University of Chicago Press.
- Lakoff, G., and Johnson, M. (1980). *Metaphors we live by*. Chicago and London: University of Chicago Press.
- Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason*. Chicago: University of Chicago Press.
- O'Keefe, J. (1990). A computational theory of the hippocampal cognitive map. In O. P. Ottersen and J. Strom-mathisen (eds.), *Understanding the brain through the hippocampus*, 287-300. Progress in brain research, vol. 83. Amsterdam: Elsevier.
- Piaget, J. (1952). *The origins of intelligence in children*. New York: Internal Universities Press.
- Rogers, T. T., & McClelland, J. T. (1999 – in preparation). *Semantics without Categorization*.
- Whitehead, A.N. (1933). *Adventures Of Ideas*. Cambridge: Cambridge University Press
- Wang, D. (1996). Primitive auditory segregation based on oscillatory correlation. *Cognitive Science*, 20, pp. 409-456