Convergence Zones and Conceptual Structure

George Lakoff

Department of Linguistics
University of California at Berkeley
Berkeley, California 94720

Commentary

This is an invited commentary on Antonio Damasio and Daniel Tranel's "Nouns and Verbs Are Retrieved With Differently Distributed Neural Systems" for the *Proceedings of the National Academy of Sciences.*

INTRODUCTION

In principle, the study of conceptual structure and neuroscience should go hand-in-hand, but this has rarely been the case. Researchers on conceptual structure come from the cognitive sciences (Linguistics, Psychology, and Anthropology) and study the nature of concepts in scrupulous detail, as they are used by normal people in various cultures. Neuroscientists use other approaches. They may, for instance, look at brain-damaged patients and to correlate their lesions with whatever cognitive deficits they have. The kind of research done in the two fields is so different in character that it is rare that they come together.

THE CONVERGENCE ZONE HYPOTHESIS

One of the most interesting things about the convergence zone hypothesis (CZH) of Antonio and Hanna Damasio is that it fits fundamental results about conceptual structure so well that it has the possibility of providing a common framework for both fields. Before explaining why, however, I must first say a few things about what convergence zones are.

A Convergence Zone (CZ) is a hypothesized neural control structure-- a neural ensemble that integrates operations performed elsewhere, in a number of different locations. The integration is done via temporal binding -- control over the timelocked activation of neural patterns in those locations. Such binding permits the integration of features computed in various other places into a single gestalt. CZs with the same neural architecture control different functions, depending on where in the brain they are located. CZs are hierarchically and heterarchically linked. It is hypothesized that cognitive

deficits associated with brain damage arise because lesions damage convergence zones or connections between them.

To get a feel for what convergence zones are, consider what it takes to name a familiar face.[1]

1. The face must be perceived as a whole, with various features (eyes, nose, mouth, chin, hair) integrated in just the right way to form a single gestalt. There are patients who have difficulty integrating parts of a figure into a whole. For such patients, the CZH says that the CZs integrating low-level facial features detected elsewhere are disrupted.

2. The face, perceived as a whole, must then activate knowledge about a particular person. There are patients who can perceive a face as a whole visually, but not connect it with knowledge about the person, though they can identify the face when either voice or gait is added. There are other patients for whom voice or gait perception does not help face recognition. Here the CZH says that the CZs connecting the face-gestalt with knowledge is disrupted.

3. Having processed the face as a gestalt and activated knowledge to know whose face it is, one must then be able to assign a name to the face-as-recognized. There are patients who can recognize a familiar face and give lots of information about the person (who may be a close relative or a celebrity), but not supply the name. Here the CZH says that the CZ integrating proper names with recognized faces is disrupted.

In each case, there are convergence zones integrating information computed or stored in other places. The CZs form hierarchies. The face-gestalt CZs (call them CZ-1) are linked to knowledge about a particular person, to voice perception, and to gait perception via a person-recognition by a CZ (call it CZ-2). And the person-recognition CZs (CZ-2) are linked to knowledge of proper

those lower-level activation patterns so that the patterns to be integrated fire in synch. In the case of a hierarchy of CZs, one CZ would control the firing in other CZs, which control the firing in other CZs, and so on until the lower-level parts of the brain where the component features are computed is reached.

The CZH thus makes sense of commonplace PET scan data. When a person is performing a single relatively simple task, it is common for PET scans to reveal activation not just in one portion of the brain but in many. On the CZH, the parts of the brain activated are linked via convergence zones that are organized hierarchically and heterarchically. The idea is that a task of any complexity at all is not performed in just one place, but rather involves a cascade of activation from higher association cortex to low-level regions in the visual, somatosensory, auditory, or other cortices.

The CZH can be understood somewhat better if one thinks what it is contrasted with. Here are some other possible theory types:

(1) Direct-connection theories, in which there are no intermediate ensembles of neurons that serve as control structures, but instead there are reciprocal connections directly among those areas.

(2) Intermediate representational structures, which have structures like convergence zones but with different functions: instead of controlling the activation in lower-level regions, they form representations at a higher level with input from lower-levels.

(3) Local modularity theories, which claim that a cognitive ability lost due to a brain lesion is localized in the very portion of the brain where the lesion occurs.

To my knowledge, the Damasios do not argue explicitly against these kinds of theories, though their papers do contain implicit arguments. Let us look first at their implicit arguments against local modularity theories.

Consider again the task of naming a familiar face. Suppose someone with a brain lesion cannot perform the task. A simple-minded local modularity theorist would say that face-naming is performed in the area where the lesion is. The Damasios' data contradicts such a simple-minded view. Someone could fail to name a familiar face for at least three different reasons: (1) a gestalt-formation deficit; (2) a recognition-of-identity deficit; or (3) a naming deficit. Such deficits correspond to lesions in three different parts of the brain: (1) the inferior occipital lobes; (2) the mid inferior temporal lobes; and (3) the left anterior temporal lobe. Face-naming is thus not a unitary capacity, but a hierarchical capacity -- it involves hierarchically organized abilities. Wherever the Damasios cite hierarchically organized abilities with deficits corresponding to each, they are implicitly arguing against local modularity theories. It is hard to imagine any serious neuroscientist being such a simple-minded local modularity theorist these days.

The Damasios do argue against intermediate representational theories. The argument is a kind of biological Ockham's razor argument: Intermediate representations would duplicate information stored elsewhere. Moreover, each level of intermediate representations to be stored would be a multiplicative function of the representations at lower levels. Thus, as one gets to higher and higher levels, more and more information would need to be stored. Yet neural structures thin out at the higher levels of cortex, which makes it unlikely that ever more information is stored as one moves toward higher levels.

As for direct-connection theories, they do not contain a mechanism to account for gestalt formation (for example, the integrating of low-level visual features to form a face), nor do they contain a mechanism to account for hierarchical capacities and the corresponding deficits. That does not mean that no future direct connection theory could emerge that could account for such things, but at present it is hard to imagine how a direct connection theory could do what is necessary. This should be seen as a challenge to any direct-connection theorists to show how all the Damasios' data could be handled.

If one rules out direct connection theories, intermediate representation theories, and local modularity theories, the only theory I know of that is left is the convergence zone hypothesis.

There is, however, a note of caution that should be sounded about the CZH. At present, no computational models of convergence zones exist. We do not know exactly what the computational properties of such structures would be, nor do we know that hierarchies of convergence zones could in fact function computationally in the way the Damasios would like them to. And even if there were such computational models, we do not know that real biological neural structures would have such biological properties.

The good news is that there is some impressive research that has been done on neural structures that control the timelocked activation of other neural structures. It has not been done on exactly the kinds of structures involved in hierarchies of convergence zones, but what has been done is quite hopeful. The research has been done by Lokendra Shastri and his students in the Computer Science Department of the University of Pennsylvania. Shastri has developed a computational model in which the binding done in predicate logic is accomplished via timelocked activation

(called ``temporal synchrony'') in a connectionist model.[3] Shastri has developed techniques for representing the binding of a variable to a referent by timelocked activation. This is basically what is needed to bind a low-level facial feature to a slot in a general face schema. Shastri is aware of the Damasios' work on convergence zones and is enthusiastic about the possibility that his computational techniques can be applied to model convergence zones. That, however, would be a major undertaking. Nonetheless, it would now be possible, were funding available, to start the computational modeling of convergence zones.

One last note of caution: It is only a theory that binding is achieved by timelocked activation. It is not known that it is true. To my knowledge, there is not even a theory as to why timelocked activation should result in binding. One might imagine why timelocked activation might be *necessary* for binding: after all if representations of a `slot'' in a schema and a `filler' for that slot were to be bound, they ought to be activated at the same time. But it is by no means clear why timelocked activation should be *sufficient* to achieve binding.

CONVERGENCE ZONES AND CONCEPTUAL STRUCTURE

Why would a cognitive scientist who studies conceptual systems be interested in convergence zone research? Let us start with the study of prototype structure. Eleanor Rosch, in the 1970's, demonstrated that not all category members are on a par: some are more typical than others. For example, if someone tells you there is an animal in your front yard, you would probably expect a medium-sized animal with four legs, fur, a tail -- probably something like a raccoon, fox, or possum. You would probably not

expect a giraffe or an elephant. If a giraffe or elephant were in your front yard, most people would identify them as such. This is the sort of asymmetry that Rosch noticed -- when superordinate category name ``animal'' is used, one tends to expect an animal of a typical shape, rather than an animal with a very unusual shape like a giraffe or elephant. In a wide variety of experiments, Rosch showed that prototypical members of categories are processed differently than nonprototypical members, thus demonstrating the cognitive reality of prototypes. [4,5] The Damasios and their co-workers [6] have provided evidence that the prototype-nonprototype distinction is not just cognitively real but physically realized. They describe patients with damage to the inferior temporal cortices (areas 21, 20, 36, 37) who cannot conjure up knowledge about animals with prototypical shapes (raccoons, foxes, wolves, etc.) but have no trouble at all with elephants and giraffes. They believe that access to knowledge of animals with similar physical shapes (that is, the prototypical shape: four legs, medium size, fur, normal neck, ears, no tusks, etc.) involves the inferior temporal cortex, while access to knowledge about animals with shapes quite distinct from those of other animals (they call them ``outliers''; cognitive scientists would call them ``nonprototypical'') does not involve inferior temporal cortex.

The claim is striking and important. Prototype theory was a major advance in the understanding of conceptual structure. It had startling consequences, among them, that the classical theory that categories are defined solely by a list of necessary and sufficient conditions was wrong -- and hence so was classical logic and every discipline that depended on it. In subsequent years, many types of prototypes were discovered [5], each with different associated inference patterns. The full brunt of prototype theory has

not yet been felt in the academic world, and so the suggestion that there is a biological basis for it is rather important.

Yet the hardnosed cognitive scientist, while rejoicing for a short while in biological justification, must nevertheless ask some tough questions. Rosch's research did not just show that there was a binary prototype/nonprototype split; rather there are degrees of prototypicality. How can a theory that claims that prototypes are accessed through the inferior temporal cortices, while nonprototypes are not, account for degrees of prototypicality?

There might be another explanation of the data. Connectionist modelers have shown that connectionist networks can learn prototypical examples and a cline of less prototypical examples. In such a net, similar examples are stored in similar patterns of weights, dissimilar examples in dissimilar patterns of weights. It might be the case that ``damage" to such a network, say, knocking out a certain proportion of neural units or connections, might make it impossible to distinguish among the similar prototypical cases (which are stored in the same portion of the network) while still allowing for relatively unique cases to be distinguished. I do not know if this is true, but it is conceivable. It is also testable via modeling experiments.

Of course, traditional PDP connectionist modeling does not contain convergence zones. A ``hidden layer" where such information would be stored would be storing a representation, not a pattern for controlling the timelocked action some other set of neural ensembles. To know if this was a real possibility within the theory of convergence zones, one would have to study the kind of neural modeling that Shastri is engaged in, not garden variety connectionist models. It is nonetheless conceivable that this would be true in a Shastri-style model of a convergence zone, though the experiment

could not be made until such models were explictly constructed. Nonetheless, suppose it is true. Then there is another possible explanation for the data: the knowledge about raccoons, foxes, and wolves would be stored in the same portion of the brain as the knowledge about elephants and giraffes. ``Damage" might not be total destruction of a network, but sufficient degradation to make it impossible to sort out subtle differences while making it possible to sort out gross differences.

In short, the Damasio's data is interesting. It still supports a biological distinction between prototypical and nonprototypical cases. But until such an alternative is ruled out, we cannot be sure that knowledge about elephants and giraffes is not accessed through the inferior temporal cortex. One of the reasons that computational models are useful is that they generate alternative biological hypotheses.

A cognitive scientist looking at the Damasios' data would want to test out everything that was known about all the different kinds of prototypes in all the categories that have been studied, something that is not always practically possible with brain-damaged patients and the kinds of tasks used in lesion studies.

There is a brief and unprofound moral here: the relationship between neuroscience and cognitive science is important but hard to establish.

A STARTLING CASE

The existence of basic-level categories is one of the major discoveries of cognitive science. The discovery was made by Brent Berlin and Eleanor Rosch, based on earlier observations by Roger Brown[5]. Before their discovery, it was assumed that all human categories had essentially the same structure,

that categories were defined by inherent properties of their members, and that the internal structure of a category had nothing to do with where it stood in a category hierarchy. What Rosch and Berlin discovered was: (1) Categories at a certain middle-level of categorization have a different structure than categories at higher levels.(2) These categories are defined not objectively by the properties of the category members, but interactively, by the nature of certain kinds of interactions between people and the category members. (3) Those interactions are (a) gestalt perception; (b) mental image formation; (c) motor interaction; and (d) extensive knowledge.

To get a feel for the distinction between a basic-level category and a superordinate category, consider the difference between the basic-level CHAIR category and the superordinate FURNITURE category. (a) One perceive chairs in terms of a single overall shape; but there is no overall shape for pieces of furniture in general -- that is, no overall shape that encompasses chairs, tables, beds, lamps, etc. (b) Correspondingly, one can form a mental image of a chair, but one cannot form a mental image of a general piece of furniture (neutral among chairs, tables, beds, etc.); (c) We have special motor programs for interacting with chairs, but no motor programs for pieces of furniture in general (not distinguishing among chairs, tables, beds, lamps, etc.); and (d) we know a lot about chairs more than we know about generalized furniture (knowledge that doesn't distinguish among kinds of furniture).

These were the fundamental criteria discovered by Rosch, who mostly studied artifacts and implements. Berlin, who mostly turned his attention to plants and animals, did not include condition (c), motor interaction. What Rosch called ``basic-level" categories, he called ``generic-level" categories, because they corresponded to the level of the biological genus. That is the

level at which oaks are distinguished from redwoods, and the level at which elephants are distinguished from giraffes, raccoons, foxes, possums, and so on. Brown, however, had observed that we have special motor programs for interacting with certain domestic animals like cats and dogs --and for many people, horses.

One of the remarkable things discovered by the Damasios and their coworkers is that there are patients with face agnosia who also cannot recognize certain basic-level categories -- though not all of them. These are patients with lesions in the inferior temporal lobes who cannot integrate a percept of a particular face with the information that goes with that face, information that would permit recognition. Some of the same patients also cannot distinguish among prototypical basic-level animal concepts. That is, they cannot tell a raccoon from a wolf or a fox, and refer to them as just animals, though they can recognize nonprototypical animals like elephants and giraffes. Such patients can, however, recognize basic-level implement and artifact categories: knives, forks, chairs, tables, and so on.

Why should an inability to recognize particular faces accompany an inability to recognize categories of animals with prototypical shapes? The answer the Damasios give is that both involve (a) the ability to make fine distinctions among images that involve the complex integration of visual features, and (b) the ability to pair such distinctions with specific knowledge. In the CZH, both disabilities involve disruptions of convergence zones that integrate visual images with other knowledge and that, consequently would be located in the same region of the brain.

And what about the fact that such patients can recognize other basic-level categories, those for implements and artifacts? Their explanation is that those categories involve the information not only of visual images and

specific information, but also of somatosensory (tactile) and motor information. Thus, knowledge of those categories would have to be accessed via convergence zones controlling somatosensory and/or motor information, which would have to be located elsewhere in the brain in a place where inferior temporal lesions would not affect them. Examples of such locations include the posterior and lateral temporal cortex, some parietal cortices and frontal cortices.

But if this is so, there should be patients with lesions in those areas who cannot recognize implements and artifacts, but can recognize faces and prototypical animals. Indeed, there are such patients, as the Damasios report [1] That is, there is a double dissociation: Inferior temporal lesions, but not more dorsal temporal or parietal lesions correspond to a loss of face recognition and prototypical basic-level animal categories. Dorsal temporal or parietal lesions can do the reverse. In all cases, superordinate categories (animal, object) are retained.

Indeed, there is one more remarkable finding in this data. Consider the case of cats and dogs -- prototypical animals for whom most of us have motor programs (e.g. for petting) and horses, for whom many people, especially in rural areas, have motor programs (for petting, feeding, and riding). With the patients tested, these motor-program-animals patterned just like implements, not like other animals. Patients with inferior temporal lesions did not lose the ability to recognize these animals. In brief:

If categories are defined by just the integration of visual images and other knowledge, they will be accessed via convergence zones in one sector of cortex.

If categories are defined by the integration of visual images, other knowledge, and motor programs, they will be accessed via convergence zones in another sector of cortex.

NOUNS AND VERBS

This brings us to the paper in this volume on nouns and verbs. Linguists know a lot about nouns and verbs, and they are fairly picky about them. We know that names for concrete objects are prototypical nouns and that names for concrete actions are prototypical verbs. But we also know that there are many more kinds of nouns and many more kinds of verbs. Concrete nouns include names for animals and implements -- dog, cat, raccoon, fox, elephant, giraffe, knife, fork, pencil, plate, and so on. The Damasio-Tranel paper is about these nouns, and not at all about other nouns, nouns like illness, action, thought, government, or umbrage as in "take umbrage at," or restitution as in "make restitution". There are thousands of nouns of various sorts that this paper is not about, as the authors are well aware. In addition, this paper is not about nouns that are formed from verbs, for example, run, as in "a run through the park" or mix as in "cake mix" or mixing as in "the mixing of ingredients."

Corresponding distinctions need to be made about verbs. The Damasio-Tranel paper discusses verbs that name concrete actions, like run, hit, kick, swim, ride, jump and so on. The paper is not about other kinds of verbs, like think, know, see, have, weigh, try, resemble, govern, regret, be and thousands more. The paper is also not about verbs that are directly formed from nouns, like hand as in "hand me the gun," or club as in "club him over the head," or head as in "he heads the committee." Nor is the paper about verbs that are

derivationally formed from nouns by the addition of a prefix or suffix, like behead or deflower or classify.

There is a good reason why linguists are picky about things like this. There are linguistic theories that claim that grammatical categories like nouns and verbs have an existence that is independent of semantic categories like animals or implement or physical actions. To claim that there are differently distributed neural systems that govern access to grammatical categories like noun and verb is thus a very different and far more radical claim than the claim that there are differently distributed neural systems that govern access to names for concrete objects and names for concrete actions. The Damasio-Tranel paper makes the latter, much less radical claim. That does not mean it is uninteresting, but it is not about the grammatical categories noun and verb. It is instead about the naming of concrete objects and concrete actions.

Given that proviso, the paper presents a remarkable result about the naming of categories. Recall that category recognition is distinct from category naming. Two of the patients, AN-1033 and KJ-1360, have no problem with category recognition, but only have naming deficits. The third, Boswell, has both recognition and naming deficits. Recall also that concepts involving tactile and motor activities (implement concepts , physical actions) are accessed via regions related to somatosensory and motor processing. On the other hand, concepts involving only visual images and other knowledge, but not motor activity, are accessed via the inferior temporal lobes. Recall as well that the language areas, which provide access to phonological forms, are located in the left hemisphere. One would therefore expect that convergence zones for naming objects where concepts depend on visual images but not motor involvement would be in the left inferior temporal cortex.

On the other hand, one would expect convergence zones for naming manipulable objects to be in the posterior temporal and inferior parietal cortices. The reason is that in those regions CZs can access not only visual knowledge but also somatosensory knowledge and, indirectly, motor patterns of the hand.

What about the locations of CZs for naming fruits? We would expect them to vary with the particular fruit, depending on how much visual, tactile, or motor pattern knowledge characterizes our interactions with that fruit. For example, consider a banana. It is a fruit that we learn a special motor program for, a program involving the hands. The CZH would lead us to expect that the banana concept would be stored in a CZ with access to not only visual knowledge, butalso somatosensory and hand-based, but not full-body, motor knowledge. Finally, we would expect concepts for physical actions such as running or swimming to be accessed through still other CZ sites, since the knowledge characterizing such actions involves both visuo-motor and somatosensory patterns of whole body actions, not just hand actions.

The hypothesis also generates a further expectation. Lesions are not strictly demarcated. Serious damage in one zone typically entails at least some lesser damage in nearby zones. Thus, where there is major damage to the inferior temporal lobes, it would not be at allstrange to find lesser damage to such a nearby area as the posterior temporal region. Such differing amounts of damage would lead us to expect corresponding differences in deficits.

On this line of reasoning, we might expect KJ-1360, who has left premotor damage, to have a naming deficit for physical actions involving motor activity over the whole body (swimming, running).But we would not expect him to have a naming deficit for objects and other concrete entities

like animals, since they involve visual and tactile information, but not whole-body motor information.

On the other hand, Boswell and AN-1033, who have left inferior temporal lesions, would be expected to have deficits in naming animals, but not in naming physical actions. It would not be surprising for them to have lesser damage in the nearby posterior temporal region, and hence a lesser deficit in naming tools and implements. This is essentially what Damasio and Tranel report.

Consider their Table 1. Boswell successfully named 92% of the physical actions, but only 24% of animals and 25% of fruits and vegetables, which is consistent with his expected deficit. He named 76% of tools and utensils, a minor deficit which is consistent with a lesser degree of damage in the nearby posterior temporal region. Incidentally, despite his problems with other fruits, he had not problem naming bananas.

Now consider KJ-1360. He got only 53% of the physical action names -- an expected deficit given his premotor damage. His uniformly high scores otherwise are consistent with his lack of damage in other areas.

AN-1033 also fits the theory. He has the expected deficit in the non-motor-involved animal and fruit/vegetable categories: 51% and 54%, which indicates less serious damage than Boswell, but significant nonetheless. He has the expected lack of a deficit in the motor-involved physical action category: 96%. And his 70% performance on tools and utensils is, like Boswell's 76%, consistent with lesser damage in nearby areas. And, like Boswell, though he had severe problems naming other fruits, he had no problem with bananas.

A COMMON FRAMEWORK ?

The CZH is interesting for the predictions it generates. But, I find it even more interesting for another reason. The CZH leads to an expectation about the nature of conceptual systems. CZs are structures that bind information occurring in two or more other neural ensembles. The binding can be of two types: (1) the formation of a gestalt assembled from various ensembles, as in categories that are characterized in terms of visual, tactile, and motor information; or (2) the formation of correspondences across two ensembles, as in the association of lexical items with concepts. I will call structures of type (1) ``gestalt structures'' and structures of type (2) ``mapping structures''.

Since CZs can create these two kinds of binding structures, we would expect a conceptual system to be made up of structures of these kinds. Since they can occur in different parts of the brain, we would expect such general types of structures to recur with different subject matter. This is, indeed, just what occurs in conceptual and linguistic systems. Here are some examples of each type:[5]

GESTALTS

Basic-level concepts: As we have seen, these require the integration of information about visual, tactile, motor, and knowledge.

Complex image-schemas: Take, as an example, the word ``into'', which requires binding the goal on a path schema (designated by ``to'') with the interior of a container schema (designated by ``in'')
to form a single gestalt.

Conceptual schemas (or `frames'): A semantic field is a collection of systematically related words like ``buy", ``sell", ``goods", ``price". These are defined relative to a conceptual schema characterizing commercial events in general. The schema is a gestalt consider of four entities (buyer, seller, goods, money) and an action scenario of three parts: (1) The precondition: buyer has money; seller has goods; buyer wants goods; seller wants money. (2) Buyer and seller exchange money and goods. (3) Final state: Buyer has goods; seller has money. Words like ``buy", ``sell", ``price", ``goods", ``cost," and so on, only make sense when defined relative to such a schema. Such schemas form gestalts, where the parts must be bound together to form a whole, and where, for example, the buyer in the three parts of the above scenario must be characterized via binding as the same entity.

Grammatical Constructions: Recent research in cognitive linguistics has revealed that grammars of languages are made up of constructions -- pairings of conceptual schemas with syntactic schemas. An example is the deictic there-construction.[5] Consider the sentence ``There goes Harry with a red hat on." This is a special construction of English with the following syntactic schema: the subject follows the verb, the deictic locative pronoun (``there") comes first, and a short phrase (``with a red hat on") comes at the end). The syntactic schema is paired with a conceptual schema, the pointing-out schema, in which a speaker is directing the attention of a hearer to a location in the visual field where an entity with certain properties is located or is moving. The binding between the conceptual schema and the syntactic schema is as follows: The location is bound to the deictic locative pronoun (``there"); the predicate of motion or location is bound to the verb (``goes"); the entity is bound to the subject (``Harry"); and the properties of the entity are bound to the short phrase (``with a red hat on").

MAPPINGS

Mental spaces[7]: Consider a sentence like ``In that painting, the girl with the brown eyes has green eyes.'' Here we understand a brown-eyed girl in the real world as having a counterpart in a painting, where the counterpart has green eyes. In the general theory of such cases, the real world and the painting are called *mental spaces* and a correspondence exists between an entity in one space (the brown-eyed girl of the real world) and its counterpart in another space (the green-eyed girl of the painting). Mental spaces are commonplace -- movies, books, dreams, and more commonly, the mental spaces of beliefs, desires, and hypotheses. Correspondences across mental spaces are not necessarily one-to-one. Consider the sentence, ``People used to think that the morning star and the evening star were two different stars, but now we know that they are the same entity, the planet Venus.'' Here there is a former belief-space containing two different entities, which both correspond to a single entity in the space characterizing reality as we now understand it. In that case, there is a two-to-one mapping. Here is an example with a one-to-two mapping: ``If you had been born twins, the two of you would have competed with one another.'' In this case, there is one entity in the reality space and two in the hypothetical space.

Mental space phenomena require mappings across such spaces. CZs are postulated as having the ability to accomplish temporal binding across neural ensembles. Such bindings should be able to accomplish mappings of the sort needed to characterize mental space phenomena.

GESTALTS + MAPPINGS

Conceptual Metaphor:[8,5] This is a phenomenon that requires mappings across conceptual domains from one schema to another. It thus requires both gestalt-formation and mapping capacities. An example would be the Conduit Metaphor, through which communication is conceptualized as a form of sending. We have a type of sending schema in which a person places an object in a container, the container is sent to another person, who then takes the object out of the container. The Conduit Metaphor maps this schema into a communication schema via the following mapping: The Object corresponds to an Idea; the Container corresponds to a Linguistic Expression; Sending corresponds to Communication; Grasping the Object in the Container corresponds to Understanding the Idea. This conceptual metaphor gives rise to many linguistic expressions:

Did I get the idea *across to* him? Did I *put* the idea *into* the right words? Do you think the idea will *go over* his head? Your words are *hollow*. His words *carry* little meaning. You can't just *stuff* ideas *into* a sentence any old way. There isn't much *content* in this paragraph. The *channels* of communication *between* them are *blocked*.

Moreover, the logical structure of the concept of communication makes use of much of the logical structure of the concept of sending: the idea of an origin and destination, a path which may be clear or blocked, the idea of a vehicle of communication (language) that the ideas are in, and the idea that that the same content that is sent is received in successful communication. In short, much of our knowledge about the physical domain of SENDING maps onto knowledge of the abstract domain of COMMUNCATION. The conceptual system underlying English contains thousands of such mappings of schemas from one conceptual domain to another.

Metonymy[8]: Imagine one waitress saying to another, "The ham sandwich left without paying." Here "the ham sandwich" is being used to refer to the customer who ordered the ham sandwich.In the restaurant schema, there is a customer and the dish he orders.A metonymy is a mapping from one element of a schema (in this case,the dish) to another (in this case, the customer). Whereas metaphor involves two conceptual domains and a mapping across them, metonymy involves one conceptual domain and a mapping within it, from one element of a schema to another.

These are the basic conceptual entities that conceptual and linguistic systems are made of. Without the Convergence Zone Hypothesis, this list might seem like a random collection of phenomena that just happen to provide the basis for conceptual and linguistic structure. The CZH, however, makes sense of this list. Each of these phenomena is the kind of thing one would expect convergence zones to be able to characterize: bindings, gestalts, and mappings. The CZH explains why phenomena like this should exist and why our conceptual and linguistic systems should be made up of structures like these.

ACTIVATION PATTERN VARIATION AND RADIAL CATEGORIES

Each CZ is a neural ensemble. A conceptual entity 'stored' in each ensemble is a pattern of synaptic weights that can give rise to an activation pattern. Once patterns are stored in an ensemble, additional patterns are stored not independently of those already there, but as variations on the patterns already stored -- variations that make maximal use of the information already there. When one activation pattern is a minor variation

on another, that means that there is a large area where the activation patterns overlap and small areas where they do not.

When we translate this into the symbolic terms in which linguists describe their results, a central pattern corresponds to a central member of a **radial category,** which is a category with central and peripheral members. A variation on the central pattern of activation has two parts: an overlap and a nonoverlap with the central pattern. In symbolic terms, the overlap translates into properties shared with the central case (called "inherited" properties) and nonoverlap translates into properties not shared with the central case (called "overrides"). The CZH thus predicts that one should find radial categories characterized by inheritance-with-overrides in both conceptual systems and language. Radial categories do occur in both. [5]

Consider the concept of a mother. It involves four conceptual schemas: (1) A birth schema, where the person giving birth is the mother. (2) A nurturance schema, where the person who nurtures and raises the child is the mother. (3) A genetic schema, where the female who provides half of the child's genetic material is the mother. And (4) a marriage schema, where the wife of the father is the mother.

In the central case, these four schemas form a single whole. The prototypical mother meets all four conditions. But there are mothers who meet fewer than those four conditions: birth mothers, genetic mothers, step mothers, surrogate mothers, and adoptive mothers. Thus, the concept of a mother has a radial structure, with a central case and a number of peripheral cases that are similar to the central case in various respects. The peripheral cases each 'inherit' some portion of the meaning of the central case, and override some other portion.

Grammatical constructions also form radial categories. For example, the deictic there-construction described above has a number of variations. For example, there is `presentational' construction, as in ``There, wearing an outrageous costume, was standing a well-known public official.'' In this variation, the short phrase (``wearing an outrageous costume'') can come before the verb and the verb may have a full auxiliary (``was standing''); pairedwith this syntactic variation is a pragmatic condition: it must be used in announcements or narratives. Another variation is the `delivery'construction, as in ``The.e.e.e.ere's you're pizza!'' or ``He.e.e.e.eres Johnny!'', where the vowel of the deictic locative pronoun is lengethened (``He.e.e.e.ere'') and the construction is paired with a schema indicating delivery. Aside from the indicated variations, all the other properties of these variational constructions are shared with (that is, ``inherited from'') the central construction. Such cases of constructions having central cases and variations are commonplace.

Without the CZH, the existence of such radial categories in both conceptual systems and grammars might be seen as simply another random phenomenon. The CZH explains why such radial categories should exist. They arise because neural ensembles optimize to form activation patterns that have a variation structure of a form that corresponds to radial categories.

CO-NAMING AND TEMPORAL BINDING

According to the CZH, binding is achieved through timelocked activation.[10] That aspect of the CZH has a consequence for linguistic structure: It makes sense of the phenomenon of co-naming. There are two types of co-naming. The first is polysemy, where a given word has a number

of systematically related meanings. For example, the word ``grasp'' means not only to hold securely, but also means to understand. This is not an arbitrary fact, but is part of the general Conduit Metaphor through which we conceptualize communication. When there is a conceptual metaphorical mapping, it seems natural that words for source domain concepts also be used for target domain concepts. But why does it seem natural? That remains to be explained.

The second kind of co-naming is productive rather than conventional. For example, when a waitress refers to a customer as ``the ham sandwich'' it isn't because ``ham sandwich'' means customer. It is due to the fact that there is a conventional conceptual link (describable as a mapping) between dishes and customers in the restaurant schema. Whenever there is such a link, we can use the name for a dish to refer to a customer. Again, that seems natural, but why?

Co-naming also coccurs with mental space mapping. For example, take the case of a movie. Consider the sentence, ``In Basic Instinct, Sharon Stone seduces a detective.'' Now ``Sharon Stone'' is not the name of a character in Basic Instinct; it is the name of the actress in the real world who plays that character. But, because there is a mapping between actors in the world and characters in movies, it seems natural to use the name of the actor to refer to the character played by that actor. It is a fact that it happens and that it seems natural, but why?

The temporal binding property of convergence zones offers a possible explanation for the existence of such cases of co-naming.First, consider simple naming. The CZH describes naming in terms of the following CZ structure: phonological CZs (CZ-phon), conceptual CZs (CZ-con), and lexical CZ s (CZ-lex) that link them. In the CZH, the pairing of a phonological form with a

concept is described as follows: an activation pattern in CZ-lex controls the pairing of an activation pattern in CZ-phon with an activation pattern in CZ-con via temporal binding: the patterns in CZ-phon and CZ-con are timelocked; they fire at the same time. For example, the activation pattern for the phonological form ``grasp'' in CZ-phon would be timelocked to fire at the same time as the concept GRASP in CZ-con.

Suppose now that there is a mapping (either metaphoric, metonymic, or mental space) between two concepts (or schemas) within CZ-con. Such a mapping would, presumably, be decribed in the CZH as follows: CZ-con has two subensembles of neurons, CZ-source and CZ-goal, and there are other convergence zones (CZ-map) that characterize the mapping by temporal binding. For example, a concept like GRASP in CZ-source might be mapped onto a concept like UNDERSTAND in CZ-goal as follows: CZ-map contains an activation pattern that makes the activation pattern for GRASP in CZ-source fire in synch with the activation pattern for UNDERSTAND in CZ-goal.

Suppose we put together the effects of CZ-lex and CZ-map. CZ-lex guarantees that the pattern for the word ``grasp'' is timelocked with the pattern for the concept GRASP. CZ-map guarantees that the pattern for the concept GRASP is timelocked with the pattern for the concept UNDERSTAND. Co-naming will occur when the pattern for the phonological form ``grasp'' is timelocked not only with the concept GRASP but also with the concept UNDERSTAND. If the timelocking is transitive, then co-naming will occur.

Timelocking is not necessarily transitive. The firing pattern for the concept GRASP could be a complex wave form decomposable into two parts, one of which fired in synch with the word ``grasp'' and the other of which

fired in synch with the concept UNDERSTAND. Were this to happen, the word "grasp" would not name both GRASP and UNDERSTAND. That is possible, but it is not the least energy solution for this simple system. The least energy solution is the one in which timelocking is transitive, that is, in which the word "grasp", the concept GRASP and the concept UNDERSTAND are mutally timelocked, and when therefore "grasp" co-names both GRASP and UNDERSTAND. The least energy solution may not always occur, but it is natural. Thus, according to the CZH, co-naming is natural, but by no means necessary, when there is a mapping between two concepts. Therefore, the CZH provides a natural explanation for the phenomenon of co-naming.

CONCEPTUAL GROUNDING

Finally, the CZH carries with it an implicit hypothesis about the nature of human reason. According to the CZH, most if not all higher-level functions are carried out through the coordination of lower-level functions via convergence zone hierarchies. This seems to imply that abstract reason, certainly a 'higher-level function,' is actually a product of lower-level functions and the operation of convergence zones. This views fits closely with the view of abstract reason that has emerged in cognitive linguistics. There it has been discovered that most (if not all) abstract concepts are the products of conceptual metaphorical mappings from more concrete, typically physical, concepts.[5,8,9] The Conduit Metaphor, in which the abstract concept of COMMUNICATION is understood in terms of the physical concept of SENDING is a typical example. Our best guess at present is that abstract

inferences are sensorimotor inferences under metaphorical mappings. This is entirely consistent with the CZH.

SUMMARY

The Convergence Zone Hypothesis has explanatory force in two very different domains: In neuroscience, it explains why patients with brain damage in certain regions should have the detailed deficits that they have. In cognitive science, it explains why conceptual systems and languages should have the kinds of structures they have.

The CZH is just a hypothesis. The anatomical facts are that, in a given brain region, say a few square centimeters in area, many of the kinds of neural structures that the Damasios' call ``convergence zones' ' are physically present. What is hypothesized is that they function as timelocking control networks, rather than, say, as representational or other computational networks. The CZH is, at present, so hypothetical that we have no computational models showing that such a control structure is even computationally tractable (although the earliest work in the area is promising). Yet, to my knowledge, there is no other hypothesis at present that accounts for the same range of data, both in neuroscience and in cognitive science. That does not mean that there can be no such alternative hypothesis. But it does place a very tough constraint on any competing hypothesis -- to account for the same range of data from both lesion studies and the study of conceptual systems and grammar.

References

1. Damasio, A.R. , Tranel,D., and Damasio, H. *Annu. Rev. Neurosci.* (1990) 13:89-109.

2. Damasio, A.R. and Damasio, H. (1992) Sept. *Scient Am* 267, 88-95.

3. Shastri, L. and Ajjanagadde, V. (1992) From Simple Associations to Systematic Reasoning. Behavioral and Brain Sciences.

4. Rosch, E. (1977) Human Categorization, In N. Warren, (ed.) *Studies in Cross-Cultural Psychology.* (Academic,London).

5. Lakoff, George. (1987) *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind* . (University of Chicago Press, Chicago).

6. Damasio, A.R., Damasio, H., Tranel, D., & Brandt, J.P. In Cold Spring Harbor Symposia on Quantitative Biology (Cold Spring Harbor Laboratory Press), Vol. LV, pp. 1039-1047.

7. Fauconnier, G. (1985) *Mental Spaces.* (M.I.T. Press., Cambridge).

8. Lakoff, G. and Johnson, M. (1980) *Metaphors We Live By* (University of Chicago Press, Chicago).

9. Johnson, M. (1987) *The Body In The Mind.* (University of Chicago Press, Chicago).

10. Damasio, A.R. (1989) Neural Computat 1, 123-132.

Afterthought

Because this paper was getting a bit long for the National Academy Proceedings, I left out a section that I thought was too detailed for most neurobiologists, but which I think is worth including for linguists and cognitive scientists who will be getting copies of this paper via informal distribution mechanisms. Here it is:

Asymmetric Mappings

Most mappings, especially metaphoric and metonymic mappings are asymmetric. For example, we conceptualize UNDERSTANDING in terms of GRASPING, but not conversely. Correspondingly, the word ``grasp" is used to mean UNDERSTAND, but the word ``understand" is not used to mean GRASP, nor is is possible for it to come to mean GRASP.

If mappings are to be represented by temporal bindings, then those bindings would have to be asymmetric in order to characterize mappings that are asymmetric (which includes all metaphorical mappings). Now, it might appear, on the face of it, that temporal binding is inherently symmetric: `A and B are firing simultaneously' would seem to be an inherently symmetric relation. How can the asymmetry of binding be represented, and how can one explain which mappings are asymmetric?

Observe first that asymmetric timelocking is a possibility with the CZH. Consider again a mapping architecture, with CZ-map controlling the the timelocking of CZ-source and CZ-goal, and take as an example the UNDERSTANDING IS GRASPING metaphor. According to the CZH, the activations for UNDERSTANDING and GRASPING are independent of one

another when they are not timelocked, that is, when the activation pattern for the metaphor is not actually activated in the convergence zone, CZ-map. CZ-map can make them timelocked, but it doesn't always do so.

When timelocking occurs there are three possibilities, given that GRASPING and UNDERSTANDING are otherwise activated independently of one another. (1) Both change to a new timelocked activation. (2) The activation of GRASPING remains the same and the activation of UNDERSTANDING is timelocked to it. (3) The activation of UNDERSTANDING stays the same and the activation of GRASPING is timelocked to it. (1) is symmetric timelocking, but (2) and (3) are asymmetric: in both, one activation pattern remains fixed and the other adjusts to it.

The alternative that fits the co-naming data is (2). Consider how (2) works with respect to co-naming: As before, CZ-lex links CZ-phon to CZ-con, in such a way that ``grasp" names GRASP and ``understand" names UNDERSTAND. That is, there is an activation pattern in CZ-lex that timelocks ``grasp" to GRASP and another pattern that timelocks ``understand" to UNDERSTAND. Under alternative (2), the GRASP pattern remains unchanged and the UNDERSTAND pattern is timelocked to it. Thus, ``grasp" still names GRASP, but now it also names UNDERSTAND. But since the activation pattern of UNDERSTAND has changed, ``understand" no longer names UNDERSTAND, and hence it cannot name GRASP.

Now, why should alternative (2) occur? The reason has to do with grounding. The concept GRASP is grounded in the sensorimotor system, via a cascade of convergence zones down to the lowest levels of that system. Each link down the chain of convergence zones involves timelocking. The activation pattern for GRASP cannot change without a lot of other changes taking place. The activation pattern for GRASP can be thought of as being `anchored' by the

chain of convergence zones that grounds it. The least energy solution is for the activation pattern for GRASP to stay in place while the activation pattern for UNDERSTAND, which is not anchored by a sensorimotor grounding, assimilates to the pattern for GRASP. The generalization is that grounded concepts tend to be source domains for ungrounded concepts. This is exactly what we find in metaphor systems in general: abstract concepts (the ones without sensorimotor grounding) tend to be understood in terms of concrete concepts (the ones with sensorimotor grounding).

Thus, there is a fully natural explanation for the asymmetry of metaphoric mappings and the asymmtery of co-naming. The CZH also permits an explanation for another important aspect of metaphor phenomena -- the distinction between conventional metaphorical language and novel metaphorical language. To see the difference, compare two examples: (1) the use of ``grasp' to mean UNDERSTAND, which is a conventional part of the English lexicon, and (2) a novel expression such as ``I had it for a minute but it got away from me,'' said of an idea that someone is trying to communicate.

In the model proposed, the activation pattern for UNDERSTAND is shifted in phase to be timelocked with the GRASP pattern. This shift can be learned by CZ-lex, as an extension of the pattern for the naming of GRASP by ``grasp.'' In other words, CZ-lex, in order to conventionally name both GRASP and UNDERSTAND by the word ``grasp'' , would have to acquire an activation pattern -- call it the ``grasp''-GRASP-UNDERSTAND pattern -- which overlaps with the ``grasp''-GRASP pattern. Such a general mechanism would be needed to account for conventional metaphorical language.

But novel metaphorical language that made use of conventional conceptual metaphorical mappings -- such as case (2) -- would arise naturally whenever a target conceptual domain was timelocked to a source domain. Then

any source domain expression could be used to name the target domain expression -- but it would be a novel, not a conventional, use. of language.

Incidentally, this account also permits degrees of conventionality for lexical metaphors. Since an activation pattern in a convergence zone is physically instantiated by synaptic weights, those weights can be acquired gradually, increasing a bit at a time until full conventionality is reached.

In the area of metaphor, the CZH has remarkable descriptive and explanatory power. Within the CZH there is (a) a description of asymmetric mapping using timelocking; (b) an explanation for the asymmetry of metaphoric mappings in terms of the way grounding works in the CZH, and (c) an explanation for why names for source domain concepts in metaphors also name target domain concepts, but not conversely, and (d) a natural distinction between conventional and novel metaphorical language, (e) a natural way to represent the gradual acquisition of conventional lexical meaning.

I should end on a cautionary note. As mentioned above, there are no computational models of chains of convergence zones. We cannot possibly know at present whether convergence zones could, even in principle, do the kinds of things that I have suggested would follow from the CZH plus general properties of neural networks. Until the computational properties of chains of convergence zones are thoroughly investigated, all the above must remain theoretical speculation. However, the prospects for possible explanation of the properties of conceptual and linguistic systems are so interesting that I think that the enterprise of modeling convergence zones would be well worth the effort.