Compositionality and Dynamics in Neural Network Representations[1]

Tim van Gelder

1. Introduction

This paper considers a problem that arises at the intersection of three very broad ideas. They are (1) that higher cognition needs compositionally structured representations ; (2) that cognition is essentially a dynamic phenomenon ; and (3) that neural networks provide the best modeling environment for the study of cognition. Each of these ideas is quite appealing, though for independent reasons. If they are all true, then, together, they pose an inevitable problem : from the perspective of dynamics, can we understand how neural networks can handle compositionally structured representations ? That is, there must be, within neural network modeling, a thoroughly dynamical way of implementing compositionally structured representations. What will these representations look like ?

The claim that higher cognition demands compositionally structured representations is a key ingredient of mainstream orthodoxy in cognitive science. It dominates the practice of most traditional cognitive modeling, and has been vigorously defended by practitioners and philosophers alike. The basic idea can be crudely expressed by saying that mental representations must be like *sentences* ; that is, they must be complex entities that are systematically constructed out of tokens of a limited set of basic types. A variety of arguments have been put forward in favor of the view that higher cognition requires compositional representations, but the most plausible seems to be just that we have no good ideas about how higher cognitive processes such as reasoning, planning, and language comprehension and production might be conducted except on the basis of the manipulation of compositionally structured representations. Such processes appear to depend on large amounts of topical and background knowledge about the world, at least in any interesting applications. How can that knowledge be stored, retrieved and deployed, except in the form of sentence-like representations ?

[1] *A version of this paper was read at the Conference «Naturalistic Approaches to Representation and Meaning», Zentrum für interdisziplinäre Forschung (ZiF), Universität Bielefeld, 1993. Portions were presented at the conferences COMPCOG I (1991) and II (1992), Abbaye de Royaumont, France, and at the A. A. A. I. Spring Symposium Series, Stanford University, 1991. Section 2 contains joint work with Robert Port ; a version of these ideas is forthcoming in [van Gelder & Port, in press].*

This argument for the dependence of higher cognitive processes on compositional representations is often extended to the conclusion that cognitive systems must have a so-called "Classical" cognitive architecture (e.g., [Pylyshyn, 1984]]. But this extension can be resisted, as long as we can find some way to incorporate compositionally structured representations in *non-Classical* cognitive systems. This is why solving the problem of the intersection of compositionality, dynamics, and neural networks is an important exercise.

The second broad idea is that cognition is essentially a *dynamic* phenomenon. This has two sides. The first is that cognition is fundamentally a matter of cognitive processing, and that cognitive processes are natural processes occurring in real biological systems. As such, they take place in *real time*, where this means something more than simply that they happen "fast enough to keep up with what is going on in the environment". It means, also, that the cognitive processes themselves are changes taking place in time, which is a continuous quantity best modeled by means of the real numbers.

This rather basic point becomes interesting when we note that most traditional computational models of cognition standardly abstract away from the temporal nature of cognitive processes. The fine temporal structure of real cognitive processing is seen as a mere implementation detail ; computational models allegedly map the abstract information processing operations that the cognitive system must go through in performing a given task. Insofar as such models incorporate a notion of time at all, time is merely a matter of order or sequence of operations, and so is adequately modeled by means of the integers, which comprise the simplest and most well known ordering. Consequently, there is a whole dimension of modeling adequacy — tracking the fine temporal detail of real cognitive processes — that virtually all traditional computational models entirely ignore.

The inherently temporal nature of cognitive processing suggests that providing adequate scientific descriptions will require a mathematical and conceptual framework capable of dealing with continuous change in real time. The obvious candidate here is traditional dynamical modeling using differential equations, and its modern extension, dynamical systems theory. The second side to the idea that cognition is essentially a dynamic phenomenon, then, is that it is *dynamics* (broadly speaking) that provides the right mathematical and conceptual tools for understanding cognitive processes. Cognitive science has always been a diverse enterprise, and one of the most consistently active sub-branches, if not the most prominent, has been that of dynamical modeling. Pioneers such as [Ashby, 1952], [Grossberg, 1988] and [Thom, 1975] have been applying dynamics to the study of cognitive processes for decades. In recent years, following the

famous explosion of theory and applications of nonlinear dynamical systems theory in the 70s and 80s, rapidly increasing numbers of cognitive scientists have been applying dynamical techniques and concepts across a wide range of areas. Indeed, if connectionism was the most dramatic theoretical revolution of the 1980s, it appears that dynamics is the connectionism of the 1990s [Port & van Gelder, eds., forthcoming, 1995].

The third broad idea is that neural networks provide the most appropriate medium for the construction of models of cognitive processes. In the light of recent debates, this idea needs little elaboration (see, e.g. [Smolensky, 1988]), and so I will pass directly to the implications accepting these ideas.

I am not presupposing that these ideas are true ; merely that they are broad empirical hypotheses which, given our current state of knowledge, look attractive quite independently of one another. Now, one way of evaluating their acceptability is to test their compatibility with each other. If they turn out to be incompatible, then presumably one or more must be rejected. The problem, in short, is to reconcile compositionality, dynamics and neural networks.

Reconciling dynamics and neural networks is no great achievement. Neural networks standardly are dynamical systems governed by differential equations, and increasingly, dynamical systems theory is the mathematical framework of choice for describing their behavior. The naturalness of the fit between neural networks and dynamics has been exemplified in Grossberg's models of various aspects of cognition. Reconciling neural networks and compositionally structured mental representations is also not particularly problematic. Plenty of connectionist models of language processing have constructed complex representations of entities such as sentences by combining, by one means or another, patterns corresponding to the constituents of those entities. The infamous Fodor & Pylyshyn's criticism of connectionism [Fodor & Pylyshyn, 1988] was not that connectionists could *not* incorporate compositional representations in their models ; rather, it was that they *can*, and in order to be adequate to cognition they *must*, but in doing so their models become mere implementations of the Classical approach. However, subsequent work has demonstrated beyond any question that connectionist models can utilize compositionally structured representations *without* amounting to mere implementations (see, e.g. [van Gelder, 1990]). How this is possible turns on the analysis of the concept of compositionality, an issue addressed in more detail below.

The trouble is that there is very little overlap between that neural network research which is truly dynamical, and that research which handles compositional representations. Those neural network researchers who bring the concepts and tools of dynamics to bear on the study of cognition with any degree of sophistication are not focusing on those

problems of higher cognition which appear to demand sentence-like representations, and *vice versa*. The real difficulty, then, appears to be reconciling compositionality and dynamics within a neural network framework. In a way, this should not surprising. Dynamics offers a vast conceptual and mathematical armory, but concepts such as *representation* and *syntactic structure* are not part of it. What we are asking, in effect, is that some central concepts of one large research framework — that of classical computationalism — be somehow reconstructed within another very different framework.

In the final part of this paper I will briefly discuss one broad approach to handling this problem that has been under independent development by the French philosopher Jean Petitot and the American connectionist linguist Robert Port (and perhaps others). The compositional representations that are constructed within this approach, however, are radically different from the more typical examples found in orthodox cognitive science. Consequently, before describing this method, which can be termed "attractor chaining", I will explore the concept of compositionality itself, and highlight some of the very different ways in which compositionality can be implemented.

2. Kinds of Compositionality

Consider some of the obvious differences between a printed sentence and a spoken utterance of the very same sentence. Both contain the same words, but in the printed sentence they are static ink configurations, while in the utterance they are temporally extended sound patterns. In the printed case, words are combined by juxtaposition in space ; in the utterance, by juxtaposition in time. The printed words are combined in a very discrete way — there is, quite literally, space between each — and all occurrences of a given word are effectively identical, whatever words happen to surround them. In the utterance, by contrast, juxtaposition is not discrete and context free ; words flow into each other, and their physical shape is affected by their neighbors.

Printed and uttered tokens of the very same sentence, then, are physically *constructed* or *built up* in quite different ways. One way to put this is that the two tokens exhibit very different kinds of *compositionality,* despite being syntactically and semantically identical. (Note that this use of the term "compositionality" is not to be confused with another very common use, in which the term is used to refer to the kind of situation in which basic constituents make approximately the same semantic contribution in every context in which they appear.)

This point has two very important consequences. First, compositionality itself comes in at least two, and possibly many, different

kinds. Second, the *concept* of compositionality can be a subject of study, quite independently of syntax and semantics. Whereas syntax and semantics focuses on a language in the abstract, the study of compositionality focuses on particular, concrete *implementations*. How many kinds of compositionality are there ? What are the fundamental issues ? And what kinds of compositionality are typically found in mainstream artificial intelligence, in neural network research, and in natural cognitive systems ?

A representation is compositionally structured when, roughly, it is systematically constructed out of tokens of a limited set of basic compoundable units. Technically, it is more appropriate to say that a representation is compositional if it stands in certain *constituency relations* ; that is, there is :

(a) some finite set of *primitive* types, realizable by actual physical tokens ;

(b) a possibly unbounded number of compound types, realizable by actual physical tokens ; and,

(c) a set of abstract constituency relations defined over these primitive and compound types.

Any particular token of a representation is compositional just in case, by virtue of belonging to a certain type, it stands in appropriate constituency relations, i.e., can be said to have constituents. (For more detailed discussion of this approach to compositionality, see [van Gelder, 1990].)

The point of specifying the concept of compositionality in this abstract way is to distance us as much as possible from the standard examples of compositionality, paradigms of which are symbolic structures such as printed sentences and LISP expressions. The printed sentence and the spoken utterance of the same sentence are both compositional representations, since they are both tokens of the same type, and that type stands in abstract constituency relations ; however, they realize that compositionality in very different ways. To understand *syntax* is to understand the abstract hierarchical structure of constituency relations among types ; to understand *compositionality* is to understand the various different ways in which concrete *tokens* actually realize those types and their constituency relations.

There are at least six key issues in understanding compositionality. It is possible to think of these metaphorically as the basic *dimensions* of an abstract space of possible kinds of compositionality. These six issues are :

1. Static *vs.* Dynamic

2. Digital *vs.* Analog

3. Arbitrary *vs.* Non-Arbitrary

4. Mode of Combination : Concatenative *vs.* Non-Concatenative

5. Static *vs.* Temporal Combination

6. Syntactic Conformity.

Ideally, these dimensions carve up the space of possible kinds of compositionality in the most theoretically revealing way ; to extend the metaphor, they should be something like the "principal components" of that space.

The first three issues concern properties of the tokens of the primitive types, which I will call *symboloids,* in order to emphasize that they include, but are not restricted to, standard symbols such as LISP atoms or printed words. The second three concern the way in which these basic tokens are actually combined in order to form tokens of complex types.

2. 1. Properties of Symboloids

Consider an individual word token from a particular printed sentence of English. It is a paradigm example of *symbol* (i.e., a particular kind of symboloid). Some of its obvious properties include the fact that it sits still on the page for a long time ; that it is straightforward to recognize the symbol *type* to which it belongs ; and that it has *meaning* or semantic significance, yet its physical structure bears no interesting relationship to that meaning. These relatively obvious properties correspond to the first three dimensions of compositionality.

It is worth pausing to mention that it can often be quite difficult to pin down the primitive types of a given compositional scheme of representation. For example, linguists, AI theorists and high school teachers all usually assume that individual words are the primitives of English. This is satisfactory as a first approximation. But Bolinger among others has frequently pointed out how difficult it is to nail down the actual list of the particular items that have just the right properties to count as the primitives [Bolinger, 1975]. Natural language is not just a set of morphemes plus rules of syntax. Speakers use and *know* linguistic fragments that come in many sizes : from submorphemic ideophones to words, idioms, and Bolinger's "collocations", as well as *clichés* and even entire sentences and paragraphs of boilerplate (as in genres like wills and academic recommendation letters). The nature of many of these mysteriously constrained yet flexible lexicalized *units* lies well beyond the grasp of current representational schemes employed in linguistic theory.

Nevertheless, in what follows I will be proceeding on the assumption that the basic types have been already adequately identified for any compositional scheme under consideration.

(a) Static *vs.* Dynamic

Any representation, and any symboloid, is a concrete physical item (an ink mark, a sound wave, a collection of electronic bits, a pattern of neural firing, etc.). It counts as a token of a particular type in virtue of some kind of distinctive physical configuration or *shape*. We can think of this distinctive configuration as a matter of change along key physical aspects of the representation. Thus, a printed token of the word *cat* exhibits change in the way ink is placed on the page, change that would be registered by a scanner as it moved along the page. It is precisely the details of this spatial change which makes it different than a token of the word *dog*. Notice that, by contrast, an instance of *cat* typically exhibits no significant change in physical configuration from one moment of time to the next. This is what I mean by saying that it "sits still", or is a static token.

Compare this with an uttered token of the same word. In this case, it is variation in frequency and amplitude of a sound wave over time which determines its type-identity ; change over space is irrelevant. This is an example of a dynamic primitive. The key feature of a dynamic symboloid or representation is that change over time is essential to type-identity. It follows that, in order to determine to which type the entity belongs, you have to wait long enough for its distinctive change to unfold. In the case of a static symboloid or representation, by contrast, you have (in principle, at least) enough information at any instant to determine type-identity.

Printed English words, then, are static symboloids, while spoken words are dynamic symboloids. One way to summarize the difference is to say that dynamic symboloids, but not static symboloids, happen in time.

(b) Digital *vs.* Analog

A key feature of most standard compositional schemes of representation is that they are digital. By this I mean that it is possible to produce symboloids in the scheme, and determine the type identity of any given symboloid with complete and unambiguous success. Another way to put this is that the symboloids are such that the reading and writing processes are positive [Haugeland, 1985].

A classic example of a process that can succeed positively is scoring in basketball ; the ball either goes through the hoop or it doesn't. There has never been a semi-basket in the history of basketball. Similarly, in standard symbol systems, such as LISP machines, the most basic production, identification and transformation processes can be carried out with unquestionable success. Barring malfunction, the system can always

tell whether the symbol in its buffer is *foo* or not ; there is simply no question of its being *somewhat foo*. Obviously, printed English is also a digital scheme, given printing presses or laserwriters to produce the words and normal human readers to type-identify them.

An analog scheme is one that is not digital ; analog symboloids are ones that cannot be produced and identified positively. This is not to say that they cannot be successfully produced and identified. Consider trying to determine who won a javelin throw by visual inspection. Most of the time, the test will indicate a winner, but sometimes the difference between the throws will be so small that it will be difficult to choose. Unlike basketball scoring, javelin competition is analog. Similarly, spoken English is analog, since, although it is usually possible to produce an identifiable token of any given word, success is always *more or less* rather than perfect. It makes sense, for example, to say that a given uttered sound was the word hardest because it sounded *rather more* like hardest than hottest.

Analog compositional schemes are, increasingly, cropping up in connectionist work. Consider, for example, Pollack's RAAM architecture [Pollack, 1990]. If one trains such a network to represent many sequences, the representations of various stack states become so closely packed in the activation space of the hidden units that the network cannot positively distinguish one representation from another very close to it in hidden unit activation space. "Attractor chaining" style compositional representations, to be discussed below, are schemes in which the symboloids themselves are analog rather than digital.

Intuitively, whether a set of entities is digital or analog in this sense has much to do with whether or not they are *discrete* rather than *continuous*. Thus it seems natural to say that printed words are discretely different, while the range of forms of spoken words is more continuous. But what does this actually amount to ? Roughly speaking, in this context, a set of entities is continuous if, *between* any two entities in the set, there is another entity which also belongs to the set. A discrete set is one such that there are large *gaps* between entities. Put differently, a continuous set is densely packed into the space of possible entities of that kind, while a discrete set is sparsely packed into it. However, it turns out that it is difficult to give any more precise formulation of this intuitive idea which really does effectively distinguish those schemes of compositional representation which intuition counts as discrete and those which it regards as more continuous. For this reason, together with the fact that digital *vs.* analog is clearly more significant from the point of view of the functional properties of a scheme of representation, I choose not to emphasize discreteness *vs.* continuity as a dimension of symboloids.

(c) Arbitrariness

I have noted already that individual symboloids belong to the type that they do in virtue of their distinctive physical configuration or makeup. The next dimension is a matter of how symboloids relate to each other. Are symboloids related in their physical configuration, or are they completely arbitrary relative to each other ? A particularly useful way to ask this question is : does fixing the physical configuration of one kind of symboloid in the scheme in any way constrain the configuration of any others ?

Most words of English, written or spoken, are basically arbitrary in this sense. Thus, a printed token of the word *cat* bears no interesting relationship, in its physical configuration, to a printed token of the word *dog*, over and above the fact that they are both ink marks constituted by smaller marks corresponding to letters. In particular, fixing the physical configuration of one word leaves the physical configuration of the other almost completely unconstrained (which is just to say that we could have used a quite different word in its place ; the choice is completely arbitrary). Likewise, the sound [k^hæt] bears no interesting physical relationship to the sound [dɔg].

The issue of the arbitrariness or non-arbitrariness of symboloids gets particularly interesting when a further complicating factor is introduced : the meaning of the individual symboloids. Symboloids standardly have at least two fundamentally different kinds of properties. On one hand, there are their physical properties. On the other, there is the meaning or semantic significance which each symboloid has and which contributes to the meaning of the compound representation that is formed out of it. Consequently, we can ask whether there is any interesting relationship between these two kinds of properties. Does the physical makeup of a given symboloid in any way reflect its semantic properties ? If it does, that symboloid is not arbitrary with respect to its meaning ; and since meaning is constraining physical configuration, different symboloids with different meanings must exhibit corresponding differences in their physical configuration, and hence the symboloids are not arbitrary with respect to each other.

Clearly, virtually all written or spoken words of English are semantically arbitrary ; a printed token of *cat* bears no interesting relationship to feline mammals over and above the semantic one. The same is true of symbols in any standard computational system. This is one deep source of the anxiety that many people feel about so-called "grounding" in computational systems. If the basic symbols are essentially semantically arbitrary, how can they have any *real* or *intrinsic* meaning ? Surely, it is often thought, symboloids in real cognitive systems must somehow reflect their meanings in their formal configuration more directly.

Some connectionist compositional schemes do utilize semantically non-arbitrary symboloids. For example, in McClelland and Kawamoto's model of case role assignment [McClelland & Kawamoto, 1986], input sentences are built out of vectors corresponding to distinct words, and those vectors themselves are essentially just lists indicating the presence or absence of microfeatures in the designated objects. Therefore, since the actual vector that is used to represent a given object (e.g., a hammer] is constrained by the features of hammers, these vector symboloids are semantically non-arbitrary. (Of course, the vector-elements corresponding to microfeatures are semantically arbitrary. This just shows that it is possible to construct non-arbitrary representations out of arbitrary components — which is exactly what happens in natural and formal languages when non-arbitrary sentences are constructed out of arbitrary word components.)

2. 2. Properties of the Manner of Combination

Compositional representations are obtained by combining symboloids to obtain compound wholes in accordance with the abstract constituency relations. It is natural and useful (though sometimes a little misleading) to think of this in terms of a mechanical process, in which actual individual symboloids are put in one end and out the other comes a compound representation. We can then ask : how does this process work ? What does it actually *do* in combining symboloids to obtain compound representations ?

(d) Mode of Combination

An obvious feature of printed sentences of English is that they are made up out of individual words, and that each word appears in the sentence exactly as it would if it were alone on the page. Words are taken *off the shelf*, as it were, and appear totally unaltered in the resulting sentence. A less obvious feature of *spoken* sentences of English is that this is not, in general, true : when a word appears in a sentence, it typically ends up with a somewhat different shape, depending on the words it is surrounded by, than it has when spoken alone. For example, observe the subtle changes in the pronunciation of the [t] — and hence of the whole word "cut" — in *cut Paul, cut some and cut out.*

Further, somewhat surprisingly, it is possible to systematically generate compound representations out of symboloids in a way that not only changes, but apparently completely *destroys* those symboloids.

An appropriate term for the kind of combination characteristic of printed sentences is *concatenation.* Roughly, a mode of combination is concatenative just in case it preserves the constituent symboloids in the resulting compound representation ; if, in other words, one can literally point to the constituent symboloids in the compound whole. Since this is true for both printed and uttered sentences of English, we need to go on to distinguish two different kinds of concatenation :

(a) *Pure* concatenation, in which the process of combination leaves symboloids utterly unchanged. All standard computer languages and representational schemes in *classical* cognitive science utilize pure concatenation.

(b) *Context-sensitive* concatenation, in which each constituent symboloid is recognizably present in the resulting compound, but its *shape* has been altered by the very process of combination, typically in a way that systematically reflects its context.

A *non-concatenative* compositional scheme is one in which symboloid constituents do *not* literally appear in the compound representations, even though they may have been input into the combination process. Pollack's RAAM architecture, already mentioned, is one relatively well-known kind of connectionist model which forms compound representations by non-concatenative combination.

In models with this architecture, basic symboloids take the form of vectors on the input layer, and they are systematically combined by the network to form compound *(stack)* representations in the hidden layer. The symboloid constituents are not literally present in the compound representation ; if you examine the activation vector which is the compound representation, you will not find any instance of the vector which was the input symboloid. Nevertheless, the compound representation systematically reflects, in its formal structure, its actual constituency relations. This latter point can be shown in two ways. First, further processing in the RAAM network can actually recreate the symboloid constituents. Second, the compound representations can be used in further processing which is systematically appropriate to its constituent structure (e.g. [Chrisman, 1991]).

(e) Static *vs.* Temporal Combination

In printed sentences of English, combination is by spatial juxtaposition, and all constituent symboloids are present simultaneously for the entire period of the existence of the whole printed sentence. In *uttered* sentences, by contrast, symboloids are combined temporally ; one follows another, and none are present at the same time. The first kind of

combination, in which all constituent symboloids are simultaneously present, is called *static* combination, while combination by temporal succession is *dynamic*.

Note that dynamic symboloids and dynamic combination are two quite independent properties of compositional representations. The first concerns the nature of the symboloids themselves, the second concerns the way symboloids are combined. Though it is quite natural for dynamic symboloids to be combined dynamically (as in speech) and for static symboloids to be combined statically (as in print), other combinations are possible ; thus, dynamic symboloids can be combined statically (a musical *chord),* and *vice versa* (as when printed words appear sequentially on a screen ; it is their temporal succession which forms a sentence).

(f) Syntactic Conformity

When basic tokens are combined to form compound wholes, to what extent are the syntactic rules of the scheme observed ? A standard assumption is that representations in a compositional scheme are, by definition, grammatically well-formed, i.e., constructed in strict accordance with certain syntactic rules. Any combination of symboloids which violates the syntactic rules is junk or *symbol salad.* However, it is often more useful to regard certain combinations as representations belonging to a given compositional scheme even if they violate some syntactical rules. Many if not most utterances of everyday spoken English are not grammatically well-formed (in the sense that their utterers would wince and then edit them if given an opportunity to examine their own speech closely), yet they are still adequate for representational and communicative functions. Their utility does, however, seem to presuppose at least some general or loose conformity to the syntax of English. The claim is that there is a rough *spectrum* of cases between those symboloidal schemes in which *strict* conformity is observed (e.g., programming languages, printed English), from schemes in which only *loose* conformity is sufficient (e.g., spoken English). It is possible that, insofar as there are compositional representations underlying general cognitive performance, these representations will, like spoken language, exhibit only loose conformity to any relevant syntactic rules.

The theoretical utility of the six dimensions outlined so far is illustrated by the following table which compares five different kinds of compositional representation : sentences of spoken and printed natural language, the representations found in a paradigm classical AI program, Hearsay 2 [Erman *et al.*, 1980] , and two kinds of connectionist schemes, RAAM networks and Port's dynamic auditory recognition model (to be described on the opposite page).

	Spoken Natural Language	Printed Natural Language	Hearsay 2	RAAM [Pollack, 1990]	Dynamic recognition [Port, 1990]
Static *vs.* Dynamic Symboloids	Dynamic	Static	Static	Static	Dynamic
Analog *vs.* Digital	Analog	Digital	Digital	Analog	Analog
Arbitrary	Arbitrary	Arbitrary	Arbitrary	Non-arbitrary	Non-arbitrary
Mode of Combination	Context-Sensitive	Pure	Pure	Non-concatenative	Context-sensitive
Static *vs.* Temporal Combination	Temporal	Static	Static	Static	Temporal
Syntactic Conformity	Loose	Strict	Strict	Strict	Loose

Table 1. A comparison of five different schemes of compositional representation along six dimensions of compositionality.

Note that each dimension separates these five kinds roughly evenly into different groups. Note also that printed English and the representations in the classical AI system, Hearsay 2, share exactly the same properties, and that spoken natural language and the representations in Port's dynamic recognition model (columns 1 and 5) are also highly similar. This suggests that these two clusters of implementations of compositionality are in fact natural kinds. When compositional schemes of representation are developed, it is natural on one hand to have static, digital and arbitrary symboloids combined statically by pure concatenation in strict syntactic conformity ; on the other hand, it is natural to have dynamic, analog symboloids temporally combined by context-sensitive concatenation and in loose syntactic conformity. I will describe the first kind of compositionality, characteristic of printed natural language, as *symbolic*, and the second kind, characteristic of spoken language, as *dynamic*.

As mentioned above, these six properties of implementations of compositionality can be thought of as dimensions of a large space of possible kind of compositionality. While research in cognitive science has sampled representational formats from various regions of this space, clearly it is generically symbolic implementations of compositionality that have been dominant. In the third section of this paper I will be suggesting

that compositionality, dynamics, and neural networks might best be reconciled by means of *dynamic* compositionality.

3. Dynamic Compositionality in Neural Networks

Paul Smolensky once characterized the problem of connectionist representation as that of finding an appropriate mapping from the set of entities that one wishes to represent into a vector space [Smolensky, 1990, p. 168]. If the entities to be represented are symbolic structures, then the problem is to find a mapping from the set of those symbolic structures into the vector space. The vectors that are the output of this mapping can then be seen as compositionally structured representations themselves.

Why is this mapping into a *vector* space ? Well, obviously, because these vectors correspond to the pattern of activity over the neural units of some connectionist network at a given time. Note that a pattern of neural activity, in this sense, is a *static* representation. Change over time is not essential to the type-identity of this representation. Only the particular distribution of activity values over the units *at a time* determine what representation you have. The changes that occur *over* time as the network states evolve are *processing* of representations ; they are transitions from one representation to the next.

If we consider neural networks as dynamical systems, then a pattern of activity across the network is the *state* of the system at a given point in time. The notion of the *state* of a system is the most elementary concept of dynamical systems theory. Indeed, many, perhaps most connectionists who have conceptualized the problem of representation as that of finding a mapping simply into *states* of the system have then gone on to study those representations and their transformations in largely non-dynamical way, paying only lip-service to the idea that their networks are dynamical systems.

From the dynamical systems perspective, the more general problem of neural network representation is to find an appropriate mapping from the entities to be represented into some set of dynamical objects existing in a neural network system. At the outset, these objects might be anything specified by means of the conceptual repertoire of dynamical systems theory. In practice, proposals have dealt only with system *states*, *trajectories*, and *attractors*.

A well-known alternative to taking network representations to be simply system *states* is to take them to be fixed point attractors. Hopfield networks are a standard example ; the local *minima* of the energy function of such networks are fixed point attractors, and the network comes to represent a given input by taking that input as its initial conditions and settling into the attractor into whose basin of attraction those initial

conditions fall [Hopfield, 1982]. This approach may be adequate for certain kinds of problems, but there is a very serious difficulty confronting anyone wanting to use this approach to develop a scheme of *compositional* representations. The problem is the essentially *productive* nature of compositional schemes. The point of having a compositional scheme is to provide a *potentially unbounded* set of representations that nevertheless bear systematic structural similarities to one another. If representations are to be fixed point attractors of a network dynamical system, then we need a network whose dynamics is sufficiently complex that it already has an arbitrarily large number of distinct fixed point attractors, one for each compound representation. As far as I know, there are currently no practicable proposals for such a system.

The deep problem here, in short, is to find a way to incorporate the *productive* nature of compositional schemes of representation into a genuinely dynamical network framework.

A radical approach is to reject the standard connectionist paradigm in which representations are conceived as points in network activation space, and to think of them instead as *trajectories* : that is, as essentially temporally extended, dynamic entities. Further, one should not suppose that the network dynamics has every possible compound representation somehow already pre-configured ; rather, the network dynamics must provide the *resources* by means of which arbitrary compound representations can be constructed.

One method of doing this is what Jean Petitot calls "attractor syntax" [Petitot, forthcoming, 1995] and is also known as "attractor chaining". Consider a dynamical system with certain state variables and parameters. Change in system state is governed by its evolution equations which depend on system parameters. Another way to put this is that the *dynamics* of the system is fixed by the current specification of the system parameters. We can thus think of the parameters as *controlling* the behavior of the system. For any given set of parameter specifications the system possesses a certain landscape of attractors. As control parameter specifications vary, this landscape changes also. Significant qualitative changes in the attractor landscape are known as bifurcations. The essence of attractor chaining is to take the representation of a complex structure to be the trajectory that results as the state of the system evolves under the influence of a changing attractor landscape. Compositional structure is imposed on the trajectory by bifurcations that occur as control parameters vary.

Currently the most mathematically sophisticated treatment of the attractor chaining approach is that under development by Jean Petitot. A dramatic change in system state as the system undergoes a bifurcation is known as a *catastrophe*. René Thom developed the mathematical theory of catastrophes and suggested the application of this theory in many domains including linguistics [Thom, 1975, 1983]. Petitot is engaged in

combining the mathematics of catastrophe theory with Langacker's *cognitive grammar* [Langacker, 1987] in a neural networks framework.

A more standard connectionist approach is work by Bob Port and his students on dynamic recognition for auditory sequences ([Anderson & Port, 1990], [Port & Cummins, forthcoming, 1995], [Port, 1990]). In this work, a simple recurrent network is trained to recognize auditory sequences. The network is a dynamical system which, at any given time, has a single global point attractor whose location is determined by the current setting of the control parameters. The system state is always heading in the direction of this point attractor ; if there is no change in the parameters, it will settle into that point. Things get interesting when the location of the current global attractor is changed by variation in the control parameters. The state of the system is pulled first in one direction, then another. The resulting *trajectory* reflects in its temporal structure the sequence of fixed point locations. Distinctive trajectory shapes correspond to each such sequence (see **Figure 1**, below).

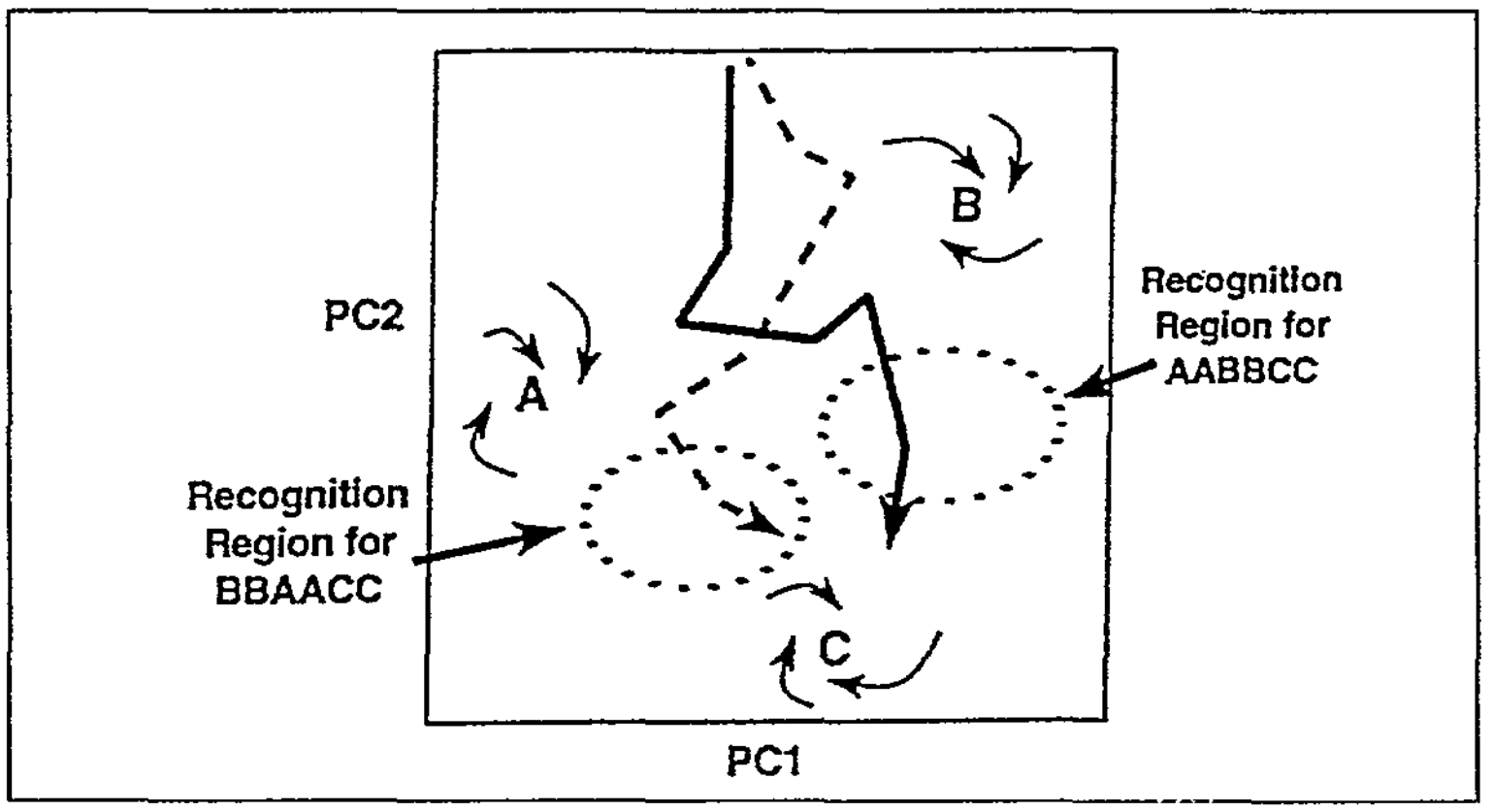

Figure 1

Figure 1 caption
A schematic illustration of a trajectory formed under the influence of a changing dynamical landscape. The points A, B and C represent the location of a single global attractor under corresponding parameter settings. The system state drifts in the direction of the global attractor as long as it is active. The temporally extended trajectory itself is the compositional representation. The axes are the first and second principal components of the system state-space. From [van Gelder & Port, in press]. For more detailed explanation, see [Port & Cummins, forthcoming, 1995].

How are these trajectory-forming parameter changes actually implemented ? The Port recurrent network model has an input layer (see **Figure 2**, on the opposite page). A pattern of activation across this layer corresponds to the acoustic spectrum as the network is exposed to another element in the auditory sequence. In gaining a truly dynamical understanding of how these networks work, however, it is crucial *not* to think of the relationship between the input layer and the recurrent network in traditional connectionist terms. It is a mistake to see the input pattern as transformed into another pattern in the recurrent network. Rather, the activation levels of the input layer units are parameters which subtly influence the *dynamics* of the recurrent network considered as a stand-alone dynamical system. The input pattern does not become another

pattern across the network units ; rather, it shapes the way the state of the recurrent network evolves by molding the dynamical landscape.

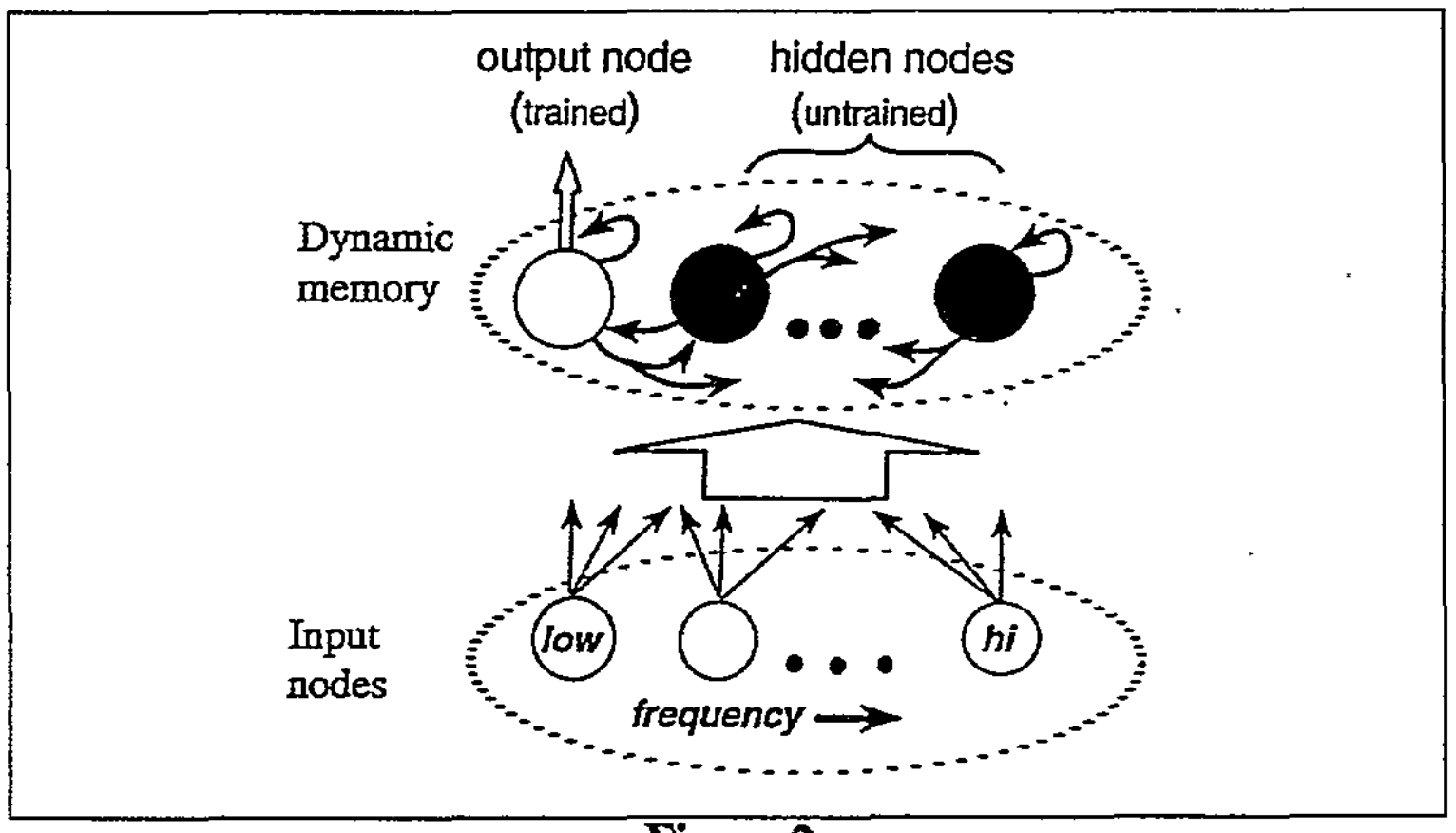

Figure 2

Figure 2 caption
The architecture of the Port dynamic auditory recognition model. Each tone in an auditory sequence is input as a pattern of activation across the input nodes. Activation levels at this layer function as parameter settings controlling the dynamics of the main system, a fully recurrent network. As a sequence of tones is input, the attractor landscape in the main system changes; the activation trajectory then reflects the structure of this sequence ; from [van Gelder & Port, in press]. For more detailed explanation, see [Port & Cummins, forthcoming, 1995].

The Port work has focused on perception of simple auditory sequences. It is not yet anything like a practical general solution to the problem of incorporating compositional representations in network dynamical systems. It is however a modest implementation of the attractor chaining approach to constructing complex representations. It therefore provides a useful prototype for the construction of general schemes of dynamical compositional representation. In this prototype, representations are trajectories systematically constructed out of basic compoundable units. Those basic units are nothing like the static, context-free symbols of standard implementations of compositional representation. Rather, they are themselves trajectory segments — stages of processing in which the system state evolves in the direction of a particular global attractor. The compositional representation is the context-sensitive, temporal concatenation of these trajectory segments.

In general, the kind of compositional representation found here is much more similar to spoken natural language than it is to printed natural language or to *LISP* structures. In order to see these entities as compositional representations — and, more generally, in order to understand how dynamics, compositionality and neural networks might be combined — we must have a suitably enriched conception of compositionality and how it can be implemented.

Research School of Social Sciences
Australian National University (Canberra ACT 0200 Australia)
tvg@coombs.anu.edu.au

References

ANDERSON (S.) & PORT (R.)
1990, "A Network Model of Auditory Pattern Recognition", *Technical Report* (Indiana University), n° 11.

ASHBY (R.)
1952, *Design for a Brain*, London, Chapman and Hall.

BOLINGER (D.)
1975, *Aspects of Language*, New York, Harcourt Brace Jovanovich Inc.

CHRISMAN (L.)
1991, "Learning Recursive Distributed Representations for Holistic Computation", *Connection Science*, n° 3 (n° 4), p. 345-366.

[Erman *et al.*]
ERMAN (L. D.) & HAYES-ROTH (F.) & LESSER (V. R.) & REDDY (D. R.)
1980, "The HEARSAY-II Speech Understanding System : Integrating Knowledge to Resolve Uncertainty", *Computing Surveys*, n° 12, p. 213-253.

FODOR (J. A.) & PYLYSHYN (Z. W.)
1988, "Connectionism and Cognitive Architecture : A Critical Analysis", *Cognition*, n° 28, 1-2, p. 3-71.

GROSSBERG (S.)
1988, *Neural Networks and Natural Intelligence*, Cambridge (MA), MIT Press.

HAUGELAND (J.)
1985, *Artificial Intelligence : The Very Idea*, Cambridge (MA), MIT Press.

HOPFIELD (J.)
1982, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities", *Proceedings of the National Academy of Sciences* (USA), n° 79, p. 2554-2558.

LANGACKER (R. W.)
1987, *Foundations of Cognitive Grammar*, vol. 1, Stanford University Press.

McCLELLAND (J. L.) & KAWAMOTO (A. H.)
1986, "Mechanisms of Sentence Processing : Assigning Roles to Constituents of Sentences", *in Parallel Distributed Processing : The Microstructure of Cognition*, J. L. McClelland and D. E. Rumelhart eds., Cambridge (MA), MIT Press.

PETITOT (J.)
forthcoming, 1995, "Morphodynamics and Attractor Syntax", *in* PORT (R.) & van GELDER (T.), eds.

POLLACK (J. B.)
1990, "Recursive Distributed Representations", *Artificial Intelligence*, n° 46, 1-2, p. 77-105.

PORT (R. F.)
1990, "Representation and Recognition of Temporal Patterns", *Connection Science*, n° 2, p. 151-176.

PORT (R.) & CUMMINS (F.)
forthcoming, 1995, "Modeling Auditory Recognition Using Attractor Dynamics", *in* PORT (R.) & van GELDER (T.), eds.

PORT (R.) & van GELDER (T.), eds.
forthcoming, 1995, *Mind as Motion : Explorations in the Dynamics of Cognition*, Cambridge (MA), MIT Press.

PYLYSHYN (Z. W.)
1984, *Computation and Cognition : Toward a Foundation for Cognitive Science*, Cambridge (MA), MIT Press.

SMOLENSKY (P.)
1988, "On the Proper Treatment of Connectionism", *The Behavioral and Brain Sciences*, n° 11, p. 1-74.

1990, "Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems", *Artificial Intelligence*, n° 46, p. 159-216.

THOM (R.)
1975, *Structural Stability and Morphogenesis* (Fowler, D.H., Trans.), Reading (MA), W. A. Benjamin Inc.

1983, *Mathematical Models of Morphogenesis*, Chichester, Ellis Horwood.

Van GELDER (T.)
1990, "Compositionality : A Connectionist Variation on a Classical Theme", *Cognitive Science*, n° 14, p. 355-384.

Van GELDER (T.) & PORT (R.)
in press, "Beyond Symbolic : Prolegomena to a Kama-Sutra of Compositionality", *in Symbol Processing and Connectionist Network Models in Artificial Intelligence and Cognitive Modelling : Steps Toward Principled Integration*, V. Honavar and L. Uhr eds., San Diego, Academic Press.