From continuous dynamics to symbols[1]

Herbert Jaeger
GMD, St. Augustin
herbert.jaeger@gmd.de

February 28, 1997

[1]to appear in the proceedings of the 1rst Joint Conference on Complex Systems in Psychology, "Dynamics, Synergetics, Autonomous Agents", Gstaad, Switzerland, March 1997

Abstract

This article deals with mathematical models of discrete, identifiable, "symbolic" events in neural and cognitive dynamics. These *dynamical symbols* are the supposed correlates of identifiable motor action patterns, from phoneme utterances to restaurant visits. In the first main part of the article, models of dynamical symbols offered by dynamical systems theory are reviewed: attractors, bifurcations, spatial segregation and boundary formation, and several others. In the second main part, *transient attractors* (TA's) are offered as yet another mathematical model of dynamical symbols. TAs share with ordinary attractors a basic property, namely, local phase space contraction. However, a TA can disappear almost as soon as it is created, which could (not very rigorously) be interpreted as a bifurcation induced by quickly changing control parameters. Such "fast bifurcation sequences" standardly occur in neural and cognitive dynamics.

1 Introduction

This paper is about symbols, viewed as identifiable events in neural dynamical systems.

The paper is not about symbols in general. That would be impossible. The empirical phenomenology of symbols is too rich, and the term "symbol" is used with too many intentions, to allow a comprehensive treatment. Compare, e.g., the multiple roles of symbols (i) as mathematical objects, amenable to a set-theoretic reconstruction; (ii) as signs or signals, which induce physical or mental reactions in humans (in semiotics and certain schools in linguistic semantics [17] [40]); (iii) as aesthetical objects in graphical arts; (iv) as physical, identifiable states in computer circuitry, which can be manipulated algorithmically.

The latter view on symbols has had a constitutive influence on artificial intelligence and cognitive science. In one of its strong versions, it has become known as the "physical symbol systems hypothesis" [44]. This hypothesis has been fervently criticised by some philosophers and psychologists, who found that experiential aspects of a symbol's "meaning" had been lost. The ensuing debate of the "symbol grounding" [27] problem has grown into an entangled mesh of claims and counterarguments [50] [2] [13] [6].

As a side line, there flamed a debate on whether connectionist models can sustain symbolic reasoning [21] [53] [12]. The attacks of "classical" symbolicists tickled connectionists so sorely that within a short period they came up with dozens of connectionist models for variable binding, the buildup of representational hierarchies, and other "symbolic" mechanisms which had been claimed inaccessible to connectionist modeling.

Many of these connectionist architectures relied on dynamical phenomena in recurrent networks [19] [41]. These developments helped an increasing number of researchers in artificial intelligence and cognitive science to open up for ideas from biocybernetics, neuroscience, and artificial neural network research. It now becomes apparent that neural dynamics can be quite directly related to high-level properties of cognitive processes. A much-cited example for the insights afforded by a neural dynamics for cognitive-level processes are chaotic neural attractors in classification of sensoric stimuli and concept representation [60] [5]. A wealth of other neurodynamical phenomena relevant for cognition is documented, e.g., in the handbooks edited by Gazzaniga [23] and Arbib [3], in the annual Computation and Neural Systems proceedings [8], or in the Behavioral and Brain Sciences journal.

A related, recent trend in cognitive science and psychology is to view cognitive systems as dynamical systems, without necessarily dealing with the underlying brain processes [51] [57] [35] [56] [34]. I need not say more about this to the participants of the Gstaad workshop. In this article, I will frequently use the term "neural/cognitive dynamics" when referring to matters relevant on both levels of description.

All of these debates and strands of research form the background for the present article. I will investigate the topic of symbols as discrete, identifiable phenomena in neural/cognitive dynamics. I will pursue this investigation from a purely dynamical systems point of view, ignoring most of the deeper epistemological questions. In particular, I will not touch the question of a symbol's meaning.

The article has to main sections. In section 2, I motivate why it is natural

to assume that in neural/cognitive processes there emerge discrete, identifiable phenomena, which I will call "dynamical symbols". I will then review several candidate mechanisms offered by dynamical systems theory which mathematically describe the nature and the emergence of dynamical symbols: attractors, bifurcations, spatial segregation and boundary formation, and others.

In the second section, I describe a kind of discrete, identifiable phenomenon in non-autonomous dynamical systems which can amply be termed "transient attractors" (TA). TAs share one crucial property with ordinary attractors, namely, local phase space contraction. However, a TA can disappear almost as soon as it is entered, which could (not very rigorously) be interpreted as a bifurcation induced by quickly changing control parameters. Such "fast bifurcation sequences" occur standardly in neural and cognitive dynamics.

2 Dynamical symbols

In this section, I shall first clarify the notion of "dynamical symbols". Then I shall review some of the mechanisms offered (or not yet offered) by dynamical systems theory for modeling dynamical symbols.

Humans behave, and their behavior can be observed by other humans. Very generally speaking, the behavior exhibited by a human is a continuous process in many variables. Sometimes in this "stream of behavior" there appear phenomena which (i) can be singled out by observers, and (ii) which can be more or less reliably classified as an instance of a particular kind of event. Examples of such discrete, identifiable events are

1. Having a meal in a restaurant.
2. Blinking one's eyes.
3. Saying "sun".
4. Producing the sound [s].

These events differ from each other in many ways. They have different temporal extensions. Some of them are sub-events of others. Some are more variable than others (there are many different ways of how the restaurant visit "script" [49] can unfold, while an eye blinking is stereotyped). Some proceed in silence, others are accompanied by oral utterances, and still others *are* oral utterances. And so on.

Despite this diversity, all of these events can be isolated and classified by human observers. Isolatability and classifiability is certainly a matter of degree – a drunken person's utterances can be slurred to the point of becoming unintelligible. For the present purposes, however, the fringe fuzzyness of behavioral event categorization is irrelevant. All we shall make use of is the fact that a human observer *often* can isolate and classify (and therefore, name) a behavioral event without much doubt.

A crucial observation is that the isolatability and classifiability of those events is to some degree non-arbitrary. In the complex processes of blinking or vocalizing [s], there is something *intrinsic* which leads observers to isolate just *these* events, and which leads different observers to the same kind of isolation and classification judgements. It would be, in some way, "unnatural" to

isolate from the observed facial dynamics of another person an "event" which starts when an eye-blink is 70 per cent finished, and extends 50 ms after the eyeblink.

Thus, there must be *something* in the high-dimensional trajectory of a human's stream of behavior which enables observers to isolate, and classify, particular periods (and particular subsets of behavioral variables), due to intrinsic features of the process which are expressed in those periods. Loosely speaking, we must expect some kind of "flavored lumps" to exist in the process: lumps there must be, since there are some entities which can be isolated, and flavoured these lumps must be, since they can be classified.

Since much in this article hinges on the notions of intrinsic isolatability and intrinsic classifiability, I will try to explain these notions a bit more. An event in some complex, ongoing process is *intrinsically isolatable* if the event itself yields information about when it occurs – about its onset and about its end. This information must be not (at least not completely) relative to arbitrary conventions made by the observer. Different observers, who do not know of each other, must find it likewise natural to isolate roughly the same event from the "process background". An example for intrinsic isolatability would be a steep rising flank in some variable which indicates the onset of "something". A non-example would be the mere crossing of a threshold value of some variable, since this way of indicating an onset would depend on the entirely conventional fixing of a numerical value.

In a first approximation, *intrinsic classifiability* means that each event carries with itself enough qualitative information to enable the observer to classify it within a (typically huge) classificatory system. Somehow, each event must display enough "features" to allow its classification. Again, these "features" must not be merely conventional. A non-example for intrinsic classifiability would be to use the first five binary digits of a numerical measurement as five features – this being an arbitrary way of classification. A positive example would be to use geometrical features from the shape of a chaotic attractor – they are, in some sense, "proper properties" of the attractor event.

After this attempt at getting two intrinsically vague concepts clearer, let us return to the main line of argument.

We know that the high-dimensional overt behavior of a human is accompanied by neural/cognitive processes in that human's brain/mind. The internal process must be, in some sense, at least as "rich" as the externally visible motor behavior, since the motor behavior is in some sense controlled by neural/cognitive processes. However, from a dynamical systems perspective, the internal dynamics differ tremendously from the external behavior, and a direct identification or even comparison of internal with external dynamics seems out of the question. However, it seems reasonable to expect that for every (or most) of the intrinsically identifiable and classifiable events in the externally observable motor behavior, there exists an accompanying event in the neural dynamics which is also intrinsically identifiable and classifiable (given suitable observation techniques for neural/cognitive dynamics). In other words, we expect flavored lumps in the neural/cognitive dynamics, too.

These latter flavored lumps I shall call *dynamical symbols*. In a nutshell, thus, dynamical symbols are any kind of intrinscally isolatable, intrinscally classifiable events in neural/cognitive dynamics which correlate with likewise isolatable and classifiable events in overt motor behavior.

This is, of course, a very narrow framing of a symbol concept. Still, narrow as it is, this specific outlook on symbols leads to interesting questions concerning the mathematical modeling of neural/cognitive information processing.

I will now review briefly some of the known mathematical candidates for dynamical symbols, comment on their shortcomings, and on the way, clarify further what I mean by isolatability and classifiability (or, "lumpiness and flavor").

One of the most widely used ways to extract discrete events from continuous dynamics is via *partition cells*. The basic recipe is to define some volume cells $(c_i)_{i\in I}$ in phase space, label them with symbols $(E_i)_{i\in I}$, and when the system trajectory passes through cell c_j, say that the event E_j has occured.

This way of transforming a continuous trajectory into a symbol sequence is constitutive for ergodic theory [46] and (chaotic) symbol dynamics (e.g. [16] [15]). It also is a common strategy in the interpretation of recurrent neural networks (e.g. [24]) or the theory of qualitative resoning in classical AI (e.g. [39]).

However, defining discrete events via the trajectory's passing through a partition cell yields no model for dynamical symbols, since these events are neither intrinsically isolatable nor intrinsically classifiable.

The delimiting coordinates of a particular volume cell stem from arbitrary converntions. They are extrinsic to the process.

Likewise, volume cells per se are not "flavored". If we only know that the trajectory passes through c_{13} now and through c_2 next then we have no information whatsover to tell us what *kind* of event we have been witnessing. Mere hitting-a-volume-cell events are not intrinsically classifiable.

Often, of course, the observer will have some extra clues telling him to delimit volume cells in a particular way, and these clues may come from particular dynamical phenomena that are exhibited when the trajectory passes through these cells. Then, the events might be intrinsically isolatable and classifiable, albeit only due to the involvment of some extra, clue-giving phenomena.

Another quite common approach to picking discrete events from continuous dynamics is to use *point attractors*. The general scheme is to report an event whenever the system has relaxed into a stable equilibrium [29].

This approach is popular with artificial recurrent neural networks used for classification (e.g. [52]) or constraint satisfaction problems (e.g. [1]). Recently, even logical inferences have been rigorously re-interpreted as fixed-point relaxation of neural networks [22]. Altogether, it seems quite natural to equate discrete cognitive units (symbols, concepts) with point attractors, and related cognitive processes (constraint satisfaction, classification, inferences) with relaxation dynamics.

Point attractors are intrinsically isolatable: stable equilibria are system properties, not observational conventions.

One obvious shortcoming of point attractor models is that a true point attractor (like any attractor) *terminally* captures the system trajectory. If one wants to describe neural dynamics which exhibit a *sequence* of point attractor events, one has to introduce extra mechanisms to kick the trajectory out of attractors. Such extra mechanisms are, e.g. noise (popular in Hopfield networks), or input-induced bifurcations. I shall treat the issue of bifurcations and attractors extensively below.

A less obvious deficiency of point attractor models is that point attractors per se are hardly intrinsically classifiable. Like partition cells, they have almost no "flavor". By this I mean that there are no obvious, non-conventional features by which point attractors might gain individuality. Theoretically, one might characterize different point attractors by the magnitude of their Lyapunov exponents, by the size of their basin of attraction, and other such measures. But this repertoire of distinguishing features seems to be quite small, too small in any case to account for the enormous variability of dynamical symbols.

Currently the most prominent candidate for dynamical symbols is attractors with complex periodic or semi-periodic orbits, and chaotic attractors. Such *complex attractors* have been detected and induced both in artificial and biological neural systems [60] [5] [28].

Complex attractors are intrinsically isolatable, like any kind of attractor. Their great charm lies in the fact that they are also intrinsically classifiable. Two chaotic attractors typically "look" quite different, even to different observers who do not know of each other. Freeman's et al. graphical representations of chaotic attractor states in the olfactory bulb, and the way they geometrically change due to sensory input, are deeply inspiring.

Thus, are complex (in particular, chaotic) attractors good candidates for dynamical symbols?

I am sceptical about the ultimate value of chaos for the practical modeling of neural/cognitive phenomena, basically because identifying a high-dimensional chaotic attractor in an empirical time series typically requires more data than can be gathered while the attractor is extant (for more detailed criticism, cf. [47] [43], for an enlightening case study cf. [48]). I am afraid that in live brains under real-life conditions, chaotic attractors cannot be monitored long and/or precisely enough to certify their existence.

I would stick out my neck even further and question that chaotic attractors are the right mathematical metaphor at all for what we would like to observe. What I find dubious is the idea of a high-dimensional, complex *attractor* in the first place. The work of Freeman, Babloyantz and others has opened our eyes for the extreme richness, subtlety, and flexibility of (assumedly chaotic) activity in recurrent neural systems. Babloyantz and her colleagues in particular have put emphasis on the hypothesis that it is the fine-grained dynamical variants of chaotic attractors which hold promise as models for conceptual memory (i.e., as models for certain dynamical symbols). Now, having fine-grained, subtle, high-dimensional chaotic attractors also means only marginal stability (which has benefits for swift and flexible reactions, as has been pointed out by the cited researchers). Marginal stability means for an attractor that it is easily disrupted by noise, and that it takes long for the trajectory to settle even in the absence of noise. Biological brain subsystems are noisy; they are driven with strong signals from sensors and other subsystems; and they are highly adaptive and learning, i.e. a brain subsystem does not stay "itself" very long. All of these conditions render a brain subsystem a hostile environment for marginally stable, subtly complex attractors. I doubt that in a live brain and under real-life working conditions a chaotic attractor ever really has a chance to stabilize.

Thus, I fear that chaotic *attractors* are more a myth than a reality in live, situated brains. However, it seems undisputable that the investigations of Freeman, Babloyantz and others have touched on something fundamental, and that this fundamental thing is somehow connected with chaos. From this perspec-

tive, a promising route would be to investigate neural activity under the auspices of chaos, but without relying on attractors. This implies that chaos has to be defined in a novel way, which works for randomly driven dynamics. Such a definition is in fact available [11].

Partition cells and attractors are probably the most common, but by no means the only candidates for dynamical symbols that dynamical system theory can offer. The candidate that I am going to describe next will turn out to be inherently classifiable, but not inherently isolatable. This combination of properties is remarkable, since attractor models of dynamical symbols may easily make one believe that isolability can be considered an implicit consequence of classifiability. Since this topic touches basic aspects of the dynamical systems outlook on neural/cognitive systems, I will explain this point in some more detail.

One basic mechanism for explaining how a system trajectory can get caught in a sequence of different attractors, is bifurcations. The trajectory is released by one attractor due to the fact that the attractor itself vanishes, and is caught by the next because that attractor newly comes into existence. Two complex attractors, which are separated in time from each other by bifurcations of the entire system, will typically have different topological features. This implies that they cannot be smoothly "morphed" into each other. The bifurcation that occurs between them marks a singularity in the "reshaping" of the system's phase portrait. Thus, in this case, inherent classifiability (granted by the attractor's topolgical features) implies inherent isolatability (since topologcial features can appear only in "catastrophes", which are markers for isolation of the newly appearing attractor).

For a long time, I believed that these observations reflected a deeper, general law: namely, that a *qualitative* change of a system's dynamics cannot occur "smoothly", or expressed more casually, that different complex processes cannot be morphed[1] into each other. (Mis-)guided by the fundamental phenomenon of bifurcations, I believed that a dynamical system's qualitative type of dynamics can change into another qualitative type (another phase portrait) only through some kind of "catastrophic" transition. To me this seemed an extremely valuable insight, because it seemed to point to a fundamental necessity in continuous nature to produce discontinuities. From here, the road seemed paved toward understanding how dynamical symbols arise in neural/cognitive dynamics. The hope was, that a sufficiently rich dynamics would by necessity show some sort of discrete "klicking and ratcheting".

The candidate for dynamical symbols which I am going to describe now demonstrates that this "insight" was false. Qualitatively different *stochastic* processes *can* be smoothly morphed into each other. This means that we have inherent classifiability without inherent isolatability.

Discrete-time, discrete-value stochastic systems are convenient mathematical tools for modeling the dynamics of cognitive and neural processes. There are many variants of such systems: Markov models, hidden Markov models, stochastic automata of various kinds, stochastic cellular automata, to name but a few. I have added to this multitude myself by introducing dynamical symbol systems [31] and observable operator models [37] (a generalization of hidden

[1] In computer graphics, the smooth transformation of a picture into another is sometimes called "morphing"

Markov models with nicer mathematical properties). Systems of this kind make for coarse-grained, transparent, and often computationally efficient models of continuous systems [26] [15], among them recurrent neural networks [55]. In the programming of mobile robots, they are widely used as learnable memory modules for representing temporal experiences, in particular in navigation [4] [38].

A simple example of such a system is given in fig. 1. The figure shows a two-state stochastic transition graph, which generates stochastic sequences of a's and b's, as follows. At any time t, where $t = 0, 1, 2, \ldots$, the system is either in state a or in state b. If it is in state a, then it jumps to state b at time $t + 1$ with probability $1 - p$. If it is in state b, it jumps to state a with probability 1. A sequence of states produced that way is a system trajectory.

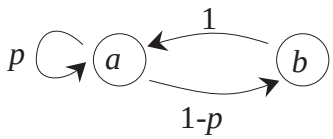

Figure 1: A simple stochastic transition graph.

Now interpret p as a control parameter. If we set $p = 1$, we observe a sequence $aaaaa\ldots$ which consists entirely of a's, with a possible leading b as an inital transient. The other extreme would be to set $p = 0$, which would result in an alternating sequence $abababa\ldots$. Intermediate settings of the control parameter would yield all sorts of *mixtures*.

If a dynamical systems theorist would be offered for analysis just the two time series $aaa\ldots$ and $ababa\ldots$, he would probably suspect to have been presented with a coarse version (derived by partitioning a phase space into two cells a and b) of a dynamical system that has undergone a period doubling bifurcation between the two observed time series!

This presumable spontaneous reaction of a system theorist demonstrates that in some intuitive sense, the sequences $aaa\ldots$ and $ababa\ldots$ are "qualitatively" different. This view is, however, very much nourished by the interpretation that a familiar period doubling bifurcation has given rise to the two sequences; in this view, intermediate mixtures between the two sequences would not be possible.

If, by contrast, the stochastic transition graph is known to be the generating system, the judgment that $aaa\ldots$ and $ababa\ldots$ are "qualitatively" different gets shaky. After all, the two sequences are just the two extremes on a continuum of processes with intermediate phenomenologies.

A similar morphing can occur in continuous-time, continuous-valued processes. For instance, consider a dynamical system governed by a control parameter γ, where there is a bifurcation between $\gamma = 0$ and $\gamma = 1$. Let the system run, and while running, stochastically switch γ between 0 and 1 with varying average frequency and with varying average relative duration of 0 vs. 1. In the two extreme cases ($\gamma \equiv 0$ and $\gamma \equiv 1$, i.e. zero switching frequency) we will observe the two "clean" bifurcative variants of the system. Depending on the settings of average switching frequency and of relative duration, however,

all kinds of intermediate dynamics will be observable, too.

In order to preclude a possible misunderstanding, let me emphasize that in the transition graph example it is not the state symbols a and b which are the candidates for dynamical symbols. Rather, the candidate for a dynamical symbol would be the qualitative sequence pattern altogether. This is completely analogous to the complex attractors discussed previously, where the dynamical symbol is the overall pattern of the trajectory.

I believe that the existence of intrinsically classifiable but not isolatable dynamical entities is not just an academic mathematical peculiarity. The occurence of such "entities" (but are they entities, if they are not isolatable?) has to be expected whenever a dynamical system is driven by stochastic input, which appears to be the standard case for neural and cognitive (sub-)systems.

This situation calls for the development of new mathematical concepts. What seems needed here is a "dynamical mixing theory" which can give us a clearer picture of what it means for temporal patterns to mix. Is it possible, in a stochastic process, to somehow factor out "pure" dynamical subprocesses, of which the observed process is a mixture? I am working toward such a theory, but it is too early to report results.

Besides the relatively prominent candidates pointed out so far, there is an unfathomable wealth of others, lesser known ones, only few of which have yet been explored as models for dynamical symbols. I shall proceed in a more summary fashion.

A basic property of biological neural structures is their spatial organization. Somatotopic or topographic maps abound, and primary visual cortices appear to exhibit, beneath the overall topographic representation, a fine-grained columnar pattern where the activation of columns represents specific features in the visual stimulus [45]. One striking feature of the cerebellum is the organization of its surface into "beams" which have been interpreted (among other options) as adaptive detectors of motor control signal sequences [9].

Findings of this kind suggest *localist*, or more generally *spatial* models of dynamical symbols. A dynamical symbol would correspond to the activation of some spatially defined collection of neurons. It would be intrinsically isolatable if the activated neural collective would exhibit a non-arbitrary "boundary" of some kind. This boundary could be anatomically defined through discontinuities in neural connectivity (as in columns or beams). Boundaries of some sort can also arise in homogeneous neural substrates by nonlinear, competitive spatiotemporal "neural field" dynamics [20]. Reaction-diffusion type of dynamics would be another possibility.

A batch of active neural tissue can be intrinsically classifiable for many reasons: e.g., by its anatomical structures, or by its being localized in a particular part of the brain. Furthermore, its activation dynamics may be complex enough to distinguish it from other batches.

Another exciting challenge for modeling dynamical symbols is *spatiotemporal dynamics*. Unfortunately, a satisfactory qualitative mathematical theory of such dynamics is presently beyond our reach. Only a few phenomena we have yet learnt to discern, e.g., solitons, or spiral patterns in reaction-diffusion dynamics.

A glimpse into the future can be cast through Bingham's article on the perception of spatiotemporal patterns by humans [7]. The experimental scheme described by Bingham is to present subjects with dynamical patterns of (a few) white dots on a black background. The patterns are derived from filmed

sequences of natural dynamical scenes, e.g. a field of high grass swaying in the wind, or honey oozing out of a jug. The white dots which are shown mark selected tips of grass or particles floating with the honey, etc. Subjects can correctly recognize these empoverished stimuli. Their performance can only be explained when one assumes that humans have rich models of *spatio*temporal patterns. This contrasts starkly with what mathematicians can yet reconstruct.

The observations reported by Bingham do not directly pertain to dynamical symbols. I have mentioned them here because his article indicates research directions for dynamical systems theory, which are equally relevant for the qualitative phenomenology of spatiotemporal neural dynamics.

The last, and possibly most elusive, candidate for dynamical symbols is *species*. Species are intrinsically isolatable and classifiable entites which arise in evolutionary processes.

Cognitive processes have been described in some detail building on the idea of "concepts = species" [10]. However, in that work only the short-term (in evolutionary perspective) *population dynamics* [59] is used to model cognitive phenomena, treating species as givens. I should also point out the inspiring young research strand of evolutionary linguistics, where the very emergence of language is modeled with concepts from evolution theory [54]. This approach sheds a bright light on the genesis of phonemes, words, and grammar and should not be missed by anyone interested in the nature of symbols. Finally, the best known attempt to tame evolutionary dynamics for modeling cognitive dynamical systems, and the evolution thereof, is classifier systems and genetic algorithms [30] [25].

Sadly, the mathematical theory of evolutionary dynamics is still in its infancy, Eigen's and Schuster's hypercycle model [18] notwithstanding. This renown mathematical achievement "only" captures speciation in certain chemical reaction systems. Spatial segregation or the emergence of ever more complex inheritance mechanisms are not addressed. The hypercycle describes one mechanism, but biological evolution is very much a story of the open-ended generation of a plurality of mechanisms [42]. Although the theory and practice of classifier systems and genetic algorithms has been developed further and broader than that of other evolutionary models, they do not offer even a convincing model of species ([25] p. 186). This corresponds with the situation in the biological theory of evolution, where it is not at all clear on which units selection actually works – genes, or species, or symbiotic multi-species systems? [14]

I feel that our lack of understanding of evolutionary processes cannot be fundamentally remedied, since evolution is *qualitatively productive* – ever new mechanisms emerge, and even mechanisms of evolution of mechanisms evolve [58]. There is no "master mechanism", the knowledge of which would give us the multitude as a corollary.

Brains are the product of evolution, and possibly the development of cognitive systems in ontogenesis also bears some marks of evolution's qualitative productivity. Inasmuch as dynamical symbols can be interpreted as "species" (or genes, or populations, or any other kind of lumps inhabiting brains/cognitive systems), it seems fundamentally impossible that we can achieve a unified mathematical model of them.

This was a sweeping pass over some mathematical models for "flavored

lumps". It has led us from simple partition cells to the most elusive products of evolutionary dynamics. The general message I wanted to convey is summarized in the following points:

- There are many ways how "symbolic" units can arise in neural/cognitive dynamics.
- We should not look for *the* correct mathematical model. Nature loves to play, not to fill out forms.
- The mathematical modeling, and our intuitive understanding, of "symbols" has barely started. Dynamical systems theory still has to integrate stochastic, spatial, and evolutionary aspects. Exciting discoveries are waiting for mathematicians and brain/cognition researchers.

3 Transient attractors

In this section, I shall motivate and explain *transient attractors* (TA's). This mathematical object generalizes the notion of attractors. Unlike classical attractors, TA's can exist in systems driven by stochastic input, and in systems whose variables dynamically change their relative time scales. Thus, TA's can serve as models for dynamical symbols in some cases where classical attractors are not defined.

I have mentioned in section 2 an intrinsic difficulty with attractor models for dynamical symbols. Namely, an attractor by definition terminally captures the system trajectory. By contrast, dynamical symbols sequentially arise and vanish in neural/cognitive dynamics. As I have noted in the previous section, an apparent way out of this dilemma is by bifurcations which generate and destroy attractors. The control parameters which induce the bifurcations presumably are input quantities from sensors or other neural/cognitive subsystems.

A problem with bifurcations is that they are well-defined only when the dynamics of the control parameters is at least an order of magnitude slower than the time scale of the controlled system. But this is not typically the case with neural/cognitive systems! Quite to the contrary, the dynamics of input variables which "control" a subsystem is typically just as fast as the dynamics of the subsystem. With the exception of some slow somatosensory modalities (temperature, hunger, certain kinds of pain, etc.), the brain is under constant fire of fast sensory input (visual, auditory, kinaesthetic). Furthermore, different brain subsystems will often tightly interact with each other, each one giving a portion of it's own dynamics as input to the other.

From a mathematical perspective, we have to admit that the notion of bifurcation (and hence, of attractors) is no longer well-defined in such situations.

The following formal example of a transient attractor shows what it means for a "control parameter" to have a dynamics which is as fast as the "controlled" system.

Consider the system specified in polar coordinates by $\dot{\varphi} = 1, \dot{r} = r(1 - r)\sin\varphi$. Its phase portrait is characterized in the vicinity of the origin by anticlockwise, closed loops, among them a loop on the unit circle (fig. 2a).

When one follows any two trajectories (the fixed point trajectory at the origin excepted) through increasing values of φ, one finds that they come closer

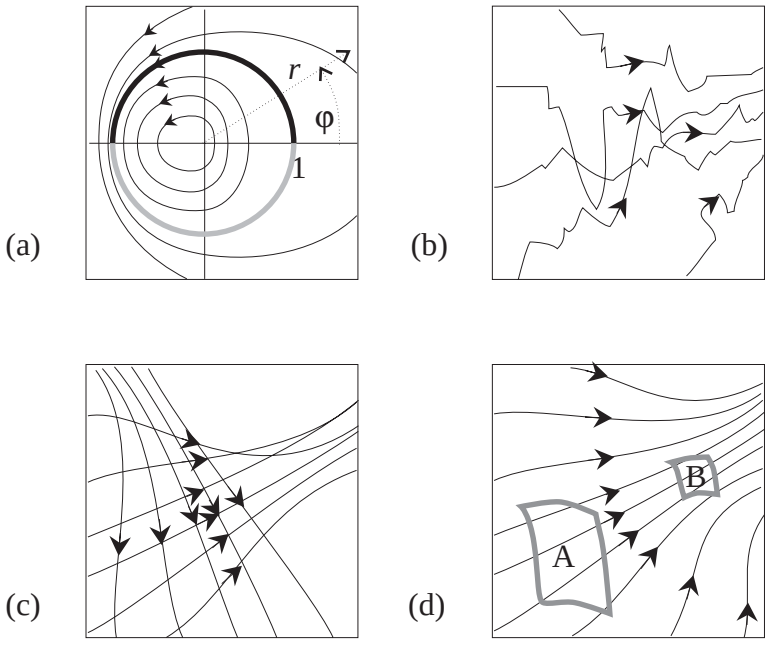

Figure 2: (a) The system $\dot{\varphi} = 1, \dot{r} = r(1-r)\sin\varphi$. (b) To be noise and not to be noise – which is which in an empirical phase portrait? (c) Crossing trajectories. (d) Phase space contraction.

to each other in the upper half of the plane (i.e., $0 < \varphi < \pi$), whereas they recede from each other in the lower half. This can be interpreted as the effect of a "fast bifurcation" induced by a fast control parameter, as follows. Re-interpret $\dot{r} = r(1-r)\sin\varphi$ as a one-dimensional system with a control parameter φ. If φ is fixed at a value between $0 < \varphi < \pi$, this system exhibits a point attractor at $r = 1$. For $\pi < \varphi < 2\pi$, the point attractor in $r = 1$ turns into a repellor. The values $\varphi = \pi$ and $\varphi = 2\pi$ mark bifurcations.

Thus, one might interpret the system shown in fig. 2a as consisting of two coupled one-dimensional subsystems (in φ and r), where one of the subsystems yields a "fast control parameter" φ for the other. Variations of this control parameter induce a "fast" creation-and-destruction cycle of an attractor – a *transient* attractor.

Examples like this have been my original motivation for the introduction and naming of transient attractors. The idea of fast bifurcations does however not lead very far in practical applications, because control parameters and the dynamics thereof are mostly unknown. The "fast control parameters" will quite often come from external input into the system, and have essentially stochastic dynamics. Even worse, typically one will not even be able to identify the relevant input parameters. But without an idea what the control parameters are, an analysis of bifurcation sequences becomes all but impossible.

Empirical phase portraits derived from neural/cognitive systems have still other properties which render the classical notions of control parameters and

bifurcations almost useless. I mention only two of these unpleasant properties. First, empirical phase portraits are noisy. But it is by and large impossible to separate noise from the "actual" system dynamics, because whatever amplitudes at whatever frequency we observe, it might be an "actual" sense-making system answer to some input, which must not be discarded as noise (fig. 2b). The second unpleasant property of empirical phase portraits is that they feature crossing trajectories (fig. 2). There are many reasons for trajectories to cross, e.g. noise; or projections of a system, which is defined on an n-dimensional manifold, on a n-dimensional subspace of the embedding space; or observation of a high-dimensional system in only a few of its variables.

Properties of these kinds force one to abandon autonomous systems ruled by differential equations, as model systems for neural/cognitive processes in which there are dynamical symbols to be found. A more general framework of stochastic processes seems adequate. Therefore, the question is: what is the intuitive "core" of TA's, when we have to abandon the descriptive tools of control parameters and bifurcations?

I suggest to use as the defining property of a TA that it lead to a contraction of phase space volume. This effect is illustrated in fig. 2d. A TA reveals its existence by trajectories which "approach each other" in time. Another way to state the same fact is to say that a TA affords us with good *local predictability* of the process. Referring to fig. 2d, if at time t_n we know (by some measurement) that the system state is in A, then if there is a transient attractor we can predict that the system will be in B at time t_{n+1}, where the volume B is smaller than A. This contraction of phase space volume corresponds to an information gain over time, and this gain is indicative for the presence of a TA.

There are many ways how this basic idea of phase space contraction can be made precise in stochastic processes. In [33] I gave a definition which I find too narrow and too complicated today. I will present a more general and more transparent definition presently. The practical use of such definitions is limited because of their high level of mathematical abstraction. Therefore, I will not put much emphasis on the formal definition, and present it with little explanation for readers who are familiar with the terminology of stochastic processes. For most readers it will be more relevant to know that a simple and transparent algorithm for detecting TA's in empirical multivariate time series is sketched in [36][2].

Now, one possible definition of TA's. Let $(\Omega, \mathfrak{A}, P, (X_t)_{t\in\mathbb{R}})$ be a stationary stochastic process with values in the observation space $(\mathbb{R}^n, \mathfrak{B}^n)$, where $\mathfrak{B}^n$ is the Borel σ-algebra on $\mathbb{R}^n$. Let $\mathfrak{A}_0$ be the sub-σ-algebra of $\mathfrak{A}$ which is generated by $(X_t)_{t\leq 0}$, i.e. $\mathfrak{A}_0$ is the σ-algebra of the processes' past up to $t = 0$.

Let us return for a moment to the situation of a classical ODE system's phase portrait. If we would want to define a mutual approaching of two trajectories T, T', we would look at the points x and x' through which they pass at time $t = 0$, and then consider their future development after they have passed there.

Very general stochastic analogues of the points x and x', and of the past of T, T' before those trajectories passed through x and x', are sets A, A' from $\mathfrak{A}_0$. Intuitively, A and A' are informations that can be gained about the system state by some observations that were made in the past up to $t = 0$.

[2] ftp'able from http://www.gmd.de/People/Herbert.Jaeger/Publications/ . The algorithm is currently being implemented for a diploma thesis.

A likewise general stochastic analog of the future of T, T' after $t = 0$ can be specified through the conditioned probability measures P_t^A and $P_t^{A'}$, which are defined by $P_t^A(B) = P[X_t \in B \mid A]$ and $P_t^{A'} = P[X_t \in B \mid A']$, where $t > 0$ and $B \in \mathfrak{B}^n$. The families $(P_t^A)_{t>0}$ and $(P_t^{A'})_{t>0}$ could amply be called "fuzzy trajectories through fuzzy points A and A'".

Now, what does it mean for two such "fuzzy trajectories" to approach each other? One would like to define some kind of "distance" between the conditioned probabilities P_t^A and $P_t^{A'}$, and then to note when this distance shrinks in time.

Consider the "distance" measure δ for two probability measures P, P' on $(\mathbb{R}^n, \mathfrak{B}^n)$ defined by $\delta(P, P') := \int\int \|x - y\| P(dx) P'(dy)$. This is actually not a distance measure in the mathematical sense, because $\delta(P, P)$ is not zero for most probability measures P (it is zero only if P is a point measure). However, for our purposes it is the right measure. In the special case of classical trajectories (i.e. where P_t^A and $P_t^{A'}$ are point measures), δ yields the ordinary metric distance. In the case of "fuzzy trajectories" (P_t^A and $P_t^{A'}$ not being point measures), δ can intuitively be interpreted as a kind of mutual information one "fuzzy trajectory" affords about the other at time t.

Now, in order to define whether the future developments of A and A' "approach" each other, we can consider the derivative of $\delta(P_t^A, P_t^{A'})$ in $t = 0$, i.e., $d\,\delta(P_t^A, P_t^{A'})/dt\ (0)$. If this number is negative, we have found an "approaching of future developments".

Having these tools ready, transient attractors can be defined as follows.

First, define a handy class of admissible observations of the past, i.e. a manageable subset $\mathfrak{C} \subset \mathfrak{A}_0$. Probably the simplest choice would be to use $\mathfrak{C}_1 = \{X_0 = x \mid x \in \mathbb{R}^n\}$, i.e. the point observations of the process in $t = 0$. A trickier but still simple variant would be $\mathfrak{C}_2 = \{X_0 = x, X_{-1} = y \mid x, y \in \mathbb{R}^n\}$, i.e. the informations about the system attainable from a point observations at the present plus a point observation one time step in the past.

Next, consider the function $ta : \mathfrak{C} \times \mathfrak{C} \to \mathbb{R}, (A, A') \mapsto d\,\delta(P_t^A, P_t^{A'})/dt\ (0)$. Then, define as a transient attractor every maximal connected region in $\mathfrak{C} \times \mathfrak{C}$ in which $ta < 0$.

Of course, this definition presupposes that $\mathfrak{C}$ has been selected in a way which allows to define a topology on $\mathfrak{C} \times \mathfrak{C}$, in order to make the notion of connectedness come to bear.

The variants $\mathfrak{C}_1$ and $\mathfrak{C}_2$ are roughly analog to describing a physical system only through its positions vs. through positions plus velocities. Another anlogue would be to describe a system by a first-order vs. a second-order Markov process. The second variant allows to disentangle transient attractors that cross each other in phase space (as in fig. 2c).

This definition is admittedly complex, not worked out in detail, and to some degree arbitrary. However, something or other in this fashion has to be fixed if we wish to rigorously work out the intuitive idea of fast generation/destruction of attractors in stochastic dynamics. I feel a bit embarassed about the current state of affair, but I cannot offer anything better (however, the practical algorithm mentioned above is much easier to understand than this abstract definition!).

4 Discussion

This article pursued two goals. First, I wished to illustrate and emphasize the phenomenological diversity of dynamical symbols. Complex attractors are an inspiring and important class of models, but a host of other yet unimaginable dynamical phenomena awaits us. Second, I took a tentative little step in that unchartered terrain, by offering the concept of transient attractors.

In former work, I have assumed dynamical symbols as givens, and developed a mathematical theory which describes how dynamical symbols can interact, build up complex "resonances", and develop into hierarchies in a self-organizing fashion. This mathematical approach, *dynamical symbol systems* [31] [32], basically describes the temporal evolution of directed graphs, where the edges are identified with dynamical symbols, and where self-organization into "resonances" comes about as self-reinforcing of cyclic subgraphs. In that work, however, I assumed only some abstract properties of dynamical symbols, and I was vague about the concrete mathematical nature of dynamical symbols themselves (all I did in this direction was to allude to chaotic attractor states). In the present article, conversely, I focussed on the mathematical nature of single dynamical symbols.

I required dynamical symbols to possess intrinsic isolatability and classifiability. It seems unlikely that a precise definition of intrinsic isolatability and classifiability can be given. I rather believe that neural dynamics, propped by billions of years of freewheeling evolution, intrinsically defies clean definitions – mathematical rigour hardly being a fitness criterion in natural selection. As a consequence, I do not think that the notion of dynamical symbols can be mathematically defined. It is more a horizon line for open-ended quest than an axiom from which to start. Whatever dynamical phenomena we will learn to see on that way, they will enrich our understanding of how walking, reasoning, speaking, reminding unfold, the dynamical everyday stuff which propels us through our lives.

Acknowledgments The ideas presented in this article grew over a long time, during which I benefitted very much from discussions with friends and colleagues, among them Andreas Birk, René ten Boekhorst, Thomas Christaller, Tim van Gelder, Dieter Jaeger, Christoph Lischka, Rafael Nuñez, Frank Pasemann, Rolf Pfeifer, Eva Ruhnau, Christian Scheier, Luc Steels and Jun Tani. I wish to thank them all. The work was sustained by a postdoctoral grant donated by GMD, Sankt Augustin.

References

[1] D.H. Ackley, G.E. Hinton, and T.J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

[2] Vera A.H. and H.A. Simon. Situated action: A symbolic interpretation. *Cognitive Science*, 17(1):7–48, 1993.

[3] M.A. Arbib, editor. *The Handbook of Brain Theory and Neural Networks*. MIT Press/Bradford Books, 1995.

[4] K. Baasye, Th. Dean, and L.P. Kaelbling. Learning dynamics: System identification for perceptually challenged agents. *Artificial Intelligence*, 72:139–171, 1995.

[5] A. Babloyantz and C. Lourenço. Computation with chaos: A paradigm for cortical activity. *Proceedings of the National Academy of Sciences of the USA*, 91:9027–9031, 1994.

[6] M. Bickhard. Representational content in humans and machines. *Journal of Experimental and Theoretical AI*, 5:285–333, 1993.

[7] G.P. Bingham. Dynamics and the problem of visual event recognition. In R. Port and T. van Gelder, editors, *Mind as Motion: Explorations in the Dynamics of Cognition*, chapter 14, pages 403–448. MIT Press/Bradford Books, 1995.

[8] J.M. Bower, editor. *The Neurobiology of Computation. Proc. 3rd Annual Conf. on Computation and Neural Systems (CNS).* Kluwer, Boston, 1995.

[9] V. Braitenberg, D. Heck, and F. Sultan. The detection and generation of sequences as a key to cerebellar function. Experiments and theory. *Behavioural and Brain Sciences*, accepted 1996, to appear.

[10] P. can Geert. Growth dynamics in development. In R. Port and T. van Gelder, editors, *Mind as Motion: Explorations in the Dynamics of Cognition*, chapter 11, pages 313–338. MIT Press/Bradford Books, 1995.

[11] M. Casdagli. A dynamical systems approach to modeling input-output systems. In M. Casdagli and S. Eubank, editors, *Nonliner Modeling and Forecasting*, volume XII of *SFI Studies in the Sciences of Complexity*, pages 265–281. Addison-Wesley, 1992.

[12] D.J. Chalmers. Connectionism and compositionality: Why Fodor and Pylyshyn were wrong. *Philosophical Psychology*, 6:305–319, 1993.

[13] W.J. Clancey. Situated Action: A Neuropsychological Interpretation Response to Vera and Simon. *Cognitive Science*, 17 (1):87–116, 1993.

[14] T.H. Clutton-Brock and P.H. Harvey, editors. *Readings in Sociobiology.* Freeman, Reading and San Francisco, 1978.

[15] J.P. Crutchfield. Semantics and thermodynamics. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting*, Santa Fe Institute Studies in the Sciences of Complexity XII, pages 317–359. Addison-Wesley, Redwood City, 1992.

[16] W. Ebeling and G. Nicolis. Word frequency and entropy of symbolic sequences: a dynamical perspective. *Chaos, Solitons & Fractals*, 2(6):635–650, 1992.

[17] U. Eco. *La struttura Assente.* Bompiani, Milano, 1968. German translation: *Einführung in die Semiotik*, UTB Fink 1972.

[18] M. Eigen and P. Schuster. The hypercycle: A principle of natural self-organization. *Naturwissenschaften*, 1977/78. 64 (1977), 541-565 (Part A); 65 (1978), 7-41 (Part B); 65 (1978), 341- 369 (Part C).

[19] J.L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[20] Ch. Engels and G. Schöner. Dynamic fields endow behavior-based robots with representations. *Robotics & Autonomous Systems*, 14:55–77, 1995.

[21] J.A. Fodor and Z.W. Pylyshin. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28:3–71, 1988.

[22] P. Gärdenfors. How logic emerges from the dynamics of information. In J. van Eijck and A. Visser, editors, *Logic and Information Flow*, Foundations of Computation, chapter 4, pages 49–77. MIT Press, 1994.

[23] M.S. Gazzaniga, editor. *The Cognitive Neurosciences*. MIT Press/Bradford Books, 1995.

[24] C.L. Giles and C.W. Omlin. Learning, representation, and synthesis of discrete dynamical systems in continuous recurrent neural networks. In *Proceedings of the IEEE Workshop on Architectures for Semiotic Modeling and Situation Analysis in Large Complex Systems*. IEEE Press, 1995.

[25] D.E. Goldberg. *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley, 1989.

[26] P. Grassberger. Toward a quantitative theory of self-generated complexity. *Int. J. of Theor. Physics*, 25(9):907–938, 1986.

[27] S. Harnad. The symbol grounding problem. *Physica*, D42:335–346, 1990.

[28] Y. Hayashi. Oscillatory neural networks and learning of continuously transformed patterns. *Neural Networks*, 7(2):219–232, 1994.

[29] J. Hertz. Computing with attractors. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 230–234. MIT Press/Bradford Books, 1995.

[30] J.H. Holland and J.S. Reitmann. Cognitive systems based on adaptive algorithms. In D.A. Waterman and F. Hayes-Roth, editors, *Pattern directed inference systems*, pages 313–329. Academic Press, New York, 1978.

[31] H. Jaeger. Dynamic symbol systems. Ph.d. thesis, Faculty of Technology, University of Bielefeld, 1994.

[32] H. Jaeger. An introduction to dynamic symbol systems. In J. Hallam, editor, *Hybrid Problems, Hybrid Solutions. Proceedings of the AISB- 95*, pages 109–120. IOS Press/Ohmsha, Amsterdam, 1994.

[33] H. Jaeger. Identification of behaviors in an agent's phase space. Arbeitspapiere der GMD 951, ftp://ftp.gmd.de/gmd/ai-research/publications/1995/jaeger.95.identify.ps.gz, GMD, St. Augustin, 1995.

[34] H. Jaeger. Brains on wheels: Mobile robots for brain research. *submitted to Theory in Neuroscience*, 1996.

[35] H. Jaeger. Dynamische Systeme in der Kognitionswissenschaft. *Kognitionswissenschaft*, 5(4):151–174, 1996.

[36] H. Jaeger. Transient attractors: Re-recognizable regularities in empirical time series. *Submitted to Theory in Bioscience*, 1996.

[37] H. Jaeger. Observable operator models and conditioned continuation representations. Arbeitspapiere der GMD 1043, GMD, Sankt Augustin, 1997.

[38] F. Kirchner. KURT: A prototype study of an autonomous mobile robot for sewerage system inspection. Arbeitspapiere der GMD 989, GMD, GMD - Forschungszentrum Informationstechnik GmbH, Sankt Augustin, 1996.

[39] C.X. Ling and R. Buchal. Learning to control dynamic systems with automated quantization. In P.B. Brazdil, editor, *Machine Learning. Proceedings of the ECML-93*, Lecture Notes in Artificial Intelligence 667, pages 372–377. Springer Verlag, Berlin, 1993.

[40] J. Lyons. *Semantics, vols. 1 & 2.* Cambridge University Press, 1977. German translation: *Semantik I & II*, C.H. Beck, München 1980.

[41] D.R. Mani and L. Shastri. Reflexive reasoning with multiple instantiation in a connectionist reasoning system with a type hierarchy. *Connection Science*, 5(3/4):205–242, 1993.

[42] J. Maynard Smith and E. Szathmáry. *The Major Transitions in Evolution.* Freeman, 1995.

[43] M. Millonas. The importance of being noisy. *The Bulletin of the Santa Fe Institute*, 9(1):22–23, 1994.

[44] A. Newell. Physical symbol systems. *Cognitive Science*, 4:135–183, 1980.

[45] K. Obermayer, H. Ritter, and K. Schulten. A principle for the formation of the spatial structure of cortical feature maps. *Proc. of the National Academy of Sciences of the USA*, 87:8345–8349, 1990.

[46] K. Petersen. *Ergodic Theory.* Cambridge University Press, 1983.

[47] H. Preißl, A. Aertsen, and G. Palm. Are fractal dimensions a good measure for neural activity? In R. Eckmiller, G. Hartmann, and G. Hauske, editors, *Parallel Processing in Neural Systems and Computers*, pages 83–86. North-Holland, Amsterdam, 1990.

[48] S.S. Robertson, A.H. Cohen, and G. Mayer-Kress. Behavioral chaos: Behind the metaphor. In L.B. Smith and E. Thelen, editors, *A Dynamic Systems Approach to Development: Applications*, pages 119–150. Bradford/MIT Press, Cambridge, Mass., 1993.

[49] R.C. Schank. *Dynamic memory. A theory of reminding and learning in computers and people.* Cambridge University Press, 1982.

[50] J.R. Searle. Minds, brains, and programs. *The Behavioral and Brain Sciences*, 3:417–457, 1980.

[51] L.B. Smith and E. Thelen, editors. *A Dynamic Systems Approach to Development: Applications.* Bradford/MIT Press, Cambridge, Mass., 1993.

[52] P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In D.E. Rumelhart and J.L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, pages 194–281. MIT Press, Cambridge, Mass., 1986.

[53] P. Smolensky. On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11:1–74, 1988.

[54] L. Steels. Synthesising the origins of language and meaning using coevolution, self-organization, and level formation. In J. Hurford, editor, *Evolution of Human Language.* Edinburgh University Press, Edinburgh, 1997.

[55] P. Tino, B.G. Horne, and C.L. Giles. Finite state machines and recurrent neural networks – automata and dynamical systems approach. Technical report umiacs-tr-95-1 and cs-tr-3396, Institute for Advanced Computer Studies, University of Maryland, 1995. To be published in *Progress in Neural Networks*, special volume on Temporal Dynamics and Time-Varying Pattern Recognition, eds. J.E. Dayhof and O. Omidvar, Ablex Publishing.

[56] T. van Gelder. The dynamical hypothesis in cognitive science. *Behavioural and Brain Sciences*, to appear.

[57] T. van Gelder and R. Port, editors. *Mind as Motion: Explorations in the Dynamics of Cognition.* Bradford/MIT Press, 1995.

[58] G. Wagner. Evolution der Evolutionsfähigkeit. In A. Dress, H. Hendrichs, and G. Küppers, editors, *Selbstorganisation: die Entstehung von Ordnung in Natur und Gesellschaft*, pages 121–148. Piper, München, 1986.

[59] O.E. Wilson and W.H. Bossert. *A Primer of Population Biology.* Sinauer Ass., Stamford, Conn., 1971. German transl.: *Einführung in die Populationsbiologie.* Springer Verlag 1973.

[60] Y. Yao and W.J. Freeman. A model of biological pattern recognition with spatially chaotic dynamics. *Neural Networks*, 3(2):153–170, 1990.