11.5.92/NOB

SYSTEMATICITY, THOUGHT AND ATTENTION IN A DISTRIBUTED CONNECTIONIST SYSTEM

Niels Ole Bernsen and Ib Ulbæk
Centre of Cognitive Science. Roskilde University, Denmark
email: nob@ruc.dk, ibu@jane.ruc.dk

Summary: The paper presents ongoing work on the processing of visual information in distributed connectionist systems. It is demonstrated that such systems are able to meet all the requirements of syntactic and semantic systematicity. The described system represent thoughts of two different kinds of complexity which are called low-level thought and higher-level thought, respectively. Higher-level thought requires temporally sequenced, discrete output for its representation. Such output may provide a crucial link in the vision-to-language processing chain within modular cognitive architectures. The simulation includes a mechanism for changing focus of attention in distributed connectionist systems. Ulbæk did the simulation and Bernsen did most of the theory.

Keywords: Distributed connectionism, cognitive architecture, systematicity, visual attention mechanisms, vision and language.

1. Introduction

In a recent paper, we have demonstrated systematicity and compositionality in distributed connectionist networks working with "real", pre-linguistic semantic information (Bernsen and Ulbæk 1992). Systematicity and compositionality had been claimed to be unique to classical syntactic AI systems (Fodor and Pylyshyn 1988). However, the capabilities of distributed networks of generalisation, abstraction and instantiation are demonstrably sufficient to account for the fact that when a system is able to represent (or think) the perceived relational fact that *aRb* it is necessarily also able to represent the fact that *bRa.* Through learning, the system acquires the abstract complex representation *xRy* which it then applies to novel, perceived individual objects in order to determine if the relation *R* holds between them. In the simulation, *R* was the spatial relation "right of". This "right-of recogniser" was a simple backpropagation network with one hidden layer. It could easily be hooked up with a set of visual modules at one end and with a synthetic speech module at the other to constitute an artificial example of the full chain from vision to speech. However, doing so would have been cheating as far as the natural language processing part of the system is concerned, as will become apparent below. The work also demonstrated the need for a painstakingly accurate semantic analysis of the task that a distributed connectionist network is actually solving. Great care should be taken in order not to confuse our own semantic capabilities with that of the simulation being analysed.

The present paper describes ongoing work on more advanced simulations using recurrent nets. The work addresses some fundamental questions about the representation of pre-linguistic complex thought dealing with perceived material in distributed connectionist systems. The first objective (1) was to address a final aspect of the systematicity (and compositionality) challenge to distributed connectionism which was not addressed in our previous paper. However, the way this challenge was actually and, it is claimed, successfully met leads to two more general ideas. The theoretical discussion leads to the hypothesis (2) that *complex thought itself is of at least two different kinds* in terms of processing and representational requirements. One class of complex thought can be handled by several existing types of distributed connectionist systems whereas a second, "higher" class of complex thought can be handled only by systems capable of using a temporal dimension of representation. *Low-level thought* consists essentially of distributed representations plus output representations of patterns. There is no need for a temporal dimension in the representation. The patterns can be rather complex such as the pattern that one spatial object is located to the right of another spatial object. But there are limits to the complexity of those patterns if we are to preserve the systematicity requirement, at least in the context of modular cognitive architectures. When the representation including the output representation has to include a temporal dimension, we speak of *higher-level thought.* Finally (3), an idea is presented as to how goal structures determining focus of attention can be realised in distributed connectionist systems dealing with static visual information.

It should be stressed that the entire argument of this paper rests on the attempt to solve concrete problems of semantic representation in distributed connectionist systems while preserving the systematicity (and compositionality) requirements. At some point when working on the issue of systematicity and compositionality, we have found ourselves forced into using temporal representations producing temporal and discrete output representations. If that is not necessary, the argument might fail, but there does not yet seem to be working alternatives in the literature. If that *is* necessary, on the other hand, the results are relevant to ongoing discussions on the hybrid character of cognitive architectures. One question then becomes whether the semantic distinction between two classes of complex thought lends plausibility to the idea that thought is itself hybrid in the sense that we have to acknowledge a kind of non-linguistic syntax in "higher" complex thought as realised through temporal and discrete output representations. In any case, the results suggest a principled solution to the problem of how to link up pre-linguistic thought with linguistic representation in distributed connectionism.

2. A note on methodology

The methodology adopted consists in (1) building minimal distributed connectionist systems which demonstrate basic principles of cognitive architectures. The systems are minimal in the sense that they work on tasks that are minimally simple for demonstration purposes. (2) The methodology is incremental in the sense that once a principle has been demonstrated in one simulation, a subsequent simulation does not necessarily have to repeat that demonstration if it can be safely assumed that a repetition lies within the power of the new system used. This, of course, may sometimes be a dangerous strategy and we have tried to be fully explicit about the shortcuts made. The systems (3) work with input which is assumed to be provided by visual modules in order to ensure that the simulations deal with undisputed semantical representations. Finally (4), we strongly adhere to the idea of the modularity of cognition. Even if distributed connectionism is on the right track in modelling cognition, it will have to involve a series of cognitive modules in order to cover, e.g., the chain from peripheral vision to language and reasoning. This implies that output representations of cognitive modules are just as important as what goes on in and between the hidden layers of a simulation. *A cognitive module and what it represents is characterised by its input, hidden representations and output as well as by the processes operating on those representations.*

The whole enterprise is reminiscent of the early days of symbolic AI. In view of the many implicit assumptions needed to do anything at all, there is a clear risk of building monuments of individual imagination rather than solid science. It is some comfort, however, to be able to relate the simulations to behavioural laboratory experiments.

3. Thoughts and other representations

One conceptual point, or perhaps it is rather a question of terminology, should be addressed right away. Cognitive science more or less universally acknowledges *representations.* Cognitive systems internally represent external and sometimes also internal states of affairs. We shall be speaking of *thoughts* in this paper without being able to offer any very precise distinction between thoughts and other representations. Thoughts are representations, but not all representations are thoughts. The working idea of a thought is the following: thoughts are internal representations of states of affairs to which we refer in explaining the actions of cognitive systems, and the kinds of thought that a cognitive system is capable of having help us characterise that system more generally. Cats have some kind of thoughts about the presence of mice. They may have thoughts about the absence of mice as well, but they don't have thoughts about nuclear energy or about protecting their eggs, and they equally don't have thoughts about their doing computations of zero-crossings in their visual system. Human beings may be capable of having thoughts about almost anything but, like cats, they don't have thoughts about their doing computations of zero-crossings in their visual system exept when theorising about the computational tasks of visual systems. In other words, the concept of thought may have a useful role to fill as designating significant output from cognitive processing modules used in explaining their actions at some relatively coarse level of detail. It may be useful, at least operationally, to conceive of thoughts as outputs from the distributed processing taking place in functionally significant cognitive modules. Obviously, thoughts don't have to be linguistic in any ordinary sense of the term.

4. Two steps toward higher-level systematicity

The system described in (Bernsen and Ulbæk 1992) masters various types of abstraction and generalisation by virtue of its ability to generalise *right of(a,b)* to *right of(x,y).* Thus, the system has learnt most of the (surprisingly rich) semantics of the general spatial relation *right of* and has successfully generalised representations of particular objects standing in that relation to the general object representations *x* and *y*. In a distributed connectionist mode of representation, it has learnt about variables and instantiation of those variables. However, whereas the system has learnt the concept of *a* particular object which could be, e.g., a bar, a triangle or a square, it still has not learnt *individual concepts* for bars, triangles or squares. The system is not able to identify or recognise a particular object present in the visual scene as being, for instance, a square. This means that, when the system correctly identifies a *right of* relation in the visual array or scene through instantiating its general concept *right of(x,y)* to what it sees there, the system instantiates *x* to *some particular spatial object* and *y* to *some particular spatial object different from the first one* but without being able to identify the specific nature (or type) of those spatial objects. In other words, the system is able to represent the fact that two different spatial objects stand in the relation *right of* to each other, but unable to represent the fact that the two objects are, say, a triangle and a square.

One way of expressing this result is to say that the system masters *low-level systematicity* (to be coupled shortly with the concept of low-level thought). In principle, given any two spatial objects, say, Paul and Mary, or Mary and Paul, or Mary and a chair, the system is able to determine whether the right-of relation holds between them. In doing so, the system has something like the thought (or intentional content) that *some spatial object is (or is not) to the right of some other spatial object.* However, the system is not able to think about the type-identities of the objects it perceives and thus does not have *higher-level systematicity* (to be coupled shortly with the concept of higher-level thought). It may have been higher-level systematicity that Fodor and Pylyshyn had in mind when they wrote their 1988-paper, but they definitely overlooked low-level systematicity.

The ability to identify and distinguish between *types* of particular object seems to be fundamental to biological systems solving such tasks as seeking food or prey. We would like to provide a distributed connectionist system with such a capability as a precondition of demonstrating that this system is capable of higher-level systematicity (or thought). Let us specify the system through a number of steps.

1. The system works on a selection of different visual 2-D objects in the input array. For convenience, each object should have a name in natural language. Let the objects be a triangle and a square, respectively. Sometimes the variables *a* and *b* will be used for brevity below in referring to those objects recognised by the system.

Given this setup, the system should learn to associate the appropriate concepts with each of these objects. It does not seem necessary for the present purpose to teach the system the general concepts of triangles and squares although this can clearly be done (it was done to a reasonable approximation in the previous system). So one might say that the system learns to associate "proper name concepts" with the objects rather than learning the corresponding abstract concepts themselves. This does not matter at this stage since the important point is that the system becomes able to identify the different objects presented.

The object-concept association should work for each object independently of its position in the input array. Whenever one of the objects are present somewhere in the scene, the system responds by telling us that the object is present.

The system described so far is a *single-object identifier*. At least two different output nodes are needed, one for each object. When one of the output nodes is active, the corresponding object should be somewhere in the input array. The system is assumed to work as an object identifier module on top of a set of visual modules which have produced the scene whose objects are identified. We explicitly *don't* say that the system learns language. What it learns is to recognise objects and hence to have thoughts to the effect that specific objects are present in the scene. Recognising objects requires concepts for those objects. We do not have to conceive of the system as having learnt linguistic names for objects.

2. Now consider the case where several objects are present simultaneously in the scene. These should be correctly identified wherever they occur in the array. In such cases, their corresponding output nodes should become active. This would give us *an object "noticer" and -identifier*, that is, a system which notices the presence of one or several objects in the scene and identifies them correctly. We may conceive of the system as one that has innate capabilities for paying attention to objects in a scene (rather than to what other information the scene makes available) and learning about their individual characteristics in order to recognise those objects whenever they re-appear in the scene. When it looks at the scene, the system pays attention to the individual objects present and recognises them.

5. The hard step

We now have the desired system with concepts of individual objects. Clearly, the system does have semantic representations (or thoughts) just like our *right of(x,y)* system had. However, in contrast to that system the new system does not yet have concepts about spatial relationships. We would like the new system to acquire such concepts. Let us again focus on the concept *right of(x,y)*.

Imagine that there is a square to the right of a triangle in the visual array. So far, the system is merely able to realise that the array contains a triangle and a square. How might the system learn to recognise the right of-relationship between these two specific objects? We will not consider this time the handling of right of-relationships in situations where more than two objects are present in the scene. This was done in the previous system and the current neglect of this task means that the new system has a weaker understanding of the semantics of *right of(x,y).* However, this does not matter for present purposes.

Since we have done it before, we assume that it will be possible for the system to learn to recognise that the general right of-relationship *right of(x,y)* obtains between objects in the scene. For this to be the case, an output node would have to be added to the system which fires if and only if there is a right of-relationship between the objects in the visual array. If the square is to the right of the triangle, the following nodes would become simultaneously active: the triangle node, the square node and the right of node. The same happens if the triangle is to the right of the square.

This, of course, is not sufficient for the system to distinguish between the two different situations in which a is to the right of b and b is to the right of a. For very good reason, Fodor and Pylyshyn (1988) leaned heavily on this point in their argument against connectionism as a general architecture for cognition. Since our previous refutation of their argument concerning systematicity and compositionality dealt with low-level systematicity, that paper did not consider the current issue which is to do with higher-level systematicity. So, assuming the feasibility of the (three) steps already described, we face the following situation: the system is able to represent the facts that two objects are present in the visual field and that right of-ness is present in the visual field. Such a system has complex thoughts (i.e., *right of(x,y)*), low-level representational systematicity and semantic compositionality. But it still lacks the ability to represent ordered spatial relationships between specific individual objects or types of such objects. The system has some thoughts but lacks others. What it lacks *might* be something which a cat does have. A cat is very likely in a laboratory experiment to be able to distinguish between the patterns*aRb* and *bRa*, where *R* is the relation *right of*. What may be less clear is whether the cat does this through representing the ordered right of-ness relationship or in simpler ways. For instance, it might be sufficient for the cat to make sure that *a is closest to something else in the scene* (e.g., to some static part of the laboratory setup) or that *a is closest to something else in the scene and b is present* in order to obtain, e.g., food rather than nothing or even electrical shock. In other words, it may not yet be clear from behavioural experiments at this point whether infralinguistic creatures can have thoughts about the ordered right of-ness relationship between a and b. The acid test is whether such creatures can provide us with behavioural output which offers convincing evidence. Obtaining this might turn out to be more difficult than expected once the semantic intricacies involved are taken into consideration.

Like so many other relational thoughts, the thought that *a is to the right of b* has an asymmetrical trajector-landmark structure (Langacker 1987, Bernsen and Ulbæk 1992). In *a is to the right of b,* a is the trajector and b is the landmark. In *b is to the right of a,* b is the trajector and a is the landmark. A system which learns the right of-relation learns the difference between these two thoughts without explicitly learning anything about trajector-landmark structure. What the system learns is to recognise situations in which a is to the right of b as being characteristically different from, but structurally (or systematically) similar to, situations in which b is to the right of a. We want an output from the system which unambiguously tells us which of these two right of-relations hold in the scene. Moreover, the output of the system should demonstrate higher-level systematicity based on distributed representations. Finally, the system should be compatible with a modular cognitive architecture capable of exemplifying the chain linking vision with natural language. So it won't do simply to have one output node firing when a is to the right of b and a different output node firing when b is to the right of a. This solution, which is certainly feasible, would imply loosing systematicity on the output side with concurrent loss of the possibility of feeding systematic output into subsequent natural language processing modules.

6. One possible solution

A possible solution is the following. Recognising the fact which we loosely describe as "a is to the right of b" can actually be done through having (at least) two different thoughts. The first thought is the one the system has when it simply realises that the right of-ness relation obtains between two objects in the scene. That is, the system realises that the abstract relation *right of(x,y)* is instantiated in the scene: there is right of-ness, there are two and only two objects involved in that relation, their identities may be known, *but their ordering in the right of-ness relationship is not known.* Let us require, in that case, that a specific right of-node becomes active. The second thought is the one the system has when it realises *which* object is to the right of which other object in cases where the abstract relation *right of(x,y)* is instantiated in the scene. Now assume that, in order to represent this second thought, *the system uses temporality in its internal representation*. The system now literally represents the right-most object *before* representing the left-most object in its output. This mode of representation requires a shift to the temporal domain. Simply thinking that right of-ness holds in the scene is an atemporal activity or an activity which, although it may be temporally extended, does not necessarily involve a specific temporal sequencing of representations whereas thinking that an identified object is to the right of a second identified object is a temporally sequenced activity.

Empirically, it does seem to be the case that we can think or realise that there is a right of-ness relationship in a scene without thinking that this right of-ness relation involves, e.g., the two specific objects a triangle and a square. For instance, we may not be in a position to be able to identify the perceived objects. Possibly, some animals can think the first but not the second when provided with the visual input which allows humans to think both.

This would mean that in order to represent *a is to the right of b -right of-ness* so as to satisfy the systematicity and modularity constraints, we need temporality in the output domain. The system should produce a temporally sequenced output string to represent what it sees. How can this be done ?

One possibility is to have an associative process from the spatial to the temporal domain. We require that the spatial pattern which we describe as "a is to the right of b" associate with the temporal pattern: "a" followed by "right of" followed by "b". We also require that the spatial pattern which we describe as "b is to the right of a" associate with the temporal pattern: "b" followed by "right of" followed by "a", and so on for arbitrary individual object combinations recognised by the system. The additional requirement is needed to guarantee higher-level output systematicity and semantic compositionality in the system. If the system were also able to handle other spatial relations such as aboveness, it would do so by again using temporal output sequencing but this time involving a different output node representing *above(x,y).*

7. An unexpected problem

In working on the first version of such a system with higher-level systematicity in its representations, we came across the following problem. Even a static visual array containing nothing but a triangle and a square is incredibly information-rich. There does not seem to be any in-principle limitations to the number of logically and semantically irreducible descriptions we can make of this array using natural language. The same, therefore, is presumably true of the number of non-linguistic thoughts a system might have concerning the information present in the array. The problem now is that all this information is present in one and the same visual array. So if a system is not able somehow to select which information it wants to pick up, it will never get started picking up information. This may not be a problem for systems which are hard-wired to picking up only specific types of information while ignoring others and which have some way of ordering their pick-up of information in those cases where several "affordances" are present. Biological systems are such systems, but ours isn't yet. Since the information that, e.g., the triangle is to the right of the square is only present in the visual array if and only if the information that a triangle and a square are present, and since the system has no hard-wiring which allows it to somehow order its handling of the information, it cannot in its present form perform both of the two desired tasks. If it tried, it would inevitably act as an organism which is hard-wired to picking up only one of those types of information.

The system therefore is clearly missing a capability for *selective attention or focusing* which would allow it to shift its attention from looking for one type of information (e.g., whether both a and b are present in the scene) to looking for another type of information (e.g., whether a is to the right of b or b is to the right of a). Biological recognisers and identifiers which are able to do that, it seems, must do so through receiving information from somewhere else in the cognitive architecture telling them what to look for, focus on or pay attention to in a static scene. If that is true, the way to equip connectionist systems with selective attention vis-à-vis static scenes is to provide them with such additional information structures which allow them to shift their attention to different types of information in, e.g., the visual array. So we are looking for a way to represent goal structures in distributed connectionist systems determining their focus of attention. How this was done is described in the next section which presents the simulation.

8. A distributed connectionist system with higher-level thoughts

The simulation used the network Tlearn (due to Jeff Elman, UCSD). When the network is in recurrent mode there is a "copy layer" in addition to the hidden layer. The copy layer is used to copy the activity of the hidden layer at time t1. At time t2 the activity of the hidden layer at t1 is fed back into the hidden layer from the copy layer. In this way the network is sensitive to earlier input activity and is able to produce dynamic, temporally discrete output based on static input.

The network was required to:

- identify object a wherever it occurs in the scene;

- identify object b wherever it occurs in the scene;

- identify objects a and b wherever they occur in the scene and independently of specifying their spatial relationship. However, a simplifying constraint was imposed in order to limit vector space complexity so that a and b always occur with a fixed distance between them;

- identify or recognise the facts that a is to the right of b and b is to the right of a wherever they occur in the scene. Again, the simplifying constraint was imposed that a and b always occur with a fixed distance between them;

Objects a and b are two significantly different 2-D spatial objects.

To solve the focus of attention problem, extra units were added to the input layer. When these units are turned on, the task of the network becomes that of thinking which of the two spatial relations hold: *a is to the right of b* or *b is to the right of a.* When these units are turned off, the task of the network becomes that of thinking whether a and b are present. We take the on/off distinction characterising the extra input nodes as representing the distinction between two different internal goal states of the network allowing its attention to change from looking a the scene in one way into looking at it in a different way.

The input scene is a two-dimensional array consisting of 10 times 10 units with 3 extra units added for signaling changes of attention. Overall, the network has 103 input units, 100 hidden units, 100 copy units (the context layer) and 2 output units. The training was run with a learning rate of 0.3 and a momentum rate of 0.9.

Each static input scene is presented in two consecutive time slices. It is the output which is time dependent or coded serially. The coding for the presence of, e.g., a anywhere in the scene is:

t1: 1 0
t2: 0 0

Coding for a and b (and b and a) is:

t1: 1 1
t2: 0 0

Coding for b is to the right of a is:

t1: 0 1
t2: 1 0

Coding for a is to the right of b is:

t1: 1 0
t2: 0 1

In the current simulation, no output node was introduced for expressing the *right of(x,y)* relation. Since the network just recognises one kind of spatial relation, such an extra output node is unnecessary. If the network is to recognise more than one kind of spatial relation (e.g., *above(x,y)* as well), units identifying the type of spatial relationship currently attended to will have to be introduced.

The training set consisted of the exhaustive set of all possible combinations of a, b, a and b, a is to the right of b, and b is to the right of a anywhere in the scene. The simplification noted above was that a and b, when occurring together, were always two units apart. Another simplification noted earlier and evident from the setup as presented, was that the network this time was not required to generalise to *spatial object(x)* and *right of(x,y).*

The input consisted of 512 time sequences (256 different scenes) and the network was trained for 900.000 epochs. The network converged nicely according to the error samples made for every 50.000 epochs. After 100.000 epochs the error measure (the total sum of squares) was 0.002 and after 900.000 epochs it was 0.0008. The objective was to verify that the network actually converged rather than to make this happen quickly and efficiently. Convergence was actually very slow.

9. Discussion

The simulation whose background and implementation has been described above has a number of implications. One is that the ability to think that, e.g., a is to the right of b, is of a rather sophisticated nature. It involves temporality of output representation and discreteness of what is represented there, that is, it involves *a temporal sequence of discrete output representations*. And obviously, if temporal and discrete output representation is required then this has to be reflected in the nature of the system's hidden, internal distributed representations. It is apparently much simpler to think that a is present in the visual field, or that a and b are present (conjunction being symmetrical), or that right of-ness holds between two non-specific individual objects than it is to think that a is to the right of b. It might be that many lower animals are able to represent the former states of affairs and all those other states of affairs which pose similar processing requirements without being able to represent, e.g., that a is to the right of b and all those other states of affairs which pose similar processing requirements. The hypothesis therefore is that low-level thought merely requires static or at most temporally un-ordered, discrete output representations from the relevant cognitive modules whereas higher-level thought requires temporal and discrete output representations from the relevant cognitive modules.

Again, this hypothesis is not about linguistic representation. Both the linguistic representation *that a is present in the scene* and the linguistic representation *that a is to the right of b* may be temporally sequenced and discrete. But it does not follow that a cognitive architecture representing one or the other of these two facts needs the same processing apparatus to do so in both cases.

Imagine a static scene with two objects in it, a triangle and a square, the square being somewhere to the right of the triangle, and the scene being framed by the borders of the visual array. Even this very simple scene contains a wealth of information which we are perfectly able to describe in some natural language. For instance, the square is to the right of the triangle, the triangle is to the left of the square, the triangle is a triangle, there are two objects in the scene, the square, the triangle and the scene each have a number of 2-D geometrical properties, the triangle is located at a specific place in the scene having specific relations to the borders of the array, the square is close to the triangle, the borders of the array are straight lines, the rest of the scene is empty, and so on. On some specific occasion we may just look at (or imagine) the scene without extracting any particular piece of information from it and thus without having any specific thoughts about it. And even if we do extract some information, we surely cannot extract all of the information present in the scene at once. Extracting increasingly more of the information present is a temporally extended process. This process requires shifts of attention. So modelling this process requires the modelling of attention and shifts of attention.

Suppose that animals and infants automatically focus their attention on some aspects of a static scene and not others. Their attention is caught by these aspects for some reason. They may have innate pattern matching capabilities which become activated once particular objects and properties are present. An infant may look at the scene and what it sees as a result of inborn pattern matching capabilities and the resulting focusing of attention is a configuration of objects. The infant may have learnt to recognise those objects. In random order, the infant recognises first one object and then a second object. It may also realise that the distribution of the objects in the scene forms a spatial pattern which it is able to recognise. When does the infant have *thoughts* about the scene ? A sensible suggestion is that the infant has thoughts once its pattern matching capabilities are operating on the scene in order to extract information such as the information just described (cf. the Introduction above). Once it extracts such information from the scene, it thinks. The infant does not need language in order to think. It thinks of what it notices or pays attention to, and that, again, is a function of its pattern matching capabilities.

Now suppose that some of the information in the scene is more complex than the rest in the following sense: in order to pick up and represent that information, a perceiver needs more complex representational capabilities. Representation of some of the information in the scene may be possible without the system's having the ability to create a temporal sequence of discrete representations whereas the representation of other information does require this ability. In principle, at least, this difference should be measurable in performance: when a system cannot possibly represent a piece of information, it cannot possibly achieve discriminative learning which presupposes the capability of representing such information. This leads to the question: is there, on empirical grounds, a distinction to be made between animals which can learn the simple types of information described above but are unable to learn the complex information described, and animals which can learn both ? Are there animals whose representational capabilities include low-level systematicity but not higher-level systematicity ? Of course, there may be no such animals or they may be so primitive that discriminative learning experiments are impossible to conduct.

A corollary of the above assumption is that the capability to represent the simpler of the types of information described and hence of low-level thought is basic to the capability to represent the more complex types of information and hence of higher-level thought. For instance, a system cannot represent the fact that a is to the right of b without being able, as a computational mechanism, to represent the fact that a and b are both present in the visual scene.

Another consequence is that the notion of a "pattern" is ambiguous. There are patterns and patterns, even in a simple, static spatial scene. In order to be able to identify patterns in a spatial scene it is not sufficient to have a "pattern matcher" since the type of pattern to be identified has to be specified first.

It is an interesting question whether and to what extent, if the hypothesis presented above is "the only game in town", the ability to produce as output a temporal sequence of discrete representations as a precondition for representing structurally complex states of affairs provides evidence, albeit within a thoroughly distributed connectionist framework, for the Language of Thought hypothesis (Fodor 1975). This question goes beyond the scope of the present paper. But if the hypothesis does provide such evidence, then the distinction between the class of (syntactically structured) thoughts which the Language of Thought hypothesis is required to account for and the class of thoughts for an account of which this hypothesis is not required is *intra-semantical*. The distinction is one between two classes of thought or internal representation, namely low-level thoughts and higher-level thoughts, respectively, rather than a distinction between what is non-semantical and merely implementational, on the one hand, and what is semantical and representational, on the other, as claimed by Fodor and Pylyshyn (1988).

Both simple and complex thoughts as described above would seem to lend themselves directly to association with linguistic items. The output of the described connectionist system can act as input to linguistic processing modules. The type of thought characterised by higher-level systematicity would seem to lend itself to being matched with linguistic syntax. Its temporal structure can be seen to be equivalent to the predicate-argument structure of standard logic. However, there is nothing in this to indicate that the temporally structured discrete outputs from distributed connectionist cognitive modules are subject to the representational and computational inadequacies of current systems of formal logic. But whereas those outputs may not be fully representable in systems of formal logic, they may be fully representable in natural language. When those outputs serve as input to natural language processing modules, their discrete temporal ordering might be transformed into the temporal ordering of words and phrases according to the grammar of some specific natural language.

10. Perspectives

What has been said above seems compatible with the following modular cognitive architecture of the vision-to-language chain:

1. *Visual image:* A static visual scene becomes represented through a set of visual modules. This representation normally contains a wealth of information about objects and relations. We are on reasonably safe grounds with respect to this point since we have a clear idea about what should be the final output of vision, namely a visual scene which normally has some structure to it. It is what the visual "channel" allows us to see when we open our eyes. Whether and how the scene representation can be produced using distributed connectionist techniques is a very different matter.

2. *Thought:* Attention becomes focused on a subset of the information in the scene. This information is extracted using various cognitive mechanisms such as those described above with the result that the system has thoughts about information present in the scene. These thoughts are of at least two general kinds, i.e., low-level thoughts that do not need a temporally sequenced representation and higher-level thoughts that do need a temporally sequenced representation. This was the theoretical suggestion above. A simple mechanism for handling shifts in focus of attention was presented and implemented in the simulation.

3. *Linguistic processing of thoughts:* The thoughts, once created, may serve as inputs to a language-producing module (or a set of such modules). Here, distributed connectionism is on safer ground than in the case of vision. It is known that distributed connectionist systems using non-symbolic microfeatural representations can exhibit (non-classical, non-syntactic) compositional structure which can be manipulated by structure sensitive operations *once they are fed with spatially or temporally structured, discrete linguistic material*. Such input can come from the reading of written words and sentences and possibly also from listening to spoken language (although connectionist systems still perform rather badly in speech

recognition). Reading has temporal structure and written words and sentences have discrete spatial structure (Sharkey 1991). Non-symbolic microfeatures are different from symbolic microfeatures in that they are individual elements that are not semantically interpretable without participating in further processing (Hinton 1981). For instance, discrete spatio-temporal input structure can be processed by simple recurrent nets like the one described above using memory for previous input information to be combined with the current input information through a feedback loop. Such nets can represent abstract grammatical structure (Elman 1989). They are "functionally compositional" (van Gelder 1990). This means that they offer general, effective and reliable processes for (a) producing an expression given its constituents, and (b) decomposing the expression back into those constituents. The representation therefore, though seemingly unstructured, carries structural information. And the representation allows structure sensitive operations such as passivisation to be performed without recourse to discrete representations (Chalmers 1990).

Or, much more simply, a thought like "the triangle is to the right of the square", and assuming that it has been discretised and temporalised as described above, may become associated with the appropriate English words and uttered as synthetic speech lacking somewhat in proper English syntax. Since the thought itself already has discrete order and semantic interpretation, it does not need further transformations. Chalmers' ideas might be used to apply structure sensitive processes to the thought in order to *infer*, e.g., that "the square is to the left of the triangle".

4. *Language understanding:* Finally, mechanisms are needed which take temporal sequences of words as input and produce either (a) structure sensitive inferential operations as just mentioned, or (b)*visual mental models of the linguistic input*. In case (a) the system needs, first, to transform a sequence of words into a thought. This is the reverse of the process described above. Second, the system has to perform structure sensitive operations on the thought in order to produce a new thought. In case (b) the system needs, first, to transform a sequence of words into a thought. Second, the system has to use the thought as input to the module which produces focused representations of visual scenes in order to create a focused visual scene which corresponds to the thought. Again, this is the reverse of the process described above.

References

Bernsen, N.O. and Ulbæk, I.: Two games in town. Systematicity in distributed connectionist systems. *AISBQ Special Issue on Hybrid Models of Cognition* Part 2, No. 79, Spring 1992, 25-30.

Chalmers, D.J.: Syntactic transformations on distributed representations. *Connection Science* 2.1, 1990.

Elman, J.L.: *Representation and structure in connectionist models*. TR 8903, CRL, Univ. of California, San Diego 1989.

Fodor, J.A.: *The Language of Thought*. New York: Thomas Y. Crowell 1975.

Fodor, J.A. og Z.W. Pylyshyn: Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition* 28, 1988, 3-71.

Hinton, G.E.: Implementing semantic networks in parallel hardware. In: *Parallel Models of Associative Memory* (eds. G.E. Hinton and J.A. Anderson) Hillsdale, NJ: Lawrence Erlbaum 1981.

Langacker, R.W.: *Foundations of Cognitive Grammar*. Stanford CA: Stanford University Press 1987.

Sharkey, N. E.: Connectionist representation techniques. *Artificial Intelligence Review* 5, 1991, 143-67.

van Gelder, T.: Compositionality: a connectionist variation on a classical theme. *Cognitive Science* 14, 1990, 355-84.