How we might be able to understand the brain

Brian D. Josephson
Department of Physics
University of Cambridge
Cambridge, UK
http://www.tcm.phy.cam.ac.uk/~bdj10

Abstract:

Current methodologies in the neurosciences have difficulty in accounting for complex phenomena such as language, which can however be quite well characterised in phenomenological terms. This paper addresses the issue of unifying the two approaches. We typically understand complicated systems in terms of a collection of models, each characterisable in principle within a formal system, it being possible to explain higher-level properties in terms of lower level ones by means of a series of inferences based on these models. We consider the nervous system to be a mechanism for implementing the demands of an appropriate collection of models, each concerned with some aspect of brain and behaviour, the observer mechanism of Baas playing an important role in matching model and behaviour in this context. The discussion expounds these ideas in detail, showing their potential utility in connection with real problems of brain and behaviour, important areas where the ideas can be applied including the development of higher levels of abstraction, and linguistic behaviour, as described in the works of Karmiloff-Smith and Jackendoff respectively.

Keywords: nervous system, brain modelling, language, hyperstructure, representational redescription, emergence.

How we might be able to understand the brain

Brian D. Josephson
Department of Physics
University of Cambridge
Cambridge, UK
http://www.tcm.phy.cam.ac.uk/~bdj10

1. Introduction

Current neuroscience deals with relationships between brain and behaviour in a piecemeal manner, using experiment or computer modelling to relate specific neural circuits to specific cognitive functions. While much detailed information concerning basic cognitive functions has been gained in this way, these achievements throw little light on more complex capacities such as those involving language, which involve a number of comparatively elementary processes working together in a coordinated way that is hard to understand in terms of information gained from the usual kinds of experiment or computer-simulation based models. Such approaches find themselves in a situation similar to trying to understand the workings of a complicated computer program without knowing the source code that describes the logical structure underlying the program s behaviour.

The essence of the problem is that the brain, as normally conceived, is a system whose behaviour is in practical terms unanalysable except in comparatively simple cases. It might be said that in the nervous system case it is difficult to see the trees (the logically significant forms) for the wood (the complicated totality): a number of such trees have been discovered, but we are not very clear what they look like and are confused by the complexity. The case of linguistic processes is of interest since here much is understood, on the basis of linguistic studies, concerning the logic, but at the level of the phenomena only. Such descriptive understanding is largely divorced from an understanding of the underlying *mechanisms*; linguists focus on modelling the phenomena and ignore the neural mechanisms, whilst conversely neuroscientists treat language essentially as a phenomenon manifesting in particular neural networks, and take little account of the details of linguistic theories.

This paper addresses these problems on the basis of a new foundation, involving a specific set of hypotheses as to the basic form and logic of the nervous system design. The concepts on which these hypotheses are based include the hyperstructure concept of Baas (1994), the dependence of design on models and abstractions (Josephson 2002), and representational redescription (Karmiloff-Smith 1992). In the following, it is shown how putting these ideas together leads to an explanation in principle of how cognitive abilities such as those associated with language can arise from a structure such as that of the nervous system. In contrast to other approaches, the specifications associated with the present approach are sufficiently constrained by the phenomena to be understood that (assuming the correctness of the basic concepts of the proposed scheme) one can envisage the details being discovered over time through a combination of experiment and analysis, just as details are discovered for other biological processes, leading to a full account of the processes underlying linguistic and other advanced cognitive skills.

2. Basic scheme

We assume in the first place that the nervous system can be characterised as a structured hierarchy whose elements are of well defined types, each type being associated with specific types of behaviour. To make this characterisation one that can be precise rather than merely qualitative, we assume also that each type has associated with it a specific abstract or formal model, in terms of which the behaviour associated with an element of that type can be understood, these assumptions being merely the translation into the context at hand of how design can viewed in general, expressed in formal terms. In the case of the nervous system it is normally not considered that such precise accounts are possible, but our hypothesis that in fact, given sufficient insight into the question of what models might be appropriate in this context (as will be discussed later), the kind of scheme proposed does apply. It is relevant to note that a structure of the kind proposed, involving a hierarchy of systems of specific types, has been found appropriate for designing computer programs required to function reliably in complicated situations, the typed systems in this case being the objects of object-oriented programming.

To apply this scheme to the actual brain, we identify the typed abstractions of the scheme with processes that can be regarded, to a first approximation, as functionally separate and approximately autonomous, such as those of balance and hand-eye coordination. The physical systems to which the specialised models refer involve the neural circuits relevant to the specified type of activity together with the environment in which they function. Intuitively, there is a specific (and specifiable) mechanism for balance, a specific mechanism for hand-eye coordination, and so on, and this is what the relevant abstract model refers to. As far as the neural-circuitry component of the model is concerned, in many cases this need not be a model in the form of a neural network, but instead a signal processing model involving an appropriate mathematical transformation between input and output signals. In the case of learning, in many cases the model can simply dictate that a consistent relationship between input and output signals, determined by trial and error to be appropriate for the successful performance of some task, be learnt , meaning that the system, after learning has taken place, applies the relevant functional relationship directly rather than having to determine the outcome by trial and error.

Each model, to be complete, needs to contain parameters that may vary, in order to take account of the variation of the details of the execution of a given task from one context to another. In general, therefore, there will be many systems for each model. As with the objects of object-oriented programming, the models need to describe not just a single process but in general a complex of processes, taking into account the way they may interfere with each other, as well as taking into account the mechanisms of learning.

3. Hierarchical aspects and hyperstructures

The thinking behind what has been proposed is that the nervous system, while very complex in itself is, from a logical point of view, constructed from comparatively simple components, and works in the way that it does as a consequence of the laws governing the behaviour of these components, which laws are in principle accessible to separate investigation and analysis. This somewhat cautious statement of the logical dependences involved anticipates our use of the hyperstructure concept of Baas (1994), which we now discuss. Normally one infers directly from behaviour at one level of a hierarchy to the next level up. Baas characterises this as *deducible emergence*. He defines in addition a process that he calls *observational emergence,* where a range of configurations are explored, with the aid of a specialised *observer mechanism*, in a search for one exhibiting some specified target behaviour. If and when the target behaviour is attained, this behaviour is learnt, in the manner previously discussed, i.e. new connections become established which permit immediate computation equivalent to the behaviour discovered earlier by trial and error. If the links concerned take time to become fully established, the computation concerned can be further tested, with the links fully established only if the tests are successful. In the context of development, a mechanism such as described does not need to know precisely what to do to obtain a target result, but can take advantage instead of knowing what field to explore in order to

be likely to achieve a target result. This choice of field to explore is one aspect of the model, and has its correlate in the physical system architecture.

It has previously been noted that the details of processes are typically context dependent, which requires a mechanism whereby different systems can be activated in different contexts, in conjunction with processes defining a context in suitable ways. In addition, the learning process can be enhanced by means of mechanisms that encourage remaining in a particular context or task situation long enough for the details of the relevant processes to be properly established.

We can characterise the hyperstructure processes as building up a hierarchy of resources, each stage providing resources needed for the next level. A process such as learning to walk can fairly obviously be seen as a process involving the generation of resources of specific types in specified ways. One of the principles involved can be characterised as a moving target principle in that the target is initially a simple one such as being able to stand up in balance, but when this target has been accomplished a new target comes into view, such as taking a step. Later on we will be discussing how processes as language acquisition can be structured in similar ways.

Such a regulated development, with each stage limited by the demand that specific targets be met, seems more likely to be able to achieve an outcome such as human language than the anything can be achieved just by adding a few more layers to the neural network concept of some network modellers (cf. Quartz and Sejnowski 1997). It also provides in principle (subject to the characterisation of the types of systems and the observer mechanisms causing one level to emerge from another) a more precise specification of how language could develop than the approach of Arbib (2000) which argues that language is the anticipated outcome of combining a certain collection of skills but is unclear about how the combination of skills comes about. In the present picture, everything is implicit in the collection of abstract descriptions, which are similar to a collection of instructions for building some multi-level structure, with diagrams of how the various parts fit together at the various levels. In other words, the physical nervous system is a realisation of an abstract scheme which is itself composed of many interrelated parts, somewhat in the way that a computer program is a realisation of the specification provided by its source code in a high-level programming language.

A variant mechanism for creating hierarchical structures involves problem-solving processes, again assumed to be associated with specialised abstractions, invoked when a process taking place meets some kind of obstacle. This kind of process differs from the hyperstructure processes in being driven primarily by environmental factors rather than developmental mechanisms. Dealing with obstacles can be facilitated by an equivalent to the *frame* mechanism used in computer programming; this involves the creation of a frame that represents the details of the problem, and the situation to which return will be made.

4. Role of the architecture

The role played by the neural architecture is implicit in the above account, and in this section we make its role explicit through examples. An example of a place where the architecture is relevant is in the idea that search processes can owe their efficiency in knowing which *field* to explore for systems to use. One device for achieving such selective investigation is to have a scheme whereby particular neural systems are assigned particular roles, and to arrange for some signal to be present which can make the relevant systems selectively available, while some other signal picks out a subset of these systems specific to the current context. More generally, any element of an abstraction stating which type of entity is required for a specific application can be implemented by an architecture that effectively assigns specific types of system to specific roles within the collection of abstractions.

5. The observer mechanism and the coordination of subsystems

Our scheme implies the presence of a large number of component subsystems, acting in accord with a range of abstractions. In the present scheme the coordination between the parts is ascribed to the observer mechanism in conjunction with the hierarchical organisation, or to problem-solving processes. Specifically, the observer mechanism tests the performance of some combination of systems, and fixes the relevant information only if the combination achieves some specified target. In a variant of this process, an abstract specification of a process (as discussed later in connection with representational redescription) is tested instead, and in a further variant, structures are created by rule and the rule tested (or it may be supplied by another person using language and assumed to be valid).

In addition, aspects of the coordination may be innate and implicit in the hierarchical organisation, for example in assigning priorities to particular activities in particular contexts. Or again, we can envisage the existence of a level of integration higher than that of a specific plan, involving an abstraction that may be thought of as a life-scheme.

6. Evolutionary aspects

There are many possible consistent constructions whose behaviour is consistent with survival, just as there are many configurations for the body that are consistent with survival, with the different constructions relating to different capacities. During the course of evolution new constructions, associated with particular abstractions, develop, associated with new capacities. One may regard abstractions as a source of power in that they enable particular things to be achieved. For example, naming things is an abstraction, associated with the existence of a connection, which may be utilised in either direction, between a name and a thing named. Naming things opens up new possibilities, whose realisation requires neural circuitry that can realise the processes concerned.

7. Application to language: preliminary concepts

The processes of language are more complex than mere naming but, according to the present proposals, arise in the same general way, on the basis of a range of abstractions concerning language (of which the relationships involved in naming form one), with corresponding processes. According to the observational emergence idea, specific systems are developed in accord with specific tests. The proposals of Jackendoff (2002), who discusses the various representations that appear to be involved in language, as well as also paths by which language could have evolved from something very basic to modern language, are consistent with this general picture, since we can consider each advance as involving a new abstraction. This will be discussed in some detail later on, but first we discuss a key concept that arises in connection with any skill involving the use of abstractions, namely that of *representational redescription*.

8. Subtler levels of behaviour: representational redescription

A concept relevant to understanding phenomena such as language is the representational redescription concept of Karmiloff-Smith (1992), the basic idea of which is that having become competent at a given process we re-represent our knowledge in a way that opens up more possibilities. An example is provided by the abstraction cat , the use of which allows one to represent direectly knowledge relating to cats in general, rather than having to treat each instance of a cat individually.

Suppose, in the present scheme, there is a certain type of system corresponding to the abstraction classes of objects . One instance of this type would consist of our example, the class of cats, which we may for the sake of argument suppose comes into existence by some built-in mechanism through exposure to a particular cat. We can build up a more complex system by creating, whenever an instance of a cat is encountered, a system combining and linking together appropriately, in a way goverened by the various applications that there may be of the class abstraction, a system for the particular cat with the system for the cat class. Such specific applications might include recognising a cat as a being a member of the class of cats (and activating the system corresponding to the class), or the process searching for a cat. There are a number of generic uses of abstractions, demanding a collection of different types of functional circuitry. In accord with the concept of observational emergence, complexes of a given type are formed only when the appropriate target condition is satisfied.

What has been discussed above involves the use and integration into activity of isolated abstractions (e.g. just cats, not relationships between cats and anything else) only. Once connections have been established between higher levels and the level of activity, *combinations* of higher level constructs with useful translations of these constructs into the level of perception and action can be created, such as assertions as to the existence of relationships between elements of a class (an abstract connection with concrete correlates). There may be a *general mechanism* to create such combinations, but one can anticipate also the existence of specific circuits for handling a range of particularly useful forms, such as anticipation and making models. Planning involves processes where activity involving abstractions and concrete activity are completely separated in time, which we hypothesise is built up step by step in a process analogous to going from walking with considerable support to walking with partial support to walking with no support (an example of the moving target concept), the equivalent to support in the planning case being confirmation of the validity of a plan by testing it against the reality. Information and rules can be tested initially in cases where the planning operations are carried out during an action and compared with activity without affecting it. Next, planning processes can be carried out separately from action, but in a situation where the outcome of planning can be tested immediately. The next stage again is where a plan is prepared in a situation that is not immediate, using signs to represent the situation concerned, further filtering being possible favouring the processes that are most reliable. Language permits yet another stage of sophistication, involving communal knowledge and practices (which may be an important factor structuring the life-scheme system postulated above).

A interesting issue arises here in connection with processes such as these. The success and reliability of, for example, a planning process is dependent on the validity of particular laws in the given context (as applies even in the context of simple processes such as taking a step). In other words, a given system applies specific rules in a given context, and the system has to be adjusted to give rise to the desired outcome. The point is that it is a good strategy in testing a generic mechanism to try out simple cases where success or failure can be quickly determined and improvements based on feedback made and tested quickly. Thus we envisage that much of intellectual development is based on systems that try out general types of rules, and quickly learn to discard the majority, leaving a residuum that can be characterised as rational.

One point needs amplification in regard to the above, namely the relationship between activity and knowledge (equivalently cognitive structure), the former being a process and the latter a state related to the process. Cognitive structures are built up, already noted, by linking together systems concerned with processes, and these links subsequently allow the process concerned to be recreated. Applying this concept to the case where the activity concerned involves abstractions rather than overt activity, we see that activity at the abstract level, which can be characterised as thought can, subsequent to a learning process related to some target being achieved, become a structure capable of regenerating the same thoughts. Such a process is the equivalent in the present scheme of the creation of semantic structures representing relationships in the discipline of artificial intelligence. Language can now be seen as a special process that connects with such structures in a very systematic way, allowing them to be manipulated at will to achieve particular ends and transferred between individuals.

9. Language

The outcome of representational redescription is the development of a partly separate abstract level of activity, mirroring activity itself (in accord with appropriate correspondence schemes), and thereby supportive of it, instantiated by its own growing collection of systems specialised for such activity. These higher level systems take a particular abstract form, and search for valid content consistent with the demands of that form. Language can then be characterised as an extension of this redescription process, a further derivative level of activity supportive of existing levels. The present approach leads us to frame the question in terms of a collection of interrelated model systems, within which the processes of language are defined. Language can be thought of as an idea made up out of many ideas, the workings of which we may, as research workers, come to understand through our analyses of how language works. Nature instead simply discovered empirically, through the mechanism of natural selection, processes implementing the ideas concerned.

Jackendoff s proposals concerning the evolution of language provide a good starting point for the application of the present approach. He argues provisionally for a particular sequence of strategies through which the present-day form of language could have evolved, examples of which are the use of symbols in a non-situation-specific manner, the use of symbol position to convey basic semantic relations, hierarchical phrase structure, and a system of grammatical functions to convey semantic relations. In addition to these specifically linguistic functions, imitation plays an important role in making language system to a first approximation a shared system rather than one deriving from an individual.

All of these processes have to be implemented by physical hardware. Proposals indicative of the general form of the architecture, including specification of the various data types involved in language, are suggested by Jackendoff, but the present scheme can support much more detailed specifications, involving a specification of all the processes associated with a particular abstraction, together with the hyperstructure and frame-creation mechanisms for creating new structure. This paper will not attempt an exhaustive analysis, but rather focus on a few important themes.

We consider what are essentially prelinguistic stages, where the fact that language makes reference has not entered into the picture. Perception and imitation are the relevant themes here. As emphasised by Arbib (2000), mirror systems play an important role in imitation, representing the perceived behaviour of others in a format suitable for making a copy of the observed behaviour, which representational system as far as spoken language is concerned can be identified with the phonology box in Jackendoff s scheme. However, there is no reason for this component of the system to be linked exclusively with speech, since imitation is equally relevant for written language and for sign languages. Each modality must have its own forms of representation in the mirror system, but the target process viz. a match between the product of one s own activity and the perceived activity of another, is the same in every case. This well illustrates our theme that it is the abstraction that counts: imitation involves a common system (as well as systems dependent on the modality), which may be used in a standard way in higher level activities. A typical use of the phonological system (or equivalent in other forms of language) is the development of systems to represent language in the standard forms of the language, i.e. words or other standard units.

One of the targets of the language learner is to build up such a system for representation and imitation. The move towards use of these representations in connection with reference and meaning provides an example of the moving target concept, the scenario addressed moving beyond perception and imitation to the connections between signs and their referents in a communicative context. While a linguistic sign could function as a trigger for action, the link is more typically indirect, which we interpret as saying its effects go via an abstract level of representation, allowing greater flexibility in meaning, in particular Jackendoff s non-situation-specific use of signs. The hyperstructure scheme demands that any tentative links, e.g. based on observing correlations, be tested by usage, so that for example when a familiar linguistic unit is perceived the corresponding referent system is activated, confirmation being given by some positive consequence of this activation, this being the

main scenario assumed to be implicated in learning sign-referent connections and determining the design of the neural systems involved.

As in previous cases, the target moves on to more advanced possibilities once a given kind of target has been achieved. The basic scenario just discussed is one where communication involves single signs, which we identify with the instances of a specific type of system (lexicon, say). The abstract system on which the hardware is assumed to be patterned would be superseded by a series of alternative scenarios, based on particular abstractions, extending the range of usability of language. One extension is straightforward, involving performing the elementary sign generation process more than once. This leads naturally to scenarios where more than one sign is involved, but without any specific constraints on the order. This basic form opens up certain possibilities for more advanced variants, including one mentioned above, involving the use of symbol position to convey basic semantic relations. The way such a capacity can evolve is for it to be naturally present in an earlier stage in evolution in a very inefficient form, in which case circuits may evolve especially adapted to the performance of the task concerned in an efficient manner; in other words, a new abstraction starts to have an important influence on the design.

We move on now to consider the role played by syntax in the language system; which abstraction might be relevant here? A general feature of language is the relationship between syntactic form and semantic form, but we cannot simply postulate a general ability to relate the two, and neither does Arbib s (2000) complex imitation human skill especially help mediate the syntactic aspect of language. A more appropriate mechanism invokes the ability to handle competently two things happening at once: for example with a speaker discovering that an existing process fails because there is no word corresponding to some construct that is being represented verbally (e.g. an object needs to be indicated by more than a single lexical item in order to avoid ambiguity), so starting a new activity while keeping the other on hold. The listener, correspondingly, may be able to detect that two distinct processes are occurring and again try to keep track of what is happening.

We may tentatively hypothesise an abstraction conceptual unit , associated with a specific system that is active when the concept concerned is active, as an explanatory mechanism, such systems being aspects of a general information management process, the functional equivalent of a set of boxes in which information may be stored temporarily (there being possibly a connection here with the mechanism for pre-planning a grasping action, regarded by Arbib as a precursor of a component of the language system). Roughly in accord with Jackendoff s parallel architecture involving three systems, phonology, syntax and semantics, such a manipulation system may be work together both with a grammatical unit such as a phrase, and with a node in a semantic structure. In the process of attempting to connect language and meaning, such a unit could assist by forming provisional links with units in the systems that are being connected. Developments of the language system could investigate and later utilise processes combining units of this kind in various ways, thereby building up structures of organisation equivalent to syntactic structures.

Note that Arbib s hypothesised complex imitation does not feature directly in this account, which is better considered as referring to an alliance between a system for simple imitation and a system for managing concepts, the essential mechanism for developing complex language being manipulation or management abilities applied to the different problem of connecting linguistic structure with meaning. We will assume, without attempting to justify this assumption, that training mechanisms using such manipulative mechanisms can lead to the ability to interconvert between linguistic utterances and semantic structures in the way characteristic of language. The concept of support by an environment, referred to in earlier sections, is relevant in connection with discovering relationships between language and meaning in a language system: just as one s ability to create a viable plan can be assisted by testing the plan against the reality, one s ability to decode linguistic messages can be assisted by the referents being ready to hand to act as confirmation that one is applying the correct process.

Next we discuss the typing (e.g. noun phrase, verb phrase) characteristic of universal grammar descriptions of language. In combination with the above discussion we identify this with processes that assign types to the

nodes of syntactic structures. Such a mechanism could arise if a pre-existing system assigning types to elementary signs became allied with the system concerned with conceptual units. As is well known, grammatical types do not necessarily correlate with semantic categories, so we may infer that evolution (though possibly the evolution of languages rather than biological evolution) found it more useful to work with a model where the inherited types had the role of conventional markers rather than indicators of meaning, which can be indicated in other ways.

Now a comment on what is special about human use of language. As far as we know, the content of human discourse is special in that it is not tied to the immediate situation. This we can interpret as a development from an enhanced ability to recall past events, stabilised by applying linguistic processes to the recollections and integrating them into plans. The use of language to make present events that are not present appears to depend both on the neural mechanisms underpinning the processes indicated and a cultural practice of applying linguistic capacities in such a way.

This section will close with a general perspective on language. We can think of a language system as having power, in the same kind of way that an axiom system has power related to what can be proved within that system. There may well be no general theorems as to what language can do; rather, users constantly improve language by discovering ways to minimise confusion and ambiguity, extending the usages of existing vocabulary and creating new vocabulary where necessary. Users also discover where the limits of using their languages lie and keep within these limits, shaping existence so that language remains an ever-ready tool.

The above discussion of language has been very speculative, and is intended mainly as an indication of how the present approach, with its emphasis on specific abstractions, and systems that make use of these abstractions, can be applied. Systems whose design follows these principles are already attuned to a range of natural regularities, some actual and some possible by means of creative processes, and can develop rapidly in the corresponding directions.

10. Conclusions

The above is not a theory of the brain, but rather a framework upon which, it is proposed, theories to address specific issues can be built. The framework is of a very abstract character, involving a number of ideas rather unfamiliar in the field. However, the validity of each individual concept, e.g. observational emergence or representational redescription, argues for the validity the scheme as a whole, though this does not in itself prove that the ideas apply to the actual nervous system. What the scheme can do is help to impose a clear structure on a very unclear problem, in the same sort of way that concepts such as molecules, binding sites, chemical reactions, catalysis and so on impose a structure for the understanding of the processes of the organism. What can be demonstrated on a purely qualitative basis as in the present paper is limited, and it is to be hoped that those equipped to do so will apply these concepts to begin the task of structuring the complexity of the nervous system and its functions in a more rigorous manner than has been possible in the present work.

Acknowledgements

I am grateful to Prof. Nils Baas for discussions of the hyperstructure concept, which plays an important role in the above thinking.

References

Arbib, Michael (2000), The Mirror System, in *Imitation and the Evolution of Language*, in Imitation in Animals and Artifacts, Nehaniv, C. and Dautenhahn, K. (a more recent version of Arbib s ideas can be found in a commentary document for Brain and Behavioral Sciences, at http://www.bbsonline.org/Preprints/Arbib-05012002/Referees/)
Baas, N.A. (1994); *Emergence, Hierarchies and Hyperstructures*; Artificial Life III (ed. C.G. Langton, Addison-Wesley (pp. 515—537).
Jackendoff, R. (2002) *Foundations of Language*, Oxford, Oxford.
Josephson, B.D. (2002) *Abstractions and the Brain*, SEED **2**, 28—35
(e-print at http://www.library.utoronto.ca/see/SEED/Vol2-2/2-2%20resolved/Josephson_abstract.htm)
Karmiloff-Smith, A. (1992); *Beyond Modularity: a Developmental Perspective on Cognitive Science*, MIT.
Pinker, S. (1994); *The Language Instinct: the New Science of Language*; Penguin
Quartz, Steven R. and Sejnowski, Terrence J. (1997) *The neural basis of cognitive development: a constructivist manifesto*, Behavioural and Brain Sciences **20(4)**, 537—556.