Context-free versus context-dependent constituency relations: A false dichotomy

Francisco Calvo Garzón

Department of Philosophy, University of Murcia
Campus de Espinardo
Murcia, 30100 - Spain
Email: fjcalvo@um.es

Abstract

In this paper I articulate a notion of constituency orthogonal both to classical and connectionist approaches. I shall consider three "structure-in-time" connectionist networks (Simple Recurrent Networks, Long Short-Term Memory Models, and Non-classical Connectionist Parsers). I shall argue that explaining compositionality by means of any of these models drives us to an information-processing blind alley. In my view, human combinatorial behaviour must be grounded in sensori-motor activity and the parameter of time. A dynamical notion of constituency will thus be offered.

Introduction

Can "structure-in-time" connectionist networks represent compositionality? That is, can they implement recursive combinations of constituents? To answer this question, (i) the notion of constituency must first be properly defined; and (ii) a particular "structure-in-time" connectionist model must be chosen. One part of Fodor and Pylyshyn's (F&P; 1988) well-known challenge to eliminative connectionism states that connectionist networks can only support compositionality, and account for systematic and productive behaviour, by implementing a classical model. Unfortunately, F&P's argument has had the indirect effect of limiting the range of empirical hypothesis to be assessed in relation to what the correct architecture of cognition is. That range is bounded at one end by classical architectures (Newell and Simon, 1972), and by standard connectionist ones (Rumelhart, McClelland et al., 1986), at the other end. Being at present an open empirical question what the correct architecture of cognition is, I shall try to articulate a framework that imports a notion of constituency that differs from the range of answers traditionally offered by orthodoxy along the classical-connectionist spectrum.

Constituency relations

Connectionist researchers have rejoined in one form or another to F&P's challenge, but the basic position, namely that we do exploit a combinatorial syntax and semantics, that allow us to behave systematically, has been granted almost by all sides. The key difference is that classicism and connectionism differ in the way they represent constituency. They assume that the building blocks of thought are context-free and context-dependent, respectively. Fodor (1987) notes that the constituent 'P' in the formula 'P' is a token of the same representational type as the 'P' in the formula 'P&Q', if 'P' is to be a consequence of 'P&Q'. Mental representations are formed out of context-independent constituents in such a way that constituents appear in different thoughts as syntactically identical tokens. By contrast, in a multilayer feedforward network or in a simple recurrent one (SRN; Elman 1990), we cannot equate discrete parts of the hidden units' activation pattern with particular components of the sentences being processed. In a SRN, for example, grammatical structure is reflected by coding grammatical variations as slight dynamical deviations in the relevant activation patterns through state space. The syntactic contribution each word makes to the sentence is measured by the word's own level of activation, as encoded in hidden state space. In this way, connectionist constituency is context-dependent, and constituents appear in different thoughts as syntactically idiosyncratic tokens (Calvo Garzón 2000).

"Structure-in-time" connectionist networks

Let us consider three "structure-in-time" connectionist networks that have already been used, or may be used in the future, to rebut to parts of F&P's challenge: (a) Elman's (1990) SRN; (b) Schmidhuber's (2002) LSTM ("Long Short-Term Memory") model; and (c) NCCP (non-classical connectionist parsers) — Calvo Garzón (2004).

Simple recurrent networks Elman (1998) trained a network to answer a criticism, along the lines of F&P, put forward by Hadley (1992), and Marcus (1998). The challenge is to explain how connectionist networks can account for strong systematicity; a sort of generalization in which the network must generalize to previously unencountered grammatical roles. Elman trained a SRN on a corpus in which words had different probability of occurrence in the grammar. Although 'boy' never appeared in direct object position for any verb in the training data set, the network does predict 'boy' as a direct object after presentation of the verb 'talk-to'. The reason is that 'boy' has already been fed to the network during training in other contexts together with other human words—e.g., girl, man, woman. So, for example, only human words appear in subject position with verbs such as 'eat', 'give', or 'transfer'. On the other hand, neither 'boy' nor other human words appear in object position with verbs such as 'terrify' or 'chase'. In short, even though the network never sees 'boy' in an object position, it is trained on roles that 'boy' shares with other human words (more than it does with other types of words). This "behaviour-based similarity" between 'boy' and other human words (see Elman 1998, for the details) is what allows the SRN to generalize to previously unencountered syntactic roles. Given that the activation patterns for, say, 'girl', 'man', or 'woman' are very similar to the one encoding for 'boy', there will be a high correlation between weight changes and activation patterns for tokens of these word-types.

"Long Short-Term Memory" models Schmidhuber (2002) offers a connectionist architecture called LSTM ("Long Short-Term Memory") as a model that manages to solve one of SRNs weak points: namely, the fact that SRNs' short-term memory can sometimes be too short for the sequential task in question. A LSTM model consists of clusters of units arranged in the following way. A linear recurrent unit (with a 1.0 weight recurrency loop) sums up the incoming signals from surrounding non-linear units. Clusters of such non-linear units with linear ones in the middle join up to constitute the architecture of the network. The critical point is that while the non-linear units allow the network to keep track of non-adjacent statistical dependencies, the linear units in between can maintain a memory of any arbitrary number of time steps—see Schmidhuber (2002) for the details.

Non-classical connectionist parsers Finally, elsewhere (Calvo Garzón 2004) I have used the label "non-classical connectionist parsers" (NCCP) to refer to the class of models that have different combinations of pattern associator/autoassociative memory/competitive network topologies (Rolls and Treves 1998), with bi-directional connectivity and inhibitory competition (kWTA), and that employ combined Hebbian and activation-phase learning algorithms (McClelland's 1994 GenRec). Error-driven learning makes use of an activation-phase algorithm that, via bi-directional connectivity and symmetric weight matrices, permits the network to alter the knowledge acquired in the weights by computing the difference between an initial phase where the networks activations are interpreted as its "expectation" of what's to happen, and a later phase where the environment provides the output response to be taken as the teaching signal. Unsupervised Hebbian learning, on the other hand, makes its contribution by representing in hidden space the first-order correlational structure of the data pool. Combined Hebbian and activation-phase learning drives "structure-in-time" networks to a better performance on generalization than the backpropagation algorithm does (O'Reilly and Munakata 2000). Moreover, thanks to one-shot Hebbian learning (Rolls and Treves 1998), a few event co-ocurrences can contribute to fast recall.

The information-processing blind alley

All (a), (b), and (c) networks can represent syntactic structures in a purely exemplar-based fashion, avoiding the exploitation of symbolic resources. The models employ neither grammatical classical constituents, nor is the processing sensitive to the syntax of such constituents. The behaviour of the networks remains rooted in representations which are context-dependent. Considerations, nevertheless, that have to do with some or all of the above models should urge us to look for a different framework. For one thing, LSTM models, for example, fall short in terms of biological plausibility. Note that the network can extend its short-term memory only thanks to the exploitation of unbiological linear units. In this respect, SRNs and NCCPs would have an advantage in terms of full-fledged non-linearity. But, more importantly, humans don't behave the way LSTM networks do. We simply don't implement arbitrarily long sequential behaviour. Forgetting is precisely part of the trick for retrieving superposed context-dependent information. If compositionality is to be explained, crucially, it will have to depend on other sort of resources, than memory resources *per se*.

Unfortunately, in my view, were we to account for compositionality by means of models (a), (b), or (c) above, we'd be in a blind alley (Calvo Garzón, forthcoming). The classical/connectionist dichotomy that I'm calling into question amounts to the context-free v. context-dependent constituency relation dichotomy that the algebra v. statistics debate in Cognitive Science involves. Syntactic processing requires constituents to be context-free, and classical syntactic computationalism is manipulation of symbols according to algebraic rules. According to this classical hypothesis, algebraic rules may correspond to GOFAI algorithms or to any ecological form that respects the syntax of context-free representations (e.g., Marcus 2001). Computations in connectionist networks, on the other hand, have to do with vectorial transformations (Churchland and Sejnowski 1992). Computationalism amounts to manipulation of subsymbols according to statistical rules. As we've seen, the difference is that content in the connectionist guise is subsymbolic and constituents are context-dependent. Granting this landscape, the friends of classical orthodoxy will keep searching for cognitive abilities that, defying statistical explanations under the poverty of the stimulus lens, embarrass their foes. Rule-following sceptics (Calvo and Colunga 2003) will rejoin by finding ecological data that (1) can be exploited statistically and (2) allow connectionist networks to remain computationally adequate. Put bluntly, the connectionist's overall strategy is to show that stimuli are not so poor after all!

It is noteworthy, however, that both hypotheses, the classical and the connectionist, fall neatly within the information-processing paradigm. The architecture of cognition has been questioned, but assumptions about its computational underpinnings remain unchallenged under (a), (b), and (c) models. Perhaps we are stuck in a never-ending dialectic of positing challenges to connectionism, and then trying to account for them statistically, forever and ever (see Peña et al. 2002; and Calvo Garzón, under review, for an illustration of this dialectical sequel). In view of this scenario, I contend, we may need to consider turning to questions concerning the role that potential contenders, such as Dynamic Systems Theory (DST) (Thelen and Smith 1994) may play in the future.

Dynamic Systems Theory

Unlike information-processing-based frameworks, DST tries to model and explain the behavior of concrete systems by identifying them with sets of variables that change continually over time. A dynamical system, in this way, can be analyzed in terms of the differential equations that contain the quantitative variables whose interdependencies describe the laws that govern the behavior of the system. DST has proved extremely useful in the physical sciences (e.g., Kelso 1995; Thelen et al. 2002). Crucially, cognitive activity cannot be accounted for without taking into account the perceptual and motor apparatus that facilitates in the first place the agent's dealing with the external world. The working hypothesis is that the same mathematical toolkit of differential equations can be put to the use of describing and explaining combinatorial behaviour. My underlying assumption is thus that human combinatorial behaviour must be grounded in sensori-motor activity and the parameter of time.

There is, however, a tension manifested in DST's attempt to put forward models of higher level forms of cognition and its commitment to respecting neurophysiological, bodily, and environmental constraints. That tension, in my view, can only be resolved by keeping our dynamic modeling, ontologically speaking, at the level of its pre-94 sensory-motor origins. The line of research championed by Thelen, Smith and colleagues in motor control primarily targeted symbol-based explanations of motor activity. Those uneasy with the first steps of the dynamic motor approach would argue that the level of description of the systems was too abstract, failing thus to throw light upon implementing physiological structures. In my view, basic neuroscience should be able to understand compositional behaviour non-computationally; that is, at the neurophysiological level.

Constituency as basins of attraction

Granting this stance for argument's sake, my proposal involves the endorsement of a dynamical form of constituency. Classical and connectionist constituents would be replaced with dynamical basins of attraction. For one thing, such an approach differs from connectionism in that neural networks have traditionally tried to account for higher-level phenomena, ignoring their ecological underpinnings. The grounding (in Harnad's 1990 sense) of the subsymbols deployed by those networks is not justified, becoming merely "thinner" symbols. But there are more important differences. Connectionist networks can be seen as universal function approximators (Hornik et al. 1989). Universality which strictly computationally speaking can be sold as a virtue, may also be seen as a vice. Models (a), (b) and (c) above may be suspicious of doing to much! (i.e., of being underconstrained). By contrast, dynamical models, because of the exploitation of constraints in terms of embeddiment and embodiment, do not behave universally, but rather incorporate ecologically-grounded parametric restrictions.

In this respect, DST constitutes an interesting departure from orthodoxy. Given that the temporal dimension can be processed implicitly by means of recurrency loops, some researchers would be willing to interpret (a), (b) and (c) as dynamical networks (cf. Port and van Gelder 1995; for a review, see Elman 2004). Nevertheless, although important progress has been made in understanding temporal changes in SRNs (e.g., Rodríguez et al. 1999), connectionist theory still hasn't got an explanation of embodied real-time dynamics. In my view, no clear-cut criterion in the market allows us to decide exactly what it is that differentiates connectionist networks from dynamical systems (although see Spencer and Thelen 2003).

On the other hand, there's no guarantee in connectionist modelling that an optimal solution will be found for several different corpora. However, this is not a problem for dynamicism. Once we have the correct equations with an adequate set of variables chosen, two different corpora will be treated equally.

Multicausality and environmental reliability

It may be argued that basins of attraction are nothing but the constituents that Fodor and Pylyshyn demand, although hidden in the dynamical jargon of bifurcations, stabilities, and the like. The answer is "no". As a matter of fact, dynamical attractors cannot even be equated with connectionist ones, since they are more idiosyncratic, in the sense of being more rooted in the here-and-now. Connectionist attractors are stable in that they imply a higher level of abstraction as they lack real-time embodiment. However, note that for the connectionist, the fact that information gets recoded in hidden layers idiosycratically is fully compatible with the positing of an inner cognizer. Thus, whereas connectionist models put the stress on the hidden space trajectories understood as cognitive representational changes, dynamicism focuses on the joint contribution of brain, body and environment and, therefore, on multicausal instabilities. This should not be interpreted merely as a difference in emphasis. Under both the classical and the connectionist hypotheses, compositionality requires information to be represented *within* the system. For classicism, it is obviously so. Regarding connectionism, Elman comments with regard to model (a) above that "(the) fundamental suggestion of the present proposal is to treat words as stimuli, whose "meaning" lies in the *causal* effects they have on the mental states." (2004; emphasis added).

The present proposal implies a departure from Elman's perspective on two fronts: Dynamical embeddiment means that the cognitive system can economize resources. Information need not be present constantly within the system. It's already out there! Note that it's easier for the system to induce a certain degree of reliability in the environment; reliability that can be couched in terms of stability, for instance (although see below). On the other hand, the emphasis should be placed not on the one-dimensional cause-effect chain between input and output within the system, but rather on endogenous plus exogenous coupled multicausality.

A final caveat will help clarify a potential source of misundertandings as to what dynamical constituency implies. Although the emphasis of dynamicism is on the stability of states of the world, it would be a mistake to think of those states as the reason why *enduring* classical or connectionist constituents are not needed. Cognition must be explained without the resource of enduring states, not because some external aspect remains intact, but rather because whatever states happen to be critical for solving compositionality, they are all *reliable*; those that remain intact as well as those that do not. The picture I'm offering is one in which constituents are not hidden states that we manipulate, but rather states that change continuously as they're coupled with the environment.

Acknowledgments

The author was supported by a *Ramón y Cajal* research contract (Spanish Ministry of Science and Technology) during the preparation of this manuscript. Thanks to Esther Thelen and Josep Miró for helpful discussion.

References

Calvo Garzón, F. 2000. A Connectionist Defence of the Inscrutability Thesis. *Mind and Language* 15:465-80.

Calvo Garzón, F. 2003. Connectionist Semantics and the Collateral Information Challenge. *Mind and Language* 18:77-94.

Calvo Garzón, F. 2004. Non-classical Connectionism Should Enter the Decathlon. *Behavioral and Brain Sciences* 26(5):603-604.

Calvo Garzón, F. Rules, Similarity, and the Information-Processing Blind Alley. *Behavioral and Brain Sciences*. Forthcoming.

Calvo Garzón, F. Statistics can induce regularities in a discontinuous speech stream. Under Review.

Calvo, F.; and Colunga, E. 2003. The Statistical Brain: Reply to Marcus' The Algebraic Mind. In *Proceedings to the Twenty-Fifth Annual Conference of the Cognitive Science Society*, 210-215. Hillsdale, NJ.: Lawrence Erlbaum Associates.

Churchland, P., and Sejnowski, T. 1992. *The Computational Brain*. Cambridge, Mass.: MIT Press.

Elman, J.L. 1990. Finding Structure in Time. *Cognitive Science* 14:179-211.

Elman, J.L. 1998. Generalization, Simple Recurrent Networks, and the Emergence of Structure. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*. Mahwah, NJ.: Lawrence Erlbaum Assoc.

Elman, J.L. 2004. A Different View of the Mental Lexicon. *Trends in Cognitive Sciences* 8:301-306.

Fodor, J. .1987. *Psychosemantics*. Cambridge, Mass.: MIT Press.

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

Hadley, R.F. 1992. Compositionality and Systematicity in Connectionist Language Learning. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Mahwah, NJ.: Lawrence Erlbaum Assoc.

Harnad, S. 1990. The Symbol Grounding Problem. *Physica D* 42:335–346.

Hornick et al. 1989. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks* 2: 359-366.

Kelso, J.A.S. 1995. *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, Mass.: MIT Press.

Marcus, G. 1998. Symposium on Cognitive Architecture: The Algebraic Mind. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*. Mahway, NJ: Lawrence Erlbaum Associates.

Marcus, G. 2001. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, Mass.: MIT Press.

McClelland, J. 1994. The Interaction of Natura and Nurture in Development: A Parallel Distributed Processing Perspective. In *Current Advances in Psychological Science*. eds. Bertelson, P., Eelen, P., and D'Ydewalle, G. Lawrence Erlbaum Associates.

Newell, A., and Simon, H. 1972. *Human Problem Solving*. Englewood Cliffs, NJ.: Prentice Hall.

O'Reilly, R. 1998. Six Principles for Biologically Based Computational Models of Cortical Cognition. *Trends in Cognitive Sciences* 2:455–462.

O'Reilly, R., and Munakata, Y. 2000. *Computational Explorations in Cognitive Neuroscience*. Cambridge, Mass.: MIT Press.

Peña, M.; Bonatti, L.; Nespor, M.; and Mehler, J. 2002. Signal-Driven computations in Speech Processing. *Science* 298:604-607.

Port, R., and van Gelder, T. 1995. *Mind as Motion*. Cambridge, Mass.: MIT Press.

Rodriguez P.; Wiles, J.; and Elman, J.L. 1999. A Recurrent Neural Network That Learns to Count. *Connection Science* 11:5-40.

Rolls, E.T., and Treves, A. 1998. *Neural Networks and Brain Function*. Oxford: Oxford University Press.

Rumelhart, D.E., McClelland, J.L., et al. 1986. *Parallel Distributed Processing.* Cambridge, Mass.: MIT Press.

Schmidhuber, J.; Gers, F.; and Eck. D. 2002. Learning Nonregular Languages: A Comparison of Simple Recurrent Networks and LSTM. *Neural Computation* 14(9):2039- 2041.

Thelen, E., and Smith, L. B. 1994. *A Dynamics Systems Approach to the Development of Perception and Action*. Cambridge, Mass.: MIT Press.

Thelen, E.; Schöner, G.; Scheier, C.; and Smith, L. 2001. The Dynamics of Embodiment: A Field Theory of Infant Perseverative Reaching. *Behavioral and Brain Sciences* 24:1-86.