Connectionist Synthetic Epistemology: Requirements for the Development of Objectivity

Ron Chrisley & Andy Holland
School of Cognitive & Computing Sciences
University of Sussex

November 29, 1994

Abstract

A connectionist system that is capable of learning about the spatial structure of a simple world is used for the purposes of *synthetic epistemology*: the creation and analysis of artificial systems in order to clarify philosophical issues that arise in the explanation of how agents, both natural and artificial, represent the world. In this case, the issues to be clarified focus on the content of representational states that exist prior to a fully objective understanding of a spatial domain. In particular, the criticisms of (Chrisley, 1993) that were raised in (Holland, 1994) are addressed: how can we determine that a system's spatial representations are more objective than before? And under what conditions (tasks, training regimes, environments) do such increases in objectivity occur? After analysing the results of experiments that attempt to shed light on these questions, the study concludes by comparing and contrasting this work with related research.

1 Synthetic epistemology: Philosophy and AI/ALife

Sometimes in order to clarify the theories and concepts one would like to use to explain a natural system, it can be of great assistance to try them out on a simple, artificial system, which allows greater control and clearer analysis. Just as one might more readily come to a clear understanding of the principles of aerodynamics by studying a simple, artificial glider than by studying the particularities of the feathers and muscles of sparrows, so one might also see more readily the general structure of a proper psychology of real systems by first attempting to apply it to a simple, artificial agent.

Thus, to clarify some new ideas being proposed for the explanation of natural intentional systems, it seems a promising idea to turn to *synthetic epistemology*: the creation and analysis of artificial systems in order to clarify philosophical issues that arise in the explanation of how agents, both natural and artificial, represent the world.

Synthesis can thus be justified as an approach to understanding epistemology in the same way that it can be justified as an approach to understanding intelligence (AI), or biology (ALife):

> Artificial systems which exhibit lifelike behaviors are worthy of investigation on their own rights, whether or not we think that the processes they mimic have played a role in the development or mechanics of life as *we* know it to be. Such systems ... expand our understanding of life as it *could* be. By allowing us to view the life that has evolved here on earth in the larger context of *possible* life, we may begin to derive

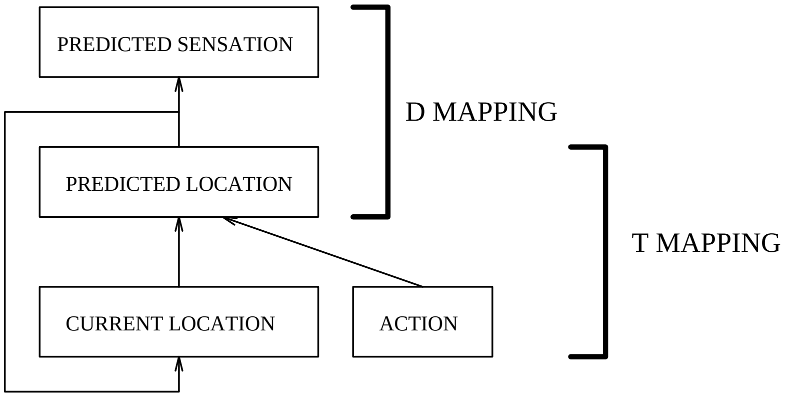


Figure 1: The PDP architecture of the predictive map (*locations* × *actions* ↦ *sensations*) formed by composing a topological mapping T (*locations* × *actions* ↦ *locations*) with a descriptive mapping D (*locations* ↦ *sensations*). Arrows indicate directed, full inter-connection between layers of units.

states) and a descriptive mapping D (from locations to sensations). In actual use, the location output of the T mapping, after a given action, is used as the location input to the T mapping for the next action.

Thus, if a constantly north-facing robot considers moving forward and then moving right, it can use the map to predict what sensations it would have after those moves by calculating $D(T(T(a,$**move-north**), **move-east**)), where a is a location representation corresponding to the robot's initial location before the actions, and **move-north** and **move-east** are action representations with the intuitive interpretation.

Given the iterative nature of the T mapping, the predictive map must be a recurrent network; in the experiments discussed here, it is implemented as a *simple* recurrent network (Elman, 1990).

2.2 The experimental setup

The experimental situation used here (roughly the same as the one used in (Chrisley, 1993) and (Holland, 1994)) is a deliberately impoverished one: a developing (learning) agent moving through a simulated "grid world"; the part of the world simulated has only 81 cells or locations (9 by 9). Each location has a 4-bit vector associated with it, which can be understood to be the sensations the agent has when at that location (see figure 4, in section 7 below). As in its normal use, the CNM is to provide a means for this agent to improve its navigation of its space (and thus increase the objectivity of its ways of representing that space) through sensory prediction.

The agent has eight actions available at any location, those of moving into each of the adjacent locations (orientation is not modeled:

2.3 Learning regime

When the agent returns home, it iteratively learns the route, not by actually moving, but by reviewing the remembered route in the following manner:

- First, generate a training set:

  1. Assume some arbitrary representation or code for the initial location ("home"). Store this code, and the code for the first action taken, as an input pattern; store the sensa-

work typically involve *fewer* states with only one or two transitions possible from or to a state, and have no notion of the sensory properties of a state that may be shared with another state, and be sensed by the agent. We believe that by adding the complexity found in the CNM world, one begins to justify talk of learning spatial representations, instead of mere arbitrary grammars. But even if that assumption is illicit, the CNM paradigm should still be valuable, at least within the finite state machine learning paradigm.

Furthermore, coarse-grained sensations actually support the intended spatial interpretation of the CNM's activity. Since there are so few (i.e., 16) types of sensory properties a location might have, the CNM cannot rely, in achieving its predictive aims, on merely recording the superficial sensory contingencies, but rather is forced to learn the more abstract spatial structure of its environment. To exaggerate this effect, we did not even let the CNM use the current sensations as an input to its predictive map, but rather forced it to use only its own representations.

It could still be objected that this, too, is unlike human cognition. It could be claimed that the way that humans and other animals achieve most of their navigation is by learning associations between actual detailed sensations, and not by developing some more abstract topological representation. That is, organisms predict what comes next by *looking* and seeing where they are.[1]

It would be a mistake to think that because we are interested in understanding how cognizers are able to make transitions from less objective to more objective ways of representing the world, that we somehow think that the majority of cognition involves representations that are at the extreme objective end of this scale. In fact, we agree that there are many kinds of cognitive interactions with the world that *require* relatively unsystematic, pre-objective ways of representing, if they involve any representation at all. Furthermore, it may be impossible for any embodied, finite system to ever achieve total objectivity or total systematicity. Nevertheless, we do think that there are interactions for which the ability to increase systematicity is a cognitive virtue, and spatial navigation is one of these.

In order to pump your intuitions concerning these matters, consider the kinds of mistakes 11990t0.67(olv)9s

> **Definition:** A system represents a location l systematically if there is a representation a such that:
>
> 1. whenever the system uses a, or a representation very functionally similar to a, it does so to represent l and not some other location l'; and
> 2. whenever the system needs to represent l, it is capable of using a, or a representation very functionally similar to a, to do so.

For the case at hand, these requirements boil down to:

> The CNM represents a location l systematically if there is a location code a such that, normally, a is active on the "current location" units if and only if the agent is currently at l.

Often, when speaking about the CNM's representations, we use expressions like "the same representation" or "different representations", when, strictly speaking, there is no such relevant issue of representational *identity*, but rather only representation *similarity*, in particular functional similarity. Thus, the above requirement is that normally all the codes a that the CNM has active on the "current location" units when at A are functionally very similar, and the CNM never has a code b, that is functionally very similar to one of the a, active on the "current location" units when the CNM is at a place other than A. Thus, there are at least three ways in which systematicity is a matter of degree:

1. the greater the number of different ways of getting to the place A that yield a code functionally equivalent to a, the greater the systematicity;
2. the greater the number of ways of getting to places other than A

a path it had never taken before. That is, form of spatial generalization occurred; the CNM was shown to be more than just a means of memorising a list of action/sensation sequences.

Such cases of the emergence of systematic codes

- $F_2(x,y) = \frac{\sum_{m=1}^{A} ||D(T(x, action_m)) - D(T(y, action_m))||}{A}$;
- $F_1(x,y) = ||D(x$

may be completely wrong in those predictions. The connection with correctness will be captured in two ways in the experiments that follow: the criterion that the net learn until it correctly predicts all sensations on its route; and the generalization that will naturally result in the cases of high systematicity.

In the experiments that follow, we give an example of the cases that justify the introduction of these functional equivalence measures: cases in which Euclidean distance/clustering would suggest a functional equivalence that is not present, and cases in which Euclidean distance/clustering would suggest functional divergence that is not present. This aspect of the research, then, can have a relatively broad application, even if one is not interested in synthetic epistemology, connectionist navigation or the development of objectivity.

5 A hypothesis concerning the requirements for the development of systematic representation in the CNM

In order to address these issues concerning the requirements for the development of objectivity, a hypothesis was formed concerning the conditions under which this style of representation will arise in the CNM, and experiments have been conducted to test this hypothesis.

Given the definition of systematic representation in section 4, the central hypothesis of this paper can be stated thus:

> **Hypothesis:** The CNM will only develop a systematic representation of a location l if its encounters with l, and with locations that resemble l, are so structured as to make such a form of representation a useful means of minimizing the error of its predictions.

The plausibility of the hypothesis is a consequence of the CNM's non-symbolic form of representation, as discussed in section 4.1. The holistic, as opposed to atomistic, nature of representation in the CNM implies that systematic representation will not be the default. Since what *primarily* determines whether the two location codes used at two different points in a route are similar is the similarity of the sensory predictions that such codes are required to produce (and not the identity of the two locations in question), the CNM will tend to violate the first of the two requirements for systematicity. Thus, it is only *likely* to satisfy the first requirement if its routes through its environment which generate its training regime are structured in particular ways.

The hypothesis itself doesn't have much force without some specifics concerning what kinds of structure the CNM's encounters must have in order to make the hypothesis true. If one prefers, one can rephrase the hypothesis into a question: what kind of spatial behaviours, if any, compel the CNM to form systematic spatial representations?

We attempted to answer this question by considering it for each of the two components of the working definition of systematic representation:

the same (or very similar) outputs on the D mapping, then there will be a tendency for it evolve weights such that the codes that are active in those two contexts, a and b, are functionally equivalent, even if the CNM is at different (albeit sensorily similar) locations in those two contexts. Thus there is a tendency to violate the first of the two requirements for systematic representation. In what situations, if any, can this tendency be overcome, such that systematic representations *are* developed?

But this is only one example of how the predictive demands placed on the CNM constrain the kinds of representations used. Another example is that making the same move at two different parts of the route will tend to produce similar codes for the location after those moves. The representational demands of a recurrent network are extremely holistic, with the "optimal" representation for the current situation being determined both by what it will give rise to in the arbitrarily distant future, and by what what gave rise to it in the arbitrarily distant past, in addition to the constraints of the present. Not only does the code that is used for the current location have to be mapped to the current sensations via the D mapping, but it needs to give rise to a code that can lead to the right predictions for the next step in the route, and it needs to be such that it can be the product of inputting the last code and action into the T mapping.

6 Principles & Predictions

To make substantive the hypothesis of the previous section, we used it to make some predictions concerning the conditions under which systematicity would and would not develop.

First, we noted four principles that we take to characterize the holistic interdependence of CNM representations (i.e., the aspects of the CNM that make it non-symbolic, as discussed in sections 4.1 and 5):

1. same inputs tend to produce same outputs
2. different inputs tend to produce different outputs
3. same outputs tend to require same inputs
4. different outputs tend to require different inputs

3. $D(a_{-1}) = D(b_{-1}) \rightarrow a = b[3]; D(a_{-1}) \neq D(b_{-1}) \rightarrow a \neq b$ [4];

4. $ma = mb \rightarrow a \neq b[4]; ma \neq mb \rightarrow a = b$ [3]

5. $D(a_{+1}) = D(b_{+1}) \rightarrow a = b[3]; D(a_{+1}) \neq D(b_{+1}) \rightarrow a \neq b$ [4];

where:

- "=" means "similar" for movement and sensation vectors, but means "functionally equivalent" for location codes; and
- "→" means "tends to make true".

Postulate 4 requires some explanation, since it does not hold unconditionally. In general, the similarity or difference of moves made from a and b has no implication in itself for the functional equivalence of the codes. But it does have implications when interacting with other contexts. In particular, if $D(a_{+1}) = D(b_{+1})$, then $ma \neq mb \rightarrow a \neq b$. This is because differences in a and b will be required in order to cancel out the differences in ma and mb in order to have a constant result.

Conversely, if $D(a_{+1}) \neq D(b_{+1})$, then $ma = mb \rightarrow a \neq b$, by principle 4. To see why, first note that principle 4 implies that $D(a_{+1}) \neq D(b_{+1}) \rightarrow a_{+1} \neq b_{+1}$. Next, note that there will be an even stronger push (via principle 4 again) for $a \neq b$ than there would be based on prediction 5 alone, since the similarity in the moves ma and mb must be compensated for by greater differences in a and b in order to achieve a comparable difference in a_{+1} and b_{+1}. There will be no special tendency produced by $ma \neq mb$.

In stating these tendencies, our use of "=" and "≠" suggests that we are once again assuming either completely equivalent or maximally different description vectors. But in fact, the relevant description and movement vectors may be more or less similar or different. These differences should affect the functional equivalence of the relevant location codes accordingly, but given a random distribution on sensation vectors and moves, we believe these additional modifying factors can be ignored in our analysis.

In light of these postulates, we defined 7 (non-exhaustive) types of route, or scenarios, that we thought might generate a large variation in the degree of systematicity of the representations the the CNM develops for two locations that are sensorily equivalent. The situations are listed in figure 2.

Using the five principles, we predicted the following rough ordering of these situations with respect to the degree of systematicity that they impose on the CNM's representations for the two locations, from most systematic to least:

SIDO These scenarios should yield the best systematicity, since because functional divergence between the codes for different places is fostered by exploring the different sensory surround of the two locations, yet each of the two locations is entered via a constant approach, providing a basis for the development of very similar codes for the same place. Within this group SIDOD should be more systematic than SIDOS, since the differing ways in to the two locations will add the the divergence between their location codes.

DIDO This should be next best with respect to systematicity, because although the lack of a common approach to the locations will yield a divergence between the codes used for the same place, there will be a greater divergence between the codes used for the two different places, due to the exploration of their different sensory surround.

DIDO : Different ways in, different ways out. The route that the CNM takes approaches each places from several different directions, and leaves from each place in several different directions.

SISO : Same way in, same way out. There are four possible sub-cases:

SS both the single direction in and the single direction out are the same for the two locations

SD the single way in is the same, but the single directions out are different for the two locations

DS the single ways in are different, but the single direction out is the same for the two locations

DD both the single ways in and the single directions out are different for the two locations

DISO : Different ways in, same way out. For each location, the CNM's route approaches from several different directions, but always leaves by the same direction. There are two sub-cases:

S the single way out is the same for both locations

D the single way out is different.

SIDO : Same way in, different ways out. For each location, the CNM's route approaches from one direction only, but leaves by several different directions. There are two sub-cases:

S one in which the single way in is the same for both locations, and

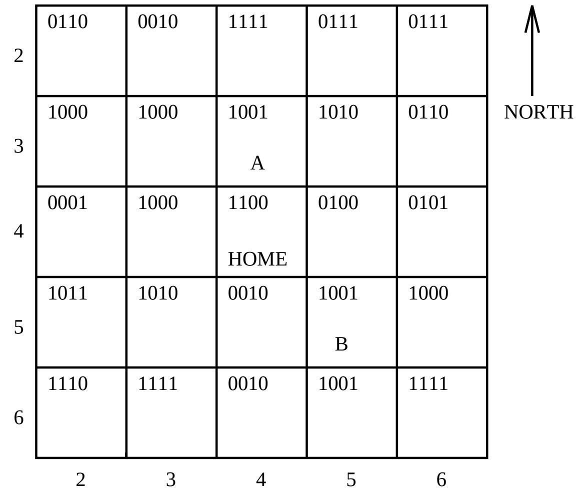
D one in which the single way in is different.

Figure 2: The classification of routes used in the experiments.

SISO These should yield poor systematicity, due to the lack of exploration of the two locations' different sensory surrounds. However, SISODS and SISODD should be more systematic than SISOSS and SISOSD, since the single moves in are not the same between the two locations, thus causing *some* functional divergence between the codes for the two places. SISODS should be slightly more systematic than SISODD, and esthe

1. **DIDO:** N; W; S; SE; E; N; N; W; S; E; S; S; NW; N; NW; E; N; S; E; SE; S; W; W; SW; NE; E; E; NW; W.
2. **SISOSS:** N; W; S; SE; SE; N; W; N; N; W; NE; SE; SE; S; SW; N; W; NE; W; N; W; SE; S; SE; N; W; W; NE; N; W; S; S; SE; E; N; W; N.
3. **SISODD:** N; E; S; S; W; NW; E; N; E; E; S; W; S; W; S; NW; NE; N; E; SW; E; S; W; N; N; E; S; S; W; N.
4. **DISOS:** N; W; SE; E; S; W; NW; N; E; W; SW; SE; E; E; W; NE; N; W; W; S; SE; S; E; N; W; N; NW; NE; S; W; SE; E; SE; W; W; N.
5. **DISOD:** N; E; S; S; N; N; W; E; SW; S; E; N; N; NW; SW; E; E; SE; S; W; N; W; NW; NE; S; E; S; SW; SE; N; N; W.
6. **SIDOS:** S; E; N; W; NW; E; N; SE; SW; S; E; E; NW; W; NW; E; E; SW; S; E; S; NW; N; NW; E; S; S; E; W; N; NW; E; W; SE.
7. **SIDOD:** S; E; E; NW; W; N; W; SE; S; E; S; NW; N; N; E; S; SW; E; W; N; N; S; S; E; N; W; N; N; SW; SE.

Figure 3: The route types used in the experiments, and the particular move sequences that realized them



2	0110	0010	1111	0111	0111
3	1000	1000	1001 A	1010	0110
4	0001	1000	1100 HOME	0100	0101
5	1011	1010	0010	1001 B	1000
6	1110	1111	0010	1001	1111
	2	3	4	5	6

Figure 4: The region of the grid world used in the experiments. The four-bit binary vector at each location indicates the description or sensation vector associated with that location.

7 Experiments & results

To test these predictions, we had the CNM learn 7 routes, each route realizing a different route type (see figure 3). The particular environment that was used is shown in figure 4. The CNM converged on a solution with no errors within, on average, 16330 epochs of training.[6] The learning rate was 0.01, and the momentum was 0.5.

As we surmised (cf section 4.2), the standard Euclidean measure of distance (and attempts at functional analysis based on it, such as cluster analysis) is an unreliable measure of functional equivalence. The non-linear nature of networks means that sometimes codes that are geometrically close will have different functional properties, and sometimes codes that are relatively geometrically distant will be functionally equivalent. An example of this was found in the codes (C_{29}, C_{24} and C_{33}) the CNM learned for the DIDO route (for moves 29, 24, and 33; see figure

[6] In a few of the simulations, there were a few prediction errors (at most 2 on any route) with respect to the learned route, but none of the errors involved the two locations under scrutiny nor their immediate neighbours.

Code 1	Code 2	Distance	F_b	F_p	F_d
C_{29}	C_{24}	0.87	87.50%	62.50%	-1.056
C_{29}	C_{33}	0.74	84.38%	50.00%	-1.517

Figure 5: An example of Euclidean similarity and functional equivalence coming apart.

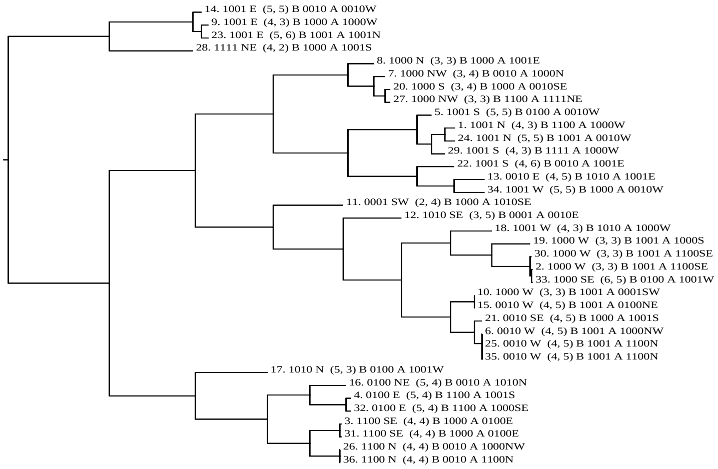

14. 1001 E (5, 5) B 0010 A 0010W
9. 1001 E (4, 3) B 1000 A 1000W
23. 1001 E (5, 6) B 1001 A 1001N
28. 1111 NE (4, 2) B 1000 A 1001S
8. 1000 N (3, 3) B 1000 A 1001E
7. 1000 NW (3, 4) B 0010 A 1000N
20. 1000 S (3, 4) B 1000 A 0010SE
27. 1000 NW (3, 3) B 1100 A 1111NE
5. 1001 S (5, 5) B 0100 A 0010W
1. 1001 N (4, 3) B 1100 A 1000W
24. 1001 N (5, 5) B 1001 A 0010W
29. 1001 S (4, 3) B 1111 A 1000W
22. 1001 S (4, 6) B 0010 A 1001E
13. 0010 E (4, 5) B 1010 A 1001E
34. 1001 W (5, 5) B 1000 A 0010W
11. 0001 SW (2, 4) B 1000 A 1010SE
12. 1010 SE (3, 5) B 0001 A 0010E
18. 1001 W (4, 3) B 1010 A 1000W
19. 1000 W (3, 3) B 1001 A 1000S
30. 1000 W (3, 3) B 1001 A 1100SE
2. 1000 W (3, 3) B 1001 A 1100SE
33. 1000 SE (6, 5) B 0100 A 1001W
10. 1000 W (3, 3) B 1001 A 0001SW
15. 0010 W (4, 5) B 1001 A 0100NE
21. 0010 SE (4, 5) B 1000 A 1001S
6. 0010 W (4, 5) B 1001 A 1000NW
25. 0010 W (4, 5) B 1001 A 1100N
35. 0010 W (4, 5) B 1001 A 1100N
17. 1010 N (5, 3) B 0100 A 1001W
16. 0100 NE (5, 4) B 0010 A 1010N
4. 0100 E (5, 4) B 1100 A 1001S
32. 0100 E (5, 4) B 1100 A 1000SE
3. 1100 SE (4, 4) B 1000 A 0100E
31. 1100 SE (4, 4) B 1000 A 0100E
26. 1100 N (4, 4) B 0010 A 1000NW
36. 1100 N (4, 4) B 0010 A 1100N

Figure 6: Cluster analysis of all location codes used in the DISOS route. Labels indicate the move number that produced the code, the description vector for the location, the move made, the coordinates of the location, the description vector of the previous place, the description vector of the following place, and the move taken to get there.

5). Although the distance between C_{29} and C_{33} was less than than the distance between C_{29} and C_{24}, the functional equivalence of the former pair was less than that of the latter pair, on all three of our measures of functional equivalence.

7.1 Qualitative analysis

One can use cluster analysis to get a rough idea of the different degrees of systematicity developed in learning the different types of routes. Figure 6 shows the cluster analysis of the location codes developed in learning the DISOS route. Note how the codes corresponding to (5, 5) are found in several parts of the tree, suggesting low functional equivalence between them. The same applies to the codes for (4,3). Note also that codes for (5,5) and (4,3) are often clustered together, suggesting a high functional equivalence between them. Both of these factors indicate a very low degree of systematicity.

In contrast, the cluster analysis of the codes developed for the SIDOD route (figure 7) suggests a high degree of systematicity. The codes for (5,5) are all clustered together, as are the codes for (4,3), and the (4,3) and (5,5) codes are in different (albeit neighbouring) sub-clusters, suggesting that they might be functionally divergent, despite the sensory equivalence of the two locations.

```
                         13. 1100 N  (4, 4) B 0010 A 1001N
                         20. 1100 N  (4, 4) B 0010 A 1001N
                            5. 1100 W  (4, 4) B 0100 A 1001N
                            26. 1100 W  (4, 4) B 0100 A 1001N
                               4. 0100 NW  (5, 4) B 1000 A 1100W
                               16. 0100 S  (5, 4) B 1010 A 0010SW
                             25. 0100 N  (5, 4) B 1001 A 1100W
                          28. 1111 N  (4, 2) B 1001 A 1000SW
                   17. 0010 SW  (4, 5) B 0100 A 1001E
                        12. 0010 NW  (4, 5) B 1001 A 1100N
                        19. 0010 W  (4, 5) B 1001 A 1100N
                         1. 0010 S  (4, 5) B 1100 A 1001E
                         9. 0010 S  (4, 5) B 1100 A 1001E
                         23. 0010 S  (4, 5) B 1100 A 1001E
                           15. 1010 E  (5, 3) B 1001 A 0100S
                                    11. 1001 S  (5, 6) B 1001 A 0010NW
                                          3. 1000 E  (6, 5) B 1001 A 0100NW
                                          7. 1000 W  (3, 3) B 1001 A 1100SE
                                        29. 1000 SW  (3, 3) B 1111 A 1100SE
    18. 1001 E  (5, 5) B 0010 A 0010W
    2. 1001 E  (5, 5) B 0010 A 1000E
    10. 1001 E  (5, 5) B 0010 A 1001S
    24. 1001 E  (5, 5) B 0010 A 0100N
       14. 1001 N  (4, 3) B 1100 A 1010E
       21. 1001 N  (4, 3) B 1100 A 1100S
        6. 1001 N  (4, 3) B 1100 A 1000W
        27. 1001 N  (4, 3) B 1100 A 1111N
  8. 1100 SE  (4, 4) B 1000 A 0010S
  22. 1100 S  (4, 4) B 1001 A 0010S
30. 1100 SE  (4, 4) B 1000 A 1100S
```

Bit-based

Pattern-based

The systematicity results using pattern- and bit-based functional equivalence measures are shown in figure 8. We also calculated the systematicity of the 7 scenarios using the distance-based measure, shown in figure 9 (note that this is not a measure of the Euclidean distance between the codes, but a measure of the distance between the sensations that two codes predict).

8 Discussion

The data are fairly univocal. The SIDO scenarios produce the most systematic codes, the SISO scenarios to a lesser extent, and the DISO scenarios even less. This agrees with our predictions, and thus supports the postulates, principles and hypothesis of sections 5 and 6.

However, two aspects of our predictions were not borne out.

10 Future work

In addition to the future work already mentioned (cf sections 3, 8 and 9), some other possibilities should be mentioned.

The generalization exhibited by the CNM so far only involves different combinations of transitions that it has made before. Another important kind of generalization is has is

Cussins, A. (1992). The limitations of pluralism. In Lennon, K. and Charles, D. (editors)