Connectionist Semantic Systematicity

Stefan L. Frank*

Willem F.G. Haselager

Iris van Rooij

Radboud University Nijmegen

Donders Institute for Brain, Cognition and Behaviour

P.O. Box 9104, 6500 HE Nijmegen

The Netherlands

Abstract

Fodor and Pylyshyn (1988) argue that connectionist models are not able to display systematicity other than by implementing a classical symbol system. This claim entails that connectionism cannot compete with the classical approach as an alternative architectural framework for human cognition. We present a connectionist model of sentence comprehension that does not implement a symbol system yet behaves systematically. It consists in a recurrent neural network that maps sentences describing situations in a microworld, onto representations of these situations. After being trained on particular sentences-situation pairs, the model can comprehend new sentences, even if these describe new situations. We argue that this systematicity arises robustly and in a psychologically plausible manner because it depends on structure inherent in the world.

Keywords: Systematicity; Connectionism; Sentence comprehension; Semantics; Analogical representation

*Corresponding author. Address: Institute for Logic, Language and Computation, University of Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands. Tel.: +31 20 5256054. E-mail address: S.L.Frank@uva.nl

1 Introduction

Human language is systematic to a considerable degree, which is to say that "the ability to produce/understand some sentences is *intrinsically* connected to the ability to produce/understand certain others" (Fodor & Pylyshyn, 1988, p. 37). For example, somebody who can understand the sentences *Charlie plays chess inside* and *Charlie plays hide-and-seek outside*, will also be able to understand *Charlie plays chess outside* and *Charlie plays hide-and-seek inside*.

Ever since Fodor and Pylyshyn (1988) argued that neural networks cannot display systematicity, except by implementing a classical symbol system, this issue has been fiercely debated. This debate is of considerable importance to cognitive science, for if it is indeed true that neural networks offer no explanation for the systematicity observed in language and thought, some would argue that connectionism has little (if any) value as a representational theory.

In this paper, our first objective is to present a connectionist model of sentence comprehension that does not implement a symbol system. Second, we investigate the model's ability to behave systematically, and compare this to different claims about systematicity in human sentence comprehension. Third, we set out to show that the model comes to display systematicity by capitalizing on structure present in the world, in language, and in the mapping from language to events in the world. Our connectionist explanation of systematic language comprehension takes into account that the structure of the world is reflected in the training input to which neural networks adapt. During training, external structures become internalized and, therefore, systematicity does not need to be inherent to the system. It is conceivable that this holds not only for neural networks, but also for the human cognitive system.

1.1 Semantic systematicity

To investigate connectionist systematicity, we need an operationalization that allows for the quantification of the systematic abilities of connectionist models. Hadley (1994a) operationalized systematicity by putting it in terms of learning and generalization. A neural network generalizes if it can successfully process inputs it was not trained on. That is, during training for the ability to process particular inputs, it also acquires the ability to correctly process others. This shows the two abilities to be "intrinsically connected", as desired by Fodor and Pylyshyn (1988, p. 37). Therefore, a network is systematic to some extent when generalization occurs, and displays higher levels of systematicity if it generalizes to new items that differ more strongly from the training examples. Since neural networks often do show at least some generalization without instantiating a classical system, the issue is not whether connectionist systematicity is possible *at all*, but whether neural networks can be as systematic as people are. We

return to this issue in Section 6.3 of the Discussion.

Hadley (1994a, 1994b) argued that, for neural networks to truly model human language performance, they should display *semantic systematicity*: the ability to construct correct representations of the meaning of novel sentences. There have been only a few attempts to demonstrate connectionist semantic systematicity, and none of these were very convincing. Two related models by Hadley and Hayward (1997) and Hadley and Cardei (1999) take as input sentences from a simple language and give as output a network representing their propositional structure. These models are quite different from most connectionist systems in that they were explicitly provided with structured representations. As argued by Aizawa (1997a), this results in a system that is actually classicist rather than connectionist. Likewise, Hadley, Rotaru-Varga, Arnold, and Cardei (2001) point out that the two models use classical, "combinatorially pre-disposed" (p. 74) representations. Consequently, these models do not instantiate true counterexamples to Fodor and Pylyshyn's (1988) claim.

A similar criticism applies to the sentence-comprehension model by Miikkulainen (1996). Its systematic capabilities result from three 'control units' that are trained to control the network's behavior at particular points in the input sentence. This means that the training input did not only consist of input-target (i.e., sentence-meaning) pairs, but also included *procedural* instructions on how to parse the sentences. As Miikkulainen admits, this is not realistic. More seriously, the control units serve as connectionist implementations of symbolic rules,[1] basing the model's systematicity on symbolic, not connectionist, computation.

Bodén and Niklasson (2000) trained a set of three Recursive Auto-Associative Memories (Pollack, 1990) to encode a very small number of propositions, such as is-a(ernie, bird), is-a(bo, fish), can(ernie, fly), and can(bo, not-fly). Next, one of the networks was trained to encode the fact that the new entity jack can fly. As it turned out, the internal representation of the token jack ended up closer to that of ernie than to bo. Bodén and Niklasson claim that this constitutes the inference that is-a(jack, bird), demonstrating connectionist semantic systematicity. Hadley (2004), however, argues strongly against this. According to him, complexities of the training procedure render the single test item not truly novel. Moreover, he argues that the network's representations lack semantic content because there is no possibility to associate a statement's representation to some state of affairs in the world that would

[1] For example, the 'push' control unit learns to activate a special memory network whenever the current input is a relative pronoun. This implements the rule 'if the input is a relative pronoun, then push the current sentence representation on the stack memory'. Miikkulainen (1996) claims that his model does not implement a symbol system because the memory network shows graceful degradation as its load increases, which is not how a symbol system would behave. Although this might be true, it only goes to show that the model's *memory* does not (perfectly) implement a symbolic memory. Its systematicity, however, is mainly due to the control units.

make the statement true. As we explain next, this problem does not occur in our model because its representations of statements *also* represent the described state of affairs in the world.

1.2 Sentence comprehension and mental representation

Our model differs from those discussed above in that it is rooted in recent psycholinguistic theories (e.g., Zwaan, 2004) according to which understanding a sentence does not (just) consist in the construction of its propositional (predicate-argument) structure, as has traditionally been assumed (e.g., Kintsch & van Dijk, 1978). Instead, a statement is only fully understood if the reader or listener has constructed a mental representation (or 'simulation') of the situation the sentence describes. This idea is comparable to Johnson-Laird's (1983) theory that mentally representing the meaning of a proposition comes down to representing one or more concrete situations (which he called 'mental models') that are consistent with that proposition.

This view of understanding as mental simulation has gained considerable experimental support. For example, Stanfield and Zwaan (2001) provide evidence that readers mentally represent objects' orientations when these are implied by (but not stated in) a sentence. They had subjects read sentences like *John put the pencil in the cup*, after which the subjects responded faster to an image of a pencil in vertical orientation than of a pencil in horizontal orientation. This outcome was reversed after reading *John put the pencil in the drawer*. That is, responses are faster if the orientation of the object in the presented image is congruent with the orientation implied by the sentence. Such a result is precisely what one would expect if readers mentally simulate the described situation, but difficult to explain by a purely propositional representation of the sentence. Likewise, research by Zwaan, Stanfield, and Yaxley (2002) indicates that the shape of a mentioned object forms part of the mental representation after sentence comprehension, even if this shape is neither explicitly mentioned nor relevant to the experimental task. The view of sentence comprehension as mental simulation was confirmed in several other experiments (for an overview, see Kerkhofs & Haselager, 2006).

Such findings suggest that the mental representation resulting from language comprehension strongly depends on the reader's experience with, and knowledge of, the world. For our current objectives, an important property of such representations is that they lead to direct inference: To mentally simulate a (normal size) pencil in a (normal size) cup *is also* to represent the pencil being (more or less) upright because, in our experience, pencils only fit in cups in an upright position. More in general, if (according to our knowledge) the world is such that some property or event a implies that b, a representation of a has the property of direct inference if it also represents b (see also Haugeland, 1987). That is,

relations between events in the world are reflected in relations between the mental representations of these events. A representation's form is thereby analogous to its meaning. Barsalou (1999) referred to representations that are analogical and modal as 'perceptual symbols' but, following Peirce (1903/1985), we will restrict our use of the word 'symbol' to refer to tokens with an arbitrary relation between form and meaning. Symbolic representations do not allow for direct inference: Getting from in(pencil, cup) to orientation(pencil, vertical) requires an inference process that works on these representations, because nothing in the representations themselves suggests how the represented situations might be related.

In spite of the evidence that understanding a sentence involves more than the construction of a propositional form, many computational models (e.g., Budiu & Anderson, 2004; Chang, 2002; Desai, 2007; Dominey, 2005; Hadley & Cardei, 1999; Hadley & Hayward, 1997; Mayberry, Crocker, & Knoeferle, in press; St.John & McClelland, 1990) represent sentence meaning as a structural combination of symbols, corresponding to the proposition expressed by the sentence. Contrary to this, Frank, Koppen, Noordman, and Vonk (2003) developed a non-symbolic representational scheme for the meaning of declarative sentences. In their Distributed Situation Space (DSS) model of story comprehension, each event or situation in a world is represented by a vector. As we explain in detail in Section 2.2, similarities among these vectors mirror dependencies among the represented events. If, in the world, the occurrence of some event a implies that event b also occurs, the vector representing event a is such that it also represents b. Clearly, this representation is analogical rather than symbolic (i.e., a vector's form and meaning are not separable) and provides a basis for direct inference. DSS vectors capture the analogical nature of Barsalou's (1999) perceptual symbols, albeit not their modality.

In the connectionist model of sentence comprehension we present here, sentence meaning is represented by vectors like those in the DSS model. The process of understanding a sentence that describes a particular event in the world, is simulated as the transformation of the sentence into the vector representing that event. As will become clear, these analogical representations are vital for reaching surprisingly high levels of systematicity in this model.

1.3 Overview

The models discussed in Section 1.1 were based on neural networks that are not exposed to anything like the structure that is inherent in (part of) a realistic world. If, as we argue, systematicity in thought is derivative from systematicity in the world, access to a world that provides sufficient structure is necessary to obtain semantic systematicity. Therefore, the first step in our simulations was to design an appropriate 'microworld'. As described in detail in Section 2.1, the microworld is populated by three people, who

can engage in several activities, be in different places, etcetera, giving rise to a variety of events that can (co-)occur in the microworld. Section 2.2 explains how all these events are assigned analogical vector representations, encoding knowledge about the microworld.

These vectors serve as the targets for the sentence-comprehension model: If the model receives as input a sentence referring to a particular event, it should give as output the vector representing that event. The input sentences come from a 'microlanguage' that was designed for describing microworld events. Section 3 presents the lexicon and syntax of this language, as well as its semantics, that is, the mapping from sentences to microworld events. The transformation of microlanguage sentences into the corresponding target vectors is performed by a recurrent neural network, presented in Section 4.1, which learns the microlanguage's semantics from examples of sentences-vector pairs.[2]

The network's ability to behave systematically is investigated by withholding four specific groups of sentences during its training phase, after which it is tested on some of these sentences. Each of these groups allows us to test for a particular level of systematicity, ranging from learning that synonyms can be interchanged, to understanding a sentence with a novel combination of concepts. In Section 4.2, we provide details of these test groups and explain what the network needs to have learned to comprehend new sentences of each group. The precise manner in which systematicity was rated is defined in Section 4.3.

The results presented in Section 5 show that the model indeed generalizes to new sentences, even when these sentences describe events that are not observed during training. This, we argue, indicates that the network displays relevant levels of semantic systematicity. In the same section, we clarify how the network accomplished this and explain to what extent it depends on the use of analogical representations that capture the microworld's structure.

In the Discussion (Section 6) we evaluate the model in relation to four common critiques of connectionist systematicity: that such models implement a symbol system; that connectionist demonstrations are not explanations; that the degree of systematicity does not compare to that of humans; and that the models do not scale up.

2 Representing a microworld

2.1 The microworld

Here, we describe the microworld that forms the basis of subsequent simulations. It is structured in the sense that there are (probabilistic) constraints on co-occurrences of events. As will be shown in

[2] A preliminary model, dealing with a much smaller world and language, was presented in Frank and Haselager (2006).

Section 2.2, this structure is captured by the events' representations that will be used by the sentence-comprehension model.

2.1.1 Concepts and events

In the microworld, there are two girls (called sophia and heidi) and one boy (charlie). As shown in Table 1, they have access to three toys and four places. Also, there are three games, which can be played and won in different manners.

Table 1: Concepts (entities and predicates) in the microworld.

Class	Variable	Class members (concepts)	#
People	p	charlie, heidi, sophia	3
Games	g	chess, hide&seek, soccer	3
Toys	t	puzzle, ball, doll	3
Places	x	bathroom, bedroom, playground, street	4
Manners of playing	m_{play}	well, badly	2
Manners of winning	m_{win}	easily, difficultly	2
Predicates	—	play, win, lose, place, manner	5

By applying each of five predicates, 44 basic microworld events can be constructed. These are listed in propositional form in Table 2. It is important to bear in mind that these propositional forms are only used for notational purposes. Within the model itself, the representations of microworld events do not contain anything like concepts, relations, or variables. Instead, as will be explained in Section 2.2, each of the 44 possible basic events is represented by one vector whose components are not related to the concepts playing a role in the event.

2.1.2 Co-occurrence constraints

Occurrences of microworld events are not mutually independent. Some events are likely to co-occur, some combinations are unlikely or even impossible, while other pairs of events are related by implication. It is only these co-occurrence relations that give any kind of 'meaning' to the events. For instance, a justification for calling a particular event win(heidi) is that it never co-occurs with lose(heidi) while it must co-occur with play(heidi, g) (with g being some game). Likewise, a justification for calling these other two events lose(heidi) and play(heidi, g) lies in a similar set of co-occurrence relations. Therefore,

Table 2: Construction of basic events from microworld concepts. Variables refer to those in Table 1.

Event name		#
play(p, g)	$3 \times 3 =$	9
play(p, t)	$3 \times 3 =$	9
win(p)		3
lose(p)		3
place(p, x)	$3 \times 4 =$	12
manner(play(p), m_{play})	$3 \times 2 =$	6
manner(win,m_{win})		2
	Total	44

the presence of co-occurrence constraints (i.e., structure) in the microworld is crucial for infusing the events with meaning.

For clarity, we divide the constraints into four groups: those concerning personal characteristics, games and toys, being-at-a-place, and winning and losing.

Personal characteristics Each of the three people in our microworld has a 'specialty': a game that (s)he usually and more easily wins. As can be seen from Table 3, a person's name and specialty sound conveniently alike. Also, charlie, heidi, and sophia differ in preferred toy (the one most often played with) and places most often visited. For example, play(charlie, puzzle) occurs more often than either play(charlie, ball) or play(charlie, doll), and win(sophia) is more likely to co-occur with play(sophia, soccer) than with either play(sophia, chess) or play(sophia, hide&seek)

Table 3: Personal specialties and preferences.

		Preferred	
Person	Specialty	toy	places
charlie	chess	puzzle	bathroom, bedroom
heidi	hide&seek	doll	—
sophia	soccer	ball	street, playground

Games and toys As listed in Table 4, there are restrictions on the places where each game and each toy can be played (with), as well as the number of people that can play a particular game or with a particular toy at any one time. Each person can only play one game or with one toy at a time. Someone who plays soccer, plays with the ball, but no other combination of game and toy is possible. Someone who plays well or badly, must play a game.

Table 4: Restrictions on games and toys.

Game/toy	# players	Possible places
chess	$0, 2$	bedroom, playground
hide&seek	$0, 2, 3$	bedroom, bathroom, playground
soccer	$0, 2, 3$	street
puzzle	$0, 1$	bedroom
ball	$0, 1, 2, 3$	street, playground
doll	$0, 1, 2, 3$	bedroom, playground

Being there Everybody is at exactly one place. If someone plays hide&seek in the playground, all players are in the playground. The two players of a chess match are in the same place. The girls tend to hang out at the same place, while charlie avoids them.

Winning and losing One cannot both win and lose, nor can two people win at the same time. If someone wins, all other players lose, and if there is a loser, there must be one winner. Someone who wins or loses, plays a game. Someone who plays well is more likely to win, and whoever plays badly is more likely to lose. Winning is usually done easily by someone who plays well and difficultly by those playing badly.

2.2 Representation

2.2.1 Objective

Our objective is to find a representational scheme for events that implements an important property of mental representations: direct inference (see Section 1.2). This is accomplished when co-occurrence relations among the 44 microworld events are apparent in relations among their representations. More precisely, using *only* the representations of any pair of microworld events a and b, it should be possible to accurately estimate the conditional probability that a occurs in the microworld given that b does (i.e.,

$\Pr(a|b)$). This conditional-probability estimate, denoted $\tau(a|b)$, is called the *belief value* of a given b, since it indicates the extent to which a might be believed to be the case, given that b is the case. If belief values indeed approximate the probabilities in the microworld, that is, if $\tau(a|b) \approx \Pr(a|b)$, then a representation of b is also a representation of anything that depends on b in the microworld, so direct inference occurs. In Section 3.2, we explain how these representations of events serve as representations of sentence meaning in our sentence-comprehension model.

Note that we are concerned with representing only events and not concepts. The concepts that appear in an event's propositional form do not even directly affect the event's representation. Take, for instance, the three basic events $\mathsf{play}(p, \mathsf{soccer})$, $\mathsf{play}(p, \mathsf{ball})$, and $\mathsf{play}(p, \mathsf{puzzle})$, for any person p. Looking only at propositional forms, the three differ from one another to the same extent: They are identical except for their second argument. The case is very different, however, if we consider the state of affairs in the world described by these propositions: $\mathsf{play}(p, \mathsf{soccer})$ implies that $\mathsf{play}(p, \mathsf{ball})$, so the two will often co-occur, while $\mathsf{play}(p, \mathsf{soccer})$ excludes $\mathsf{play}(p, \mathsf{puzzle})$, so the two never co-occur. The representations of these three events should encode their co-occurrence relations rather than their conceptual relations in that $\tau(\mathsf{play}(p, \mathsf{ball})|\ \mathsf{play}(p, \mathsf{soccer})) \approx 1$ while $\tau(\mathsf{play}(p, \mathsf{puzzle})|\mathsf{play}(p, \mathsf{soccer})) \approx 0$.

2.2.2 Situation space

For their DSS model, Frank et al. (2003) developed a representational scheme that has exactly the properties we desire. In that model, each microworld event a is assigned a *situation vector* $\mu(a) = (\mu_1(a), \ldots, \mu_n(a)) \in [0, 1]^n$, that is, a point in *situation space*. The vector's individual components $\mu_i(a)$ are not generally interpretable. Situation vectors represent events by virtue of encoding the events' probabilities in the microworld. As explained in detail below, both prior and conditional probabilities of events can be estimated from the events' representations. Moreover, a vector representation of any boolean combination of microworld events (called a *complex event*) can easily be computed from the vectors representing the events involved.

Due to situation vectors having real values, there are infinitely many of them. In contrast, there are only a finite number of basic or complex events. We will use the term 'situation' (or 'microworld situation') for anything that is represented by some vector in situation space. This means that basic and complex events are themselves situations, but that most (i.e., infinitely many) situations are not events.

It is important to note that situation vectors are not compositional: They do not have parts representing the concepts of Table 1. This means that any systematicity cannot be explained by resorting to the classical idea of compositionality. Also, situation vectors are not functionally compositional in the sense of van Gelder (1990), that is, they cannot be computed from representations of concepts, simply

because *there exist no such representations.*

Computing belief values First, the prior probability that event a occurs is estimated from its vector representation by the average value of the vector's components:

$$\tau(a) = \frac{1}{n} \sum_i \mu_i(a) \approx \Pr(a), \tag{1}$$

which is called the prior belief value of a. Second, $\Pr(a \wedge b)$, the prior probability of the occurrence of the conjunction $a \wedge b$ (with $a \neq b$) is estimated by:

$$\tau(a \wedge b) = \frac{1}{n} \sum_i \mu_i(a)\mu_i(b) \approx \Pr(a \wedge b). \tag{2}$$

For $a = b$, we define that $\tau(a \wedge a) = \tau(a)$, since $\Pr(a \wedge a) = \Pr(a)$. This is different from Frank et al. (2003) where, in general, $\tau(a \wedge a) \neq \tau(a)$.

Given situation vectors for which Equations (1) and (2) hold, an expression for belief values $\tau(a|b)$ follows directly. By definition, $\Pr(a|b) = \Pr(a \wedge b)/\Pr(b)$, so the conditional probability is estimated by:

$$\tau(a|b) = \frac{\tau(a \wedge b)}{\tau(b)} = \frac{\sum_i \mu_i(a)\mu_i(b)}{\sum_i \mu_i(b)} \approx \Pr(a|b). \tag{3}$$

Representing complex events Vector representations of negations and conjunctions of (basic or complex) events are computed as is common in fuzzy logic:

$$\begin{aligned} \mu(\neg a) &= 1 - \mu(a) \\ \mu_i(a \wedge b) &= \mu_i(a)\mu_i(b) \quad \text{for } a \neq b. \end{aligned} \tag{4}$$

Furthermore, we define $\mu(a \wedge a) = \mu(a)$. It is easy to see that these operations retain the relations between vectors and probability estimates, as expressed by Equations 1, 2, and 3. The belief value of a negation is $\tau(\neg a) = 1 - \tau(a)$, in accordance with the fact that $\Pr(\neg a) = 1 - \Pr(a)$. Also, combining Equations 1 and 4 indeed yields the expression for $\tau(a \wedge b)$ of Equation 2.

A well-known fact from propositional logic is that any boolean combination of propositions can be expressed using only the operators for negation and conjunction. Therefore, our definitions of negation and conjunction lead to a representation for *any* complex event. For example, a disjunction is defined by $a \vee b \equiv \neg(\neg a \wedge \neg b)$. Therefore, $\mu_i(a \vee b) = 1 - ((1 - \mu_i(a))(1 - \mu_i(b))) = \mu_i(a) + \mu_i(b) - \mu_i(a)\mu_i(b)$.

2.2.3 Organizing situation space

As the above discussion makes clear, once we have basic-event vectors such that Equations 1 and 2 hold, we can compute the vector for any microworld event and estimate the probabilities of any event given

any situation vector. The question remains how to find such vectors. Following Frank et al. (2003), we do this by automatically generating a large number (25 000) of 'observations' of states-of-affairs in the microworld. In each of these observations, each basic event is either the case or not the case. More formally, an observation takes the form of a 44-dimensional binary vector S_k, the components of which indicate the status of all basic events at one instant k: If basic event a occurs at that instant, then $S_k(a) = 1$. If a does not occur, $S_k(a) = 0$.

Microworld constraints are apparent in these examples. For instance, play(charlie, soccer) implies that ¬play(charlie, chess), so if S_k(play(charlie, soccer)) = 1 then S_k(play(charlie, chess)) = 0. Also, win(sophia) is more likely when manner(play(sophia), well), so there is a positive correlation between the values of S_k(win(sophia)) and S_k(manner(play(sophia), well)) over all k.

Maximum likelihood estimates of the probabilities of basic events and conjunctions are easy to compute from the observation vectors $S_1, \ldots, S_K$ (where K is the number of observations):

$$\Pr(a) \approx \frac{1}{K} \sum_k S_k(a) \tag{5}$$

$$\Pr(a \wedge b) \approx \frac{1}{K} \sum_k S_k(a) S_k(b). \tag{6}$$

Comparing Equations 5 and 6 to Equations 1 and 2, respectively, it is obvious that taking $\mu(a) = (S_1(a), \ldots, S_K(a))$ leads to basic-event vectors with the desired properties, but only if K is large enough. Unfortunately, taking a very large number of observations, like $K = 25\,000$ as used here, makes the number of situation-space dimensions unpractically large. Reducing K to a more manageable level, on the other hand, would reduce the quality of the probability estimates. Therefore, a dimensionality-reduction technique is applied to transform the observation vectors S into situation vectors μ that have a more reasonable number of dimensions. Note that this is not intended to simulate the psychological process of developing event representations. That is, it is merely a tool to obtain compressed representations. Also, we do not make any cognitive claims about how people perceive (co-)occurrences of discrete events in the world, but simply assume that they can reliably perceive such (co-)occurrences.

As illustrated in Figure 1, the observation vectors S are used as training input to a self-organizing system called a Competitive Layer, consisting of n units. Each of these units is associated to 44 values, corresponding to the 44 basic microworld events. During training, these values are adapted to the observations in an unsupervised manner reminiscent of the well known Self-Organizing Map (Kohonen, 1995).[3] A description of the training algorithm is provided in Appendix A. The result is a vector

[3]The difference between a Competitive Layer and a Self-Organizing Map is that the latter creates a topological mapping of the input. Since the task at hand does not require such a mapping, a Competitive Layer is preferred over the Self-Organizing Map used by Frank et al. (2003).

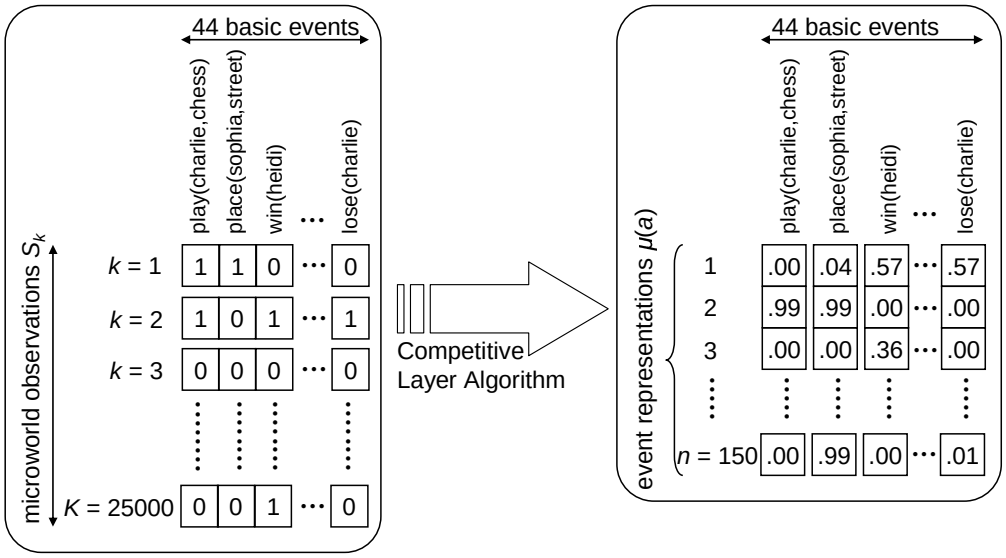

Figure 1: Transforming microworld observations into representations of basic events. A value of 1 in the observations (row vectors S_k) denotes the occurrence of a basic event at a particular moment in the microworld, while 0 denotes non-occurrence. Individual values of basic-event representations (column vectors $\mu(a)$) are between (and usually close to) 0 and 1, and are not interpretable.

$\mu(a) \in [0,1]^n$ for each basic event a, where n (the dimensionality of situation space) can be freely chosen prior to training. The quality of these vectors is investigated by comparing the true (conditional) probabilities in the microworld to the corresponding belief values. If the coefficient of correlation between them is close to 1, the vectors accurately encode probabilities in the microworld. As it turns out, larger n generally gives better results. For $n = 150$, results are very good ($r \geq .996$; see Appendix A) and they hardly improve for larger n. Therefore, we set n to 150.

3 The microlanguage

Events in the microworld can be described by sentences in a microlanguage. Below, we present this language's lexicon and grammar, and informally describe its semantics.

3.1 Words

The microlanguage's 40 words are listed in Table 5. It is generally straightforward how content words refer to the concepts in Table 1. For instance, the word *charlie* refers to the concept charlie. Note that some word pairs are synonymous, that is, the two words refer to the same concept. These word pairs

are: {*charlie, boy*}, {*soccer, football*}, {*puzzle, jigsaw*}, and {*bathroom, shower*}. Some other content words, such as *girl, inside*, and *toy*, affect sentence meaning without referring to a single concept. For instance, a statement about *girl* describes the disjunction of all such statements about individual girls, that is, sophia and heidi.

Table 5: Lexicon of the microlanguage.

Class	Words	#
proper nouns	*charlie, heidi, sophia*	3
(pro)nouns	*boy, girl, someone, chess, hide-and-seek, soccer, football, game, puzzle, ball, doll, jigsaw, toy, ease, difficulty, bathroom, bedroom, playground, shower, street*	20
verbs	*wins, loses, beats, plays, is, won, lost, played*	8
adverbs	*well, badly, inside, outside*	4
prepositions	*with, to, at, in, by*	5
	Total	40

3.2 Sentences

Words can be combined into 13 556 different sentences according to the grammar in Table 6. As an additional constraint (not shown in the grammar), a sentence never describes the case of someone beating or losing to him/herself (which would violate the microworld constraints). That is, sentences of the form p_1 *beats* p_2 and p_1 *loses to* p_2 are not allowed if $(p_1, p_2) \in \{$(*charlie, charlie*), (*charlie, boy*), (*boy, charlie*), (*boy, boy*), (*heidi, heidi*), (*sophia, sophia*)$\}$.

Each sentence has one meaning, corresponding to a basic or complex event. Table 7 lists some typical sentences and the propositional notation of the event to which they refer (those of other sentences can be extrapolated from the ones listed). To find the situation vector representing the event described by a sentence, we take the propositional form (as in Table 7), the representation(s) of the basic event(s) involved, and (if needed) compute the vector for the described complex event by applying Equation 4.

Table 6: Grammar of the microlanguage (see text for additional constraints). Variable $n \in$ {person, game, toy} denotes noun types; $v \in$ {play, win, lose} are verb types. VP = verb phrase; APP = adverbial/prepositional phrase; PP = Prepositional phrase. Items in square brackets are optional.

S	→	N_n $\mathrm{VP}_{n,v}$ $\mathrm{APP}_{n,v}$
$\mathrm{N}_{\mathrm{person}}$	→	*charlie* | *heidi* | *sophia* | *someone* | *boy* | *girl*
$\mathrm{N}_{\mathrm{game}}$	→	*chess* | *hide-and-seek* | *soccer* | *football* | *game*
$\mathrm{N}_{\mathrm{toy}}$	→	*puzzle* | *ball* | *doll* | *jigsaw* | *toy*
$\mathrm{VP}_{\mathrm{person,\ play}}$	→	*plays*
$\mathrm{VP}_{\mathrm{person,\ win}}$	→	*wins* | *beats* $\mathrm{N}_{\mathrm{person}}$
$\mathrm{VP}_{\mathrm{person,\ lose}}$	→	*loses* | *loses to* $\mathrm{N}_{\mathrm{person}}$
$\mathrm{VP}_{\mathrm{game,\ play}}$	→	*is played*
$\mathrm{VP}_{\mathrm{game,\ win}}$	→	*is won*
$\mathrm{VP}_{\mathrm{game,\ lose}}$	→	*is lost*
$\mathrm{VP}_{\mathrm{toy,\ play}}$	→	*is played with*
$\mathrm{APP}_{\mathrm{person,\ play}}$	→	[$\mathrm{N}_{\mathrm{game}}$] [Manner] [Place] | $\mathrm{PP}_{\mathrm{toy}}$ [Place] | Place $\mathrm{PP}_{\mathrm{toy}}$
$\mathrm{APP}_{\mathrm{person,\ win}}$	→	[$\mathrm{PP}_{\mathrm{manner}}$] [$\mathrm{PP}_{\mathrm{game}}$] [Place] | $\mathrm{PP}_{\mathrm{game}}$ $\mathrm{PP}_{\mathrm{manner}}$| Place $\mathrm{PP}_{\mathrm{game}}$
$\mathrm{APP}_{\mathrm{person,\ lose}}$	→	[$\mathrm{PP}_{\mathrm{game}}$] [Place] | Place $\mathrm{PP}_{\mathrm{game}}$
$\mathrm{APP}_{\mathrm{game,\ play}}$	→	[Manner] [$\mathrm{PP}_{\mathrm{person}}$] [Place]
$\mathrm{APP}_{\mathrm{game,\ win}}$	→	[$\mathrm{PP}_{\mathrm{manner}}$] [$\mathrm{PP}_{\mathrm{person}}$] [Place]
$\mathrm{APP}_{\mathrm{game,\ lose}}$	→	[$\mathrm{PP}_{\mathrm{person}}$] [Place]
$\mathrm{APP}_{\mathrm{toy,\ play}}$	→	[$\mathrm{PP}_{\mathrm{person}}$] [Place] | Place $\mathrm{PP}_{\mathrm{person}}$
Manner	→	*well* | *badly*
Place	→	*inside* | *outside* | $\mathrm{PP}_{\mathrm{place}}$
$\mathrm{PP}_{\mathrm{place}}$	→	*in bathroom* | *in shower* | *in bedroom* | *in street* | *in playground*
$\mathrm{PP}_{\mathrm{person}}$	→	*by* $\mathrm{N}_{\mathrm{person}}$
$\mathrm{PP}_{\mathrm{game}}$	→	*at* $\mathrm{N}_{\mathrm{game}}$
$\mathrm{PP}_{\mathrm{toy}}$	→	*with* $\mathrm{N}_{\mathrm{toy}}$
$\mathrm{PP}_{\mathrm{manner}}$	→	*with ease* | *with difficulty*

Table 7: Examples of microlanguage sentences and the propositional form of the described event. c = charlie; h = heidi; s = sophia.

Sentence	Semantics
charlie plays chess	play(c, chess)
chess is played by charlie	play(c, chess)
girl plays chess	play(h, chess) $\vee$ play(s, chess)
heidi plays game	play(h, chess) $\vee$ play(h, hide&seek) $\vee$ play(h, soccer)
heidi plays with toy	play(h, puzzle) $\vee$ play(h, ball) $\vee$ play(h, doll)
sophia plays soccer well	play(s, soccer) $\wedge$ manner(play(s), well)
sophia plays with ball in street	play(s, ball) $\wedge$ place(s, street)
someone plays with doll	play(c, doll) $\vee$ play(h, doll) $\vee$ play(s, doll)
doll is played with	play(c, doll) $\vee$ play(h, doll) $\vee$ play(s, doll)
charlie plays	play(c, chess) $\vee$ play(c, hide&seek) $\vee$ play(c, soccer) $\vee$ play(c, puzzle) $\vee$ play(c, ball) $\vee$ play(c, doll)
heidi wins	win(h)
heidi loses at chess	lose(h) $\wedge$ play(h, chess)
chess is lost by heidi	lose(h) $\wedge$ play(h, chess)
sophia wins with ease	win(s) $\wedge$ manner(win, easily)
charlie wins inside	win(c) $\wedge$ (place(c, bedroom) $\vee$ place(c, bathroom))
charlie wins outside	win(c) $\wedge$ (place(c, street) $\vee$ place(c, playground))
soccer is won	(win(c) $\wedge$ play(c, soccer)) $\vee$ (win(h) $\wedge$ play(h, soccer)) $\vee$ (win(s) $\wedge$ play(s, soccer))
charlie loses to sophia	win(s) $\wedge$ lose(c)
charlie beats someone	win(c) $\wedge$ (lose(c) $\vee$ lose(h) $\vee$ lose(s))
sophia beats charlie at chess	win(s) $\wedge$ lose(c) $\wedge$ play(s, chess)

4 Simulations

4.1 The network

The sentence-comprehension model consists in a Simple Recurrent Network (SRN; Elman, 1990) that transforms microlanguage sentences into situation vectors. Here, we describe the network's architecture, a measure for the extent to which input sentences are understood, and details of the training method. The architecture is the most basic form of a SRN (e.g., there were no additional hidden layers) and the

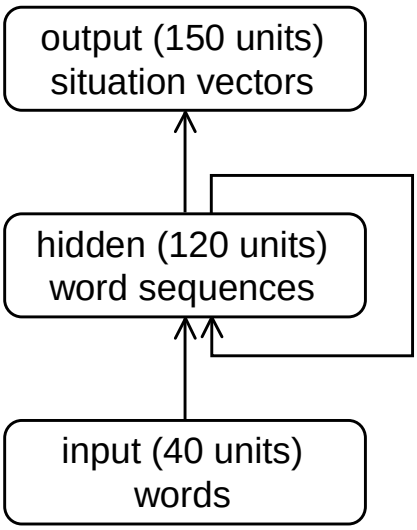

Figure 2: Simple recurrent network for transforming word sequences into situation vectors. Arrows denote connections from each unit in one layer to all units in the next.

training regime and algorithm are as simple as possible (e.g., the learning rate is constant). Although increasing the complexity of the network or the training regime may improve performance, we wanted to make sure that any systematic behavior that is observed would not critically depend on such complexities.

4.1.1 Network architecture

The SRN has three layers of units, as shown in Figure 2. The input layer has 40 units, each corresponding to one word of the microlanguage. Words enter the network one at a time. The activation from the unit representing the current input word is sent to the 120-unit hidden layer[4] that receives, through recurrent connections, its own previous activation pattern as additional input and thereby comes to represent the word sequence so far. The activation pattern over the 150-unit output layer, constituting the situation vector constructed by the network, ideally represents the event described by the input sentence.

4.1.2 Rating the output

Ideally, the model transforms all sentences describing some (basic or complex) microworld event a into its vector representation $\mu(a)$. In practice, the model's *actual* output situation vector $\mu(z)$ is at best similar to $\mu(a)$. Given some output vector $\mu(z)$, we obtain information about the represented situation z

[4]In preliminary simulations, we experimented with hidden-layer sizes between 40 and 150 and found that larger networks generalize better (the same was found by Frank & Haselager, 2006) For reasons of training efficiency, we settled for a hidden-layer size of 120.

by looking at the belief values $\tau(b|z)$ for different events b. In particular, the extent to which the model has understood the sentence describing event a is apparent from $\tau(a|z)$.

More formally, the *comprehension score* is a value between -1 and $+1$ that is computed from belief values $\tau(a)$ and $\tau(a|z)$. If the model has simulated sentence comprehension even minimally, the belief value of the described event a in situation z should be larger than the prior belief value, that is, $\tau(a|z) > \tau(a)$. Ideally, $z = a$ so $\tau(a|z) = 1$. If the network 'misunderstood', then $\tau(a|z) < \tau(a)$. In the worst possible case, $\tau(a|z) = 0$. The comprehension score is the attained fraction of the maximum possible increase (or decrease) in belief value of a, as expressed by Equation 7 below. Positive values indicate some level of correct comprehension, while negative values indicate comprehension errors.

$$\text{comprehension} = \begin{cases} \frac{\tau(a|z)-\tau(a)}{1-\tau(a)} & \text{if } \tau(a|z) > \tau(a) \\ \frac{\tau(a|z)-\tau(a)}{\tau(a)} & \text{otherwise.} \end{cases} \tag{7}$$

Some complex events violate the microworld's constraints, for example, $\mathsf{win}(\mathsf{charlie}) \wedge \mathsf{lose}(\mathsf{charlie})$ can never occur, nor can $\mathsf{win}(\mathsf{heidi}) \wedge \neg\mathsf{win}(\mathsf{heidi})$. We shall call such events (as well as sentences describing them) *unlawful*. Ideally, $\tau(a) = 0$ for unlawful a because such a never occurs in the world (i.e., $\Pr(a) = 0$). Perfect comprehension means that $z = a$ so $\tau(z) = 0$, in which case $\tau(a|z)$ is not defined. To prevent this problem, we leave comprehension scores undefined for unlawful a. In practice, we are not interested in comprehension of unlawful sentences anyway.

4.1.3 Network training

Ten networks, differing only in their initial random connection weights, were trained twice, once for each of two sets of training sentences (as presented in Section 4.2). All training sentences from a set were presented in random order, and the standard backpropagation algorithm was used for adapting the network's connection weights. Initial connection weights were taken randomly from a uniform distribution between ± 0.15. The backpropagation's learning rate parameter was fixed at .02, and no momentum was used.

After processing each word of a training sentence, the network was trained to give as output the vector representing the event described by the complete sentence. Although this is similar to the task of a language learner who perceives simultaneously a situation in the world and an utterance describing that situation, we stress that the model is not intended to simulate human language acquisition.

Training was repeated until the average comprehension score (see Equation 7) on training sentences reached .5. On average, 659 presentations of the training set were needed to reach this criterion. Training up to an average comprehension score of .5 might not seem like much, but it should be taken into account

that the training set (and, thereby, the average comprehension score) is dominated by long sentences that describe highly complex events. For example, *sophia beats charlie easily at chess in bedroom* describes a conjunction of as much as five basic events (i.e., win(sophia) ∧ lose(charlie) ∧ manner(win, easily) ∧ play(sophia, chess) ∧ place(sophia, bedroom)) and a comprehension score close to 1 would require this complete conjunction to be understood nearly perfectly. Test sentences are generally shorter and describe simpler events than training sentences. Consequently, they often result in comprehension scores close to 1, as we shall see in Section 5.

4.2 Training and test sentences

Two sets of training sentences were constructed, containing on average 9 534 sentences (i.e., 70.3% of all possible sentences). All sentences that are missing in one set are present in the other, making sure that the results we find do not crucially depend on the exclusion of some very particular set of sentences during training. Since the choice of training set had no significant qualitative effect on model performance, we will usually collapse over the two training sets. That is, when referring to a sentence as a 'training sentence' or 'test sentence', we leave implicit which of the two training sets was used.

The sentences that are excluded from a training set are divided into four groups, called the Word, Sentence, Complex Event, and Basic Event groups. We shortly present the rationale behind these groups. Each group came in two versions, one for each of the two training sets. After training, the network is tested on novel sentences from these four groups. As explained below, sentences from each group afford testing for a particular level of systematicity, and model performance is expected to decrease when testing consecutively with sentences from the Word, Sentence, Complex Event, and Basic Event groups.

4.2.1 Word group

All sentences in the Word group contain two words that have a synonym in the microlanguage. More precisely, the first training set has no sentences containing both *charlie* and *soccer*, nor any sentence containing both *boy* and *football*. In the other set, these word combinations are reversed: It has no sentences containing either *charlie* and *football*, or *boy* and *soccer*.

When the network is tested, Word group sentences can be understood by simply generalizing the use of one word of a synonym pair to contexts in which only the other synonym has been seen. For instance, to correctly understand the test sentence *charlie plays soccer*, a sufficiently trained network only needs to have learned that *charlie* is the same as *boy*, or that *soccer* is the same as *football*. This, we expect, will be accomplished easily because two synonymous words often occur in the same sentence context and such sentences describe identical situations.

4.2.2 Sentence group

The Sentence group contains sentences with phrases of the form p_1 *beats* p_2 and p_1 *loses to* p_2, where the words denoted by p_1 and p_2 depend on the training set. The following combinations are excluded from the first training set: $(p_1, p_2) \in \{$(*charlie, heidi*), (*boy, heidi*), (*heidi, sophia*), (*sophia, charlie*), (*sophia, boy*)$\}$. In the second training set, there are no sentences in which $(p_1, p_2) \in \{$(*charlie, sophia*), (*boy, sophia*), (*heidi, charlie*), (*heidi, boy*), (*sophia, heidi*)$\}$.

The test sentences in the Sentence group (like those in the Word group) describe events that also appear in training sentences. For example, the training sentence *heidi loses to charlie* describes the same event as the test sentence *charlie beats heidi*. To understand such a test sentence, the network needs to generalize to the new sentence but not to a new event, that is, it must construct a situation vector that it learned to construct during training. Therefore, we expect these test sentences to be processed relatively well compared to sentences that do require generalization to a new event.

4.2.3 Complex Event group

The previous two groups were defined by particular combinations of words. For the Complex Event group, on the other hand, sentences describing particular complex events are selected: The two training sets contain no sentences describing particular conjunctions of games and place. In particular, sentences in the first training set never describe events in which hide&seek is played anywhere inside (i.e., in bathroom or bedroom), nor any event in which chess is played outside (i.e., in street or playground). For the second training set, these combinations of games and places are reversed.

To understand a new sentence from this group, the network must construct a complex event on which it was not trained. For example, to process the test sentence *sophia plays chess in playground*, the network has to construct the situation vector of the novel conjunction play(sophia, chess)$\wedge$place(sophia, playground). Because of the systematic relation between $\mu(a), \mu(b)$, and the conjunction $\mu(a \wedge b)$, as expressed by Equation 4, such generalization is possible in principle. Nevertheless, Complex Event group test sentences are expected to lead to lower comprehension scores than test sentences from the Word and Sentence groups because generating an output vector that was never a target during training is likely to be challenging for the network.

4.2.4 Basic Event group

All sentences in the Basic Event group describe one of three basic events. To be precise, the first training set contains no sentences stating that play(charlie, doll), play(heidi, ball), or play(sophia, puzzle). In the

Table 8: Test sentences frames and number of test sentences per group. See text for constraints on variable instantiation.

Group	Sentences	#
Word	*p plays g*	8
	g is played by p	
Sentence	p_1 *beats* p_2	20
	p_1 *loses to* p_2	
Complex Event	*p plays g [in] x*	80
	g is played by p [in] x	
Basic Event	*p plays with t*	20
	t is played with by p	
	Total	128

second training set, no sentence describes play(charlie, ball), play(heidi, puzzle), or play(sophia, doll).

To correctly process test sentences from the Basic Event group, the network needs to construct the representation of a basic event on which it was not trained. For instance, it may never have learned to produce the output vector μ(play(heidi, ball)). It seems impossible for this network to correctly process the test sentence *heidi plays with ball* since μ(play(heidi, ball)) is not computable from tokens for heidi, doll, or play. To understand this sentence, the network cannot take advantage of any systematic relation between sentences of the form *p plays with t* and situation vectors $\mu(\mathsf{play}(p, t))$, because there is no such systematic relation. In a classical symbol system, precisely such a relation is responsible for systematic behavior. According to the classical view, our network should therefore not be able to understand Basic Event group test sentences.

4.2.5 Specification of test sentences

So far, we have only presented examples of test sentences. In total, there were 128 different test sentences, which were all the lawful non-training sentences that can be formed by taking the sentence frames from Table 8 and instantiating the variables by words from the following sets: $p \in \{$*charlie, boy, heidi, sophia*$\}$, $t \in \{$*ball, doll, puzzle, jigsaw*$\}$, $g \in \{$*hide-and-seek, chess, soccer, football*$\}$, and $x \in \{$*inside, outside, bathroom, shower, bedroom, playground*$\}$.

4.3 Rating systematicity

When a test sentence describes some event a, the comprehension score for a should be positive. However, this is not always sufficient to conclude that the sentence was understood properly. In the Sentence and Complex Event test groups, a is a conjunction of two basic events, and these should individually have positive comprehension scores too. Take, for instance, the sentence *charlie beats heidi*, which states that $\mathsf{win}(\mathsf{charlie}) \wedge \mathsf{lose}(\mathsf{heidi})$. If the network has understood *only* $\mathsf{win}(\mathsf{charlie})$ this will already lead to a positive comprehension score for the conjunction, because the information that $\mathsf{win}(\mathsf{charlie})$ makes it more likely that $\mathsf{win}(\mathsf{charlie}) \wedge \mathsf{lose}(\mathsf{heidi})$. Conversely, positive comprehension scores for both basic events individually should not be mistaken for a positive comprehension score for their conjunction, because wrongly believing that *either* $\mathsf{win}(\mathsf{charlie})$ *or* $\mathsf{lose}(\mathsf{heidi})$ would also lead to positive comprehension scores for these two basic events, even though their conjunction is excluded. For sentences describing a conjunction, it is therefore important to look at comprehension scores for both the conjunction and the basic events it comprises.

Even if test sentences are understood to some extent, this need not indicate semantic systematicity. Take again the test sentence *charlie beats heidi*. It is possible that the network understands nothing more than the information that there is 'beating' going on, that is, there is a winner and there is a loser. This in itself suffices for positive comprehension scores for $\mathsf{win}(\mathsf{charlie})$, $\mathsf{lose}(\mathsf{heidi})$, and their conjunction, that is, for precisely the events stated by the test sentence. However, it also leads to positive comprehension of basic events that are inconsistent with the sentence, namely $\mathsf{lose}(\mathsf{charlie})$, $\mathsf{win}(\mathsf{heidi})$, and $\mathsf{win}(\mathsf{sophia})$. To warrant the conclusion that the network behaves systematically, such 'competing events' should have comprehension scores that are negative, or at least significantly smaller than those of the described events.

To summarize, processing a test sentence should result in positive comprehension scores for the described basic event(s) and (if applicable) their conjunction, and significantly smaller (ideally, even negative) comprehension scores for competing events. Table 9 lists which basic events we regard as described or competing for test sentences of the four groups. A competing event is always inconsistent (given the constraints of the microworld) with the described situation, but can be described by a superficially similar sentence.

5 Results and explanations

Figure 3 plots the average comprehension scores for described and competing events, resulting from processing test sentences and matched training sentences from each of the four groups. The training

Table 9: Described and competing events for test sentences in each group of Table 8. Within each group, identical variables have identical values and variables with different indices have unequal values.

Group	Described event(s)	Competing events
Word	play(charlie, soccer)	play(charlie, hide&seek)
		play(charlie, chess)
Sentence	win(p_1)	win(p_2)
	lose(p_2)	win(p_3)
		lose(p_1)
Complex Event	play(p, g_1)	play(p, g_2)
	place(p, x_1)	place(p, x_2)
Basic Event	play(p_1, t_1)	play(p_1, t_2)
		play(p_2, puzzle) (only if t_1 = puzzle)

sentences that gave rise to these results were the same as the test sentences, because all networks trained on one training set were tested using sentences from the other training set, and vice versa. Since there is no reason to expect the test sentences to be understood better than matched training sentences, comprehension scores on tests should be assessed relative to the scores on the corresponding training items.

Sentences of the Word group are comprehended very well: Comprehension scores are large and positive for described events, and strongly negative for competing events. Test sentence scores are close to those of training sentences, indicating that test sentences are comprehended as well as could be reasonably be expected.

As we move from the Word to the Sentence and Complex Event groups, comprehension scores resulting from test sentences decrease in absolute value, while remaining the same for training sentences. This effect of test group was expected considering the differences in the required level of systematicity (see Section 4.2), but it should be taken into account that sentences from different groups differ in many other aspects as well.

Training sentences from the Basic Event group are understood remarkably poorly. Presumably, this is because these sentences, which are all about playing with toys, occur much less frequently than the sentences making up the other three groups, which are about playing games. As a result, the networks might not have been sufficiently exposed to Basic Event group sentences. Interestingly, even in this group, test sentences are understood to some extent: Average comprehension scores are positive for described

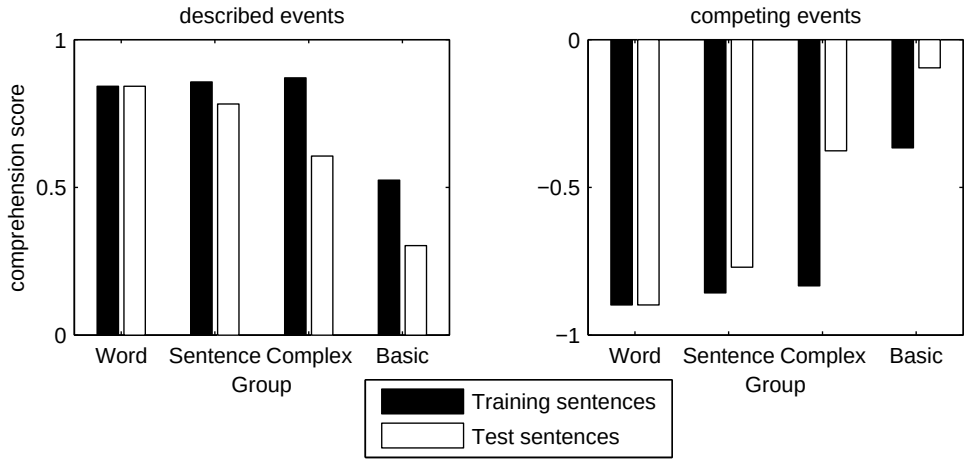

Figure 3: Average comprehension scores of described (left) and competing (right) events, after processing training (black bars) or test sentences (white bars) from each of the four groups.

events and negative for competing events (sign tests showed these values to significantly differ from zero: $N = 200; z = 14.1; p \approx 0$ and $N = 560; z = 3.0; p < .003$, respectively). This is noteworthy because, as explained in Section 4.2.4, no sign of systematicity is expected here from the classical viewpoint.

An error occurs whenever a described event has a negative comprehension score, or when a competing event has a positive comprehension score. There were no errors on described events, except in just 1.9% of the cases for Complex Event group test sentences. For competing events, error rates increase strongly when testing consecutively with sentences from the Word, Sentence, Complex Event, and Basic Event groups, as shown in Table 10.

Table 10: Percentage of cases in which a competing event erroneously receives a positive comprehension score.

	Sentences	
Group	Training	Test
Word	0.0%	0.0%
Sentence	0.0%	6.7%
Complex Event	3.9%	24.9%
Basic Event	21.2%	56.4%

We will now look at comprehension scores in each of the groups in more detail. Tables 11 to 14 display comprehension scores after processing the test sentences from each of the different groups, averaged over

all trained networks. In these tables, scores in bold are comprehension scores of described events, that is, these should be positive. The other scores are comprehension scores of competing events, and should be negative. No results are presented for events that are neither described nor competing, which is why empty cells appear in Tables 12 and 14.[5]

5.1 Word group

5.1.1 Results

As Table 11 shows, test sentences from the Word group, all stating that play(charlie, soccer), are processed very well. Especially passive test sentences are understood close to perfectly, as is apparent from the comprehension scores for the described event being close to 1. Also, the network does not wrongly believe charlie to play some other game. To understand the test sentence *charlie plays soccer*, the network must have learned that *soccer* and *football* have the same effect on the situation vector under construction, that is, that they are synonymous. The same holds for the synonym pair *charlie* and *boy* in passive test sentences.

Table 11: Relevant comprehension scores after processing test sentences from the Word group (c = charlie).

	comprehension score of		
Test sentence	play(c,soccer)	play(c,chess)	play(c,hide&seek)
charlie plays soccer	**.79**	−.88	−.78
charlie plays football	**.75**	−.85	−.78
boy plays soccer	**.75**	−.85	−.78
boy plays football	**.79**	−.88	−.78
soccer is played by charlie	**.92**	−.98	−.98
football is played by charlie	**.91**	−.97	−.95
soccer is played by boy	**.91**	−.97	−.95
football is played by boy	**.92**	−.98	−.98

[5]These results, and all other data, are available upon request.

5.1.2 Explanation

It is not hard to explain how connectionist systematicity in the Word group comes about. In short, systematicity arises because synonymous words receive highly similar representations during training. The vector of connection weights originating from an input unit can be viewed as the network's representation of the word corresponding to that unit. Obviously, if two words have identical representations, the effect of their occurrences will be identical, that is, they are perfect synonyms.

There are many training contexts in which both halves of a synonym pair occur, for example, *heidi plays soccer* and *heidi plays football* are both in the training set. The target output for these two training examples is the same, namely $\mu(\mathsf{play}(\mathsf{heidi}, \mathsf{soccer}))$. As a result, the training algorithm changes the network's connection weights in the same direction in both cases. This means that the weights of connections from input units converge if the units stand for synonymous words. Synonyms thereby receive highly similar representations and, therefore, have similar effects on the network, independent of the context in which the words appear.

Whether or not this explanation holds can be investigated by directly observing the word representations, or rather, differences between these representations. For this analysis, we used the network that showed best performance on test sentences from the Word group, for each of the two training sets. We measured the Euclidean distance between all pairs of word representations. Averaged over all word pairs, the distance was 20.5 (sd = 5.69), while the average distance between the words of the four synonym pairs was only 0.697 (sd = 0.099). Indeed, the representations of synonymous words are much more similar than those of other word pairs.

5.2 Sentence group

5.2.1 Results

Most test sentences from the Sentence group are understood quite well. As can be seen in Table 12, comprehension scores for the two described basic events and their conjunction are strongly positive. In general, the network has learned that sentences of the forms p_1 *beats* p_2 and p_2 *loses to* p_1 refer to the event $\mathsf{win}(p_1) \wedge \mathsf{lose}(p_2)$. However, the network does make a few errors in the sense that some competing events receive positive comprehension scores. For example, after processing *charlie/boy loses to heidi*, the output situation vector results in a positive comprehension score for $\mathsf{win}(\mathsf{sophia})$, even though this is clearly inconsistent to the information in the sentence, which states that heidi wins. Note, however, that this score of .08 is only marginally significantly different from 0 ($t_{19} = 1.79; p < .09$).

The reason for this error is that every training sentence starting with *charlie/boy loses to* describes

an event in which it is indeed sophia who wins (except when the winner is ambiguous, as in *charlie loses to someone*). That is, the network has learned that after the sentence fragment *charlie/boy loses to*, the output should be a situation vector representing (among others) win(sophia). This is difficult to undo fully when the sentence's last word turns out to be *heidi*. Importantly, however, the comprehension score for the described event win(heidi) is much larger than for the competing win(sophia) (.65 and .08, respectively). This means that the output vector more strongly encodes the intended microworld situation. If forced to give one winner, the information in this vector would provide the correct answer: It is heidi who wins. In general, the model makes no errors if we take the basic event with the highest comprehension score to be its response in a forced-choice task. This is analogous to an experimental setting in which subjects provide only discrete responses although their internal representations are probabilistic in nature(cf. Spivey, 2007).

Table 12: Relevant comprehension scores after processing test sentences from the Sentence group.

		comprehension score of					
Test sentence	Test event	win(charlie)	win(heidi)	win(sophia)	lose(charlie)	lose(heidi)	lose(sophia)
charlie/boy beats heidi	**.63**	**.90**	−.93	−.91	−.92	**.63**	
heidi loses to charlie/boy	**.69**	**.71**	−.94	.01	−.89	**.90**	
charlie/boy beats sophia	**.63**	**.87**	−.84	−.83	−.76		**.64**
sophia loses to charlie/boy	**.85**	**.88**	−.64	−.98	−.95		**.92**
heidi beats charlie/boy	**.62**	−.88	**.85**	−.76	**.65**	−.81	
charlie/boy loses to heidi	**.64**	−.97	**.65**	.08	**.91**	−.84	
heidi beats sophia	**.83**	−.98	**.90**	−.82		−.84	**.85**
sophia loses to heidi	**.83**	−.59	**.85**	−.85		−.83	**.90**
sophia beats charlie/boy	**.80**	−.93	−.86	**.88**	**.82**		−.90
charlie/boy loses to sophia	**.68**	−.94	.05	**.68**	**.90**		−.85
sophia beats heidi	**.76**	−.95	−.70	**.87**		**.76**	−.74
heidi loses to sophia	**.87**	−.77	−.89	**.90**		**.91**	−.89

5.2.2 Explanation

It is noteworthy that the model's analogical representations are vital for successful processing of Sentence group test sentences. As explained above, the absence of training sentences like *sophia loses to charlie*

results in a very strong learned association between the word sequence *sophia loses to* and the event $\mathsf{win}(\mathsf{heidi})$. The network will also have learned that sentences ending with *loses to charlie* refer to situations in which $\mathsf{charlie}$ wins. When tested on *sophia loses to charlie*, why then would the network not conclude both $\mathsf{win}(\mathsf{charlie})$ and $\mathsf{win}(\mathsf{heidi})$? After all, the test sentence provides evidence for both these events, considering its meaning and its similarity to the training sentences. Indeed, a model that uses symbolic representations might construct as output $\mathsf{win}(\mathsf{heidi}) \wedge \mathsf{win}(\mathsf{charlie})$, even though the microworld does not allow for this complex event. However, our model uses analogical representation that encode the structure of the microworld. As a result, a representation of $\mathsf{win}(\mathsf{charlie})$ is also a representation of $\neg\mathsf{win}(\mathsf{heidi})$. The network cannot represent a situation in which both $\mathsf{win}(\mathsf{heidi})$ and $\mathsf{win}(\mathsf{charlie})$ are very likely. It needs to choose between the two and, as it turns out, it usually chooses correctly. This shows the model's representation of microworld structure to be crucial for its systematicity.

There are two routes by which the network can come to display this level of systematicity. First, the interpretation of a test sentence may depend on its superficial similarity to particular training sentences, that is, the similarities between the literal word sequences. In that case, the network comprehends a test sentence of the form p_1 *beats* p_2 by its superficial similarity to the two training sentences p_1 *beats someone* and *someone beats* p_2. The first of these states that $\mathsf{win}(p_1)$, while the second says that $\mathsf{lose}(p_2)$. The conjunction of these two basic events is exactly the complex event described by the test sentence. We will call this route the 'conjunction route', because test sentences are understood as the conjunction of two basic events described by training sentences. Note that this strategy works similarly for test sentences of the form p_1 *loses to* p_2.

The second route, which we will call the 'inversion route' does not involve any combining of basic events. If the network takes the inversion route, it has learned that any training sentence of the form p_3 *beats* p_4 refers to the same event as its 'inverse' p_4 *loses to* p_3, namely $\mathsf{win}(p_3) \wedge \mathsf{lose}(p_4)$. Now if the network receives the *test* sentence p_1 *beats* p_2, this is interpreted by inversion as equivalent to the *training* sentence p_2 *loses to* p_1, which the network learned to map to the target output $\mu(\mathsf{win}(p_1) \wedge \mathsf{lose}(p_2))$. In this way, the network can produce the same output vector for the test sentence. As was the case for the conjunction route, it works similarly for test sentences of the form p_1 *loses to* p_2.

The network does not need to choose between one route or the other. It is quite possible that both are followed simultaneously to different degrees, or that different routes are taken for different test sentences. An analysis of the network's output, presented in Appendix B, revealed that the conjunction route is usually preferred to some extent, but that there are a few instances in which the inversion route was taken.

5.3 Complex Event group

5.3.1 Results

Results on test sentences from the Complex Event group (see Table 13) are not unlike those of the Sentence group: Comprehension scores are positive for the two basic events described, as well as their conjunction. In general, the network has learned that sentences of the forms *p plays g in x* and *g is played by p in x* refer to the complex event $\mathsf{play}(p, g) \wedge \mathsf{place}(p, x)$, even though that particular conjunction of basic events was never a target output during network training. However, there are a few instances of positive comprehension scores on competing events (i.e., errors). In particular, after processing sentences about playing hide&seek in the bedroom, the competing event play(p, bathroom) receives a comprehension score of .14. The described event play(p, bedroom) scores slightly higher but the difference is far from significant ($N = 80; z = .35; p > .7$ in a Wilcoxon matched-pairs signed-ranks test). The other two cases of positive comprehension scores for competing events are not as problematic because the corresponding described event scores much better.

Table 13: Relevant comprehension scores after processing test sentences from the Complex Event group, averaged over $p \in \{\mathsf{charlie}, \mathsf{heidi}, \mathsf{sophia}\}$. Note: $\mathsf{place}(p, \mathsf{in}) \equiv \mathsf{place}(p, \mathsf{bathroom}) \vee \mathsf{place}(p, \mathsf{bedroom})$ and $\mathsf{place}(p, \mathsf{out}) \equiv \mathsf{place}(p, \mathsf{playground}) \vee \mathsf{place}(p, \mathsf{street})$.

		comprehension score of						
Test event	Test event	play(p,chess)	play(p,hide&seek)	play(p,soccer)	place(p,bathroom)	place(p,bedroom)	place(p,playground)	place(p,street)
play(p, hide&seek) ∧ place(p, in)	**.68**	−.77	**.89**	−.99	**.39**	**.06**	−.51	−.97
play(p, hide&seek) ∧ place(p, bathroom)	**.50**	−.97	**.94**	−.99	**.46**	−.45	−.19	−.96
play(p, hide&seek) ∧ place(p, bedroom)	**.35**	−.66	**.86**	−.99	.14	**.16**	−.22	−.96
play(p, hide&seek) ∧ place(p, out)	**.35**	−.35	**.50**	−.02	−.20	−.81	**.37**	−.18
play(p, hide&seek) ∧ place(p, playground)	**.42**	−.25	**.76**	−.98	.07	−.71	**.51**	−.98
play(p, chess) ∧ place(p, out)	**.55**	**.80**	−.63	−.55	−.98	−.29	**.55**	−.70
play(p, chess) ∧ place(p, playground)	**.67**	**.87**	−.54	−.99	−.99	−.42	**.73**	−.99
play(p, chess) ∧ place(p, in)	**.50**	**.83**	−.53	−.98	−.72	**.41**	.07	−.96
play(p, chess) ∧ place(p, bedroom)	**.58**	**.87**	−.72	−.98	−.89	**.52**	−.03	−.95

5.3.2 Explanation

For comprehending test sentences from the Complex Event group, the network cannot take the inversion route described in Section 5.2. This is simply because there are no training sentences that describe the same event as the test sentences. These test sentences can therefore only be understood by taking the conjunction route, that is, test sentences *p plays g in x* (and their passive-voice counterparts) are understood by their superficial similarity to the training sentences *p plays g* and *p plays in x*, which provide play(p, g) and place(p, x), respectively. The network is able to combine these by conjunction, giving an output vector similar to $\mu(\text{play}(p, g) \wedge \text{place}(p, x))$.

5.4 Basic Event group

5.4.1 Results

Basic Event group sentences are understood more poorly than those from the other groups. Table 14 shows that described events receive positive comprehension scores, but that the same is true for many competing events. Nevertheless, described events are encoded more strongly than competing events. So, after processing test sentences describing heidi playing with the puzzle, the average comprehension score is larger for play(heidi, puzzle) than for play(sophia, puzzle). Although the difference is small, a Wilcoxon matched-pairs signed-ranks test showed that it is statistically significant ($N = 40; z = 2.11; p < .04$).

Table 14: Relevant comprehension scores after processing test sentences from the Basic Event group. c = charlie; h = heidi; s = sophia.

	comprehension score of								
Test event	play(c,doll)	play(c,ball)	play(c,puzzle)	play(h,doll)	play(h,ball)	play(h,puzzle)	play(s,doll)	play(s,ball)	play(s,puzzle)
play(c, doll)	**.20**	−.05	.01						
play(c, ball)	−.20	**.49**	−.41						
play(h, ball)				−.42	**.55**	−.56			
play(h, puzzle)			.05	.11	−.25	**.18**			.15
play(s, doll)							**.15**	−.18	.09
play(s, puzzle)			−.03			.13	.06	−.37	**.29**

This is remarkable, because all training sentences containing *heidi* describe events in which she was *not* playing with the puzzle, and all training sentences containing *puzzle* describe events in which it was

not heidi playing with the puzzle (except when the toy or player was not mentioned, as in *heidi plays with toy* and *puzzle is played with*). Therefore, a network that uses solely the learned associations between sentence fragments and situation vectors would give an output vector representing $\neg\mathsf{play}(\mathsf{heidi}, \mathsf{puzzle})$ after processing *heidi plays with puzzle*. The fact that our model does *not* display such behavior, is clear evidence that the network has learned more than mere associations between the inputs and targets in the training data.

5.4.2 Explanation

If systematicity can only result from compositional, symbolic representations, Basic Event group test sentences would not be understood correctly because there is no systematic mapping between sentences of the form *p plays with t* (or *t is played with by p*) and vectors $\mu(\mathsf{play}(p, t))$. Also, a vector for $\mathsf{play}(p, t)$ cannot be computed on the fly from play, p, and t, because the smallest meaningful unit in the model is not the concept but the basic event. Nevertheless, we do observe signs of systematicity here.

How can this be explained? Figure 4 shows the comprehension scores for several informative events, resulting from processing test sentences from the Basic Event group. For this analysis, we used the networks that performed best on these test sentences, for each of the two training sets.

First, let us look at the outcome for test sentences describing a person playing with the ball, in the center panel of Figure 4. The correct output for test sentences p_1 *plays with ball* is $\mu(\mathsf{play}(p_1, \mathsf{ball}))$, but considering the superficial similarity to the training sentences p_1 *plays with doll/puzzle*, we might expect such test sentences to incorrectly lead to output situations in which an inconsistent event $\mathsf{play}(p_1, \mathsf{doll})$ or $\mathsf{play}(p_1, \mathsf{puzzle})$ is the case. However, this not what we find: The comprehension scores for $\mathsf{play}(p_1, \mathsf{doll})$ and $\mathsf{play}(p_1, \mathsf{puzzle})$ are negative after processing p_1 *plays with ball*. Instead, the test sentence seems to be understood as sharing its meaning with training sentences about someone else playing with the ball: After processing p_1 *plays with ball*, the comprehension score for $\mathsf{play}(p_2, \mathsf{ball}) \vee \mathsf{play}(p_3, \mathsf{ball})$ is larger than that of $\mathsf{play}(p_1, \mathsf{ball})$. This might seem like a major error, but keep in mind that it is indeed very likely that p_2 or p_3 plays with the ball, given that p_1 does. A similar pattern can be seen for test sentences about p_1 playing with the doll, which are not interpreted according to their superficial similarity to training sentences about p_1 playing with another toy (which would be inconsistent with the described event), but are considered as referring to the same events as training sentences about someone else playing with the doll.

Test sentences about p_1 playing with the ball or doll are superficially similar to training sentences about p_1 playing with another toy, and to training sentences about someone else playing with the mentioned toy. However, the described event $\mathsf{play}(p_1, t_1)$ (with $t_1 = \mathsf{ball}$ or $t_1 = \mathsf{doll}$) is similar to $\mathsf{play}(p_2, t_1)$ but

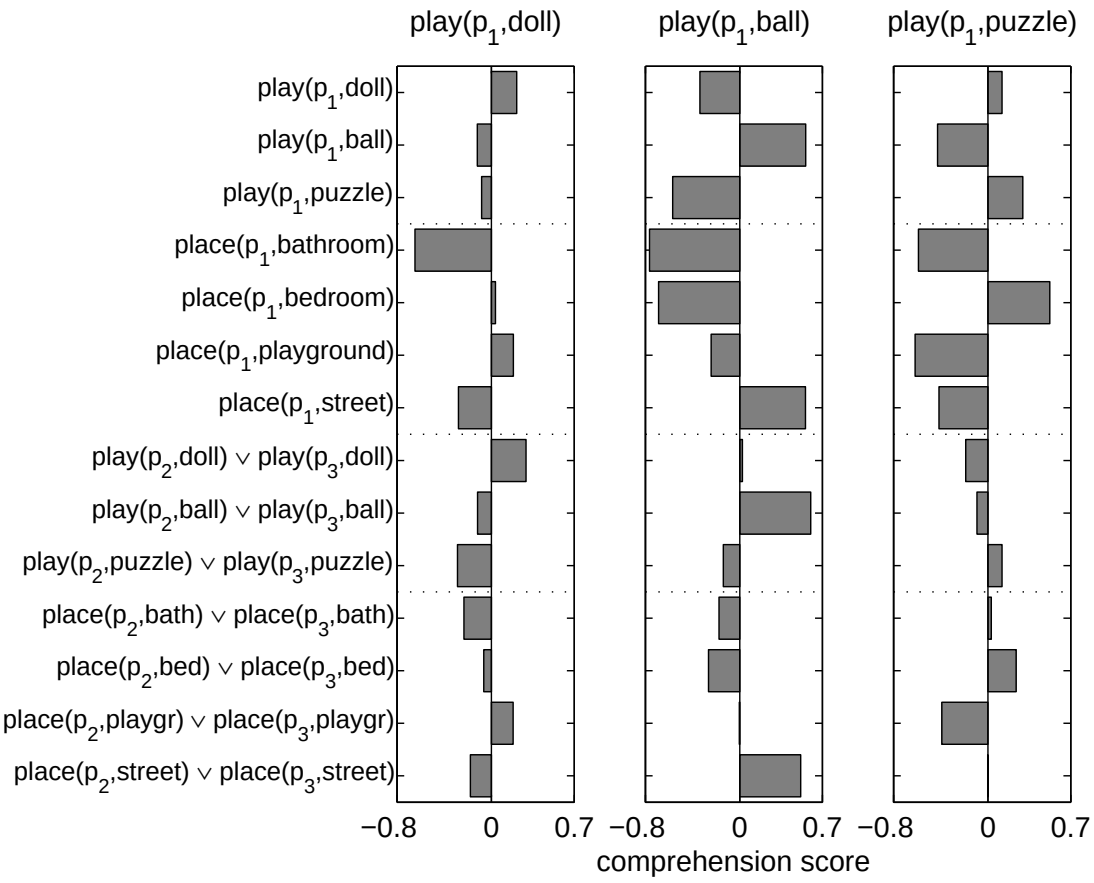

Figure 4: Averaged comprehension scores for several relevant events, resulting from processing test sentences describing play(p_1, doll) (left), play(p_1, ball) (middle), or play(p_1, puzzle) (right). The person mentioned in the test sentence is denoted p_1, the other two are p_2 and p_3. Abbreviations: bath = bathroom; bed = bedroom; playgr = playground.

not to play(p_1, t_2) (with $t_1 \neq t_2$). This is because in the microworld, two or three people often play with the ball or doll, but the same person cannot play with two different toys at the same time. Given that the network cannot directly construct the correct situation vectors for test sentences of the Basic Event group (as argued above) it does the next best thing: Interpret these test sentences by using their superficial similarity to training sentences that describe compatible events. This results in the desired outcomes because, in the microworld, the situation in which p_2 plays with the ball or doll is quite a lot like (i.e., often co-occurs with) the situation in which p_1 plays with the ball or doll, as described in the test sentence. Note that this correct performance would not have been possible without access to knowledge about the microworld, as encoded in the situation vectors.

But what if play(p_1, t) is *not* like play(p_2, t)? This is the case when $t =$ puzzle because two people

cannot play with the puzzle at the same time. As a result, test sentences p_1 *plays with puzzle* (and their passive-voice counterparts) cannot be properly understood by superficial similarity to training sentences p_2 *plays with puzzle*. Indeed, we find such test sentences to be understood much more poorly than those involving *ball* or *doll*. For example, the comprehension scores for $\mathsf{play}(p_1, \mathsf{doll})$ and $\mathsf{play}(p_2, \mathsf{puzzle}) \vee \mathsf{play}(p_3, \mathsf{puzzle})$ are slightly positive, even though these events cannot co-occur with $\mathsf{play}(p_1, \mathsf{puzzle})$. These errors result from the superficial similarity between the test sentence and training sentences.

Nevertheless, the comprehension score for the described event is clearly larger than for the incompatible events, which is remarkable considering that nearly all training sentences that are similar to this test sentence describe incompatible events. A possible explanation for this positive finding is provided in the right panel of Figure 4. Quite noticeable is the large comprehension score for $\mathsf{place}(p_1, \mathsf{bedroom})$ resulting from processing test sentences about p_1 playing with the puzzle. Indeed, someone who plays with the puzzle must be in the bedroom. However, the fact that the comprehension score for that basic event is larger than any other suggests that p_1 *plays with puzzle* is mainly interpreted as meaning $\mathsf{place}(p_1, \mathsf{bedroom})$. Given that p_1 is in the bedroom, it is indeed likely that (s)he plays with the puzzle, which explains the positive score of $\mathsf{play}(p_1, \mathsf{puzzle})$.

When processing test sentences p_1 *plays with puzzle*, the network is only minimally distracted by the superficial similarity to the training sentences p_1 *plays with ball/doll* and p_2 *plays with puzzle*, which describe incompatible events. Instead, the test sentences are correctly interpreted as referring to an event in which p_1 is in the bedroom. Again, we find that the model does quite well considering that representations of individual concepts do not exist and Basic Event group test sentences can therefore not be understood directly.

As before, we find that the analogical nature of situation vectors is crucial for this positive effect to occur. Had the representations of $\mathsf{play}(p_1, \mathsf{puzzle})$ and $\mathsf{place}(p_1, \mathsf{bedroom})$ been symbolic, the test sentence p_1 *plays with puzzle* would not have resulted in an output representing $\mathsf{place}(p_1, \mathsf{bedroom})$, and even if it would have, this output would not also encode the increased likelihood of $\mathsf{play}(p_1, \mathsf{puzzle})$.

6 Discussion

The model we presented shows that a neural network with a standard architecture can display semantic systematicity under a relatively unconstrained training regime. In part, this is accomplished by relying on the structure of the microworld, as reflected in the model's analogical representations. Additional structure is of course present in the microlanguage. The network uses these external structures to discover systematicity in the mapping from sentences to event representations. That is, the systematicity

originates externally rather than being inherent to the network.

There are a number of standard counterarguments against connectionist claims of this kind. First, the neural network may be an implementation of a symbol system. Second, the simulations may be mere demonstrations rather than providing an explanation of systematicity. Third, the model's degree of systematicity may not be comparable to that of people. Fourth, the simulations may not scale up to worlds of realistic size. In the following four subsections, we discuss how these critiques relate to our model.

6.1 Implementation of a symbol system

Fodor and Pylyshyn (1988) admit that a neural network can be systematic if it implements a classical symbol system. However, this would not constitute a *connectionist* explanation of systematicity since it would be the implemented symbol system, rather than the underlying network, that does the explaining. As discussed in the Introduction, several earlier proposals for semantically systematic connectionist models (Hadley & Cardei, 1999; Hadley & Hayward, 1997; Miikkulainen, 1996) indeed depended on symbolic representations or mechanisms.

To uphold our claim that we have successfully dealt with Fodor and Pylyshyn's challenge, it is vital that our model does not merely implement a symbol system. As we have argued before, our model accomplishes systematicity without implementing a compositional representational system, as is clear from the fact that situation vectors do not have constituent structure. It is therefore not a symbol system in the sense of Fodor and Pylyshyn, who take the presence of combinatorial syntax and semantics to be one of the hallmarks of a symbol system.

Marcus (1998b), on the other hand, defines a symbol system as one that (among other things) performs operations over variables. A connectionist model that performs such operations could therefore be considered an implementation of a symbol system. For instance, the Recursive Auto-Associative Memory (RAAM) models proposed by Chalmers (1990) and Niklasson and van Gelder (1994), consist of two networks: The first learns to encode the possible instantiations of a variable, and the second performs transformations over the resulting representations. This architecture, so Marcus (1998b) argues, 'precisely parallels the division between encoding and computation in standard symbolic models' (p. 270) and is therefore not a relevant counterexample to Fodor and Pylyshyn's (1988) claims. Contrary to this, our network, being a standard SRN, does not make such a distinction between instantiations of variables and operations over variables, so it is not a symbol system in the sense of Marcus (1998b).

One could argue that our model's separation between situation-space development and SRN training

comes down to a division between encoding and computation, similar to that in RAAM models. However, even if the occurrence of a particular situation vector is considered to be the instantiation of a variable, our model does not perform any operations over this variable. Instead, the situation vector becomes 'instantiated' by the SRN during sentence comprehension.

6.2 Demonstration versus explanation

Demonstrating that a specific connectionist model can display systematicity is not enough for *explaining* systematicity because the apparent structure-sensitive behavior of a network might simply be the result of a specific arrangement of the network's representations or architecture. For example, Frank (2006) noted that the models by Bodén (2004) and Hadley et al. (2001) only managed to behave systematically because they were specifically tailored for that purpose. Likewise, the systematic connectionist model proposed by Niklasson and van Gelder (1994) was criticized by Hadley (1994a) and Phillips (1998) for depending on hand-crafted input representations that explicitly encoded syntactic class information. In addition, Haselager and van Rappard (1998) remarked that the model required a very extensive and carefully arranged training regime.

Such counterarguments illustrate that assessing the explanatory value of connectionist examples of structure-sensitive processing is far from straightforward. The matter of distinguishing real systematicity from prearranged performance comes to the fore in the discussion about Fodor's repeated claim that merely providing examples of connectionist systematicity is far from sufficient to show that connectionism can deal with systematicity in a completely satisfactory way. As he says, it is a *law* that cognitive capacities are systematic (Fodor & McLaughlin, 1990; Fodor & Pylyshyn, 1988).[6]

According to a relatively early interpretation of the law-requirement (Butler, 1993; see also Aizawa, 1997b), the idea is that it is not enough to merely show that systematicity is *possible* on the basis of a connectionist architecture; It must be indicated why systematicity is *necessary* given the architecture. Likewise, Butler says, a theory of planetary motion that merely allowed for the possibility of elliptical orbits of planets would be considered as insufficient. To really count as an explanation, it would have to show that the nature of such orbits necessarily followed from the theory. Similarly, connectionists have to demonstrate that systematicity necessarily follows from the architecture.

Aizawa (1997b, 2002) has taken the debate a step further by indicating that the requirement that the explanans must necessitate the explanandum is not formulated sufficiently exact. As he says, the Ptolomean theory of planetary motion *does* necessitate the observed trajectories of the planets. The

[6]Fodor's claim about the lawfulness of systematicity has been questioned (e.g., Dennett, 1991; McNamara, 1993; Sterelny, 1990; Wilks, 1990). See also Note 8.

problem is that it does this in an ad hoc or prefabricated way (i.e., by the use of several, not independently well-motivated additional hypotheses, such as epicycles). Formulated in the context of systematicity:

> once you have LOT [Language of Thought], you automatically get the systematicity of thought. There are no arbitrary hypotheses in the explanation. ... If a network can as easily generate a set of systematic representations as not, then there must be in Connectionism some arbitrary hypothesis. (Aizawa, 1997b, pp. 120–121)

So the question becomes, what would count as arbitrary, as distinct from well-motivated, non-arbitrary, hypotheses? The history and philosophy of science do not, as Aizawa (2002) notes, provide a definitive answer to these questions. We cannot address this issue fully here, but instead merely try to indicate in general terms why our additional hypotheses are not to be considered as arbitrary.

Traditionally, connectionist solutions to the problem of systematicity are sought in architectural constraints, combined with specifics of training data. Such an approach is unlikely to succeed in our opinion, because the specifics of the architectures and training procedures appear to be chosen to achieve the desired results rather than being independently motivated. Moreover, the results are obtained by limiting the robustness of the network. If the performance of a network is overly dependent on the details of its architecture and/or training regime, it cannot be a satisfactory model of natural cognition that, after all, displays systematicity under a wide variety of circumstances. As Chalmers (1993) suggests, networks need not only have an appropriate architecture but also have to display systematicity under many different learning conditions. This emphasizes the fact that merely demonstrating a network to be systematic is not sufficient, since the performance achieved might be an artificial result of the specific characteristics of the network and the training and test data. In developing our model, we therefore aspired to make it as simple and general as possible, and refrained from using a sophisticated architecture, training algorithm, training regime, or search for optimal parameter settings. Also, our results do not seem to depend crucially on the particular microlanguage, microworld, or network architecture: Frank and Haselager (2006) present similar findings using a simpler language and world, and different architecture. Also, they show their results to be highly robust to differences in parameter setting.

Of course there is *something* additional that helps to generate the systematicity displayed by our model. Systematicity does not come about for free. Still, we would like to argue that we did not invoke anything *arbitrary*. To explain this, we refer back to Simon's (1969/1996) classical example of the ant on the beach. The ant's behavior looks complicated and difficult to describe. Yet the complexity may not reside within the ant, but could arise out of the complexity of the surface of the beach. The same, Simon suggests, might be true for human beings: "Human beings, viewed as behaving systems, are quite

simple. The apparent complexity of our behavior over time is largely a reflection of the complexity of the environment in which we find ourselves" (p. 53).

This suggestion, we submit, could very well apply to systematicity as well. Because of the systematic features of the *environment*, a very general connectionist architecture under a very unrestricted training regime can develop systematicity. The world does not consist of an arbitrary set of unrelated events, and the representational resources that cognitive systems are endowed with might be sufficiently equipped to be able to pick up this 'worldly degree' of systematicity under an appropriately wide variety of circumstances. Contrary to the demand that systematicity should follow necessarily from the architecture, that is, that the representational system in itself should be intrinsically systematic, the suggestion we present here is that the displayed systematicity derives from the *interaction* between the architecture and its environment. It may well be that the systematicity of human cognition depends more on only 'weakly' representational resources combined with a largely systematic world, than on the cognizers having somehow a built-in intrinsically systematic representational system. A representational system capable of reflecting the systematicity in the environment could suffice for displaying a psychologically plausible degree of systematicity.

This idea of combining internal and external constraints to model or generate specific behavioral and cognitive phenomena is of course not new. Bechtel and Abrahamsen (1991) follow (among others) Rumelhart, Smolensky, McClelland, and Hinton (1986) in suggesting that "networks may develop the capacity to interpret and produce symbols that are external to the network. ... In the externalist approach to symbol processing the focus is turned from symbols in their mental roles to symbols in their external roles" (Bechtel & Abrahamsen, 1991, pp. 248–249). This use of external structures could, they argue, provide a connectionist means of obtaining systematicity. From the late 1980s and early 1990s onwards, the idea that cognition is embedded in the world has gained support (e.g., Brooks, 1991; Chiel & Beer, 1997; Clancey, 1997; Clark, 1997; Thelen & Smith, 1994, to name but a few). From this perspective, cognitive phenomena should be modeled not on a purely internalist basis, but explicitly taking external factors into account, among which the systematicity found in the world.

Our hypothesis that it is the structure inherent in the world that allows a connectionist model to display systematicity is not arbitrary, but rather well-motivated. In continuation of Butler's (1993) and Aizawa's (1997b, 2002) analogy with planetary motion, invoking features of the environment to explain systematicity is comparable to explaining the earth's trajectory by positing the existence of the sun. This does add an extra hypothesis to the laws of astronomy, but an explanatory relevant and empirically justified one.

6.3 Degree of systematicity

It is difficult to judge whether the model displays the same degree of systematicity as does the human cognitive system. Even if it could somehow be established how systematic people are, it is unclear how this might be compared to the model's performance. After all, the model learns a very simple language and receives minimal information about a tiny world, whereas people have a full-blown language and rich knowledge of a highly complex world. Therefore, we would not expect the model to reach the same levels of systematicity as people do.

Nevertheless, to uphold our claim that connectionist systematicity is possible, certain aspects of semantic systematicity that can be observed in people, should also be available to the model. We have already demonstrated that the model comprehends the occurrence of synonyms in new contexts, as well as new combinations of phrases, even if these refer to new complex events. Moreover, we found the model to be able to deal with new combinations of concepts (to be more precise, of people and toys), which is remarkable considering that the model's representations hold no meaningful content at a more fine-grained level than the basic event.

These four degrees of systematicity, corresponding to the four groups of test sentences, seem easily manageable by people as well. Despite these successes, however, it may be argued that the model's level of systematicity does not suffice. This raises the question which level of generalization a network needs to reach in order to be considered 'systematic enough'. Since a network's degree of systematicity corresponds to the level of input novelty it can tolerate (Hadley, 1994a), the question becomes how strongly the test sentences need to differ from the training examples.

Frank and Čerňanský (2008) argue that, at the very least, one or more specific groups of sentences should be excluded from the training set, as was the case in our simulations. This prevents the distribution of the training sample from accurately reflecting the true distribution, making it impossible for the network to correctly process the withheld sentences by simple interpolation from the training examples. Instead, the network needs to have learned about the system that generated the training and test sentences. In the connectionist sentence-comprehension models by Desai (2007), Miikkulainen and Dyer (1991), and St.John and McClelland (1990), test sentences are unlikely to differ strongly from training examples because each sentence is randomly assigned to either the training or the test set. As a result, the generalization displayed by these models does not indicate any systematicity.

Other authors have came up with stricter definitions of sufficient systematicity. Below, we discuss how two of these relate to our model.

6.3.1 Words in novel grammatical roles

According to Hadley (1994a), a neural network exhibits so-called 'strong systematicity' in sentence processing if it handles test sentences with words in "syntactic positions" (p. 249) they did not occupy during training. In practice, this means that the grammatical subjects of training sentences are objects in the test sentences (and vice versa).[7] Hadley (1994a) argues that people display strong systematicity, unlike the connectionist models proposed by Chalmers (1993), Elman (1990), Pollack (1990), and St.John and McClelland (1990).

If our model is to be strongly systematic, it should understand test sentences of the form p_1 *beats* p_2 without being trained on *any* sentence containing the verb phrase *beats* p_2 or *loses to* p_2.[8] This is trivially achieved when p_2 is *charlie* or *boy* because, as we have shown in Section 5.1, synonymous words have almost identical effects on the network. Therefore, even if no training sentence contains *beats boy* or *loses to boy*, the network can process these phrases correctly if it was trained on *beats charlie* and *loses to charlie*. As Hadley and Cardei (1999) remark, however, restricting strong systematicity to words with a synonym "would certainly violate the spirit of the definition of strong systematicity" (p. 218).

At first glance, it may seem unlikely that the network can comprehend a test sentence with *heidi* or *sophia* in object position if it has not been trained on *any* such sentence. This is because there is no systematic relation between verb phrases *beats p* and event vectors $\mu(\mathsf{lose}(p))$, nor between *loses to p* and $\mu(\mathsf{win}(p))$. Without training exposure to a particular phrase-vector pair, that phrase can therefore not be processed correctly. Nevertheless, the network might be able to exhibit strong systematicity to some extent. Be reminded from Section 5.2 that test sentences p_1 *loses to* p_2 are occasionally processed by their systematic relation (in both form and meaning) to training sentences p_2 *beats* p_1. In principle, this allows for comprehension of p_1 *loses to* p_2 even if p_2 never appeared as object in training sentences.

We investigated the model's potential for strong systematicity by training ten networks again, but using an adapted Sentence group: The training set contained no sentence with *charlie* or *boy* in object position. After training, each network was tested on the four sentences *heidi/sophia loses to charlie/boy*, all stating that $\mathsf{win}(\mathsf{charlie})$. The results, presented in Table C7 of Appendix C, show that the compre-

[7] As an additional requirement for strong systematicity, sentences should have embedded clauses containing words in new syntactic positions. Since our microlanguage's sentences do not have embedded clauses, we shall not discuss this requirement.

[8] Note that, in our microlanguage, strong systematicity is only relevant to sentences of the form p_1 *beats* p_2 and p_1 *loses to* p_2, that is, the Sentence group. This is because 'sentences' like *toy plays with girl* and *boy is played by game* are meaningless and, therefore, do not need to be generalized to. Incidentally, this observation raises doubts about the validity of unrestricted assertions concerning systematicity, such as Fodor and McLaughlin's (1990) claim that "it is a law of nature that you can't think aRb if you can't think bRa" (p. 203).

hension scores for win(charlie) are positive. This is remarkable since all training sentences beginning with *heidi/sophia loses to* (except those in which the object is *someone*) describe events in which win(charlie) is *not* the case. Therefore, this result is indicative of strong systematicity.

However, the comprehension scores for one of the inconsistent events win(sophia) and win(heidi) is positive, while it should be negative. The problem here is that the microworld only has three people. Since charlie is never mentioned as object, all training sentences of the form *heidi loses to p* describe events in which sophia wins (except when $p = someone$), creating a strong association between the phrase *heidi loses to* and the event win(sophia). Similarly, the phrase *sophia loses to* becomes associated to win(heidi). If the microworld held more than three people, many of the training sentences *sophia loses to p* would not state that heidi wins. As a result, the test sentence *sophia loses to charlie* would not lead to a large comprehension score of win(heidi). Nevertheless, even with our three-person microworld, we found promising signs of strong systematicity, in that win(charlie) correctly received a positive comprehension score.

6.3.2 Generalizing outside the training space

Marcus (1998a, 1998b, 2001) argues that neural networks cannot generalize to items that lie 'outside the training space', meaning that they contain input values that were not present in any training example. A well-known example is the following: a SRN is trained to predict the next word at each point of sentences like *A rose is a rose* and *A lily is a lily*. After training, it is tested with the input *A blicket is a* ____, where *blicket* is a novel word. People invariably respond that the next word will be *blicket*, but the SRN produces *rose* or *lily* (or something in between). It is not difficult to see why this is so: The weight of the connection from the input unit representing *blicket* has never been updated, because the word never occurred during training. When the new word does finally occur, the network's best guess is to predict *rose* or *lily* again, as it learned to do after the words *is a*.

Our network, too, would not be able to understand sentences containing a word that did not occur during training. When given the test sentence *heidi plays with blicket*, it could not construct a situation vector representing play(heidi, blicket). Importantly, however, people will also have difficulties imagining heidi playing with a blicket if that concept is completely new to them. The model's failure to represent play(heidi, blicket) is appropriate considering that it takes mental simulation, and not the construction of a predicate-argument structure, as the cognitive process relevant to sentence comprehension. Moreover, generalization outside the training space does seem possible for neural networks trained on next-word prediction: Altmann (2002) shows that SRNs can generalize to novel input items if they have enough prior exposure to sequential structure.

6.4 Scalability

The extent to which our model scales up remains to be investigated. However, it is important to note that the issue of scalability is orthogonal to that of systematicity. Fodor and Pylyshyn (1988) did not argue that only small-scale connectionist models can display systematicity, and none of their arguments against connectionist systematicity are restricted to large-scale models. So, even if our model turns out to suffer from scalability problems, it still challenges Fodor and Pylyshyn's claims.

Having said that, we do recognize that scalability is necessary for any model, connectionist or symbolic, to be cognitively plausible (i.e., functional in a realistic world). When applying connectionist models to domains of real-world size and complexity, two problems of scalability can arise: First, the size of networks required to implement the modeled capability may grow out of bounds (Parberry, 1994). Second, the time required for the network to learn the required connection weights may become unrealistically long (Judd, 1990).

Let us first consider the network's size, which depends in large part on the size of its output layer. This, in turn, depends on the size of the microworld. One concern may be that a 150-unit output layer does not suffice to represent larger worlds because the number of required units grows with world size. Although this intuition is likely to be correct, one should keep in mind that what matters for the size of the vector representations is not so much the size of the world but rather the number of independent events in the world. As the world gets larger, there will be more dependencies among events, so the number of necessary situation-space dimensions may grow slower than the number of basic events. This expectation is consistent with the finding that our situation space had the same number of dimensions as Frank et al.'s (2003), even though their microworld was much simpler (having only 14 basic events). Moreover, our belief values estimated the microworld's co-occurrence probabilities more accurately than did theirs (compare our Figure A5 to their Figure 3).

As for the scalability of network training, it is hard to predict how learning time will increase for larger and more complex worlds and languages. It is known that backpropagation learning in general is NP-hard[9] (Šíma, 1996), which may encourage pessimism about the scalability of the learning algorithm. However, the intractability of general backpropagation learning does not mean that scalability is impossible for backpropagation in general. The NP-hardness result merely means that not *all* backpropagation learning is efficient. Whether or not our network's weights can be efficiently trained —by backpropagation or otherwise— is an open question.

[9]If a computation is NP-hard then it cannot be computed in a practicable (i.e., polynomial) time, unless a conjecture most mathematicians conjecture to be true (i.e., $P \neq NP$) turns out to be false (see, e.g., Garey & Johnson, 1979 for more details).

Whereas network size and learning time are the important scaling factors for connectionist models, inferential time is the bottleneck in symbolic models. In analogical models (such as ours), inference is direct, but in symbolic models, the time required to unpack and compute the implications of representational changes can easily become prohibitive for larger domains (Ford & Pylyshyn, 1996; Haselager, 1997; Pylyshyn, 1987). In practice, almost all cognitive models of sufficient power and generality are plagued by computational intractability (Bylander, 1994; Cook, 1971; Cooper, 1990; Levesque, 1988; Roth, 1996; van Rooij & Wareham, 2008). This, to us, signals that intractability cannot at present be used as an argument for one modeling framework or another (see also van Rooij, 2008). Be that as it may, the scalability of our account of systematicity does need to be established. In this paper we demonstrated the *in principle* possibility of connectionist semantic systematicity. We hope that future research may establish its practical feasibility as well.

7 Conclusion

We have presented a connectionist model of sentence comprehension that displays a considerable degree of systematicity. The model simulates sentence comprehension as the transformation of a sentence into an analogical vector representation of a described situation in a microworld. Importantly, the model is purely connectionist: In contrast to several previous connectionist attempts to model systematicity, it does not implement a symbol system. Also, our simulations are more than just a demonstration of connectionist systematicity, because we did not resort to ad hoc or arbitrary assumptions to bring about the observed systematicity. Instead, the systematicity is developed robustly because it derives from the structure that is present in the world as well as the language used to described that world.

The origin of systematicity should be sought in the cognitive system's embeddedness in the world rather than in inherent properties of the system itself. By doing so, our simulations provide evidence against Fodor and Pylyshyn's (1988) claim that connectionism cannot explain systematicity.

Acknowledgements

We would like to thank Ken Aizawa, Michael Klein, Paco Calvo Garzón, and three anonymous reviewers for helpful comments on an earlier version of this paper. The research presented here was supported by grant 451-04-043 of the Netherlands Organization for Scientific Research (NWO).

A Competitive-Layer training

Each Competitive Layer unit i is associated to a weight vector $\mu_i \in [0,1]^{44}$ and a single bias value b_i. Initially, all weights are .5 and all biases are 1. Element $S_k(a)$ of microworld observation vector $S_k \in \{0,1\}^{44}$ has a value of 1 if basic event a occurs at instant k, and 0 otherwise. During each of 20 training epochs, the following is repeated for all observations S_k:

1. For every unit i, determine the cityblock distance between i's weight vector and the current observation: $\mathrm{d}(\mu_i, S_k) = \sum_a |\mu_i(a) - S_k(a)|$.
2. Determine the winner w, this is the unit with the shortest *biased* distance to the input: $w = \mathrm{argmin}_i(\mathrm{d}(\mu_i, S_k) - b_i)$.
3. Update the winner's weight vector: $\Delta\mu_w = \alpha(S_k - \mu_w)$, with α the weight learning-rate parameter.
4. Decrease the winner's bias (to a minimum of 1): $\Delta b_w = \beta b_w(1 - b_w)$, with β the bias learning-rate parameter.
5. Increase the biases of all losers: $\Delta b_i = \beta b_i$ (for every $i \neq w$).

Learning rates are initially set at $\alpha = 1$ and $\beta = 10^{-4}$. After each of the first 10 training epochs, their values are reduced linearly to end up at 10% of the initial values. Over the last 10 epochs, they remain at these levels.

Figure A5 shows the resulting similarity between individual estimated probabilities (from Equations 1 and 2) and corresponding probabilities in the microworld. For basic events, the two are virtually identical ($r = 1$). For conjunctions and conditional probabilities, the correlation is very strong ($r = .997$ and $r = .996$, respectively) and there are no extreme outliers. These results indicate that the vectors μ indeed encode the regularities in the microworld and form the desired representations of basic events.

B Processing Sentence group test sentences

As explained in Section 5.2, there are two ways by which the network can comprehend test sentence from the Sentence group: the 'conjunction route' and the 'inversion route'. We can find out to what extent one of the routes is preferred by looking at the network's output. If the output resulting from the test sentence p_1 *beats* p_2 is very similar to the result of the training sentence p_2 *loses to* p_1, then the network seems to have interpreted p_1 *beats* p_2 by analogy with p_2 *loses to* p_1, that is, it took the inversion route. On the other hand, if the network's output is very much like the conjunction of the outputs resulting from p_1 *beats someone* and *someone beats* p_2, then it took the conjunction route.

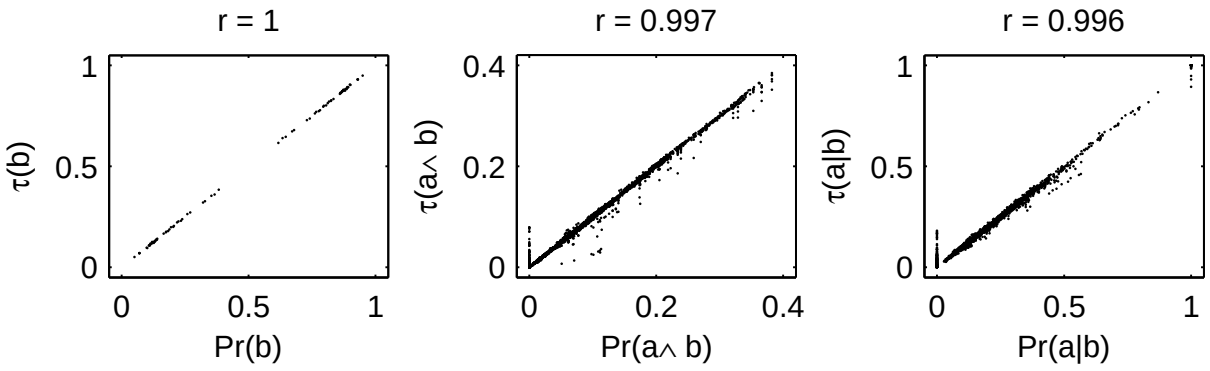

Figure A5: Scatter plot of actual (Pr) against estimated (τ) probabilities. Basic events are indicated by a, while b denotes basic events or negations thereof. Left: prior probabilities of (negations of) basic events. Middle: prior probabilities of conjunctions of (negations of) basic events. Right: conditional probabilities of basic events.

The 'conjunction preference' is a measure for the extent to which a particular sentence is processed more by the conjunction route than by the inversion route. When processing the test sentence p_1 *beats* p_2, the conjunction preference is computed as follows: The network processes the sentence and the resulting comprehension scores of all 44 basic events are computed. Let $\vec{c}_{\text{test}}$ denote the 44-element vector containing these comprehension scores. Likewise, $\vec{c}_{\text{inv}}$ contains comprehension scores resulting from processing the training sentence p_2 *loses to* p_1, and $\vec{c}_{\text{con}}$ are the comprehension scores in the conjunction (Equation 4) of the outputs resulting from training sentences p_1 *beats someone* and *someone beats* p_2. If the network would process training sentences to perfection, $\vec{c}_{\text{inv}} = \vec{c}_{\text{con}}$ because the first training sentence describes the same event as the conjunction of the latter two. In practice, however, the two vectors will be unequal.

If test sentence p_1 *beats* p_2 is processed through the conjunction route, $\vec{c}_{\text{test}}$ will equal $\vec{c}_{\text{con}}$. If the test sentence is processed through the inversion route, $\vec{c}_{\text{test}} = \vec{c}_{\text{inv}}$. Which of the two routes is preferred is measured by comparing the correlation between $\vec{c}_{\text{test}}$ and $\vec{c}_{\text{con}}$ (denoted $r_{\text{con, test}}$) to the correlation between $\vec{c}_{\text{test}}$ and $\vec{c}_{\text{inv}}$ (denoted $r_{\text{inv, test}}$). The average values for $r_{\text{con, test}}$ and $r_{\text{inv, test}}$ were .94 and .90, respectively, with minima of .72 and .63. Such high values were to be expected considering that the network simulates comprehension of both training and test sentences very well (in case of perfect comprehension, all correlations would be 1).

Formally, the extent to which the conjunction route is preferred over the inversion route is defined as

$$\text{conjunction preference} = \frac{r_{\text{con, test}} - r_{\text{inv, test}}}{2 - r_{\text{con, test}} - r_{\text{inv, test}}}.$$

The conjunction preference is positive if $r_{\text{con, test}} > r_{\text{inv, test}}$ and negative in the opposite case. In the marginal cases, where $r_{\text{con, test}} = 1$ or $r_{\text{inv, test}} = 1$, the conjunction preference is $+1$ or -1, respectively. If $r_{\text{con, test}} = r_{\text{inv, test}}$, no one route is preferred over the other, so conjunction preference equals 0.

The histogram in Figure B6, plotting data from all Sentence Group test sentences and all trained networks, shows a clear preference for the conjunction route. However, the conjunction preference is sometimes negative, indicating that the inversion route was taken.

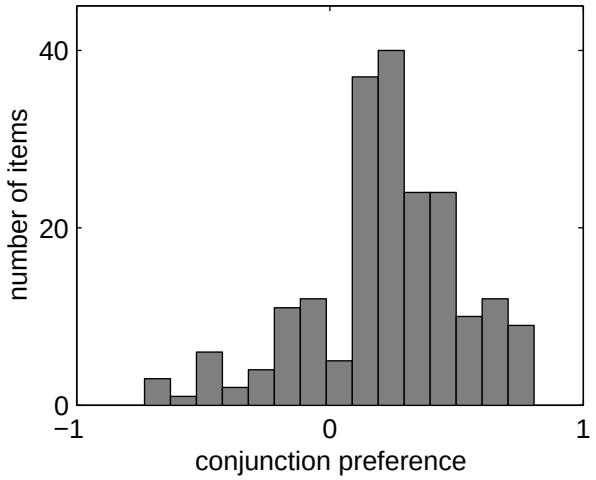

Figure B6: Histogram of preferences for conjunction over inversion route, for 200 test items.

C Strong systematicity results

Table C7 shows the results of investigating the model's ability to display strong systematicity. See Section 6.3 for details.

Table C7: Comprehension scores after processing test sentences for investigating strong systematicity.

	comprehension score of		
Test sentence	win(charlie)	win(heidi)	win(sophia)
heidi loses to charlie/boy	**.39**	−.87	.36
sophia loses to charlie/boy	**.28**	.48	−.90

References

Aizawa, K. (1997a). Exhibiting versus explaining systematicity: a reply to Hadley and Hayward. *Minds and Machines*, *7*, 39–55.

Aizawa, K. (1997b). Explaining systematicity. *Mind & Language*, *12*, 115–136.

Aizawa, K. (2002). *The systematicity arguments.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Altmann, G. T. M. (2002). Learning and development in neural networks — the importance of prior experience. *Cognition*, *85*, B43–B50.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660.

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the mind.* Oxford, UK: Blackwell.

Bodén, M. (2004). Generalization by symbolic abstraction in cascaded recurrent networks. *Neurocomputing*, *57*, 87–104.

Bodén, M., & Niklasson, L. (2000). Semantic systematicity and context in connectionist networks. *Connection Science*, *12*, 111–142.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*, 139–159.

Budiu, R., & Anderson, J. R. (2004). Interpretation-based processing: a unified theory of semantic sentence comprehension. *Cognitive Science*, *28*, 1–44.

Butler, K. (1993). Connectionism, classical cognitivism and the relation between cognitive and implementational levels of analysis. *Philosophical Psychology*, *6*, 321–333.

Bylander, T. (1994). The computational complexity of propositional STRIPS planning. *Artificial Intelligence*, *69*, 165–204.

Chalmers, D. J. (1990). Syntactic transformations on distributed representations. *Connection Science*, *2*, 53–62.

Chalmers, D. J. (1993). Connectionism and compositionality: why Fodor and Pylyshyn were wrong. *Philosophical Psychology*, *6*, 305–319.

Chang, F. (2002). Symbolically speaking: a connectionist model of sentence production. *Cognitive Science*, *26*, 609–651.

Chiel, H. J., & Beer, R. D. (1997). The brain has a body: Adaptive behavior emerges from interactions of nervous system, body and environment. *Trends In Neurociences*, *20*, 553–557.

Clancey, W. J. (1997). *Situated cognition: On human knowledge and computer representation.* Cambridge, UK: Cambridge University Press.

Clark, A. (1997). *Being there: Putting brain, body and world together again.* Cambridge, MA: MIT Press.

Cook, S. (1971). The complexity of theorem-proving procedures. In *Proceedings of the 3rd annual ACM symposium on Theory of Computing* (pp. 151–158). New York: ACM Press.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, *42*, 393–405.

Dennett, D. C. (1991). Mother nature versus the walking encyclopedia: a western drama. In W. Ramsey, S. Stich, & D. Rumelhart (Eds.), *Philosophy and connectionist theory* (pp. 21–30). Hillsdale, NJ: Erlbaum.

Desai, R. (2007). A model of frame and verb compliance in language acquisition. *Neurocomputing*, *70*, 2273–2287.

Dominey, P. F. (2005). Emergence of grammatical constructions: evidence from simulation and grounded agent experiments. *Connection Science*, *17*, 289–306.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Fodor, J. A., & McLaughlin, B. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution does not work. *Cognition*, *35*, 183–204.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, *28*, 3–71.

Ford, K. M., & Pylyshyn, Z. W. (Eds.). (1996). *The robot's dilemma revisited: The frame problem in Artificial Intelligence.* Norwood, NJ: Ablex.

Frank, S. L. (2006). Learn more by training less: systematicity in sentence processing by recurrent networks. *Connection Science*, *18*, 287–302.

Frank, S. L., & Haselager, W. F. G. (2006). Robust semantic systematicity and distributed representations in a connectionist model of sentence comprehension. In R. Sun & N. Miyake

(Eds.), *Proceedings of the 28th annual conference of the Cognitive Science Society* (pp. 226–231). Mahwah, NJ: Erlbaum.

Frank, S. L., Koppen, M., Noordman, L. G. M., & Vonk, W. (2003). Modeling knowledge-based inferences in story comprehension. *Cognitive Science*, *27*, 875–910.

Frank, S. L., & Čerňanský, M. (2008). Generalization and systematicity in echo state networks. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness.* New York: Freeman.

Hadley, R. F. (1994a). Systematicity in connectionist language learning. *Mind & Language*, *9*(3), 247–272.

Hadley, R. F. (1994b). Systematicity revisited: reply to Christiansen and Chater and Niklasson and van Gelder. *Mind & Language*, *9*, 431–444.

Hadley, R. F. (2004). On the proper treatment of semantic systematicity. *Minds and Machines*, *14*, 145–172.

Hadley, R. F., & Cardei, V. C. (1999). Language acquisition from sparse input without error feedback. *Neural Networks*, *12*, 217–235.

Hadley, R. F., & Hayward, M. B. (1997). Strong semantic systematicity from Hebbian connectionist learning. *Minds and Machines*, *7*, 1–37.

Hadley, R. F., Rotaru-Varga, A., Arnold, D. V., & Cardei, V. C. (2001). Syntactic systematicity arising from semantic predictions in a Hebbian-competitive network. *Connection Science*, *13*(1), 73–94.

Haselager, W. F. G. (1997). *Cognitive science and folk psychology: The right frame of mind.* London: Sage.

Haselager, W. F. G., & van Rappard, J. F. H. (1998). Connectionism, systematicity, and the frame problem. *Minds and Machines*, *8*, 161–179.

Haugeland, J. (1987). An overview of the frame problem. In Z. W. Pylyshyn (Ed.), *The robot's dilemma: the frame problem in Artificial Intelligence* (pp. 77–93). Norwood, NJ: Ablex.

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge, UK: Cambridge University Press.

Judd, J. S. (1990). *Neural network design and the complexity of learning.* Cambridge, MA: MIT Press.

Kerkhofs, R., & Haselager, W. F. G. (2006). The embodiment of meaning. *Manuscrito – Revista Internacional de Filosofia*, *29*, 753–764.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363–394.

Kohonen, T. (1995). *Self-organizing maps.* Berlin: Springer.

Levesque, H. J. (1988). Logic and the complexity of reasoning. *Journal of Philosophical Logic*, *17*, 355–389.

Marcus, G. F. (1998a). Can connectionism save constructivism? *Cognition*, *66*, 153–182.

Marcus, G. F. (1998b). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*, 243–282.

Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science.* Cambridge, MA: MIT Press.

Mayberry, M. R., Crocker, M. W., & Knoeferle, P. (in press). Learning to attend: A connectionist model of situated language comprehension. *Cognitive Science.*

McNamara, P. (1993). Introduction. *Philosophical Studies*, *71*, 113–118.

Miikkulainen, R. (1996). Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science*, *20*, 47–73.

Miikkulainen, R., & Dyer, M. G. (1991). Natural language processing with modular PDP networks and distributed lexicon. *Cognitive Science*, *15*, 343–399.

Niklasson, L. F., & van Gelder, T. (1994). On being systematically connectionist. *Mind & Language*, *9*, 288–302.

Parberry, I. (1994). *Circuit complexity and neural networks.* Cambridge, MA: MIT Press.

Peirce, C. S. (1903/1985). Logic as semiotics: The theory of signs. In R. E. Innis (Ed.), *Semiotics: An introductory anthology* (pp. 4–23). Bloomington, IN: Indiana University Press.

Phillips, S. (1998). Are feedforward and recurrent networks systematic? Analysis and implications for a connectionist cognitive architecture. *Connection Science*, *10*, 137–160.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*, 77–105.

Pylyshyn, Z. W. (Ed.). (1987). *The robot's dilemma: The frame problem in Artificial Intelligence.* Norwood, NJ: Ablex.

Roth, D. (1996). On the hardness of approximate reasoning. *Artificial Intelligence*, *82*, 273–302.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

Simon, H. (1969/1996). *The sciences of the artificial.* Cambridge, MA: MIT Press.

Spivey, M. J. (2007). *The continuity of mind.* New York: Oxford University Press.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, *12*, 153–156.

Sterelny, K. (1990). *The representational theory of mind.* Oxford, UK: Blackwell.

St.John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*, 217–257.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action.* Cambridge, MA: MIT Press.

van Gelder, T. (1990). Compositionality: a connectionist variation on a classical theme. *Cognitive Science*, *14*, 355–384.

van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science*, *32*, 939–984.

van Rooij, I., & Wareham, T. (2008). Parameterized complexity in cognitive modeling: Foundations, applications and opportunities. *Computer Journal*, *51*, 385–404.

Šíma, J. (1996). Back-propagation is not efficient. *Neural Networks*, *9*, 1017–1023.

Wilks, Y. (1990). Some comments on Smolensky. In D. Patridge & Y. Wilks (Eds.), *The foundations of AI* (pp. 327–336). Cambridge, UK: Cambridge University Press.

Zwaan, R. A. (2004). The immersed experience: toward an embodied theory of language comprehension. In B. Ross (Ed.), *The psychology of language and motivation* (Vol. 44, pp. 35–62). New York: Academic Press.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, *13*, 168–171.