Composition

Elie Bienenstock
Division of Applied Mathematics, Brown University, Box F
Providence, RI 02912, USA
elie@dam.brown.edu
on leave from CNRS, Paris, France *

This chapter is a composition on the theme of binding and composition. The binding problem—the investigation of the mechanisms used by the brain to bind with each other the representations of different parts or features of an object—has motivated a number of studies in recent years. In particular, attention has been given to the possible role of accurate temporal structure of multi-neuron spike trains in cortex [1–3]. The present paper uses linguistic examples to suggest that the mental material that is operated upon by the establishment of dynamical bonds largely consists, in itself, of dynamical bonds. Further, using the construction game *Lego* as a metaphor for compositional symbol systems, we stress the role of *cooperativity* among elementary bonds. We use the term *map* to refer to a *coherent* system of bonds, i.e., a system where elementary bonds cooperate with each other. We propose to view the fundamental operation of mental composition as consisting of the doing and undoing of maps between maps ... between maps. Cooperative binding in Lego can be characterized mathematically as the establishment of *commutative* mapping diagrams, and we outline an approach to language in which the construction of meaning consists of the establishment of recursively embedded commutative mapping diagrams. Drawing from these considerations, we suggest that the mechanism the brain uses to achieve recursive composition may consist of the cooperative binding of complex spatio-temporal patterns. Particularly interesting in this context is the *synfire-chain* model proposed by Abeles [4] to explain the accurately timed events in cortex reported by him and his group. This model offers a simple and plausible example of a system of recursively bound spatio-temporal neural activity patterns. Moreover, in this model, the interpretation of cooperative binding as the establishment of commutative mapping diagrams is straightforward, and consistent with the generally accepted principle of Hebbian plasticity.

*This research was supported by Office of Naval Research contract N00014-91-J-1021, National Science Foundation contract DMS-9217655, and ARL contract MDA972-93-1-0012. My thanks go to Stuart Geman and to Lee Teverow for their careful critical reading of the manuscript.

1. ROLLERBLADES AND ROSES

This chapter is a discussion of the neurobiological substrate of higher cognitive functions, in particular language. The discussion will be focused on the *microstructure* of the patterns of neural activity that may be involved in such functions, rather than on their gross-anatomical layout in the brain. It is, in essence, a discussion of the neural *representation*—sometimes called *code*—used by the brain. We shall make use of various metaphors during the course of the paper, and propose a formal framework for a theory of neural compositionality.

The discussion will be fed mostly by abstract considerations. Our approach may thus be viewed as an attempt to derive specific suggestions about neural mechanisms from 'first principles.' Sections 1 to 5 offer some reflections on mind and language, and sections 6 to 8 attempt, with the help of the *Lego* metaphor, to organize these reflections in a mathematical framework. The last two sections contain suggestions as to neural representations proper. These suggestions rely on the observations made by Moshe Abeles and his group about well-timed events in cortex, and borrow from the synfire model proposed by Abeles to account for his findings [4–6].

Rather than attempting to define in general terms the notion of a neural representation, we introduce it with a simple example.

Consider the following sample of utterances, which could occur in a conversation about the sport known as *in-line skating,* or, in practitioners' parlance, *rollerblading.*[2]

How long have you had those blades? (1)

With this new braking technology, rollerblading is a piece of cake. (2)

If you stay long enough on each skate, you get the maximum out of your stroke. (3)

With reference to these utterances and to the context—linguistic and other—in which they could have arisen, one may ask the following questions.

Can one characterize the pattern of neural activity that takes place in the brain of the speaker when he or she points at a specific pair of rollerblades while producing utterance 1?

Assuming that such a characterization can be achieved, and referring from now on to this pattern as the neural representation of rollerblades, how does this representation change when another pair of rollerblades is being referred to? Some parts or aspects of the neural representation are likely to change, in connection with changes of the visual attributes of the rollerblades, such as shape, distance, size, color, orientation in space, etc. These changes undoubtedly affect the firing of many neurons in the visual areas of the brain of the speaker currently making reference to the rollerblades.

One may conceive of a neural representation in a narrower sense, excluding everything that varies with the visual presentation of the rollerblades. That type of neural representation would not include the rates of activity of neurons in primary visual cortex. But there might exist in

[2]Note to the uninitiated: whereas the four wheels of an old-style rollerskate (or 'quad') form a rectangle, the wheels of a rollerblade are positioned along a single line. Hence *in-line* skate or roller*blade*.

the brain a pattern of neural activity, located perhaps mostly in infero-temporal cortex (IT), which would be associated with the visual perception of rollerblades in a fully invariant way.

Further, in principle, nothing seems to preclude the existence of an *amodal* neural representation of rollerblades. Such a pattern of neural activity would be elicited, reliably and reproducibly, each time the mental entity 'rollerblades' is evoked in connection with any mental operation, perceptual or not, and only then.

Is such an assumption sound? Is it not the case that no two visual occurrences of a complex object such as a pair of rollerblades are identical? Is it not the case that the number of substantially different occurrences is, for practical purposes, infinite? When carried out thoroughly, would not the elimination of all the variable aspects of the object lead one to discard virtually all that could possibly be part of a visual representation of it? Would not the same consideration hold for other modalities or aspects of the representation? If so, how is the 'rollerbladeness' of rollerblades best captured, and how could it be represented neurally?

Now one might seek an invariant representation in the region of the brain that is responsible for the production or processing of the *word* rollerblade, a symbol used by individuals within a given community to communicate information about the real object. Inasmuch as the acoustic/phonetic signal corresponding to the word rollerblade is uniquely defined, there should exist in the brain of the speaker or listener a pattern of neural activity that corresponds uniquely to this word, hence to the conceptual entity rollerblade. This pattern, which will manifest itself each time the word rollerblade is uttered or heard, will be fully invariant with respect to the visual attributes of a specific object that might be referred to at that moment.

However, it is hardly the case that the acoustic signal corresponding to a given word is unique. In fact, as has been discovered in particular by researchers in automatic speech recognition, the variability of speech signals is considerable. Therefore, there can hardly exist in the brain a single pattern of activity uniquely associated with the *speech* signal 'rollerblade.'

Moreover, as exemplified in utterance 3, the very same rollerblades can, context permitting, be referred to by using a word such as 'skate,' phonetically quite distant from the word rollerblade. Or, as illustrated in 1, the first two syllables of the word—the *roller* part—may be dropped without much harm. When this is done, context becomes all-important: the occurrence of the lexical item 'blade' in a different context might elicit a radically different meaning. Similarly, the highly ambiguous word 'stroke' receives from its context a quite specific meaning in utterance 3, and the metaphorical use of 'cake' in 2 takes this word quite far from the food category, where it originates.

What is then the mechanism whereby the meaning of symbol 'rollerblade' is constructed? Note that the morpheme 'roller' in 'rollerblade' may be viewed either as a *component* in a composite item, or, alternatively, as an element of *context* acting to disambiguate the meaning of item 'blade.' One of the recurrent themes in this paper will be the difficulty in distinguishing the *inside* of an entity from its *outside*.

Contrasting with these remarks, note now that entities as different as 'rollerblade,' 'rose,' 'route,' 'royalty,' 'rudeness' and 'rust' share the following important property (in addition to their common first letter). Each of these entities, when occurring in an appropriate context, is used, i.e., perceived, retrieved from memory, combined, communicated, or otherwise manipulated, *in an all-or-none fashion*, i.e., as a perfectly well-defined entity in its own right. When we say 'rollerblade' (or 'blade' as in 1, or 'skate' as in 3), we do not mean 'rollerskate,' or 'rollercoaster,' or 'knife blade,' or 'ski boot,' or 'mountain shoe.' Nor do we mean a hybrid, say 95% rollerblade and 1% each of the other five items. Rather, we have in mind a well-defined entity, perfectly distinct from fellow mental entities. Further, nearly all the time, we are successful in conveying to our listener that unique *grain* of meaning, using an appropriate symbol, in context.

It is the neural counterparts to such mental entities that we seek to investigate in this paper. We are concerned with the dual nature of these entities, the apparent paradox whereby each of these grains of meaning seems to enjoy an individual life like any physical object—animate or not—yet does so in the absence of a clear borderline that would separate the inside of the entity from its outside.

2. GRAINS AND BONDS

For the moment, we shall proceed with the following assumption: to each well-defined mental entity there corresponds a well-defined neural counterpart, i.e., a pattern of activity which, however simple or complex, localized in a small part of the brain or widely distributed across it, occurs each time this entity manifests itself, in perception, speech, or any other conscious mental process, and only then. Said otherwise, there exists a *mapping* from mental entities to some putative, yet to be defined, space of neural-activity-related quantities, and this mapping is one-to-one. We refer to this mapping as a *neural representation.*

The neural-representation assumption can be found, explicitly or implicitly, in much of modern neurobiological literature. And yet, the few remarks made above point at a serious difficulty in the way to the elucidation of the nature of the map. The difficulty has to do with the observation that at any level and within any modality where one attempts to *isolate* an even mildly complex entity such as a pair of rollerblades, this entity will exhibit an unpleasant tendency to elude us.

Thus, we noted, when examining the *physical* object rollerblade, that the visual presentation of the object, while perfectly unambiguous as a whole, was highly variable, hence ambiguous, in any of its parts. The same goes, in the visual modality, for virtually any physical object presenting itself as part of a natural scene. Almost without exception, natural scenes include many objects and clutter, in overlapping and mutually obscuring configurations. Object boundaries in natural images are, more often than not, difficult if not impossible to trace *locally.* Our ability to segment a scene into meaningful parts is a complex process involving the use of stored knowledge about physical objects, their parts and subparts, the interactions between these and the ways in which we interact with them. This knowledge is of a high level, i.e., *relational* and combinatorial, and we would be hard-pressed to draw a precise boundary around the collection of relationships that exactly make up an entity such as a pair of rollerblades in the visual modality.

When examining the *linguistic* manifestation of a mental entity, we found ourselves confronted with a similar difficulty: a semantically well-defined entity exists only by virtue of a complex set of relationships, which we may at times want to view as *internal* to the entity, and at other times as belonging to the *context* in which the entity is currently experienced.

In this paper, we shall attempt to reconcile these two aparently contradictory aspects of mental entities. We shall use the term *grain-like* to refer to the all-or-nothing quality of mental entities as they manifest themselves in a fully specified context. But we shall also attempt to characterize the structure of mind as *map-like.* By this we mean that mental material itself consists of a collection of maps, or correspondences. This will appear most clearly by examining some of the processes underlying the *construction* of meaning of linguistic material.

On the neurobiological side, we shall follow a suggestion made by von der Malsburg [1,7]—see also [8–10]—namely that dynamical bonds may be represented in the *fine temporal structure* of neural activity. We shall elaborate on this proposal, relying on experimental data on well-timed events in cortex collected in particular by the group of Moshe Abeles, and using the model of *synfire chain* proposed by Abeles to account for his observations.

We shall argue in favor of a neural representation that maps mental material into synfire patterns—or other complex spatio-temporal patterns possibly defined on a lower time resolution—in such a way as to (i) preserve the map-like structure of mind, and (ii) account for the grain-like quality of mental entities by the cooperative binding of a large number of patterns into a globally coherent construct.

3. MONKEYS AND ATTRACTORS

We start with an example of mental entities that exhibit, for all practical purposes, *grain-like* structure, and are of interest in neurophysiological experiments.

Consider the following situation, fairly typical in behaving-monkey experiments. A monkey has been trained to push a lever, or point in a certain direction, upon presentation of a given visual stimulus. The electrophysiologist, who records the activity of neurons in various areas of the brain of the animal while it is performing the task, focuses his/her attention on entities such as: “The second green light to the left has been on for two seconds or so,” or: “The monkey is now pointing its hand to the left,” or: “The image displayed on the monitor in front of the monkey is the 27th (in a fixed series of images used to train the monkey to perform a given task, every day of the last three months)” etc. The neurophysiologist, in an attempt to ‘read the mind’ of the monkey, is looking for a neural correlate, or *predictor*, for one of these mental entities,

In this example, the entities of interest can be construed as having *grain-like* structure. A *grain* of meaning is a simple percept or motor command, such as the perception of a green light, the perception of a given image or shape previously experienced, the pointing in a given direction, etc. The primary operation of the mind on these grains of meaning is *activation,* that is, the evoking or retrieving of an entity, in a clear-cut, all-or-none, fashion, at a given instant of time. This activation obeys some rules, associative and other, subject to various types of fast or slow changes.

Much of the physiological data collected in experiments of this type—e.g. [11]—indicates that for a given mental entity there is in the monkey's brain a well-defined collection of neurons, also called *cell assembly* [12], the activity of which *increases* during the activation of that entity. This increase of activity is a predictor for the entity under study. Depending on the type of mental entity, the neurons which form the assembly may be located in different regions, in particular cortical areas, of the monkey's brain. It is, for instance, observed that the cell assembly corresponding to the perception of a small, simple, visual stimulus displayed at a given location on the retina includes neurons at a given location of the primary visual cortex of the animal. In the case of more complex visual stimuli, the cell assembly, inasmuch as its existence can be revealed, tends to be located in more central regions of the visual system, in particular IT.

The activation of cell assemblies is believed to be an all-or-none process, reflecting the grain-like aspect of mental entities. This aspect, already emphasized by Hebb [12], is rendered most clearly in recent models, which stress properties of *error correction,* or *auto-association,* or *retrieval* of memories from incomplete data [13]. Alternatively, these models are described as relying on the principle of *computing with attractors* [14]. The attractors of the dynamics of a feedback network correspond to *grains* of meaning, through a one-to-one neural representation.

Our contention (§10) will be that this category of neural representations is too limited. Correspondences that assign grains of neural activity to grains of meaning do account for an important aspect of mental entities, namely their elicitation in an all-or-nothing fashion. However, they fail to render the 'dual' aspect of mental material, its map-like structure.

4. BANKS AND RAIN

Leaping now across species and modalities, we turn to linguistic examples. We use these examples to illustrate what we call the 'map-like' structure of mental material.

Consider the lexical item 'bank.' Consider the mechanisms—syntactic, semantic and pragmatic—whereby context specifies or alters the meaning of this lexical item. Suppose we were assigned the task of exhaustively enumerating the collection of mental entities—call this collection M_{bank}—which can *possibly* come into play in these meaning-construction mechanisms.

In attempting to establish such a list, i.e., to enumerate the set M_{bank}, we could get help from Webster's dictionary [15]. The list would likely include a few dozen acceptations of the word—as a noun and as a verb, the latter most of the time intransitive but also, as it turns out, transitive—as well as fragments of speech that illustrate these various uses.

Note that the medium used in the dictionary to convey the different meanings of an item consists itself of strings of lexical items. Since any entity is rendered in this way as a composition of entities of same nature, a dictionary is a *self-referential* affair. Moreover, it is fairly clear that any list that we may propose as an enumeration of M_{bank} will be incomplete and tentative; it will, at best, be a representative sample, since the collection of mental entities M_{bank} is open-ended.

These remarks suggest that the material of language and thought may not lend itself to a decomposition into discrete entities. The meaning of a lexical item is a highly connected object, bound to straddle many boundaries in any picture of a cognitive landscape that one may attempt to draw.

The main difficulty with a grain-like conception of mind is that there is no clear level on which grains of meaning are to be found. As we just saw, lexical items envisaged *before* context comes into play are not appropriate candidates, since their meaning is, more often than not, poorly defined at that stage. Decomposing the meaning of an item into a number of specific alternative explanations, each involving a different context, will bring little salvation, since such a decomposition can be carried out only incompletely and imperfectly, and, what's worse, will rely on composite constructs utilizing other such entities, that is, on entities that are hardly simpler than the one we are attempting to decompose.

Let us then examine in more detail the mechanisms whereby context acts to allow us to *construct* the meaning of a given string, say one that includes the lexical item *bank.*

Two examples of such strings, again from [15], are the following:

The engineers hadn't given the road enough bank. (4)

In the rainy season the clouds would bank up about midday,
and showers fall with true tropical violence. (5)

These examples once more illustrate the crucial role of context in the focusing of semantic content; before context comes into play, the word bank is highly ambiguous.

It appears that what we are doing in our mind, with remarkable speed, reliability, and effortlessness, when putting together the meaning of such strings can be roughly summarized as consisting of the following two processes (this account is partly inspired from [16]): (i) bringing to the fore, in some sort of active workspace, a number of concrete *past experiences,* as well as *domains of knowledge,* of more or less abstract nature; (ii) establishing *maps,* i.e, systems of bonds or relationships (more about 'maps' in sections 5 and 7), between these cognitive spaces.

Specifically, to flesh out the meaning of utterance 4, I may, *for instance*, retrieve the specific recollection of a narrow winding road between two fields of corn on a summer afternoon. Having done that, I may alter the time of day and weather and repaint the scene in a dark rainy night. I may place, in the field of corn on the outer side of the turn, a heavy truck lying on its side, after its driver has lost control of it because of the rain, because he was sleepy, and because *the road didn't have enough bank.* I may also repaint the image a few days later, with, in place of the vehicle which has been removed by now, two fellows with yellow helmets—the engineers who designed the road—discussing the bank or lack thereof. In the process, I may have activated a number of fragments of knowledge, such as: 'Rain makes for slippery roads'; 'A lateral incline will counteract the centrifugal force'; 'Engineers should build safe roads,' etc. The evoking of these memories and fragments of knowledge, among many others of which I am hardly conscious, is part of the process of constructing an interpretation of the sentence.

The second important aspect in the process of meaning construction is the establishment of maps, or correspondences, between all these domains, experiences, or chunks of knowledge.

Specifically, I will *instantiate* the engineers who occur in the sentence in a nondescript form (nothing is specified except their having done a poor job on this road) as two fellows with their yellow helmets; I will *place* the engineers at the edge of the field of corn to the left of the road two days after the accident; I will *chain* the propositions about roads, safety, rain, etc., into a logical sequence, which I will *apply* to that particular road.

A crucial observation here is that these maps between cognitive domains, far from being arbitrary, obey very strict constraints. To appreciate the stringency of these constraints, consider the countless *nonsensical* ways in which one could connect the same entities with each other. Thus, consider on the one hand the engineer-related domain and some of the entities and fragments of knowledge it evokes: 'engineers build roads,' 'engineers are humans and humans have heads,' 'heads of engineers often carry yellow helmets,' etc. Consider, on the other hand, the road-related domain: 'roads are outdoors,' 'I remember a road that crossed a field of corn,' 'outdoors is where precipitation occurs,' etc. Then build a picture that connects these two domains by having yellow helmets, rather than rain, descend on the road. By doing this one would have established a system of bonds between the two domains. Yet this system of bonds is nonsensical, and will not normally occur in the process of meaning construction. Clearly, only a small fraction of all *possible* systems of bonds are legitimate.[3]

5. MAPS BETWEEN MAPS

Let us then summarize, in two propositions, our discussion of the manipulation of linguistic/conceptual material.

Proposition 1: The chief mode of operation of the mind on conceptual material (construction of meaning) is *composition,* i.e., the bringing together of various parts, or regions, or fragments, or aspects, of this material into a structured map, i.e., a globally coherent pattern of bonds or relations. Composition takes place on the time scale characteristic of conscious mental processes, roughly 1 s.

Proposition 2: The process of composition is recursive. It may also, in some cases, be self-referential, that is, involve loops, or circular constructions.

The second part of Proposition 2 roughly states that if we attempt to recursively *decompose* a given entity, call it E, into constituents and these constituents into more elementary constituents and so on, we may at one point be faced with a situation where entity E shows up *again,* as one of the sub-sub...sub-parts of itself. This is not an essential tenet of our proposal. However, it would appear to follow from the above observation that any closed system such as a dictionary, when considered as a self-sufficient body of definitions for all the lexical items in a given language, has to include loops. When constructing the meaning of a sentence, or phrase, proposed in a dictionary as a definition of a given item, the reader carries out in his/her mind compositional

[3]Note, however, that one is generally free to choose from among many alternative compositions to construct a coherent meaning for a linguistic string such as 4. In particular, the past experiences elicited in the process are not only idiosyncratic to the speaker or listener but may also depend on his or her present 'state of mind.' In spite of this variability, the normal use of language is such that the global meaning of a complete string is, almost always, uniquely defined.

processes of the same nature as those outlined above for the road example.

It could be argued that dictionaries also rely at times on non-verbal knowledge such as is conveyed by illustrations, and on implicit knowledge assumed to be shared by all the users of the dictionary. In this fashion, any concept, concrete such as rollerblades or abstract such as rudeness, might, in principle, be construed as a loop-free, if complex, construction involving, at the bottom, objects that might be regarded as grains of meaning, such as green lights and bananas. The latter could conceivably be defined with the help of simple pictures, examples of real-life situations, actions, etc., without further recourse to linguistic/conceptual input.

This debate need not affect much the *practical* conception one has of the structure of mental material and of language, as it has evolved and is in use in a given community. In effect, self reference in language, if present at all, cannot easily be pinned down in a specific entity or a small collection thereof. Rather, it is a *global* property of a conceptual system. The definition of a given concept (whether taken from a dictionary, a textbook, a car owner's manual, a street conversation, etc.) would generally refer to a collection of other concepts, which themselves would refer to yet another collection and so on, until the trace of where the definition process starts and ends is lost.

What constitutes the main tenet of the proposal made here is that composition is recursive. Again, the internal/semantic and/or contextual/pragmatic structure of any entity (inasmuch as one wants to distinguish between these two aspects) appears to be, essentially, a system of maps between maps . . . between maps. Thus:

Proposition 3: The *structure* of mind/language is map-like.

In other words, the fragments of mental material which, according to Proposition 1, the mind composes with each other by the establishment of dynamical patterns of bonds, are, themselves, dynamical patterns of bonds. In the next sections, we shall make Proposition 3 more concrete by discussing a few metaphors for mind/language, all of which involve, in one way or another, the use of *symbols.* We will focus on one particular metaphor, the game of Lego, to illustrate the notion of maps between maps.

6. LANGUAGE AND LEGO

In this section we shall use the term *symbol system* to refer to any system made up of objects that *represent*, i.e., stand for or mimic in some sense, external objects or entities. For instance, the commodities traded in stock markets constitute a symbol system. An interesting phenomenon is that some of the brokers who buy and sell *derivative products,* the trade of which now dominates, by a large factor, the volume of transactions in many stock exchanges, care little, in their daily practice, about how these 'abstract' commodities relate to coffee beans, mineral oil, or semiconductor components. In the hands of these agents, the high-level commodities operate in a quasi-autonomous way, with little *practical* interaction with objects of low level. One may draw a parallel between this observation and our proposal that linguistic/conceptual items are, in practice, systems of bonds between other linguistic entities.

A family of examples of *natural* symbol systems is provided by genetic material, proteins and related biomolecules. A man-made example, which, like stock markets, has evolved millions of times faster than DNA and proteins, is a construction game such as *Lego*. All these systems are compositional: entities within the system are constructed by bringing together other entities, in a recursive/hierarchical manner.[4]

The following are two important properties of each of these compositional symbol systems. First, the system includes an infinite, or potentially infinite, collection of recursively/hierarchically constructed entities. This property is sometimes referred to as (infinite) *productivity.* Second, allowable constructions nevertheless respect specific constraints, whereby overwhelmingly most combinations are illegitimate, or meaningless. Note that in language, the *surface* structure is strictly hierarchical: phonemes bind into syllables; syllables bind into words, etc.

In this paper we focus on one particular symbol system, the game Lego. The simple, yet in some ways remarkable, binding mechanism of Lego makes it a useful metaphor for natural compositional systems such as language/mind. The Lego metaphor for language was introduced briefly in an earlier paper [21].

Our discussion of Lego will extend over several sections. This section and the next one discuss, in some detail, the mechanism of hierarchical composition in Lego. This mechanism consists of a *cooperative* interaction between the many elementary bonds that make up a composite bond between two composite Lego objects. We characterize mathematically this cooperativity principle as being equivalent to the *commutativity* of a mapping diagram. We then propose (sections 8 and 9) to use Lego, viewed as a symbol system, as a metaphor for mind/language. We suggest that there might be, in the mechanisms of neuronal interactions in cortex, an aspect analogous to the property of cooperativity of Lego bonds. In sections 10 and 11 we shall make this proposal more specific, and suggest that cooperativity of bonds in neural activity may also be, formally, a property of commutativity of mapping diagrams.

An *elementary Lego brick*—or simply brick—is a small rectangular parallelepiped of rigid plastic material with two rows of regularly spaced teeth on its upper surface and complementary concavities on its lower surface. This motif, forming a square lattice, allows a limited number of bindings between bricks. Thus, there are only four possible binding angles between two bricks. The number of allowable configurations between composite Lego constructs is also quite limited (see below). In spite of these limitations, the system is infinitely productive; Lego constructs can be made to *represent*, i.e., resemble, an unlimited class of real-life objects.[5] For simplicity, we consider a game consisting of *identical* bricks only, say with 2×4 teeth. We assume we are given an unlimited supply of such identical bricks.

We first examine the mechanism which allows multiple-brick Lego constructions to be as-

[4] The notion of compositionality is sometimes used in a narrow sense, implying that the semantic content or functional properties of a composite symbol are strictly and uniquely determined by the list of its constituents and by the specification of the mutual relationships between them. Some authors therefore hold that language is *not* strictly compositional, for, as we have seen, meaning, more often than not, depends on context. Yet, again, we do not insist on a sharp boundary between the 'intrinsic' content of a lexical item (semantics and syntax) and its context (pragmatics), so that the issue is a moot one.

[5] See a good toy store.

sembled dynamically, i.e., to be done, undone, and recomposed with each other in very many different *stable* configurations.

Consider two Lego constructs A_1 and A_2, each in the shape of a brick, i.e., a rectangular parallelepiped. Say A_1 and A_2 are both of size $20 \times 20 \times 4$: their upper and lower surface (length $\times$ width) is 20×20, while their height[6] is 4.

Consider the following two alternative situations: (i) A_1 and A_2 are bound with each other along a whole horizontal face. For instance, the bottom face of A_1 is bound to the top face of A_2. The bond between A_1 and A_2 is then of *size* 20×20, and the composite B obtained in this way is itself a parallelepiped, of size $20 \times 20 \times 8$. (ii) A_1 and A_2 are bound with each other, into composite C, through a bond of size 1, by bringing a corner brick of A_1 in contact with a corner brick of A_2.[7]

Evidently C is *unstable*: the bond between A_1 and A_2 is too weak. On the other hand, B is just as stable as each of its components A_1 and A_2. What makes B a stable construct is that it brings A_1 and A_2 together through a *composite bond,* i.e., a bond made of many bonds.

We shall say that the 20×20 elementary bonds[8] that ensure the stability of the composition of A_1 and A_2 into B *cooperate* with each other, in the following sense. If one of the bricks of A_2—call this brick β_2—is *correctly placed* vis-a-vis one of the bricks of A_1—call that brick β_1—the same holds of all pairs of bricks through which A_1 and A_2 come into contact. By β_2 being 'correctly placed' with respect to β_1 we mean that β_1 and β_2 stand precisely in one of the small number of possible spatial relationships that allow a binding between these bricks. Cooperative binding in Lego results from the symmetry and exact reproducibility of elementary bricks.

In short, all of the 20×20 elementary bonds that hold A_1 and A_2 together *agree* with each other as to the relative position of the composite constructs A_1 and A_2 in 3-D space. It is this cooperation of a large number of elementary bonds which ensures the stability of the composition of objects A_1 and A_2 into object B.

The notion of cooperation will play a crucial role in the rest of the paper, and we discuss it now in detail. Note first that there are many different ways in which an agreement between a large number of elementary bonds between A_1 and A_2 can be reached. These different bindings stand in a one-to-one correspondence with a subset $\mathcal{T}$ of the discrete symmetry group of the infinite 2-D lattice of Lego-brick teeth. For each rigid transformation of the 2-D plane $\tau \in \mathcal{T}$, there is a *stable* composite object B_τ deduced from B by applying the transformation τ to A_2 while keeping A_1 in the same position.

[6]Length, width and height are measured against the corresponding dimensions of an elementary brick. Note that Lego bricks are assembled in such a way that each brick *straddles* two or more bricks underneath it. This may result in ragged vertical faces. Hence A_1 and A_2 are not *perfect* parallelepipeds. We ignore this in the following discussion.

[7]For instance, the brick in the southwest corner of the bottom face of A_1 is brought on top of the brick in the northeast corner of the top surface of A_2. The composite C obtained in this way—not a parallelepiped—has overall size $39 \times 39 \times 8$.

[8]Because of the straddling of bricks, there are actually more than 20×20 elementary bonds, and each involves less than 8 teeth.

Next, we wish to stress, with the help of a thought experiment, that the mechanism of cooperativity in Lego, while very apparent, should not be taken for granted.

Consider a noise-corrupted Lego world, which we call *Mego*, where elementary bricks are *not* identical. Like Lego bricks, Mego bricks are exactly of same size, yet the lattice of teeth on the top face of Mego bricks is shifted by a small random amount in a random direction with respect to the rest of the brick, in particular with respect to the complementary lattice on the bottom. This shift is drawn, independently for each brick, from a probability distribution which we need not specify. On the other hand, Mego bricks are somewhat flexible, and a bond between two bricks can be established even if the match is not perfect. This is achieved by bending somewhat the bricks to be attached, and/or some other bricks in the construct. A successful Mego bond is, just like a Lego bond, dynamical and stable: it can be done and undone by pressing the two bricks together or pulling them apart. When there is too much discrepancy in the relative positions of the two bricks, the bond fails altogether.

To a certain extent, Mego 'works.' Indeed, suppose we were asked to reproduce in Mego a given N-brick Lego object A. By carefully choosing the N bricks one by one, one may succeed, for small enough N, to create a Mego version A' of A. Note that this composite A' is likely to be *warped*, in an unpredictable manner.

Hierarchical composition on the other hand fails: since Mego objects are warped each in a different way, one cannot compose them with each other to produce an unbounded hierarchy of arbitrarily complex objects as is done in Lego. It is impossible, in Mego, to bind with each other two large constructs, say two (almost-)parallelepipeds A'_1 and A'_2, which have been assembled independently of each other. Each one of them suffers from distortions of its own, making it incompatible with the other object.

The incompatibility between A'_1 and A'_2 resides in the lack of agreement between the elementary tentative bonds that would hold the two together. Each one of these bonds has a slightly different idea about what the relative position of A'_1 and A'_2 should be, making cooperative binding impossible. Of course one could assemble A'_1 and A'_2 into a construct of the C type, i.e., with only one elementary bond between A'_1 and A'_2. Yet the resulting construct would be unstable, again for lack of cooperative binding.

This thought experiment shows that the mechanism of compositionality in Lego does rely on a strong property of composite Lego constructs: they all use the same, universal (within the world of Lego) lattice of teeth and indentations, and, thanks to their symmetry, *extend* the lattice in a consistent way across arbitrarily large composite constructs. As a result, a composite construct A_1 will bind with a composite construct A_2 through a composite bond made up of a large number of *cooperating* elementary bonds.

Another way to look at this is as follows. The binding of two Lego bricks α and β with each other can be viewed as a *seed* for the binding of two larger constructs A and B, where A includes α and B includes β. In effect, due to the rigidity and exact reproducibility of Lego bricks, α is, implicitly, part of an infinite Lego lattice, and so is β. Therefore, registering α with respect to β allows us to predict accurately the registration of any brick in A with respect to any brick in B.

In the next section we shall formalize this idea, and propose to view the binding of brick α to brick β as equivalent to a rigid transformation from the entire Euclidean space R^3 onto itself. The formalism that will result is actually not necessary to study the composition mechanics of Lego, yet we shall make use of it later when discussing *neural* mechanisms of composition.

7. MAPS BETWEEN MAPS AGAIN

One of the aims of this paper is to suggest that a mechanism analogous to the cooperative binding mechanism of Lego may underlie the recursive dynamical binding of neural representations in cortex. In order to help us articulate this proposal, we shall now introduce some mathematics to the effect of characterizing the composition of composite Lego objects as a *map between maps.* The cooperativity, or mutual agreement, between the elementary bonds will be formalized as a property of *commutativity* of a mapping diagram. (Cooperativity is also an essential aspect of the binding mechanism of proteins, or nucleic acids, yet the mathematics there is likely to be less straightforward than in Lego.) We shall then conclude that, from a *functional* perspective, a composite Lego construct is, essentially, a system of *maps acting on maps.*

We start with the simplest of composites: a stable construct A made of two elementary bricks α and β. As noted in the previous section, the composite A is totally specified by a map ϕ, the rigid transformation of R^3 that takes α into β, and, in particular, the lattice of teeth of α into the lattice of teeth of β. The *composition rules* of Lego, dictated by the structure of Lego bricks, constrain this map to belong to a small discrete set of transformations of R^3. We call ϕ the *binding map* of A, and denote $A = \langle\alpha|\phi|\beta\rangle$. We also say that ϕ is an *elementary* binding map, since it binds two elementary bricks.[9]

Consider now a stable N-brick Lego construct A, with $N > 2$. Stability, a physical property of Lego constructs, may be given the following coarse definition, using the notion of a *chain.* A chain is a sequence of elementary bricks $\alpha_1, \alpha_2, \ldots, \alpha_n$, such that α_1 is directly bound to α_2, α_2 is directly bound to α_3, ..., α_{n-1} is directly bound to α_n. A Lego construct A will be deemed stable if any two disjoint large subsets of A are connected with each other, in A, through many non-overlapping chains. According to this definition—which we shall not make more precise—a $20 \times 20 \times 4$ parallelepiped as described above is stable. On the other hand, a composite of type C, consisting of two such parallelepipeds assembled through a bond of size 1, is unstable.

We now define, for any stable multi-brick construct A, the *binding diagram* $\Psi(A)$. First, this diagram is a mapping diagram, i.e., a collection of maps between *spaces*, called the *nodes* of the diagram. The binding diagram $\Psi(A)$ is defined as follows:

1. $\Psi(A)$ contains N nodes, i.e., N spaces, one for each brick in A.

2. Each node of $\Psi(A)$ is identical[10] to the Euclidean space R^3.

[9] Throughout, we assume that Lego bricks are given a fixed orientation in space. The transformation of R^3, ϕ, is then well-defined.

[10] Formally $\Psi(A)$ includes N *distinct* spaces $S_\alpha, \alpha \in A$. Each S_α is a replica of R^3.

3. For each pair of bricks that are *directly* bound in A, i.e., for each two-brick construct $\langle\alpha|\phi|\beta\rangle$ that is a *subobject* of A, there is a map in $\Psi(A)$ between the nodes corresponding respectively to α and to β. This map is the elementary binding map ϕ.

Now a mapping diagram is *commutative* if the following is true: for any two nodes X and Y—remember these nodes are spaces—such that there are several *paths* from X to Y in the diagram, the maps composed along these different paths—each of these chained maps carries X into Y—are all identical. Said otherwise, it does not matter which path one uses to go from node X to node Y.[11]

Note that a path in $\Psi(A)$ corresponds to a chain in A. Clearly, the binding diagram $\Psi(A)$ of any stable Lego construct A is commutative. To spell it out, let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be a chain of bricks within the composite object A. Let $\beta_1, \beta_2, \ldots, \beta_m$ be another chain of bricks within A with same first and last bricks as the α chain, i.e., such that $\alpha_1 = \beta_1$ and $\alpha_n = \beta_m$. Then the composition of binding maps along the α chain is identical to the composition of binding maps along the β chain.

Of course with N large, there generally are, for any two given bricks α and β in A, *many* different chains $\alpha_1, \alpha_2, \ldots, \alpha_n$ leading from α to β, i.e., such that $\alpha = \alpha_1$ and $\beta = \alpha_n$. The commutative diagram $\Psi(A)$ is thus *redundant*: the composite A is actually totally specified by a subset[12] of the binding maps in $\Psi(A)$.

The commutativity of binding diagrams of stable Lego objects is trivial: by definition, the composition of binding maps along *any* chain from α to β in a stable construct A is the rigid R^3 transformation that takes α into β.[13] The binding diagram of an *unstable* construct is also commutative, yet it does not contain 'enough' commutative loops, and this is what makes the object unstable.

While trivial, the property of commutativity should not be taken for granted. In Mego, compositionality fails precisely when the *tentative* binding diagram of a *tentative* composition fails to be commutative. On the other hand, the flexibility of Mego bricks allows us to *repair* the distortions in a binding diagram that is *almost* commutative, and render this diagram *perfectly* commutative.

To summarize, the stability of Lego construction is achieved by the cooperation of many elementary bonds within a composite bond. Mathematically, this cooperativity is a property of commutativity of binding diagrams.

We now extend the definition of a binding map, in the obvious way, to *any* pair of bricks α and β within a given stable Lego construct A: the binding map from α to β is the composite map along a chain from α to β. This definition is legitimate because $\Psi(A)$ is commutative: it does not matter which chain one uses to compute the binding map between distant bricks. We also revise our definition of binding diagrams: the binding diagram $\Psi(A)$ will from now on include the binding map between α and β for every pair of elementary bricks α and β in A. Of course $\Psi(A)$ is still a commutative mapping diagram.

[11] Commutative diagrams play an important role in many branches of mathematics.

[12] Any spanning tree.

[13] Formally, this map transforms S_α into S_β.

Up to now, binding maps were defined only for pairs of *elementary* bricks α and β, whether directly bound with each other or indirectly bound within a given stable composite A. We now extend the notion of binding map so that it applies to the case where a composite construct, say A_1, is bound with another composite construct, say A_2, within a larger object, say B.

Specifically, the *composite binding map*, or simply *binding map*, Φ, from A_1 to A_2, within a given stable composite B which includes A_1 and A_2, is defined as the *collection* of binding maps ϕ that take *any* brick α_1 within A_1 to *any* brick α_2 within A_2.[14] This is a legitimate definition because the binding diagram $\Psi(B)$ is commutative; the binding map Φ is a subset of the collection of binding maps that make up $\Psi(B)$.

If A_1 and A_2 make up *all* of B, the composite construct B is totally (in fact, redundantly) specified by the binding map Φ. We can then use the notation $B = \langle A_1|\Phi|A_2\rangle$, a straightforward extension of the notation $B = \langle\alpha|\phi|\beta\rangle$ used above for the binding of elementary bricks. Also, the binding diagram $\Psi(B)$ is then the *disjoint union* of the composite binding map Φ with the binding diagrams $\Psi(A_1)$ and $\Psi(A_2)$: $\Psi(B) = \Psi(\langle A_1|\Phi|A_2\rangle) = \Phi \bigcup \Psi(A_1) \bigcup \Psi(A_2)$. In words, the binding map from A_1 to A_2 is exactly what it takes to complete the separate binding diagrams of these two objects to a bona fide Lego binding diagram. Of particular interest is the case where, given A_1 and A_2, there exists only *one* binding map Φ that makes $\langle A_1|\Phi|A_2\rangle$ a stable construct. We then say that the bond Φ between A_1 and A_2 is of the 'lock-and-key' type,[15] and we use the shorter notation $B = \langle A_1|A_2\rangle$.

The commutativity of $\Psi(B)$ implies in particular that if ϕ_1 is a binding map between two bricks within A_1 and ϕ_2 is a binding map between two bricks within A_2, then the composition of ϕ_1 by Φ (i.e., ϕ_1 followed by Φ) is identical to the composition of Φ by ϕ_2.[16] This commutativity allows us to view Φ as *transforming* the map ϕ_1 *into* the map ϕ_2. Since this applies to any $\phi_1 \in \Psi(A_1)$ and any $\phi_2 \in \Psi(A_2)$, the binding map Φ is seen to transform the binding diagram of object A_1 into the binding diagram of object A_2. Thus, the composition of A_1 with A_2 within B is precisely a *map between maps*, i.e., a map acting on maps. Mathematically, the operation of *composing* A_1 with A_2 is equivalent to *transforming* the binding diagram $\Psi(A_1)$ into the binding diagram $\Psi(A_2)$.

Consider, finally, a stable composite $B_1 = \langle A_1|\Phi_1|A_2\rangle$ and another stable composite $B_2 = \langle A_2|\Phi_2|A_3\rangle$, such that the two bonds Φ_1 and Φ_2 can be established *simultaneously*, i.e., such that

[14] Formally, this elemantary binding map transforms S_{α_1} into S_{α_2}.

[15] A simple example of a lock-and-key pair A'_1 and A'_2 is as follows. Let A_1 and A_2 be two identical rectangular parallelepipeds, of size $20 \times 20 \times 4$ each, as in §6. Create A'_2 by binding to A_2 an additional 10 elementary bricks, at 10 different positions randomly drawn on the top face of A_2; add the same random pattern of 10 bricks on the bottom face of A_2; finally, create A'_1 by *removing* from the bottom face of A_1 a 'complementary' pattern of bricks. Clearly, A'_1 and A'_2 will bind, with A'_1 on top of A'_2, into a $20 \times 20 \times 8$ regular rectangular parallelepiped, the construct we called B in §6. This is the only *stable* construct that can be created by binding A'_1 to A'_2, since, with very high probability, any *other* composite bond between A'_1 and A'_2 will include at most 10 elementary bonds, hence be unstable, according to our informal definition of stability. (The 10 bricks on the bottom face of A'_2 prevent a stable binding with A'_1, with A'_2 on top.)

[16] As collections of maps, these are *not* identical, since their respective source and target spaces are not the same. Consider however, in each of these two collections of maps, the map that takes the source space of ϕ_1 into the target space of ϕ_2. The statement made here is that these two maps are identical.

A_1 and A_3 do not overlap in space when composed with A_2 through Φ_1 and Φ_2. Assuming that the resulting tripartite composite, say T, is stable too, we can denote it $T = \langle A_1|\Phi_1|A_2|\Phi_2|A_3\rangle$. In the case where both Φ_1 and Φ_2 are lock-and-key bonds, this simplifies to $T = \langle A_1|A_2|A_3\rangle$. The intermediate construct A_2 then acts like a binding map Φ between A_1 and A_3: $\langle A_1|\Phi|A_2\rangle = \langle A_1|A_2|A_3\rangle$. Said otherwise, A_2, for the purpose of composition with A_1 and A_3, is, formally, a binding map Φ.

We now summarize, in words, the main conclusions of our study of the 'compositional mechanics' of Lego. (Following each proposition in English we give, in parentheses, its mathematical equivalent.)

First, since all Lego bricks are identical, the *identity* of an object A_1 resides in the collection of internal spatial relationships between the constituents of this object. Simply put, a Lego object is *made up of relationships* (if $\Psi(A_1) = \Psi(A_1')$, then A_1 and A_1' are 'the same'). These relationships, due to the structure of elementary bricks, are highly *constrained*; in particular, they all agree with each other, i.e., form a coherent whole (the binding diagram $\Psi(A_1)$ is commutative).

Further, the collection of *internal* relationships that make up the identity of A_1 also determines the *function* of A_1, i.e., its interactions with *external* objects ($\Psi(A_1)$ is a diagram of transformations of the entire space R^3 into itself). Thus, the composition of A_1 with another object A_2 is an operation that acts on the internal structure of A_1 and maps it, in complete detail, into the internal structure of A_2 (in object $\langle A_1|\Phi|A_2\rangle$, the map Φ transforms $\Psi(A_1)$ into $\Psi(A_2)$).

Finally, if A_1 and A_2 are indeed compatible, i.e., if there exists a composition operation (a binding map Φ) that binds them into a stable Lego construct B, then B, as a whole, is coherent: all the relationships within B, even those which 'cross over' from A_1 to A_2, agree with each other (the diagram $\Psi(A_1) \bigcup \Psi(A_2) \bigcup \Phi = \Psi(\langle A_1|\Phi|A_2\rangle)$ is commutative). In particular, B can be further operated upon, i.e., composed with another Lego object C. That is, the *composite* structure of B can be further mapped into the inner structure of C. Thus, composition in Lego is recursive.

8. LANGUAGE AND LEGO AGAIN

In Sections 1 to 5, we argued that the mechanism of construction of meaning in language could be viewed as the recursive building of a globally coherent pattern of bonds, and we summarized our discussion with Proposition 3, stating that the structure of language is map-like. Our study of Lego, a simple symbol system with a transparent map-like structure, suggests that composition might be analyzed quite generally in terms of commutative binding diagrams.[17] What makes up the coherence of a linguistic construct is, no doubt, more subtle and complex

[17] One of the acceptations of the word 'bond' is the following [15]: 'A connection or system of connections in which adjacent parts of a structure are made to overlap so as to be tied or bound together.' Webster's lists eighteen different *types* of bond used in masonry: 'American bond,' 'blind bond,' etc. All these *systems of connections* are commutative patterns of binding maps, just like the Lego bond.

than the commutativity of the Lego bond. Yet we suggest that the notion of global coherence in language parallels that of global coherence in Lego in the sense that it is this coherence or consistency that lends its identity to a composite entity, i.e., allows it to function as a well-defined unit. In the earlier sections of this paper, we referred to this aspect of linguistic/mental constructs as the grain-like structure of mind. We shall now interpret more specifically the global coherence of a linguistic construct in terms of the commutativity of mapping diagrams.

Consider again utterance 1 (Section 1). Constructing the meaning of this sentence is a complex process, which involves, in particular, the instantiation of predicate 'have,' i.e., its binding with subject 'you' and object 'those blades.' In a first analysis, one may say that the predicate has *slots*, or *roles,* and that instantiation is a process of role filling. Thus, in 1, the subject slot is filled by the pronoun 'you,' and the object slot is filled by the noun phrase 'those blades.'

Incorporating some of the semantics of predicate 'have' will allow a more detailed analysis of the process of instantiation. We outline such an analysis, using the composition formalism developed for Lego.

The composition of the linguistic symbol 'have' with the symbol 'you' and the symbol 'those rollerblades' into string 1 will be viewed as the establishment of a *binding map* between two mental constructs, one describing the semantic content of predicate 'have,' while the other describes the concrete situation at hand, i.e., the *context* in which this symbol occurs as string 1 is being used.

We posit that the semantic content of predicate 'have,' a construct which we denote HAVE, subsumes a *collection of past experiences of situations* where this symbol was used. In each of these instances, symbol 'have' was used to express a relation of ownership between a person and an object. Thus for instance, the speaker herself, say Adina, may remember that when she was four she *had* a red bike. In Adina's mind, Adina is a construct which we denote ME, and the red bike in question is a construct which we denote RED-BIKE. Adina may also recollect that a friend of hers, say Pam, *has* a car. We posit that construct HAVE, consisting of memories or experiences of this kind, is elicited by each use of predicate 'have.'

Like a Lego construct, HAVE can be 'decomposed' in this way into a collection of mutually agreeing (see below) binding maps. However, while the recursive decomposition of a Lego object all the way to elementary bricks is straightforward—and yields a unique collection of elementary bricks regardless of the order in which the decomposition was performed—we stressed in previous sections that such a thorough decomposition is hardly possible for linguistic/conceptual entities.

Linguistic entities remain moving targets *until* they are composed—in a coherent, i.e., meaningful, way—within a specific context. Thus, in a specific enough context, the map-like structure of HAVE emerges. For instance, in the context of MOVING-THINGS-WITH-WHEELS, Adina may construe HAVE as a *binding diagram* Ψ(HAVE) containing four constructs, ME, PAM, RED-BIKE, and CAR, and four maps, Φ_1 mapping ME (i.e., Adina) to RED-BIKE, Φ_2 mapping PAM to CAR, Φ_3 mapping ME into PAM, and Φ_4 mapping RED-BIKE into CAR. In this first analysis, which, we now see, is actually more a composition than a decomposition, the diagram Ψ(HAVE) is already seen to be commutative: Φ_1 composed with Φ_4 is the same as Φ_3 composed with Φ_2. Said otherwise, Ψ(HAVE) places ME in the same *relative position* vis a vis RED-BIKE as it places PAM vis a vis CAR.[18]

Now, in Lego, relative positions between *elementary* bricks are measured by applying rigid maps in R^3, and relative positions between *composite* bricks are measured through maps between maps between … maps of R^3. When dealing with linguistic constructs we cannot, *practically*, decompose a construct all the way to the end. We do, however, decompose—actually compose—to a certain extent, i.e., specify contextual maps through which relative positions are expressed. For instance, two contextual maps that can operate on MOVING-THINGS-WITH-WHEELS are USE-FOR-TRANSPORTATION and GET-HIT-BY. With respect to these two maps, the relative position of ME and RED-BIKE is the same as that of PAM and CAR.

Other maps could be used to compose the construct HAVE, even in a specific context such as provided by string 1. Thus, for instance, one may use the map WAS-RECEIVED-AS-BIRTHDAY-PRESENT: since YOU is a PERSON, and PERSONS can HAVE objects that are BIRTHDAY-PRESENTs, it may be the case that RED-BIKE and CAR stand vis a vis ME and PAM, respectively, in the relative position obtained through the map WAS-RECEIVED-AS-BIRTHDAY-PRESENT. Another commutative mapping diagram is thereby established, providing an alternative partial construction for HAVE. The two constructions are in mutual agreement, since MOVING-THINGS-WITH-WHEELS are popular BIRTHDAY-PRESENTs. This makes the *union* of the two diagrams commutative. In this larger context, both Φ_1 and Φ_2 emerge as consisting of the union of the map USE-FOR-TRANSPORTATION with the map WAS-RECEIVED-AS-BIRTHDAY-PRESENT.

Note that we did not specify *how* constructs (i.e., maps) such as GET-HIT-BY or WAS-RECEIVED-AS-BIRTHDAY-PRESENT bind with (i.e., operate on) constructs such as ME or PAM. There is no need to do so because all the bonds involved are, with a few exceptions, of 'lock-and-key' type (see §7).

The instantiation of predicate 'have' in string 1 can now be analyzed as follows. Suppose the name of the person being asked the question 'How long have you had those rollerblades' (string 1) is Deborah. The context of this utterance has elicited, in the speaker's (i.e., Adina's) mind, a collection of partially bound constructs, including DEBORAH and ROLLERBLADES. The instantiation of 'have' in utterance 1 then consists of the dynamical binding of HAVE with DEBORAH and ROLLERBLADES, in such a way that the global binding diagram, including the maps mentioned above *as well as* USE-FOR-TRANSPORTATION, WAS-RECEIVED-AS-BIRTHDAY-PRESENT, etc. between DEBORAH and ROLLERBLADES, is commutative.

This globally coherent construct, i.e., commutative binding diagram, makes *explicit* the facts that Deborah stands vis-a-vis her rollerblades in the same position as Pam vis-a-vis her car, in a number of distinct but partly overlapping contexts. Note that one context of interest is the physical proximity between the owner and the object owned: Deborah is *on* her rollerblades and Pam was *in* her car. This incorporation of low-level perceptual—in this case visual—binding

[18]Ignoring tenses and other details.

maps into the semantics of HAVE may be seen as a step in the direction of a *de*-composition of that construct into 'primitives.'

9. LIGHT-YEARS AND HORSES

In this section we examine briefly some limitations of the Lego metaphor for language.

In §4, we attempted to define a set M_{bank} as the collection of all mental entities that make up the entity BANK. We immediately noted the futility of this attempt, in particular because M_{bank} would be open-ended. Using the notation of the previous section, we can now construe the mental construct BANK as a recursively embedded mapping diagram, which we may denote Ψ(BANK). The diagram Ψ(BANK) is indeed open-ended: depending on the context in which the lexical item 'bank' is used, BANK will assume a different form.

Thus, the structure of a mental entity is 'external' rather than 'internal.' While a Lego construct *decomposes* into a finite collection of elementary bricks, the meaning of a linguistic construct is obtained by a process of *composition*. In this sense, the structure of mind appears to be fundamentally different from that of any strictly hierarchical symbol system assembled from objects that live in the physical space R^3, such as Lego or DNA.[19]

Said otherwise, the Lego metaphor is valid only *locally*, i.e., when composition is considered a few recursive steps at a time. The composition mechanics of composite Lego objects can be studied *independently* from the mechanics of elementary Lego bricks. As noted in §6, stock exchanges provide another such example.

Consider now, at a given time t in the life of a given individual, the collection of all binding diagrams elicited before t in the mind of this person. Let us agree to term the *union* of these binding diagrams the 'mind,' $\mathcal{M}$, of that person at time t. This union is not *disjoint*: there is a considerable overlap between the individual binding diagrams that make up $\mathcal{M}$, and $\mathcal{M}$ is itself an immense diagram of maps. Mentation after time t may be viewed as the activation, on the 1-second time scale, of subdiagrams of $\mathcal{M}$.[20] One may view the emergence of 'grains' of meaning as a mere consequence of the fact that some large enough subcollections of individual binding diagrams in $\mathcal{M}$ happen to bind with each other into globally coherent, i.e., commutative, diagrams.

Again, these grains elude a rigorous definition. However, since any composite Lego construct A can be viewed, alternatively, as a binding *diagram* $\Psi(A)$ (which makes the metaphor possible) or as a *node* in the composition diagram of A with other Lego constructs, let us attempt to reconcile our approach with the notion of a 'semantic node': the semantic node BANK would be the counterpart of a very large multibrick subpart of an immense Lego construct M, the latter being the counterpart of $\mathcal{M}$.

In other words we assume, for a moment, that the Lego metaphor for language really 'works' all the way to the end: there exists, *in principle*, a mapping μ which takes the mind $\mathcal{M}$

[19] On the surface however, language is strictly hierarchical: phrases decompose uniquely into words, which decompose uniquely into syllables, etc.

[20] The process of memory then alters $\mathcal{M}$.

into a formidable Lego world M in a composition-preserving way. Thus, in M, the image μ(DEBORAH HAS ROLLERBLADES) of the $\mathcal{M}$ construct DEBORAH HAS ROLLERBLADES is a composite of the Lego constructs μ(DEBORAH), μ(HAVE), and μ(ROLLERBLADES).

If we assume that such a Lego representation (M, μ) of $\mathcal{M}$ exists, we obtain a Lego world M with rather unusual features. First, whereas the binding diagram of any familiar Lego object B is *completely connected*, this is not the case of the binding diagram of M. Indeed, for any two subparts—elementary or not—of a *real* Lego object B, the binding diagram $\Psi(B)$ includes the binding map between these two parts. In M however, consider for instance the nodes μ(HORSE) and μ(LIGHT-YEAR). It would seem quite unlikely that $\mathcal{M}$ includes a map between HORSE and LIGHT-YEAR, i.e., that these two nodes would ever have been activated together in a single experience, giving rise to a memory that assembles them in one and the same context. In effect, these two semantic nodes seem as *distant* from each other as two semantic nodes can possibly be.

This problem is perhaps not insurmountable. One may attempt to picture M as a noisy, *Mego*-style, Lego world (see §6). In such a model M, distortions, though small, would accumulate over distance, so that commutativity would eventually break down. Said otherwise, binding maps would reach only up to a certain *distance* across R^3, and—so we hope—the graph-theoretic distance induced by the binding maps on M would reflect a 'semantic distance' between nodes in $\mathcal{M}$. Thus, μ(HORSE) and μ(LIGHT-YEAR) would be very distant from each other in M.

With this picture in mind, consider the string

This is putting the cart one light-year ahead of the horse. (6)

String 6 creates a globally coherent $\mathcal{M}$ construct in which the 'semantic nodes' HORSE and LIGHT-YEAR, which were thought to be, until time t, very distant from each other, stand all of a sudden, at time t, in close contact.

This breakdown of the Lego metaphor is quite serious. Examples such as string 6, which abound in all natural languages—particularly in English—demonstrate the futility of any attempt to map $\mathcal{M}$ into the Euclidean space R^3: it seems that composition as our minds perform it makes little case of any *fixed* distance that one may want to impose on 'mental space.'

One way to pursue the Lego metaphor in the face of this *mobility* of mental material may be to give up altogether on low-dimensional Euclidean spaces and to view M as a manifold embedded in a high-dimensional space H. The manifold M would then be able to 'fold' upon itself in multiple ways (think of a 1-D chain of amino-acids folding upon itself in 3-D space), thereby establishing connections, at time t, between regions of M that were very distant from each other, in the topology of M, until time t.

The *neural* representation discussed in the next section is an attempt in this direction.

10. NEURONS AND LEGO

In sections 2 and 3 we discussed the view that mental 'entities' can be mapped, in a one-to-one fashion, into neural activity patterns. We argued in sections 4 and 5 that mind is best decribed with the help of recursively embedded systems of maps, rather than as a collection of grain-like entities. We suggested that what lends to mental material its grain-like aspect is the global consistency of the system of maps that we construct, on the fly, from a given input and context—linguistic and/or visual—and from memory.

With the help of a few examples, we attempted to characterize the notion of global consistency of linguistic constructs. Using the Lego metaphor, we argued in sections 7 to 9 that consistency could formally be viewed as a property of commutativity of a recursively embedded mapping diagram. The composition mechanics of Lego is based, in a transparent way, on this commutativity property.

We now, after this long introduction, approach the goal of this paper, which is to derive, from first principles, a compositional neural representation. To this end we shall again use the Lego metaphor, carrying it this time all the way down to neurons and synapses. We shall offer mainly qualitative arguments. The reader is referred to [10] for a quantitative treatment of some of the aspects of the model.

The basic idea [1,8] is that binding takes place in *time* rather than *space*. Reasoning from first principles this has to be so, since neurons, unlike Lego bricks for instance, do not move in space. We shall therefore posit, very roughly, that two monosynaptically coupled neurons are bound whenever their firing times agree with the *transmission delay* between them.

This shall result in a fairly straightforward correspondence between the composition mechanics of Lego and the composition mechanics of cortical activity patterns. Whereas the former takes place in the Euclidean space R^3, the latter takes place in time, a one-dimensional space. However, rather than using the entire real line R, we shall consider neural events that take place in a brief interval T, say of length 1000 ms. Thus, while Lego binding maps are defined on the entire space R^3 (the spatial coherence of Lego constructs is potentially infinite, thanks to the exact reproducibility of Lego bricks), we shall define neuronal binding maps on a limited interval T rather than on the entire time axis R.

The correspondence we propose to establish between the Lego world and the neuron world is thus *local*. The main reason for this is that commutativity would not hold for *long* neuronal constructs. But note that mental constructs are indeed short-lived. While a record of a given construct—a synaptic trace—may remain in the system for arbitrarily long times, the construct itself, defined as a collection of binding maps, i.e., firing events, dies in a matter of a few seconds.

Another difference between Lego composition and neural composition is that, while elementary Lego bricks are all identical, no two neurons are morphologically the same, and, more importantly, no two neurons are *connected* in the same way. Neural composition uses the connectivity graph of cortex as a highly specific and idiosyncratic substrate, a substrate which contains records of the past function of one particular brain.

Most of the choices to be made in constructing the correspondence between Lego and cortex are virtually forced, from elementary anatomical and physiological considerations; these choices will appear as statements.[21] Some issues however are left open, and these will appear as

[21]We do not claim that these statements are unconditionally true, i.e., that the dynamics of cortex, modulo the correspondence thus established, is identical to the dynamics of Lego. We merely argue that these are

'Questions.'

Elementary bricks. The elementary bricks in the cortical composition system are cortical pyramidal cells. The rationale for this is as follows. We argued in the last section that M, our tentative Lego model of mind, should be viewed as embedded in a high-dimensional space. The axonal and dendritic branches of stellate cells extend only locally. Therefore, their connectivity graph is of dimension[22] three. In contrast, the axons of pyramids always extend to the white matter, and often project to distant cortical areas. As a result, the connectivity graph of pyramidal cells—which is exclusively excitatory and accounts for perhaps as much as 80% of cortical synapses—is of high dimension. For practical purposes, the dimension of this graph may as well be said to be infinite. Almost equivalently, the *average synaptic distance* between two pyramidal cells—call it the 'effective diameter' of the graph—is very small, most likely less than 5, a vanishingly small number compared to the *size* of the graph, 10^{10}. In short, the graph of pyramidal cells can be said to *ignore* the three-dimensional topology of our physical world (which is a fairly unique feature for a physical system).

Binding apparatus. The binding apparatus of a pyramidal cell consists of the synapses it establishes with other pyramids, either as the pre- or as the post-synaptic element. We shall denote the synapse between presynaptic neuron α and postsynaptic neuron β (ignoring the possibility of multiple synapses) $\langle\alpha, \beta\rangle$. Note that, like a Lego brick, which has teeth and complementary concavities, a neuron is *oriented*, i.e., comes equipped with binding devices of two complementary types. Thus, for a given pyramidal cell α, there are from a few thousands to a few tens of thousands of synapses (up to about 40,000 in human cortex) of type $\langle\alpha, \beta\rangle$ and about the same number of type $\langle\beta, \alpha\rangle$.

Question: Which synapses support binding? Does neuronal binding machinery include *all* pyramid-pyramid synapses, or does it consist only of a subclass of these, defined perhaps by anatomical, morphological, or pharmacological criteria?

Elementary bond. If α and β are two cortical pyramids with a synapse $\langle\alpha, \beta\rangle$, the two cells are bound with each other whenever the spikes they emit are *paired* with each other, i.e., the relative timing of these spikes respects the total transmission delay between α and β. The synapse $\langle\alpha, \beta\rangle$ is then said to be in bound state. Note that elementary Lego bricks are all identical, hence any one of them can bind with any other one. In contrast, in our model, a pyramidal cell can bind only with monosynaptically connected cells, a large but finite set.

Question: What is the precise condition for spikes to be paired? The total transmission delay between α and β is the time it takes for a spike emitted by α to trigger a spike in β. Is this delay uniquely defined? While axonal conduction times and synaptic transmission

reasonable choices to make in a Lego model of cortex.

[22] A simple way to define the dimension of a large graph is as follows. For any node α in the graph and for any integer k, let $B(\alpha, k)$ be the ball of center α and radius k, i.e., the set of all nodes β such that $d(\alpha, \beta) \leq k$. The graph-theoretic distance $d(\alpha, \beta)$ is the length of the shortest path between α and β (in cortex, $d(\alpha, \beta)$ is the shortest synaptic distance between neurons α and β). If, when k grows, the number of nodes in $B(\alpha, k)$ grows, on average, like $C \times k^r$ for some positive C and r, then the graph is said to be of dimension r. Any *local* graph in R^3, e.g. the network of smooth stellate cells, is of dimension 3. If the number of nodes in $B(\alpha, k)$ grows faster than k^r for all r, the dimension is infinite.

delays appear to be constant and reproducible (from a fraction of a millisecond to a few milliseconds for the former and about 1 millisecond for the latter), there is reason to doubt that this is the case of somato-dendritic integration times. From what is currently known about cortical pyramidal cells, it appears that the time it takes for the membrane potential at the axon hillock of β to reach the firing threshold, measured from the time of arrival of an EPSP (Excitatory PostSynaptic Potential) delivered to β by α, depends, among other factors, on the number of other EPSPs that impinged on β close to that time. Matters may be further complicated by the relative positioning of synapses on the membrane of β. In the case of interest here, where a large volley of EPSPs impinges almost synchronously on cell β (see below), the integration time may still depend on the *size* of that packet of EPSPs [17].

Question: How stringent is the condition for pairing of spikes? Assuming that the transmission delay between α and β has been defined, how *strictly* should the temporal relationship be satisfied for two spikes to form a pair, i.e., for $\langle \alpha, \beta \rangle$ to be in bound state? It is not unreasonable to require millisecond accuracy; this, as we shall see, leads to the synfire model.

Binding is statistical pairing. Most cross-correlograms of pairs of cortical cells are flat. In some cases a small narrow peak with an offset of a few milliseconds is observed, indicating the presence of a direct excitatory connection between the two cells. Thus, at best a small fraction of all spikes, in a given presynaptic cell α, are paired with a spike in a given postsynaptic cell β. This observation is consistent with the high fan-in/fan-out intracortical connectivity, and, to many authors, is evidence that there is little accurate time structure in the multi-dimensional point process made up of the firing times of any collection of cortical cells. While we defend the opposite view, the fact remains that, from the point of view of any given synapse $\langle \alpha, \beta \rangle$, the majority of spikes produced by either α or β are unpaired according to our definition. Therefore, in the model proposed, the bond between α and β necessarily resides in a *statistical property* of the two-dimensional point process: we shall say that the synapse is in bound state if a *large enough* fraction of all spikes emitted during T are paired. One may for instance use a criterion based on comparing the frequency of observed pairs with the *expected* frequency of pairs under the hypothesis that the two processes are independent, with marginal distributions as observed. This still leaves a few questions open. In particular, are bonds *binary-valued* as they are in Lego, or is their strength gradual, depending on how many of the spikes emitted during T are paired, and on how accurate the pairing is?

Question: How are bonds stabilized? The Lego binding apparatus uses a friction mechanism, exploiting the elasticity of plastic material, to create bonds that can be undone yet are *stable*. What is the physical mechanism for the stabilization of neuronal bonds? Since a neuronal bond consists of paired spikes, stabilizing the bond means persistently generating enough paired spikes during the entire duration of T; this mechanism should be reversible. These considerations are strongly suggestive of fast, reversible, Hebbian-type synaptic plasticity, as proposed by von der Malsburg [1], and as documented in a number of studies, e.g. [18].

Competition between bonds. While an elementary Lego brick α can bind with any other elementary brick, it cannot bind *simultaneously* with more than a few. This exclusion rule, in

Lego, results from the limited amount of R^3 space available in the immediate neighborhood of any brick α. In cortex, competition, like binding, takes place on R, the time axis. Whereas in Lego the exclusion rule is strict, cortical competition is likely to take the form of a soft constraint relying on *inhibitory* interactions, such as provided, for instance, by chandelier cells, synapsing on the initial segments of pyramidal-cell axons. The precise biological underpinnings of the competition mechanism need not be discussed here. Suffice it to remark, in the spirit of our derivation from first principles, that some form of competition or exclusion is likely to be enforced on the short time scale: there must be a mechanism to prevent a large fraction of the cells that are monosynaptically coupled to a given cell α from firing simultaneously in a small interval of time around the firing of that cell α. In view of the vanishingly small 'effective diameter' of the cortical graph (see above), this is roughly equivalent to excluding epileptic states.

Elementary binding map. If α and β are two cortical pyramids with a synapse $\langle\alpha, \beta\rangle$ between them, the binding map $\phi_{\alpha\beta}$ is the shift, defined on the time interval T, which maps a spike of α onto a paired spike of β. The neuronal construct A that consists of two neurons α and β bound with each other through the binding map $\phi_{\alpha\beta}$ will be denoted $A = \langle\alpha|\phi_{\alpha\beta}|\beta\rangle$. Binding maps are thus shifts, and, for a given pair of neurons, the amount of shift is equal to the transmission delay between the two neurons. The map may be different for different pairs of neurons, and, in particular, longer if the two neurons reside in distant cortical areas. As noted above a neuronal binding map is defined on a limited time interval T rather than on the entire time axis R. This stands in sharp contrast with Lego binding maps, which are defined on the entire space R^3.

Constraints on elementary binding maps. In Lego, the binding map ϕ belongs to a small set of transformations of R^3. In cortex, the binding map $\phi_{\alpha\beta}$ for a given pair of neurons α and β is unique; this map, a simple time shift, can in principle be computed from anatomical/morphological data. We shall therefore use the notation $A = \langle\alpha|\beta\rangle$ rather than $A = \langle\alpha|\phi_{\alpha\beta}|\beta\rangle$. Thus, $\langle\alpha, \beta\rangle$ denotes the *synapse* between α and β, whether in bound or unbound state, while $\langle\alpha|\beta\rangle$ denotes the two-neuron construct consisting of α, β, and the synapse $\langle\alpha, \beta\rangle$ in bound state.

So far, we discussed the mechanism of *elementary* binding in cortex. We now extend this discussion to the recursive binding of composite constructs, following again the Lego example.

N-Neuron constructs. We consider a collection S of N pyramidal cells, with N large.[23] For some pairs of cells $\alpha \in S$, $\beta \in S$, there exists a synapse $\langle\alpha, \beta\rangle$; further, *some* of these synapses are in bound state during T, i.e., are composed into a 2-neuron construct $\langle\alpha|\beta\rangle$. A collection A of 2-neuron constructs is said to be an N-neuron construct with support S if (i) A covers all the N neurons of S, and (ii) it forms a connected graph. These two requirements are self-explanatory, but notice that A need not include *all* active bonds between pairs of cells in S.

In effect, A will generally consist of a *strict subset* of the collection of all 2-neuron constructs

[23] In Lego we did not need a *name* for a collection of bricks: since all elementary Lego bricks are interchangeable, any two N-brick collections are equivalent. Neurons however, due to their connectivity, are not interchangeable.

made from neurons of S. In contrast, an N-brick Lego construct A obviously includes all the two-brick subobjects of A. The reason why we *leave out* some bonds when defining A is that a given neuron may take part simultaneously in several distinct multi-neuron constructs. As we shall see below, the binding diagram of each construct separately will then be commutative, yet the *union* of these binding diagrams will not. This situation, which we shall examine below in more detail in the case of synfire chains, does not arise at all in Lego.

Binding diagram. As for Lego, we define the binding diagram $\Psi(A)$ in two steps. For the moment, we define $\Psi(A)$ to be the collection of all elementary binding maps $\phi_{\alpha\beta}$, for all 2-neuron constructs $\langle\alpha|\beta\rangle$ in A.

Commutativity. The binding diagram $\Psi(A)$ is commutative if the composition of elementary binding maps along different A-chains with same first and last neurons is independent of the chain. Said otherwise, the transmission delay is the same—or very nearly the same—along all A-chains that start with the same neuron and end with the same neuron. For instance, if A includes the synaptic bonds $\langle\alpha|\beta\rangle$, $\langle\alpha|\gamma\rangle$, $\langle\beta|\delta\rangle$, $\langle\gamma|\delta\rangle$, then the total transmission time from α to δ has to be the same whether one uses the β path or the γ path.

Note that commutativity, even when defined, as it is now, on binding diagrams that include only monosynaptic bonds, is not granted. There are essentially three, non-exclusive, ways in which commutativity can be violated: (i) some A-chains start with the same neuron and end with same neuron yet have different synaptic lengths; (ii) with reference to the above four-neuron example, the three neurons α, β and δ sit in the same cortical area whereas γ is located in a distant area; (iii) different synapses have very different locations on the somato-dendritic membranes.

Situation (i) can cause arbitrarily large violations of commutativity. If we pick a large collection of neurons S *at random*, chances are that the construct which includes *all* the synapses in bound state between neurons of S will be highly non-commutative. Certainly by leaving out enough bonds—which we can always do according to our definition of an N-neuron construct—we may obtain a commutative construct A. Commutativity is for instance trivially realized if the graph A is loop-free, i.e., if it is a tree. However, a construct obtained in this way need not be stable (see below). The existence of stable neural constructs is thus non-trivial. In the same way as it takes non-trivial engineering to create a Lego game that satisfies the commutativity axiom (cf. the Mego experiment in §6), it may take a specific ontogenetic development mechanism to equip the cortical network with large numbers of stable commutative constructs (see below).

Considering now a commutative construct A with support S, we extend, as in Lego, the notion of a binding map to *all* pairs of neurons in S. Since A is a connected graph, there exists a chain of monosynaptic bonds in A between any two neurons α and β in S, and we define the binding map $\phi_{\alpha\beta}$ as the composite map along this chain. Thanks to commutativity, the map is well defined. We now modify the definition of the binding diagram: $\Psi(A)$ will include the map $\phi_{\alpha\beta}$ for *any* pair of neurons α and β in S; $\Psi(A)$, which includes monosynaptic bonds as well as *polysynaptic* bonds, is commutative.

Note that a monosynaptic binding map $\phi_{\alpha\beta}$ is a *short* shift, generally of the order of 1 ms, and never much longer than 10 ms. On the other hand, a binding map $\phi_{\alpha\beta}$ between an arbitrary pair of neurons in A can be an arbitrarily long shift, which, in view of the above remark, is undesirable. In principle, we should impose a limit on the *length* of multi-neuron constructs.

Stable constructs. We give a *loose* definition of stability, following the Lego example: a construct is stable provided it contains *enough* elementary bonds that *cooperate* with each other. As in Lego, cooperativity is equivalent to the commutativity of the binding diagram.

The physiological interpretation of cooperativity of neuronal bonds is as follows. Consider again the four-neuron example, where α is connected to δ through two different disynaptic pathways, one through β, the other through γ. As noted above, commutativity of this mapping diagram means that the total transmission time from α to δ is the same whether one uses the β path or the γ path. This condition favors the impinging of two *simultaneous* spikes on δ, since the firing of α tends to cause the simultaneous firing of β and γ. This, in turn, favors the spiking of δ, since cortical neurons are triggered most effectively by simultaneous EPSPs.

However, as in Lego, commutativity *per se* does not grant stability. There must be *enough* cooperativity for the entire construct to hold together. Physiologically, two EPSPs impinging simultaneously on neuron δ will be most of the time insufficient to bring δ to fire. Estimates vary as to how many synchronous EPSPs are required to bring the membrane potential to firing, starting from resting level; a figure of 25 appears reasonable [5]. Thus, while Lego constructs need not be large to be stable, neuronal constructs cannot be stable until they reach a *critical size*; stability of binding is a collective property.

This picture is consistent with the assumption that the stabilization of monosynaptic bonds relies of fast Hebbian-type plasticity (see above). If every neuron receives, from time to time, a large packet of synchronous EPSPs, all the synapses implicated will be reinforced. This in turn will favor the occurrence of paired spikes at these synapses, i.e., stabilize all the bonds in the construct. This principle, consisting of short-term feedback between converging/diverging configurations of synaptic weights and packets of synchronous spikes, was originally proposed by von der Malsburg [1].

We shall now illustrate the mechanism of stable binding with the example of a synfire chain. We assume that the reader is familiar with the definition and properties of synfire chains [4,5,19]; for a summary, see [10]. In a synfire chain as defined in [5], all elementary transmission delays are the same, say 1 ms. If the chain is *strictly feedforward*, i.e., if a given neuron occurs only once in the chain, and if the chain operates in noise-free conditions, all spikes are paired hence all synapses, according to our definition, are in bound state. The mapping diagram $\Psi(A)$ of the chain is the following: if α is in pool i and β is in pool j, then the map $\phi_{\alpha\beta}$ is a shift of $j - i$ ms.

For such a synfire chain A, $\Psi(A)$ is evidently commutative. Stability of A as defined above is moreover equivalent to stability of synfire transmission [5]. The conditions that guarantee stability in a synfire chain have been investigated numerically and analytically by a number of authors [5,19,10,17,20]. As shown in these studies, a key parameter for stability is the *multiplicity* of the chain, i.e., the degree of convergence/divergence between successive pools; multiplicity is the synfire-chain equivalent of our general notion of cooperativity.

The definition of a synfire chain can be generalized to allow non-uniform transmission times.

Thus, in the above four-neuron example, the condition of commutativity may be met while the four individual transmission delays are all different. The resulting network, termed a *synfire braid* in [10], is again a stable construct, provided there is enough cooperativity between elementary bonds.

As documented in a number of numerical studies, synfire transmission is resistant to various types of noise, either in the firing behavior of neurons or in the connectivity. Of particular interest is the case of *chains with feedback*, obtained by allowing any given neuron to recur in several pools [5,10]. In a chain with feedback, commutativity is clearly violated. However, the definition of a construct A on a given support S (see above) allows us to *ignore* some of the bonds between neurons of S.

Thus, if we ignore the 'feedback' connections, resulting from the recurrence of neurons in the chain, we obtain a subgraph A with the two properties required for stability: (i) $\Psi(A)$ is commutative and (ii) there is enough cooperativity to ensure stable synfire transmission. Note that the bonds that have to be ignored to make a given portion of the chain commutative—we may call these bonds 'spurious' from the perspective of that portion of the chain—may be essential to the stability of transmission in other portions of the chain. Our definitions of commutativity and stability nevertheless apply *locally* in the chain. Equivalently, one may consider that different portions of one chain are different chains altogether. In this situation, several strictly feedforward synfire chains are superimposed on one given collection of N neurons; the synfire bonds that make up one of the chains are spurious, i.e., violate commutativity, from the perspective of other chains.

When kept within limits, this violation of commutativity does not interfere with the stability of each of the chains sharing these same N neurons. What allows the chains to function independently of each other while sharing the same neuronal support is that (i) the incoming/outgoing binding apparatus of each neuron can accommodate many simultaneous bonds, and (ii) binding as expressed in terms of pairing of spikes is a statistical property, i.e., does not require that all spikes, presynaptic *or* postynaptic to a given bond, be paired along that bond. Again, there is no Lego equivalent to this behavior, since it requires a *substrate* of high dimensionality (see above).

The limits within which commutativity can be violated in this way without affecting synfire transmission are investigated in [10], under a simple synfire-superposition model. With the numerical values proposed in §8 of that paper, the critical value at which stability of synfire transmission breaks down corresponds to an overlap of about 400 synfire links per neuron. This means that, on average, the proportion of spikes that are paired along a given synapse among all presynaptic (or postsynaptic) spikes 'seen' by that synapse is 1/400. While this figure is a mere order of magnitude, it is consistent with the thesis [4,1,9,5] that cortex may support exquisitely accurate time structure of a type that may hardly show in pairwise correlograms.

In sum, synfire chains, with or without feedback, provide an example of a family of stable neuronal constructs for our Lego-inspired model. This model might perhaps accommodate constructs of some other type, yet synfire chains have, at least in the hands of one group of researchers, received convincing experimental support.

Moreover, synfire chains can be shown to grow spontaneously, in an initially disordered graph, under the effect of slow Hebbian plasticity with a few additional mild conditions [21–23]. This growth mechanism may be seen as the counterpart, in the Lego metaphor, of the *design* of a regular lattice of binding devices.[24] We shall now, in effect, pursue this metaphor in the case of composite constructs, and show that synfire chains—assuming they exist in cortex in large numbers—provide a substrate for recursive construction.

Recursive binding of neuronal constructs. From now on we consider stable constructs of synfire type only. Let A_1 and A_2 be two synfire chains, of supports S_1 and S_2, with binding diagrams $\Psi(A_1)$ and $\Psi(A_2)$. Remember that these diagrams generally do not include *all* the bonds within S_1 and S_2, since the chains generally include feedback connections. A_1 and A_2 are bound into a stable construct B whenever there are enough cooperating monosynaptic bonds of type $\langle\alpha_1|\alpha_2\rangle$, with $\alpha_1 \in A_1$ and $\alpha_2 \in A_2$. As in Lego, these bonds, extended through $\Psi(A_1)$ and $\Psi(A_2)$ to polysynaptic bonds between *any* pair of neurons, $\alpha_1 \in A_1$, $\alpha_2 \in A_2$, create a composite binding map Φ, and we write $B = \langle A_1|\Phi|A_2\rangle$.

Physiologically, the mechanism of binding of the two chains A_1 and A_2 is the *synchronization* of the synfire waves that travel down each of the two chains. This mechanism has been studied in particular by [19,10,17], who showed that, thanks to the intrinsic stability of each of the two chains, a small number of bonds between them is enough to ensure their binding.

Note that, in order for A_1 and A_2 to be able to bind with each other, the collection of synaptic connections between them should include a large enough, spatio-temporally coherent, subset. This constraint, through which the composition mechanics of neuronal constructs is made to depend on development and learning, has no counterpart in Lego world.

Our formal treatment of composite binding maps for Lego (§7) however applies, *mutatis mutandis*, to neuronal constructs. In particular, the binding diagram of the composite construct B is the disjoint union of the binding diagrams $\Psi(A_1)$ and $\Psi(A_2)$ with the binding map Φ.

Further, the *neuronal* interpretation of the binding map Φ is, as in Lego, a map between maps: the commutativity of $\Psi(B)$ allows us to view Φ as transforming the diagram $\Psi(A_1)$ into the diagram $\Psi(A_2)$. Each of $\Psi(A_1)$ and $\Psi(A_2)$ is, essentially, a mapping of a collection of neurons—S_1 or S_2—onto the time axis. Thus, Φ is, in essence, a mapping of the time axis onto itself.[25]

Finally, as in Lego, we may consider bonds of lock-and-key type. The bond between the two constructs A_1 and A_2 is of lock-and-key type if there is only *one* binding map between them; we may then write $B = \langle A_1|A_2\rangle$. We may also view A_2 as equivalent to a binding map in the construct $\langle A_1|A_2|A_3\rangle$, if the two bonds in this construct are of lock-and-key type.

With neuronal constructs, bonds of this lock-and-key will arise naturally from Hebbian plasticity. Consider two synfire chains A_1 and A_2 that grew and functioned independently of each other up to a certain time. At this time, A_1 and A_2 are *not* connected with each other in a spatio-temporally coherent way, as required for binding. Synaptic connections are likely to

[24] This mechanism may also be viewed as a natural selection principle, with cooperativity playing the role of fitness.

[25] As usual, Φ is defined only *locally*, on an interval of say 1000 ms.

exist between individual neurons of A_1 and A_2, since some of these pairs of neurons belong to a common third chain (each neuron, as seen above, is shared by many, perhaps hundreds, of synfire chains); however, these connections between A_1 and A_2 are unlikely to form a coherent spatio-temporal pattern. When, in an appropriate functional context, the two chains are activated together, Hebbian plasticity will result in the development of a coherent system of dynamical bonds between them. Since this composite bond between A_1 and A_2 is the only one at that point, it is of the lock-and-key type.

11. LANGUAGE, LEGO, AND NEURONS

Commutative mapping diagrams are the key mathematical concept which, in our proposal, should help us understand mental composition in terms of neural mechanisms.

We introduced the notion of a binding map with the Lego example. We defined the binding map Φ between two complex Lego constructs A_1 and A_2 as a composite map between two system of maps—the binding diagrams of A_1 and of A_2. Φ was thus defined as a map between maps. When each of A_1 and A_2 is in turn envisaged as a composite object constructed from simpler objects, Φ becomes a map between maps between maps. This recursion, we proposed, is suggestive of the recursive construction of meaning during mentation.

However, the composite Lego construct $\langle A_1|\Phi|A_2\rangle$, is also, quite conspicuously, a map from the mere 3-D space R^3 onto itself. Specifically, after we provide a thorough description of A_1 and A_2 separately, all it takes to further specify the composite $\langle A_1|\Phi|A_2\rangle$ is merely the rigid R^3 transformation that takes any given brick of A_1 into any given brick of A_2. And since all elementary Lego bricks are interchangeable, all it takes to provide a thorough functional description of an N-brick construct A_1 in the first place is a collection of $N-1$ binding maps in R^3. Thus, if it were only for Lego, we would have no use for maps between maps.

In short, Lego provided only a convenient excuse for introducing our mathematical tools. These tools are quite unnecessary to the analysis of Lego itself.

Neuronal composition however is immensely more complex. Each and every neuron has an *identity*, associated with the graph $\mathcal{M}$ of permanent synaptic connections in which the neuron is embedded (§9). *If* we were given a thorough functional description of two complex neuronal constructs, say two synfire chains A_1 and A_2, all it would take to specify a composite $\langle A_1|\Phi|A_2\rangle$ would, indeed, be a mere shift from the time axis R (or from a 1000-ms time interval T) onto itself.

Yet what *is* a thorough functional description of a synfire chain embedded in cortex? We would certainly need to include in it a description of the identity of each neuron α in the chain. The amount of information that this would entail is gigantic: consider that we would have to specify the list of all cortical neurons presynaptic and postsynaptic to α, in a format that would allow each of these neurons to be essentially *any* of the 10^{10} pyramidal cells in cortex.

Shall we conclude that the Lego metaphor, in the end, is useless? Perhaps not. Recall our attempt, in §8, to interpret the mental construct HAVE, used in the construction of the meaning of linguistic string 1, as consisting of an embedded system of commutative mapping diagrams, paralleling our—correct if unnecessary—account of Lego mechanics. In light of our Lego-inspired model of neural composition, this attempt may be viewed as resting on the assumption that a *functional* description of a large neuronal construct A_1 can indeed be achieved much more succinctly than by listing all the external connections of each and every neuron in A_1. Such a description may for instance be one of the many possible descriptions of the mental construct HAVE, or of the mental construct BANK, or the mental construct ROLLERBLADES.

Unlike the simple unique description that can be achieved for Lego constructs, mental constructs are never uniquely nor thoroughly described. A mental construct can *only* be viewed as an embedded system of maps. Formally, however, the composition dynamics of mental constructs is identical to that of Lego constructs. The binding of two mental constructs is generally highly constrained, often of lock-and-key type, thanks to the precision of the wiring of $\mathcal{M}$. These lock-and-key constraints allow us, for instance, to refer to the construct DEBORAH HAS ROLLERBLADES as the only construct that can possibly be assembled from DEBORAH, HAVE, and ROLLERBLADES (ignoring, as above, the tense and other modifiers).

In other words, the connectivity graph $\mathcal{M}$, which, in our proposal, serves as a substrate for the composition mechanics of neuronal constructs, can be analyzed at two very different levels. When scrutinized at the neural level, $\mathcal{M}$ may conveniently be viewed as a manifold embedded in a very high-dimensional—practically infinite-dimensional—space. While this level of analysis may provide some insight into the elementary neural mechanisms of binding, it is of little use when we try to account for the specifics of composition at the cognitive level. At that level, no reference is made to individual neurons any more. Rather, with the help of what may appear to some as a leap of faith, we assume that the Lego analogy, plausible at the microscopic level, carries over to the level of highly embedded systems of maps between maps.

Our proposal leaves us, it seems, with more questions than we set out with. In addition to the numerous far-from-resolved issues, mentioned above, concerning the implementation of the composition dynamics at the neural level proper, we hardly touched upon the issue of how to represent regularities, say of linguistic nature, into composition rules for neuronal constructs.

To take but one example, imagine we view the neural representation of HAVE, used in a given linguistic context such as string 1, as the activation of a regular synfire chain A, with discrete pools numbered 1 to p. One may then, for instance, consider the following neural model for predicate instantiation. The binding of DEBORAH to HAVE with DEBORAH in the subject role will be achieved by the establishment of bonds between the synfire chain corresponding to DEBORAH and neurons in even-numbered pools of A; in contrast, odd-numbered pools of A would provide the sites of binding for a direct object, e.g. ROLLERBLADES.[26] In our current state of knowledge—and lack thereof—about brain function, this is one out of a virtually limitless number of possibilities. As such, it is almost certainly wrong, even in outline. Nevertheless, there is a possibility [25] that some issues pertaining to the mechanisms of acquisition of linguistic regularities might be fruitfully addressed within the framework proposed here, using simple organization principles related to theories of natural selection.

[26] This is a rough synfire-chain equivalent of the neural-oscillator model of predicate instantiation proposed in [24].

Finally, our proposal may be of interest in the context of the recently much debated issue of the neural mechanisms of consciousness. In §9, we proposed to relate the 'mobility' of mental material, a concept which arose from our futile attempt to map mind into a low-dimensional space, to the high *dimensionality* of the graph of cortical pyramids. Our suggestion is that this high dimensionality, a fairly unique feature of evolved brains, might be responsible for our ability to bring together mental domains which do not, *a priori,* belong together.[27]

One of the most striking capacities of the human mind is indeed that of drawing a common thread between disparate experiences. Such experiences may vary enormously in terms of sensory modality, time, context, etc., we are nevertheless remarkably apt at connecting and integrating them with each other, revising them as needed, and synthesizing from them various multi-dimensional pictures, eventually resulting in encompassing, multi-faceted, representations of the world.

It has been proposed [26] that the mechanisms underlying this capacity might rely on the synchronization of large populations of cortical neurons that oscillate more or less regularly at some frequencies near 40 or 50 Hz. While the general principle of binding through the establishment of spatio-temporal relationships [1] appears not only basically sound but almost inevitable, the role of neuronal oscillators in this context might turn out to be quite limited. Simply put, the synchronization of oscillators, while allowing the binding of any collection of entities into a small number of classes, is hardly, however one wants to look at it, a mechanism for recursive composition.

If our 'derivation from first principles' is anything near reality, our brains make full use of the high dimensionality of the connection graph of cortical pyramids. In this view, meaning is not located in 3-D space, neither in the form of cell-assembly activation, nor in the form of cell-assembly oscillation. Rather, meaning is to be sought in cortical events which are both well-timed and accurately defined in *neuronal space*, and which therefore almost completely elude current investigation techniques.

REFERENCES

1. von der Malsburg, C. 1981. *The correlation theory of brain function.* Internal report 81-2, Max-Planck Institute for Biophysical Chemistry, Dept. of Neurobiology, Göttingen, Germany.
2. Gray, C M., König, P., Engel, A. K., and Singer, W. 1989. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties, *Nature*, 338:334–337.
3. Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., and Reitboeck, H.J. 1988. Coherent oscillations: a mechanism of feature linking in the visual cortex? *Biol. Cybernetics*, 60:121–130.
4. Abeles, M. 1982. *Local cortical circuits: An electrophysiological study,* Springer-Verlag, Berlin.

[27] As a further illustration, consider the titles of the various sections of the present chapter.

5. Abeles, M. 1991. *Corticonics: Neuronal circuits of the cerebral cortex.* Cambridge University Press, Cambridge UK.
6. Abeles, M., Bergman, H., Margalit, E., and Vaadia, E. 1993a. Spatiotemporal Firing Patterns in the Frontal Cortex of Behaving Monkeys. *J. Neurophysiol.*, 70(4):1629–1638.
7. von der Malsburg, C. 1987. Synaptic Plasticity as Basis of Brain Organization. In *The Neural and Molecular Bases of Learning,* J.P. Changeux and M. Konishi, Eds., Wiley and Sons, pp. 411–432.
8. Milner, P.M. 1974. A model for visual shape recognition. *Psychol. Review,* 81:521–535.
9. von der Malsburg, C., and Bienenstock, E. 1986. Statistical coding and short-term synaptic plasticity: A scheme for knowledge representation in the brain. In *Disordered Systems and Biological Organization,* E. Bienenstock, F. Fogelman, and G. Weisbuch, Eds., Springer-Verlag, Berlin, pp. 247–272.
10. Bienenstock, E. 1995. A model of neocortex, *Network: Computation in Neural Systems*, 6:179–224.
11. Miyashita, Y. and Chang, H.S., 1988. Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature,* 331:68–70.
12. Hebb, D.O. 1949. *The Organization of Behavior.* Wiley, New York.
13. Hopfield, J.J. 1982. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci. U.S.A.,* 79:2554–2558.
14. Amit, D.J. 1989. *Modeling Brain Function.* Cambridge University Press.
15. Websters Third New International Dictionary of the English Language, 1986. Merriam-Webster, Springfield, MA.
16. Langacker, R.W. 1987. *Foundations of Cognitive Grammar, Vol. 1, Theoretical Prerequisites.* Stanford University Press.
17. Arnoldi, H.M.R., and Brauer, W. 1995. Synchronization without oscillatory neurons. *Biol. Cybernetics,* In press.
18. Zucker, R.S., 1989. Short-term synaptic plasticity. *Annu. Rev. Neurosci.* 12:13–31.
19. Abeles, M., Vaadia, E., Bergman, H., Prut, Y., Haalman, I., and Slovin, H. 1993. Dynamics of Neuronal Interactions in the Frontal Cortex of Behaving Monkeys. *Concepts in Neuroscience,* 4:131–158.
20. Diesmann, M. Gewaltig, M.O., and Aertsen, A. 1995. Characterization of synfire activity by propagating 'pulse packets.' In: Proc. of the Conference *Computation and Neural Systems '95,* in press.
21. Bienenstock, E. 1991. Notes on the growth of a composition machine, in *Proceedings of the Royaumont Interdisciplinary Workshop on Compositionality in Cognition and Neural Networks—I,* D. Andler, E. Bienenstock, and B. Laks, Eds., pp. 25–43.
22. Doursat, R. 1991. *Contribution à l'étude des représentations dans le système nerveux et dans les réseaux de neurones formels.* Ph.D. Thesis, Université Paris VI, Paris, France.
23. Bienenstock, E., and Doursat, R. *The Hebbian Development of Synfire Chains.* In preparation.
24. Shastri, L., and Ajjanagadde, V. 1993. From simple associations to systematic reasoning:

A connectionist representation of rules, variables and dynamic bindings, *Behavioral and Brain Sciences*, 16:417–494.
25. Bienenstock, E. 1992. Suggestions for a Neurobiological Approach to Syntax, in *Proceedings of the Royaumont Interdisciplinary Workshop on Compositionality in Cognition and Neural Networks—II,* D. Andler, E. Bienenstock, and B. Laks, Eds., pp. 13–21.
26. Crick, F., and Koch, C. 1990. Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences,* 2:263–275.

November 15, 1995

To appear in
Brain Theory:
Biological Basis and Computational Theory of Vision
Ad Aertsen and Valentino Braitenberg eds.
Elsevier