Note on self-organizing semantic maps

ARTICLE *in* IEEE TRANSACTIONS ON NEURAL NETWORKS · FEBRUARY 1995

Impact Factor: 2.95 · DOI: 10.1109/72.410347 · Source: PubMed

2 AUTHORS:

James C. Bezdek

University of Missouri

346 PUBLICATIONS **23,155** CITATIONS

SEE PROFILE

Nikhil Pal

Indian Statistical Institute

249 PUBLICATIONS **9,129** CITATIONS

SEE PROFILE

Biol. Cybern. 61, 241–254 (1989)

Biological Cybernetics
© Springer-Verlag 1989

Self-Organizing Semantic Maps

H. Ritter* and T. Kohonen

Helsinki University of Technology, Laboratory of Computer and Information Science, Rakentajanaukio 2C, SF-02150 Espoo, Finland

Abstract. Self-organized formation of topographic maps for abstract data, such as words, is demonstrated in this work. The semantic relationships in the data are reflected by their relative distances in the map. Two different simulations, both based on a neural network model that implements the algorithm of the self-organizing feature maps, are given. For both, an essential, new ingredient is the inclusion of the contexts, in which each symbol appears, into the input data. This enables the network to detect the "logical similarity" between words from the statistics of their contexts. In the first demonstration, the context simply consists of a set of attribute values that occur in conjunction with the words. In the second demonstration, the context is defined by the sequences in which the words occur, without consideration of any associated attributes. Simple verbal statements consisting of nouns, verbs, and adverbs have been analyzed in this way. Such phrases or clauses involve some of the abstractions that appear in thinking, namely, the most common categories, into which the words are then automatically grouped in both of our simulations. We also argue that a similar process may be at work in the brain.

1 Hypotheses About Internal Representation of Linguistic Elements and Structures

1.1 The Objective of this Work

One of the most intriguing problems in the theory of neural networks, artificial and biological, is to what extent a simple adaptive system is able to find abstractions, invariances, and generalizations from raw data. Many interesting results, e.g., in pattern recognition (artificial perception of images as well as acoustical and other patterns) have already been obtained. Extraction of features from geometrically or physically related data elements, however, is still a very concrete task, in principle at least. A much more abstract and enigmatic object of study is *cognitive information processing* that deals with elements of consciousness and their relationships; it is frequently identified with the ability to use languages. The purpose of the present paper is to study whether it is possible to create in artificial neural networks abstractions such that they, at least in a primitive form, would reflect some properties of the cognitive and linguistic representations and relations.

In particular we are here reporting new results which demonstrate that a self-organizing process is indeed able to create over a neural network topographically or geometrically organized maps that display semantic relations between symbolic data. It may be proper to call such representations *self-organizing semantic maps.*

We are also relating our results to the fundamental basis of cognition, namely, *categorization of observations.* As the connection of these ideas to fundamental theories of knowledge might otherwise remain obscure, it may be proper to commence with a short review of the philosophical background, namely, the *theory of categories* as the ultimate framework of abstractions.

1.2 On Categories and their Relation to Linguistic and Neural Representations

The most general concepts or abstractions that are needed to interprete the empirical world are called *categories*; such basic reduced elements and forms of thinking and communication can also be encountered in all languages, primitive as well as more developed.

* On leave from: Institut für Physik, Technische Universität München, James-Franck-Strasse, D-8046 Garching, FRG
Present address: Beckman Institute, University of Illinois, Urbana, IL 61801, USA

The categories have been supposed to embrace the whole domain of knowledge, and to form the basis of consciousness. *Aristotle* (384–322 B.C.) in fact already distinguished ten categories. The most common of them are: 1. Items (objects). 2. Qualities (properties). 3. States (or state changes). 4. Relations (spatial, temporal, and other).

In languages, Category 1 corresponds to nouns, Category 2 to adjectives, and Category 3 to verbs, respectively. For the representation of Category 4, different languages use, e.g., adverbs, prepositions, postpositions, endings, inflections, or syntax (order of words). Naturally many auxiliary word classes are needed to interrelate phrases and clauses, to indicate logic modalities, as well as to facilitate inductive and deductive inference.

The very deep original metaphysical meaning of "category" has to a large extent been lost in the common usage of this word. "Categories" are often simply identified with *classes* of items such as animals, plants, nationalities, professions, etc. More accurately, such classes then only constitute subcategories of Category 1.

Since representations of categories occur in all languages, many researchers have stipulated that the deepest semantic elements of any language must have a physiological representation in the neural realms; and since they are independent of the different cultural histories, it has been concluded that such representations must have been inherited genetically.

At the time when the genetic predisposition of language elements was suggested, there was no mechanism known that would have explained the origin of abstractions in neural information processing other than evolution. It was not until "neural network" modeling reached the present level when researchers began to discover *internal representations* of abstract properties of signals from the physical network models. There exist at least two classes of models with this potential: the *backpropagation network* (Rumelhart and McClelland 1984; Rumelhart et al. 1986), and the *self-organizing [topological feature] map* (Kohonen 1982a–c, 1984). Such findings indicate that the internal representations of categories may be derivable from the mutual relations and roles of the primary signal or data elements themselves, as demonstrated below.

However, it is not the purpose of this paper to contend that all representations in the biological brain would solely be acquired by learning. The adaptive principles discussed below must be regarded as theoretical frameworks, and the one-phase learning its simplest form. It is quite possible that a similar process is at work in the genetic cycles, although its explicit mechanisms are difficult to devise.

It will now be proper to approach the problem of self-organizing semantic maps using data that contain implicit information relating to the simplest categories; if the latter are then detectable automatically, we may think that a significant step towards self-organizing linguistic processing has been taken.

One aspect may still be emphasized. It is perhaps not reasonable to look for elements of *languages* in the brain. A more fundamental view is that the physiological functions are expected to reflect the *categorical* organization and not so much the detailed linguistic forms.

1.3 Examples of Neural Network Models for Internal Representations

1.3.1 Semantic Networks. For a straightforward materialization of the internal representations, the *semantic networks* have been suggested. In their original form they comprise a graph structure with nodes and links. The nodes may stand for items or concepts (sets of attributes), whereas the links usually indicate relations: the simplest binary relations represent co-occurrences of items in the observed events, whereas labeled links describe their qualified relations. The semantic networks have also been supposed to have a one-to-one counterpart in neural cells and their interconnects, whereby a searching process would be interpreted as spreading activation in such a neural network (Quillian 1968; Collins and Loftus 1975). In view of the contemporary neurophysiological data, such a degree of specificity and spatial resolution is highly improbable in biology. One also has to realize that in the semantic-network model of the brain, predisposition of semantic meaning to the nodes and links has to be *postulated*; such a "mapping" is not derived from any self-organizing process.

1.3.2 Hidden Layers in Backpropagation Networks. Whether the nowadays familiar feedforward "neural" networks with error-backpropagation learning may be regarded as biological models or not, cells or nodes in their "hidden layers" often seem to learn responses that are specific to some abstract qualities of the input information (Rumelhart and McClelland 1984; Rumelhart et al. 1986; Sejnowski and Rosenberg 1987). However, it has to be emphasized that backpropagation is crucially based on *supervised learning*. The outputs, in relation to input stimuli, are forced to given values by optimization of the internal weight parameters of the nodes of the network. In a multilevel network, with structured data, it may then happen that in order to reach the global optimum, some nodes of the innermost layers (hidden layers) become tuned to represent some kind of "eigendata" of the occurring signals, that then represent the "generalizations" or "abstractions". It has recently been demonstrated that the weight vectors of the hidden layers may converge to values that encode linguistic items according to their semantic roles. These roles were defined explicitly in the learning process (Miikkulainen and Dyer 1988a, b).

1.3.3 Self-Organizing [Topological Feature] Maps. A more genuine type of self organization is *competitive (unsupervised) learning* that is able to find clusters from the primary data, eventually in a hierarchically organized way. In a system of feature-sensitive cells (cf. Nass and Cooper 1975; Perez et al. 1975; Grossberg 1976) competitive learning means that a number of cells is comparing the same input signals with their internal parameters, and the cell with the best match ("winner") is then tuning itself to that input. In this way different cells learn different (average) aspects from their input, that then may be regarded as the simplest forms of abstractions.

The *self-organizing [topological feature] maps* (Kohonen 1982a–c, 1984) are a further development of competitive learning in which the best-matching cell also activates its topographical neighbors in the network to take part in tuning to the same input. A striking, and by no means obvious result (cf. Sect. 3) from such a collective, cooperative learning is that, assuming the "neural network" as a two-dimensional sheet, the different cells become tuned to different inputs *in an orderly fashion, defining some feature coordinate system over the network.* After learning, each input elicits a *localized response*, whose position in the sheet reflects the most important "feature coordinates" of the input. This corresponds to a non-linear projection of the input space onto the network that makes the most essential neighborhood relationships [topology] between the data elements geometrically explicit. In particular, if the data are clustered hierarchically, a very explicit localized representation of the same structure is generated.

While the self-organizing maps, as such, have been used in many applications to visualize clustered data (cf. Kohonen 1984; Miikkulainen and Dyer 1988b), a much more intriguing possibility is that *they are also directly able to create in an unsupervised process topographical representations of semantic, nonmetric relationships implicit in linguistic data*, as will be pointed out in Sect. 4.

2 Are the Information-Processing Functions in the Brain Localized? Justification of the Model

2.1 General Issues Against and in Favor of Localization

Behavioral psychology generally emphasizes the global and holistic nature of higher human information processing. Some earlier neurophysiological findings indeed seemed to support this view. Distributedness of learning results in the cell mass of the brain was discovered in Lashley's classical experiments in 1938 (Beach 1960), which for a long time were interpreted such that the brain is a "black box" with more or less equipotential components that can even replace each other. An extreme view holds all attempts to isolate and locate cognitive functions in the brain as a modern form of phrenology.

It is true that in a process that leads to a perception or action, many parts of the brain are involved, eventually in an iterative or recursive fashion. This, however, could be said of any device or machinery that has been designed to perform a particular task, and needs the cooperation of all of its components. However, it would be absurd to deny, in view of massive neurophysiological data, that the brain contains parts, areas, networks, and even single neural cells that perform specific partial functions. There exist recordings of many types of feature-sensitive cells or sites that respond to specific qualities of sensory stimuli, and the motoneurons that control particular muscles are certainly localized. The global functions obviously ensue from the cooperation of a great many components of this type. The amount of parallelism and redundancy in such processing may be enormous. The question that then remains only concerns the *degree* or *spatial acuity* of localization, as well as a possible hierarchical organization of such localized functions.

2.2 On the Techniques to Determine Localization, and their Criticism

By the end of the nineteenth century, a rather detailed topographical organization of the brain, especially cortex, was already deducible from functional deficits and behavioral impairments that were induced by various kinds of defects caused either accidentally, due to tumors, malformations, hemorrhages, or artificially caused lesions. A modern technique to cause controllable and reversible "lesions" is to *stimulate* a particular site on the cortical surface by small electric currents, thereby eventually inducing both excitatory and inhibitory effects, but anyway disturbing an assumed local function. If such a spatially confined stimulus then systematically disrupts a specific cognitive ability such as naming of objects, there exists at least some indication that the corresponding site is essential to that task. This technique has frequently been criticized for the following reason, which holds for all lesion studies. Although a similar lesion at the same site would always cause the same deficit, and the same deficit were never produced by another type of lesion, it is logically not possible to use such data as a conclusive proof for localization; the main part of the function may reside elsewhere, while the lesion may only destroy a vital control connection to it. Hughlings Jackson (1878) already stated: "To locate the damage which destroys speech and to localize speech are two different things".

Another controllable way for the determination of localization is to chemically depress or enhance the process that causes the triggering of the neurons, e.g., using small patches soaked in strychnine. This technique has earlier been used with success to map, e.g., primary sensory functions.

A straightforward method to locate a response is to *record* the electric potential or train of neural impulses associated with it. In spite of ingenious multielectrode techniques developed, this method does not detect all responses in an area, and since the neural tissue is very inhomogeneous, the coupling made to a particular neuron may be haphazard. However, plenty of detailed mappings, especially from the primary sensory and associative areas, have been made by various electrophysiological recording techniques.

More conclusive evidence for localization can be obtained by modern *imaging techniques* that directly display the spatial distribution of brain activity associated with a function, achieving a spatial resolution of a few millimeters. The two main methods that are based on radioactive tracers are: *positron emission tomography* (PET), and autoradiography of the brain through very narrow collimators (*gamma camera*). PET reveals changes in oxygen uptake and phosphate metabolism. The gamma camera method directly detects changes in cerebral blood flow. Both phenomena correlate with local neural activity, but they are not able to follow fast phenomena. In *magnetoencephalography* (MEG), the low magnetic field caused by neural responses is detected, and by computing its sources, the neural responses can directly be analyzed reasonably fast, with a spatial resolution of a couple of millimeters. The main drawback is that only such current dipoles are detectable that lie in parallel to the surface of the skull; i.e., it is mainly the sulci of the cortex that can be studied with this method.

There seems to exist no ideal technique that alone could be used to map all the neural responses. It is necessary to combine anatomical, electrophysiological, imaging, and histochemical studies (for a general overview, see, e.g. Kertesz 1983).

2.3 Topographic Maps in Sensory Areas

Generally speaking, two types of physiological maps are distinguishable in the brain: those that are clearly ordered, and those looking almost randomly organized, respectively.

Maps that form a continuous, ordered image of some "receptive surface" can be found in the visual (cf., e.g., van Essen 1985) and somatosensory cortices, in the cerebellum, and in certain nuclei. The local scale, or magnification factor of these maps depends on the behavioral importance of particular signals; e.g., images of the foveal part of the retina, and of fingertips and lips are greatly magnified in proportion to those of other parts. There is thus a "quasiconformal" mapping of the "receptive surfaces" onto the brain.

There also exist more abstract, ordered, continuous maps in many other primary sensory areas, such as the tonotopic or auditive-frequency maps (Tunturi 1950, 1952; Reale and Imig 1980); echo delay and Doppler shift maps in bats (Suga and O'Neill 1979); and the color maps in the visual area V4 (Zeki 1980). It is a common feature of such maps that they are confined to a rather small area, seldom exceeding 5 mm in diameter, whereby it is justified to use a model for them in which the whole network is assumed structurally homogeneous. Over such an area, a spatially ordered mapping along one or two important feature dimensions of the sensory signals is usually discernible.

Physiologists also use the word "map" for nonordered responses to sensory stimuli as long as these are spatially localizable, even if they are scattered randomly over an area of several square centimeters, and many different kinds of responses are found in the same area. More complex visual responses found on higher levels are mapped in this way: for instance, cells have been detected that selectively respond to faces (Rolls 1984).

2.4 Evidence for Localization of Linguistic Functions

It has been known at least for a century that sensory aphasia is caused by lesions in the posterior superior part of the temporal lobe of the brain called Wernicke's area; but even with modern imaging techniques, only a very rough location of language functions has been possible. Practically all systematic high-resolution mappings have been made by the stimulation method.

It is much more difficult to locate linguistic or semantic functions in the brain than to map the primary sensory areas. First, it is still unclear to what aspects of language the feature dimensions might correspond. Second, as noted earlier, such a map may be scattered. Third, responses to linguistic elements may only occur within short "time windows". Fourth, the experimental techniques used in animal studies, being usually invasive, cannot be applied to human beings, unless there exists indication for a surgical operation.

Nevertheless, a significant amount of experimental evidence is already available supporting the view of a rather high degree of localization of language functions.

PET imaging has revealed that during a task of processing single words, several separate cortical sites are simultaneously active. These are not all located in Wernicke's area; some parts in the frontal lobe and the associative areas may simultaneously show responses, too, especially in sites obviously associated with visual and auditory perception, articulation, and planning of the task (Peterson et al. 1988).

In order to study possible internal representations, location of sites relating to semantic processing needs a finer resolution, of the order of one millimeter, so far hard to achieve even by stimulation mapping. However, this method cannot detect any temporal peaks of activity, it can only produce reversible temporary blocking of processing in a region confined to a few square millimeters. Repeated stimulation of the same site then usually causes a reproducible kind of temporary deficit, e.g., errors in the naming of objects, or difficulty in recollection from short-term verbal memory. However, stimulation of another site only 5 mm apart may already induce a completely different type of deficit, or no effect at all (Ojemann 1983). Further, there are cases of bilingual patients where naming of the same object is impaired only in either of the two languages, dependent on the site being stimulated (Ojemann ibid.). It seems as if the language functions were organized as a "mosaic" of localized modules (Ojemann ibid.).

Other, more indirect evidence for fine-structured mapping is available from several cases of selective deficits as a result of strokes or brain injuries (Warrington and McCarthy 1987). Examples include deficits in the use of concrete (impaired) versus abstract (spared) words (Warrington 1975), inaminate versus animate words (Warrington and McCarthy 1983; McCarthy and Warrington 1988), or living objects and food (impaired) versus inaminate (spared) words (Warrington and Shallice 1984). There exist other well-documented reports on selective impairments relating to such subcategories as indoor objects (Yamadori and Albert 1973), body parts (McKenna and Warrington 1978), and fruits and vegetables (Hart et al. 1985); see also review articles on categorical impairments (Goodglass et al. 1986; Caramazza 1988).

Analysis of such data has led to the conclusion that there exist separate modules in the brain for a "visual word lexicon" and a "phonetic word lexicon" for word recognition, a "semantic lexicon" for the meaning of words, as well as an "output lexicon" for word articulation (Caramazza 1988), respectively. Each of these modules can be impaired independently.

The categorical impairments reported above seem to relate to selective damages caused to the "semantic lexicon". These observations cannot provide conclusive evidence for localization of semantic classes within this lexicon, because in all these cases it was not possible to assess the precise spatial extent of critically affected brain tissue. Nonetheless it seems justified to state that selective impairment in such a large number of cases would be very difficult to explain if the semantic organization apparent from such observations were not in some way reflected in the spatial layout of the system.

3 Representation of Topologically Related Data in the Self-Organizing Map

Any model suggested for the self-organized formation of internal representations (such as the feature-sensitive cells) should also be able to make essential relations among data items explicit. An intriguing way of achieving this is the formation of *spatial maps*, which are perhaps the most explicit local representations known.

Several years ago, one of the authors (Kohonen 1982a–c, 1984) developed a model of neural adaptation that is capable of unsupervised formation of spatial maps for many different kinds of data. This section first summarizes the (simplified) model equations and then explains how a structure-preserving map of hierarchically related data is generated by them. More detailed description of the process and its background can be found in the following original publications and some recent developments (Kohonen 1982a–c, 1984; Cottrell and Fort 1986; Ritter and Schulten 1986, 1988, 1989).

The model assumes a set of laterally interacting adaptive neurons, usually arranged as a two-dimensional sheet. The neurons are connected to a common bundle of input fibers. Any activity pattern on the input fibers gives rise to excitation of some local group of neurons. After learning, the spatial positions of the excited groups specify a mapping of the input patterns onto the two-dimensional sheet, the latter having the property of a topographic map, i.e. it represents distance relations of the high-dimensional space of the input signals approximately as distance relationships on the two-dimensional neural sheet. This remarkable property follows from the assumed lateral interactions and a very simple, biologically justifiable adaptation law. In fact, it seems that the main requirements for such self organization are that (*i*) the neurons are exposed to a sufficient number of different inputs, (*ii*) for each input, the synaptic input connections to the excited group are only affected, (*iii*) similar updating is imposed on many adjacent neurons, and (*iv*) the resulting adjustment is such that it enhances the same responses to a subsequent, sufficiently similar input.

Mathematically, the activity pattern at the input is described by an n-dimensional real input vector $\mathbf{x}$, where n is the number of input lines. The responsiveness of neuron r is specified by an n-dimensional vector $\mathbf{w}_r$, eventually corresponding to the vector of synaptic efficacies, and it is measured by the dot product $\mathbf{x} \cdot \mathbf{w}_r$. For efficiency of the process and mathematical convenience, all the input vectors are always normalized to unit length, whereas the $\mathbf{w}_r$ need not be normalized in the process explicitly; sooner or later, the process will normalize them automatically. The neurons are arranged in a two-dimensional lattice, and each neuron is labeled by its two-dimensional lattice position r. The group of excited neurons is taken to be centered at the neuron s for which $\mathbf{x} \cdot \mathbf{w}_s$ is maximal. Its extent and shape are described by a function h_{rs}, whose value is the excitation of neuron r, if the group center is at s. This function may be constant for all r in a "neighborhood zone" around s and zero elsewhere, or bell-shaped, like in the present simulations that are supposed to describe a more natural mapping. In this case h_{rs} will be largest at $r=s$ and decline to zero with increasing distance $\|r-s\|$. A rather realistic modeling choice for h_{rs} is

$$h_{rs}=\exp\left(-\frac{\|r-s\|^2}{\sigma^2}\right), \tag{1}$$

i.e. a Gaussian of the distance $\|r-s\|$, whose variance $\sigma^2/2$ will control the radius of the group. The adjustments corresponding to the input $\mathbf{x}$ shall then be given by

$$\mathbf{w}_r^{(\text{new})}=\mathbf{w}_r^{(\text{old})}+\varepsilon \cdot h_{rs} \cdot (\mathbf{x}-\mathbf{w}_r^{(\text{old})}). \tag{2}$$

Equation (2) can be justified by assuming the traditional Hebbian law for synaptic modification, and an additional nonlinear, "active" forgetting process for the synaptic strengths (Kohonen 1984). Equation (2) has the desired property of confining any adaptations to the neighborhood of neuron s and improving subsequent responses to $\mathbf{x}$ there.

We shall not present here any formal proof for that these conditions indeed lead to an ordered organization of the map (an intuitive picture is that the weight vectors $\mathbf{w}_r$ become aligned with the input signal distribution as if they would constitute elements of an "elastic surface"). The reader interested in these aspects may find them in the literature (Kohonen 1982b; Cottrell and Fort 1986; Ritter and Schulten 1988). For the present purpose it may suffice to assert that the resulting maps are nonlinear projections of the input space onto this surface with the following two main properties: (*i*) the distance relationships between the source data are preserved by their images in the map as faithfully as possible. However, a mapping from a high-dimensional space to a lower-dimensional one will usually distort most distances and only preserve the most important neighborhood relationships between the data items, i.e., the topology of their distribution. This is the driving factor for the formation of a reduced representation in which irrelevant details are ignored. (*ii*) If different input vectors appear with different frequencies, the more frequent ones will be mapped to larger domains at the expense of the less frequent ones. This results in a very economic allocation of memory resources to data items, and complies with physiological findings.

If the data form clusters in the input space, i.e. if there are regions with very frequent and at the same time very similar data, (*i*) and (*ii*) will ensure that the data of a cluster are mapped to a common localized domain in the map. Moreover, the process will arrange the mutual placement of these domains in such a way as to capture as much of the overall topology of the cluster arrangement as possible. In this way, even hierarchical clustering can be achieved, a capability frequently thought to represent one form of abstraction. Earlier demonstrations of this can be found in Kohonen (1982c, 1984) and Kohonen et al. (1984); in the present work, more refined results, relating to linguistic expressions, are given.

4 Self-Organizing Semantic Maps

4.1 *Self-Organizing Symbol Map*

In the demonstrations described in (Kohonen 1982c) and (Kohonen 1984), the self-organizing maps mainly reflected *metric distance relations* between patterned representation vectors. Such data are characteristic of most lower levels of perception. However, much of higher-level processing, in particular *language* and *reasoning*, seems to rest on the processing of *discrete symbols.* Hence we must understand how the brain might form meaningful representations of symbolic entities. In view of the localization seemingly apparent even on this level, we must in particular explain how maps of symbols can be formed in which logically related symbols occupy neighboring places.

One might think that applying the neural adaptation laws (2) to a symbol set (regarded as a set of vectorial variables) might create a topographic map that displays the "logical distances" between the symbols. However, there occurs a problem which lies in the different nature of symbols as compared with continuous data. For the latter, similarity always shows up in a natural way, as the metric differences between their continuous encodings. This is no longer true for discrete, symbolic items, such as words, for which no metric has been defined. It is in the very

		dove	hen	duck	goose	owl	hawk	eagle	fox	dog	wolf	cat	tiger	lion	horse	zebra	cow
	small	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0
is	medium	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
	big	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
	2 legs	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	4 legs	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
has	hair	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	hooves	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	mane	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0
	feathers	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	hunt	0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0
likes	run	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0
to	fly	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0
	swim	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 1. Animal names and their attributes

nature of a symbol that *its meaning is dissociated from its encoding*[1]. Hence logical relatedness between different symbols will in general not be directly detectable from their encodings and one may thus not presume any metric relations between the symbols, even when they represent similar items. How could it then be possible to map them topographically? The answer is that the symbol, at least in the learning process, must frequently be presented *in due context*, i.e. in conjunction with all or part of the attribute values of the item it encodes, or with other, correlating symbols.

The simplest system model for symbol maps assumes each data vector $\mathbf{x}$ as a concatenation of two (or more) fields, one specifying the symbol code, denoted by $\mathbf{x}_s$, and the other the attribute set, denoted $\mathbf{x}_a$, respectively.

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_a \end{bmatrix} = \begin{bmatrix} \mathbf{x}_s \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{x}_a \end{bmatrix}. \tag{3}$$

Equation (3) illustrates in vector notation that the encodings of the symbol part and the attribute part can form a vector sum of two orthogonal components. The core idea underlying symbol maps is that the two parts are weighted properly such that *the norm of the attribute part predominates over that of the symbol part during the self-organizing process*; the topographical mapping then mainly reflects metric relationships of the attribute sets. Since the inputs for symbolic signals, however, are also active all the time, memory traces from them are formed to the corresponding inputs of those cells of the map that have been selected (or actually forced) by the attribute part. *If then, during recognition of input information, the attribute signals are missing or are weaker, the (same) map units are selected on the basis of the symbol part solely. In this way the symbols become encoded into a spatial order reflecting their logic (or semantic) similarities.*

Attributes may be variables with scalar-valued discrete or continuous values, or they may attain qualitative properties such as "good" or "bad". It is simplest to assume that the identity of each attribute is clear from its position in the "attribute field" $\mathbf{x}_a$, whereby the presence or absence of a particular qualitative property may be indicated by a binary value, say 0 or 1, respectively. Then the (unnormalized) similarity between two attribute sets may be defined in terms of the number of attributes common to both sets, or equivalently, as the dot product of the respective attribute vectors[2].

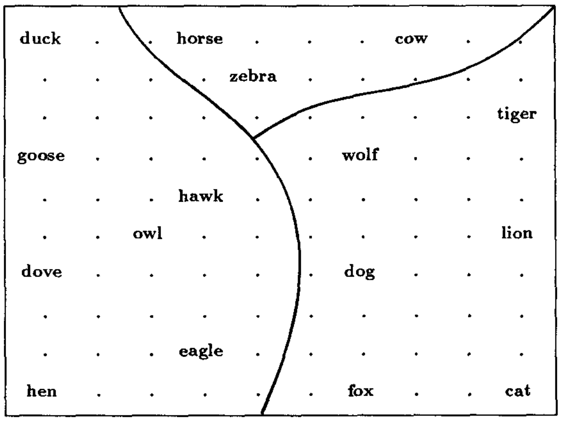
To illustrate this with a concrete model simulation, consider the data given in Fig. 1. Each column is a very schematic description of an animal, based on the presence (=1) or absence (=0) of some of the 13 different attributes given on the left. Some attributes, such as "feathers" and "2 legs" are correlated, indicating more significant differences than the other attributes. In the following, we will take each column for the attribute field $\mathbf{x}_a$ of the animal indicated at the top. The animal name itself does not belong to $\mathbf{x}_a$ but instead specifies the symbol part $\mathbf{x}_s$ of the animal. Selection of the symbol code can be done in a variety of ways. However, we now want to be sure that the encoding of the symbols does not convey any information about similarities between the items. Hence we choose for the symbol part of the k-th animal a d-dimensional vector, whose k-th component has a fixed value of a, and whose remaining components are zero. Here d is the number of items ($d=16$ in our example). For this choice, the metric distance between any two of the vectors $\mathbf{x}_s$ is the same, irrespective of the symbols they encode.

The parameter a may be interpreted as scaling the "intensity" of input from the symbol field and it determines the relative influence of the symbol part as compared to the attribute part. As we wanted the latter to predominate, we chose a value of $a=0.2$ for our simulation. Combining $\mathbf{x}_a$ and $\mathbf{x}_s$ according to (3), each animal was encoded by a 29-dim data vector $\mathbf{x}=[\mathbf{x}_s, \mathbf{x}_a]^T$. Finally each data vector was normalized to unit length. Although this is only a technical means to guarantee good stability in the self-organizing process, its biological counterpart would be intensity normalization of the incoming activity patterns.

[1] Some symbols may contain a residue of analogical representation, e.g., in some words describing sounds

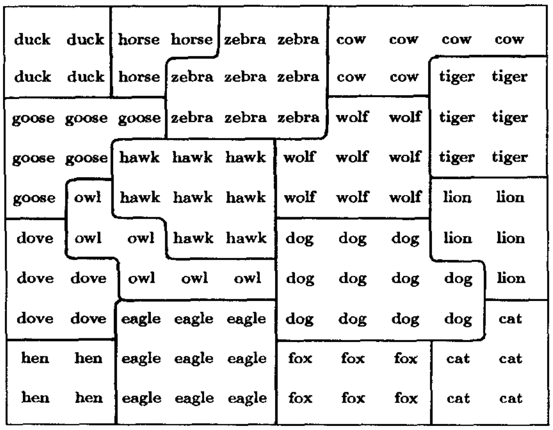
[2] It might seem more logical to use the value +1 to indicate the presence of an attribute, and −1 for its absence, respectively; however, due to normalization of the input vectors, and their subsequent comparison by the dot product, the attribute values 0 have a qualitatively similar effect as negative components in a comparison on the basis of vectorial differences

The members of the data set thus obtained were presented iteratively and in a random order to a planar network of 10 × 10 neurons subject to the adaptation process described above. The initial connection strengths between the neurons and their $n=29$ input lines were chosen to be small random values, i.e. no prior order was imposed. However, after a total of 2000 presentations, each "cell" became more or less responsive to one of the occuring attribute combinations and simultaneously to one of the 16 animal names, too. If we now test which cell gives the strongest response if only the animal name is presented as input (i.e. $\mathbf{x}=[\mathbf{x}_s, \mathbf{0}]^T$), we get the map shown in Fig. 2 (the dots indicate neurons with weaker responses). It is highly apparent that the spatial order of the responses has captured the essential "family-relationships" among the animals. Cells responding to, e.g., "birds" occupy the left part of the lattice, "hunters" such as "tiger", "lion", and "cat" gather towards the right, more "peaceful" species such as "zebra", "horse", and "cow" aggregate in the upper middle. Within each cluster, a further grouping according to similarity is discernible. Figure 3 shows the result of a "simulated electrode penetration mapping" for the same network. It differs from Fig. 2 in that now each cell has been marked by the symbol that is its best stimulus, i.e., elicits the strongest response for that cell. This makes the parcellation of the "neural territory" into domains specific to one of the input items visible. Hierarchy thereby is represented by *nested domains.* A general class (e.g. "bird") occupies a larger territory, which itself is differentiated into nested subdomains, corresponding to more specialized items ("owl", "duck", "hen" etc.). Although highly idealized, this result is very suggestive of how a self-organizing system can learn to *spatially guide* the formation of memory traces in such a way that its final physical layout forms a direct image of the hierarchy of the most important "concept relationships".



duck	.	.	horse	.	.	.	cow	.	.
.	.	.	.	zebra	.	.	.	.	.
.	.	.	.	.	.	.	.	.	tiger
goose	.	.	.	.	.	wolf	.	.	.
.	.	.	hawk	.	.	.	.	.	.
.	.	owl	.	.	.	.	.	.	lion
dove	.	.	.	.	.	dog	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	eagle	.	.	.	.	.	.
hen	.	.	.	.	.	fox	.	.	cat

Fig. 2. After the network has been trained with inputs encoding animal names together with some attributes (see Fig. 1), presentation of the animal names alone elicits maximal responses at the cell locations shown. A grouping according to similarity has emerged



duck	duck	horse	horse	zebra	zebra	cow	cow	cow	cow
duck	duck	horse	zebra	zebra	zebra	cow	cow	tiger	tiger
goose	goose	goose	zebra	zebra	zebra	wolf	wolf	tiger	tiger
goose	goose	hawk	hawk	hawk	wolf	wolf	wolf	tiger	tiger
goose	owl	hawk	hawk	hawk	wolf	wolf	wolf	lion	lion
dove	owl	owl	hawk	hawk	dog	dog	dog	lion	lion
dove	dove	owl	owl	owl	dog	dog	dog	dog	lion
dove	dove	eagle	eagle	eagle	dog	dog	dog	dog	cat
hen	hen	eagle	eagle	eagle	fox	fox	fox	cat	cat
hen	hen	eagle	eagle	eagle	fox	fox	fox	cat	cat

Fig. 3. "Simulated electrode penetration mapping" for the network in Fig. 2. Each cell is labeled by the animal name eliciting the strongest response. Cells responsive to the same animal name form domains, which are grouped according to similarity between the animals

4.2 Role-Based Semantic Maps

In the example of the animal map, the role of context was still very simple: the encoded symbol was related to a set of explicit attributes statically. In natural languages, and obviously in any natural perception, too, the items and their attributes, and obviously some state information usually occur in a temporal sequence. The concept of context then needs to be broader and span the time dimension, too. Perhaps the simplest way to do this is to define for the context of each item all those items (together with their serial order) that occur in a certain "time window" around the selected item.

In this work we will not pay any attention to the concrete physical representation of signals, i.e., whether the patterns are temporal, like in speech, or spatial, like in text. For the series-to-parallel conversion, neural networks may use paths with different delays, eigenstates that depend on sequences, or any other mechanisms implemented in the short-term memory. Here we shall only concentrate on the similarities between the expressions that arise from conditional occurrences of their parts, and simply imagine that triples or pairs of words can somehow be presented to the input ports of the system.

Languages contain very many levels of meaning. It is possible to construct cases, where the due "window" for the understanding of a word has to comprise a whole novel. On the other hand, the possibility of *forming grammars* demonstrates that a significant part of the structure of a language already manifests itself on a very low level, down to patterns of words and endings. Detection of such "short range" structure will be the focus of our interest in this section and we shall demonstrate that the inclusion of a very limited *word context* enables the basic network model (1) to form *semantic maps*, in which the word items are grouped according to semantic categories (objects, activities, qualifications etc.) and simple similarity.

For our demonstration, we used a set of randomly generated three-word sentences constructed from the vocabulary in Fig. 4a. The vocabulary comprises nouns, verbs, and adverbs, and each class has further subdivisions, such as names of persons, animals, and inanimate objects in the category of nouns. These distinctions are in part of a grammatical, in part of a semantic nature. However, for the reasons discussed in Sect. 4.1, they shall not be discernible from the coding of the words themselves but only from the context in which the words are used. In natural languages, such a context would comprise a much richer variety of sensory experiences. In this very limited demonstration, however, we will only take into account the context provided by the immediately adjacent textual environment of each word occurrence. It will turn out that even this extremely restricted context will suffice to convey some interesting semantic structures. Of course this requires that each sentence be not totally random, but obey at least some rudimentary rules of grammar and semantic correctness. This is ensured by *restricting the random selection to a set of 39 "legal" sentence patterns only*. Each pattern is a triple of numbers from Fig. 4b. A sentence is constructed by randomly choosing one of the triples and substituting each number by one of the words with the same number in Fig. 4a. This results in a total of 498 different three-word sentences, a few of which are given in Fig. 4c. (Whether those statements are true or not is not our concern; we are only interested in their semantic correctness).

Bob/Jim/Mary	1
horse/dog/cat	2
beer/water	3
meat/bread	4
runs/walks	5
works/speaks	6
visits/phones	7
buys/sells	8
likes/hates	9
drinks/eats	10
much/little	11
fast/slowly	12
often/seldom	13
well/poorly	14

(a)

Sentence Patterns:		
1-5-12	1-9-2	2-5-14
1-5-13	1-9-3	2-9-1
1-5-14	1-9-4	2-9-2
1-6-12	1-10-3	2-9-3
1-6-13	1-11-4	2-9-4
1-6-14	1-10-12	2-10-3
1-6-15	1-10-13	2-10-12
1-7-14	1-10-14	2-10-13
1-8-12	1-11-12	2-10-14
1-8-2	1-11-13	1-11-4
1-8-3	1-11-14	1-11-12
1-8-4	2-5-12	2-11-13
1-9-1	2-5-13	2-11-14

(b)

Mary likes meat
Jim speaks well
Mary likes Jim
Jim eats often
Mary buys meat
dog drinks fast
horse hates meat
Jim eats seldom
Bob buys meat
cat walks slowly
Jim eats bread
cat hates Jim
Bob sells beer
(etc.)

(c)

Fig. 4. a List of used words (nouns, verbs and adverbs), **b** sentence patterns, and **c** some examples of generated three-word-sentences

In this very simple demonstration, it was supposed that the context of a word was sufficiently defined by the pair formed by its immediate predecessor and successor. (To have such pairs also for the first and the last word of a sentence we assume the sentences to be concatenated in the random order of their production.) For the 30-word vocabulary in Fig. 4a we could have proceeded as in Sect. 4.1 and represented each of such pairs by a 60-dim vector with two non-zero entries. For a more economical encoding, however, as explained more closely in Appendix I, we assigned to each word a 7-dim random vector of unit length, chosen at the outset for each word independently from an isotropic probability distribution. Hence each predecessor/successor-pair was represented by a 14-dim code vector.

It turned out in all of our computer experiments that instead of paying attention to each clause separately, a much more efficient learning strategy was to consider each word in its *average context* over a set of possible clauses, before presenting it to the learning algorithm. The (mean) context of a word was thus first defined as *the average over 10,000 sentences of all code vectors of predecessor/successor-pairs surrounding that word*. The resulting thirty 14-dim "average word contexts", normalized to unit length, assumed a similar role as the attribute fields $\mathbf{x}_a$ in the previous simulation. Each "attribute field" was combined with a 7-dim "symbol field" $\mathbf{x}_s$, consisting of the code vector for the word itself, but scaled to length a. This time, the use of the random code vectors approximately guaranteed that the symbol fields $\mathbf{x}_s$ did not convey any information about similarity relationships between the words. As before, the parameter a determined the relative influence of the symbol part in comparison to the context part and was set to $a=0.2$.

For this experiment, a planar lattice of 10×15 formal neurons was used. As before, each neuron initially made only weak random connections to the $n=21$ input lines of the system, so that again no initial order was present.

After 2000 input presentations the responses of the neurons to presentation of the symbol parts alone were tested. In Fig. 5, the symbolic label was written to that site at which the symbol signal $\mathbf{x}=[\mathbf{x}_s, \mathbf{0}]^T$ gave the maximum response. We clearly see that *the contexts have "channeled" the word items to memory positions whose arrangement reflects both grammatical and semantic relationships*. Words of same type, i.e. nouns, verbs, and adverbs, have segregated into separate,

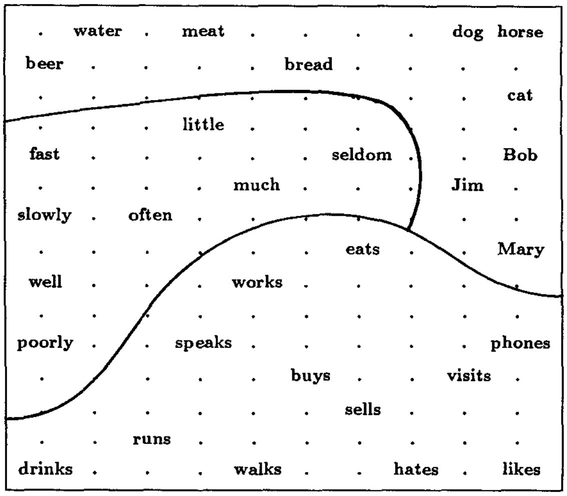


Fig. 5. "Semantic map" obtained on a network of 10 × 15 cells after 2000 presentations of word-context-pairs derived from 10,000 random sentences of the kind shown in Fig. 4c. Nouns, verbs and adverbs are segregated into different domains. Within each domain a further grouping according to aspects of meaning is discernible

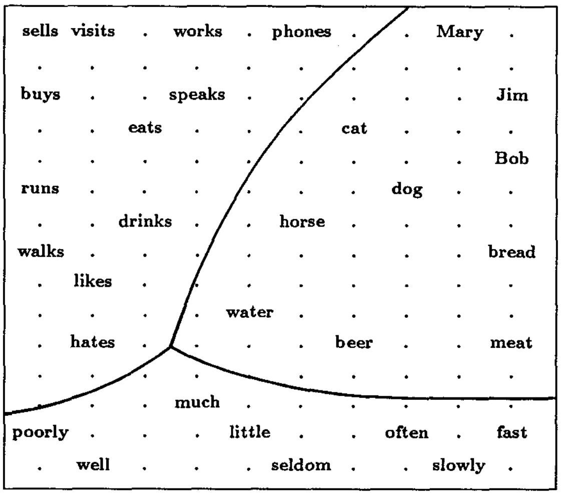


Fig. 6. This map has been obtained by the same procedure as the map in Fig. 5, but with a more restricted context that included only the immediate predecessor of each word

large domains. Each of these domains is further subdivided according to similarities on the semantic level. For instance, names of persons and animals tend to be clustered in separate subdomains of a common "noun-domain", reflecting different co-occurrence with, e.g., verbs such as "run" and "phone". Adverbs with opposite meaning tend to be particularly close together, as their opposite meaning ensures them maximum common usage. The grouping of the verbs indicates differences in the ways they can co-occur with adverbs, persons, animals, and nonanimate objects such as e.g. "food".

Figure 6 shows the result of a further computer experiment, based on the same vocabulary and the same sentence patterns as before. However, in this simulation the context of a word was restricted to its immediate predecessor only (i.e. the context now consists of a 7-dim vector). Even this very limited context proved sufficient to produce a map with roughly similar properties as in Fig. 5. This shows that the displayed regularities are fairly robust to changes in the details of the encoding as long as the context captures a sufficient amount from the underlying logical structure.

One might argue that the structure resulting in the map has artificially been created by a preplanned choice of the sentence patterns allowed for the input. However, it is easy to check that the patterns in Fig. 4b almost completely exhaust the possibilities for combining the words in Fig. 4a into semantically well-formed three-word sentences (an astute reader may notice a few "semantic borderline cases" not covered, such as "dog eats cat"). This may make it clear that all the selected sentence patterns were really determined by the constraints inherent in the semantically correct usage of the words, and not vice versa. Moreover, a significant percentage of the word neighborhoods extended across borders of the randomly concatenated sentences. As this concatenation was unrestricted, such neighborhoods were largely unrelated[3] to the grammatical and semantic structure of the sentences and constituted a kind of "noise" in the ordering process. It is important to notice that this noise does not disguise the regularities otherwise present in the clauses.

However, one important remark is due here. Any realistic semantic brain maps would need a much more complicated, probably hierarchical model. The purpose of the simple artificial model used in this work was only to demonstrate the *potential* of a self-organizing process to form abstract maps. In particular, the simulation results, as such, should not be used to serve as a reference for direct topographic comparison with brain areas. As a comparison between Fig. 5 and Fig. 6 shows, there are many almost equivalent ways, in which a set of similarity relationships can be displayed in a map. Therefore the maps generated by the model are not unique, unless further constraints, such as e.g. boundary conditions or some coarse initial ordering are imposed. These may then initially "polarize" the system that then converges to a more unique map.

[3] They may still reflect differences in the most likely position of a word within a sentence

5 Discussion

One of the biological mechanisms that up to time has been poorly understood is the ability of the brain to form abstractions from primary sensory experiences at increasingly higher levels of generalization.

It is already well known that on the lower perceptual levels, sensory information first becomes organized into topographically ordered sensory maps, and it has also already been demonstrated theoretically that such maps can be formed adaptively, reflecting mutual metric relationships and statistics of the incoming signals. This same principle has been applied with considerable success to exacting technical pattern recognition tasks such as speech recognition.

In this work we have now shown that the principle of self-organizing maps can also be extended to higher levels of processing, where the relationships between items are more subtle and less apparent from their intrinsic features, a property that is characteristic of *symbolic expressions*. Symbols, in general, do not contain metrically relatable components. Consequently, meaningful topographic maps of symbols must no longer display the intrinsic features, but instead the *logical similarities* of their inputs. It turns out, however, that organized mappings of symbolic data may still ensue from the same basic adaptation laws, provided that the symbolic input data are presented together *with a sufficient amount of context, that then defines the similarity relationships between them.* If the symbolic descriptions leave memory traces on the same neurons at which the contextual signals converge, too, the same neurons then also become sensitized to the symbolic signals in a spatial order that reflects their logical similarity.

Symbols play a particularly important role in languages. In this work we have given two simulation examples that demonstrate the self-organized formation of *semantic maps*, in which semantic relationships between words have been encoded into relative spatial positions of localized responses. Our artificial maps are parcelled into hierarchically nested domains reflecting different categories of words. This parcellation totally emerges from the co-occurrence sensory context and words. In our simulations the sensory context was restricted to simple attribute sets or adjacent words in sentences. The simple kind of clauses used in this experiment occur in all languages, even primitive ones. It is therefore also of interest to note that experimental data (Sect. 2) indicate similar organizations in brain areas related to language processing. Especially the category-specific language impairments discussed in Sect. 2 (Warrington and McCarthy 1987) seem to reflect a very similar organization on a physiological level.

In the first simulation we used rather explicit attributes, thereby assuming that some neural mechanism had already generated them. The philosophy underlying our work is that a similar self-organizing tendency must exist at all levels of processing; its illustration, however, is only possible if the signals have some meaning to us.

The term "semantic map" used in this work does not yet refer to "higher word understanding"; words have only been grouped according to their local contexts. Due to the strong correlation between local context and word meaning, however, this *approximates* the semantic ordering met in natural languages, which presumably cannot yet be generated in such a one-phase learning. It is an intriguing question whether any subsequent processing stages could create an ordering that reflects higher-level meanings – which then would facilitate full understanding of the meaning of words – by some kind of iteration of the basic self-organizing process.

Our model emphasizes the role of the spatial arrangement of the neurons, an aspect only considered in very few modeling approaches. However, we would not like to give the impression that we are opposing the view of neural networks as distributed systems. The massive interconnects responsible for lateral interactions as well as the "engrams" relating to associative memory are certainly disseminated over large areas in the network.

On the other hand, it seems inevitable that any complex processing task needs some kind of *segregation of information into separate parts*, and localization is one of the most robust and efficient ways to achieve this goal. The semantic maps offer an efficient mechanism to perform a meaningful segregation of symbolic information even on a fairly high level of semantics, and they have the further virtue of being solely based on unsupervised learning. Whether we still should consider *relative timing* of signals (cf. von der Malsburg and Bienenstock 1986) remains a further objective of study.

There are further reasons not to disregard the spatial arrangement of the processing units. For instance, the anatomy of neural circuits sets constraints to the realizable connectivity among units. Further, brain signals do not solely rest on axonal signal transmission over freely selectable distances, but often employ diffusion of neurotransmitters and neuromodulators. In all likelihood, these constraints would limit the implementation of many computational mechanisms, unless this handicap were alleviated by the efficient spatial organization provided by the maps.

From a hardware point of view it should be expected that minimization of connectivity costs would strongly favor this kind of a neural network design. This may also give a clue why topographic organization is so widespread in the brain. Other arguments for localization are that spatial segregation of representations makes them more logical, by reducing the degree of their mutual interference, and that logically similar symbolic items, being spatially adjacent, may evoke each other associatively, as expressed in the classical laws of association.

Another remark may be necessary. Our simulations should not be taken as a suggestion that each word is represented by a so-called "grandmother cell" in the brain. Each word is a complex piece of information probably redundantly encoded by an entire neuronal population (and several times in separate "lexica", cf. 2.4). Even in the highly idealized model used in our simulations, it is not a single neuron but a whole subset of cells, surrounding the most responsive one, that gets tuned to a word (cf. Fig. 3). These subsets may then be engaged in further processing, not captured by the basic model. The number of cells assigned to such a subset also depends on the frequency of occurrence of the word. This is analogous to the case that the frequency of stimulus occurrence determines the local magnification factor in a sensory map (Kohonen op. cit., Ritter and Schulten 1986). Similarly, frequent words would recruit cells from a larger neural territory and be more redundantly represented. As a consequence, the more frequent words should be less susceptible to local damage. This complies with empirical observations in stroke patients, whereby the familiar words are more likely to "survive" than the rare ones.

Finally we would like to present an intriguing philosophical notion. As earlier pointed out, there exists both biological evidence and theoretical justification for the functioning of the brain requiring representation of its input information by meaningful parts processed at spatially separated locations. *The idea about fundamental categories postulated for the interpretation and understanding of the world must obviously stem from prior formation of such representations in the biological brain itself.*

Acknowledgement. One of the authors (H.R.) would like to acknowledge support from the German Ministry for Science and Technology (ITR-8800-G9) and from the Academy of Finland.

Appendix I: Dimensionality Reduction by a Random-Projection Method

When the dimensionality of context grows, the attribute vectors soon become impracticably high-dimensional. For instance, in the example in Sect. 4.2 the context was formed of adjacent words. A straightforward encoding would require for each word as many components in the attribute field as there are words in the vocabulary. It is therefore advisable to try to map the space of all different contexts into a much-lower-dimensional vector space, thereby approximately preserving their metric relations. In this Appendix we give a random-projection solution for this.

We assume the set of all conceivable contexts to be finite (albeit very large) and accordingly label the contexts by integers $i=1,2\ldots D$. Then the "straightforward encoding" would assign to each symbol A a vector $\mathbf{x}(A)$ from a D-dimensional space V_D by

$$\mathbf{x}(A)=\sum_{i=1}^{D} p(i|A)\hat{\mathbf{e}}_i. \tag{4}$$

Here $p(i|A)$ is the conditional probability for the occurrence of context i, given the presence of symbol A. The $\hat{\mathbf{e}}_i$, $i=1,2\ldots D$ form an orthonormal frame in V_D (this is a refinement over a mere enumeration of all contexts compatible with A, taking also their different frequencies into account). Similarity between symbols is then measured by the Euclidean distances between their code vectors given by (4).

However, (4) must be regarded as a very formal expression, since in most situations of practical interest the dimension D of the required space will be unmanageably high. As a remedy, we will replace the orthonormal frame $\{\hat{\mathbf{e}}_1\ldots\hat{\mathbf{e}}_D\}$ by a set of D unit vectors ξ_i, $i=1,2\ldots D$, selected from a space of a much lower dimension $d\ll D$, and whose directions are independently chosen from an isotropic random distribution. This is formally equivalent to a random projection ϕ from the original space onto the new one, ϕ being defined by

$$\phi\left(\sum_{i=1}^{D} x_i\hat{\mathbf{e}}_i\right)=\sum_{i=1}^{D} x_i\xi_i. \tag{5}$$

In the following theorem we shall give a justification of this intuitive procedure.

Theorem. *Let $\|\cdot\|_D$ and $\|\cdot\|_d$ denote the Euclidean distance norms in V_D and V_d, respectively, and let $\langle\cdot\rangle_\phi$ be the average over all possible isotropic random choices for the unit vectors ξ_i defining ϕ in* (5). *Then, for any pair of vectors* $\mathbf{x}$, $\mathbf{y}\in V_D$, *there exists the relation*

$$\langle(\|\phi(\mathbf{x})-\phi(\mathbf{y})\|_d^2-\|\mathbf{x}-\mathbf{y}\|_D^2)^2\rangle_\phi\leqq\frac{2}{d}\cdot\|\mathbf{x}-\mathbf{y}\|_D^4. \tag{6}$$

In other words, although the original distances were distorted under the random mapping ϕ, the relative distortion in general will be small, if d is large enough. Hence we can expect that any essential structures are preserved even in the much lower-dimensional space V_d.

Proof. Let

$$\mathbf{v}:=\mathbf{x}-\mathbf{y}=\sum_{i=1}^{D} v_i\hat{\mathbf{e}}_i \tag{7}$$

and

$$\sigma^2:=\langle(\|\phi(\mathbf{x})-\phi(\mathbf{y})\|_d^2-\|x-y\|_D^2)^2\rangle_\phi. \tag{8}$$

Then (all summation ranges are $\{1\ldots D\}$)

$$\begin{aligned}\sigma^2&=\langle(\|\mathbf{v}\|_D^2-\|\phi(\mathbf{v})\|_d^2)^2\rangle_\phi\\&=\left\langle\left(\sum_i v_i^2-\sum_{ij}c_iv_j\xi_i\cdot\xi_j\right)^2\right\rangle_\phi\\&=\left(\sum_i v_i^2\right)^2-2\left(\sum_i v_i^2\right)\sum_{jk}v_jv_k\langle\xi_j\cdot\xi_k\rangle_\phi\\&\quad+\sum_{ij}\sum_{kl}v_iv_jv_kv_l\langle(\xi_i\cdot\xi_j)(\xi_k\cdot\xi_l)\rangle_\phi\end{aligned} \tag{9}$$

As ξ_i and ξ_j are normalized isotropic and independent for $i\neq j$,, $\langle\xi_i\cdot\xi_j\rangle_\phi=\delta_{ij}$ must hold. For the same reasons,

$\langle(\xi_i \cdot \xi_j)(\xi_k \cdot \xi_l)\rangle_\phi \neq 0$ requires one of the following cases to hold: (*i*) $i=j\neq k=l$, (*ii*) $i=k\neq j=l$, (*iii*) $i=l\neq j=k$ or (*iv*) $i=j=k=l$. Cases (*i*) and (*iv*) yield expectation values of unity, the other two cases yield a value of $1/d$. Hence

$$\langle(\xi_i \cdot \xi_j)(\xi_k \cdot \xi_l)\rangle_\phi = \delta_{ij} \cdot \delta_{kl} + \frac{1}{d}(\delta_{il}\delta_{jk} + \delta_{ik}\delta_{jl}) \cdot (1-\delta_{ij}). \tag{10}$$

Substituting (10) into (9) gives

$$\sigma^2 = \frac{2}{d}\sum_{i\neq j} v_i^2 v_j^2 = \frac{2}{d}\left(\|\mathbf{v}\|_D^4 - \sum_{i=1}^{D} v_i^4\right), \tag{11}$$

proving the Theorem.

A neural realisation of such encodings may be straightforward. If the different contexts occur with roughly equal probabilities, then, association of a reproducible random excitation pattern ξ_i with each context is all that is needed.

Appendix II: Simulation Parameters

The simulations in Sect. 4 were based on (1) and (2). During each simulation, the radius $\sigma(t)$ of the adjustment zone was gradually decreased from an initial value σ_i to a final value σ_f according to

$$\sigma(t) = \sigma_i(\sigma_f/\sigma_i)^{t/t_{\max}}. \tag{12}$$

Here t counts the number of adaptation steps and the parameter settings were $\sigma_i=4$, $\sigma_f=0.5$, $t_{\max}=2000$ and $\varepsilon=0.8$ for both simulations.

References

Beach FA (ed) (1960) The neurophysiology of Lashley. Selected papers of KS Lashley. McGraw-Hill, New York

Caramazza A (1988) Some aspects of language processing revealed through the analysis of aquired aphasia: the lexical system. Ann Rev Neurosci 11:395–421

Collins AM, Loftus EF (1975) A spreading-activation theory of semantic processing. Psych Rev 82:407–428

Cottrell M, Fort JC (1986) A stochastic model of retinotopy: a self-organizing process. Biol Cybern 53:405–411

Essen D van (1985) Functional organization of primate visual cortex. In: Peters A, Jones EG (eds) Cerebral cortex, vol 3. Plenum Press, New York, pp 259–329

Goodglass H, Wingfield A, Hyde MR, Theurkauf JC (1986) Category specific dissociations in naming and recognition by aphasic patients. Cortex 22:87–102

Grossberg S (1976) Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors. Biol Cybern 23:121–134

Hart J, Berndt RS, Caramazza A (1985) Category-specific naming deficit following cerebral infarction. Nature 316:439–440

Jackson HJ (1878) On affections of speech from disease of the brain. Brain 1:304–330

Kertesz A (ed) (1983) Localization in neuropsychology. Academic Press, New York

Kohonen T (1982a) Self-organized formation of topologically correct feature maps. Biol Cybern 43:59–69

Kohonen T (1982b) Analysis of a simple self-organizing process. Biol Cybern 44:135–140

Kohonen T (1982c) Clustering, taxonomy and topological maps of patterns. Proceedings of the Sixth International Conference on Pattern Recognition. IEEE Computer Society Press, Silver Spring, pp 114–128

Kohonen T (1984) Self-organization and associative memory. Springer Series in Information Sciences, vol 8. Springer, Berlin Heidelberg New York

Kohonen T, Mäkisara K, Saramäki T (1984) Phonotopic maps – insightful representation of phonological features for speech recognition. Proceedings of the IEEE Seventh International Conference on Pattern Recognition. IEEE Computer Society, Montreal, pp 182–185

Malsburg C von der, Bienenstock E (1986) Statistical coding and short term synaptic plasticity. A scheme for knowledge representation in the brain. In: Bienenstock E, Fogelman-Soulié F, Weisbuch G (eds) Disordered systems and biological organization. NATO ASI Series, vol F20. Springer, Berlin Heidelberg New York

McCarthy RA, Warrington EK (1988) Evidence for modality-specific meaning systems in the brain. Nature 334:428–430

McKenna P, Warrington EK (1978) Category-specific naming preservation: a single case study. J Neurol Neurosurg Psychiatry 41:571–574

Miikkulainen R, Dyer MG (1988a) Forming global representations with extended backpropagation. Proceedings of the IEEE ICNN 88, San Diego, vol I. IEEE Computer Society Press, Silver Spring, pp 285–292

Miikkulainen R, Dyer MG (1988b) Encoding input/output representations in connectionist cognitive systems. In: Touretzky DS, Hinton GE, Sejnowski TJ (eds) Proceedings of the 1988 Connectionist Models Summer School, CMU. Morgan Kaufmann, Los Aeros, Calif

Nass MM, Cooper LN (1975) A theory for the development of feature detecting cells in visual cortex. Biol Cybern 19:1–18

Ojemann GA (1983) Brain organization for language from the perspective of electrical stimulation mapping. Behav Brain Sci 2:189–230

Perez R, Glass L, Shlaer RJ (1975) Development of specificity in the cat visual cortex. J Math Biol 1:275

Petersen SE, Fox PT, Posner MI, Mintun M, Raichle ME (1988) Positron emission tomographic studies of the cortical anatomy of single-word processing. Nature 331:585–589

Quillian MR (1968) Semantic memory. In: Minsky M (ed) Semantic information processing. MIT Press, Cambridge, Mass

Reale RA, Imig TH (1980) Tonotopic organization in auditory cortex of the cat. J Comp Neurol 192:265–291

Ritter H, Schulten K (1986) On the stationary state of Kohonen's self-organizing sensory mapping. Biol Cybern 54:99–106

Ritter H, Schulten K (1988) Kohonen's self-organizing maps: exploring their computational capabilities. Proceedings of the IEEE ICNN 88, San Diego, vol I. IEEE Computer Society Press, Silver Spring, pp 109–116

Ritter H, Schulten K (1989) Convergency properties of Kohonen's topology conserving maps: fluctuations, stability and dimension selection. Biol Cybern 60:59–71

Rolls ET (1984) Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. Hum Neurobiol 3:209–222

Rumelhart DE, McClelland JL (1984) Parallel distributed processing. MIT Press, Cambridge, Mass

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by backpropagating errors. Nature 323:533–536

Sejnowski T, Rosenberg CR (1987) Parallel networks that learn to pronounce English text. Complex Syst 1:145–168

Suga N, O'Neill WE (1979) Neural axis representing target range in the auditory cortex of the mustache bat. Science 206:351–353

Tunturi AR (1950) Physiological determination of the arrangement of the afferent connections to the middle ectosylvian auditory area in the dog. Am J Physiol 162:489–502

Tunturi AR (1952) The auditory cortex of the dog. Am J Physiol 168:712–717

Warrington EK (1975) The selective impairment of semantic memory. Q J Exp Psychol 27:635–657

Warrington EK, McCarthy RA (1983) Category specific access dysphasia. Brain 106:859–878

Warrington EK, McCarthy RA (1987) Categories of knowledge. Brain 110:1273–1296

Warrington EK, Shallice T (1984) Category-specific impairments. Brain 107:829–854

Yamadori A, Albert ML (1973) Word category aphasia. Cortex 9:112–125

Zeki S (1980) The representation of colours in the cerebral cortex. Nature 284:412–418

Received: April 3, 1989
Accepted: April 7, 1989

Prof. Dr. Teuvo Kohonen
Helsinki University of Technology
Laboratory of Computer and Information Science
Rakentajanaukio 2C
SF-02150 Espoo
Finland