Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Discrete combinatorial circuits emerging in neural networks: A mechanism for rules of grammar in the human brain?

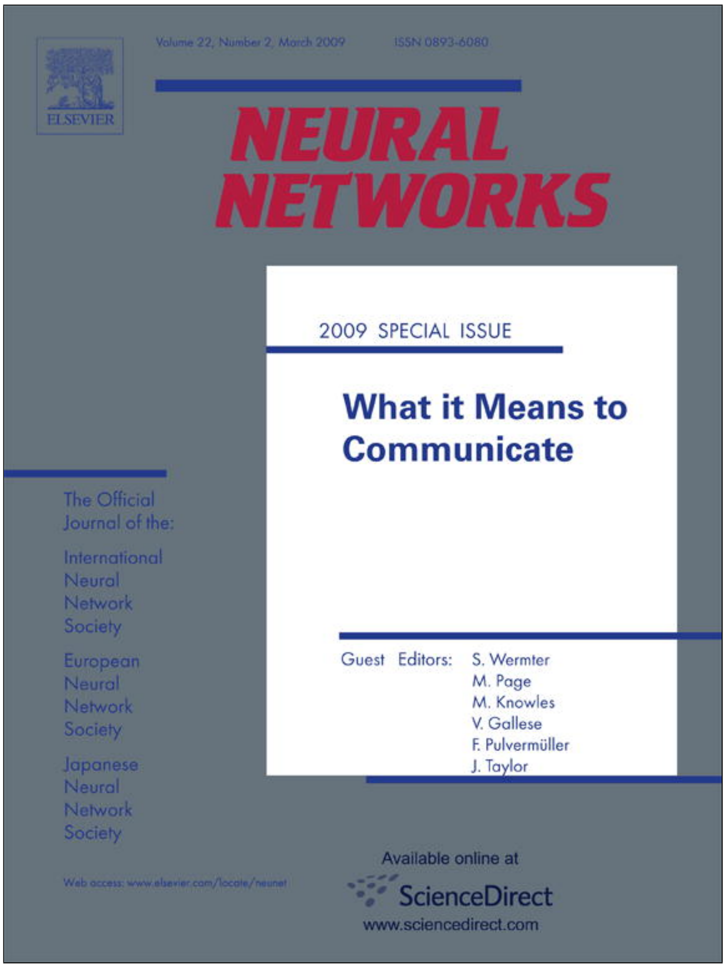
Friedemann Pulvermüller, Andreas Knoblauch

2009

Preprint:

This is an accepted article published in Neural Networks. The final authenticated version is available online at: https://doi.org/[DOI not available]

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



Volume 22, Number 2, March 2009 ISSN 0893-6080

ELSEVIER

NEURAL NETWORKS

2009 SPECIAL ISSUE

What it Means to Communicate

Guest Editors: S. Wermter
M. Page
M. Knowles
V. Gallese
F. Pulvermüller
J. Taylor

The Official Journal of the:
International Neural Network Society
European Neural Network Society
Japanese Neural Network Society

Web access: www.elsevier.com/locate/neunet

Available online at
ScienceDirect
www.sciencedirect.com

This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

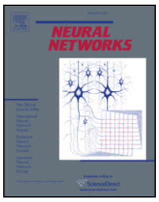
Author's personal copy

Neural Networks 22 (2009) 161–172

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



2009 Special Issue

Discrete combinatorial circuits emerging in neural networks: A mechanism for rules of grammar in the human brain?

Friedemann Pulvermüller [a,*], Andreas Knoblauch [a,b]

[a] Medical Research Council, Cognition and Brain Sciences Unit, Cambridge CB2 2EF, UK
[b] Honda Research Institute Europe, Carl-Legien-Strasse 30, 63073 Offenbach/M., Germany

ARTICLE INFO

Keywords:
Brain
Correlation
Grammar
Discrete combinatorial rule
Language
Linguistics
Neural network
Neuronal assembly
Syntax
Unsupervised learning

ABSTRACT

In neural network research on language, the existence of discrete combinatorial rule representations is commonly denied. Combinatorial capacity of networks and brains is rather attributed to probability mapping and pattern overlay. Here, we demonstrate that networks incorporating relevant features of neuroanatomical connectivity and neuronal function give rise to discrete neuronal circuits that store combinatorial information and exhibit a function similar to elementary rules of grammar. Key properties of these networks are rich auto- and hetero-associative connectivity, availability of sequence detectors similar to those found in a range of animals, and unsupervised Hebbian learning. Input of specific word sequences establishes sequence detectors in the network, and substitutions of words and larger string segments from one syntactic category, occurring in the context of elements of a second syntactic class, lead to binding between them into neuronal assemblies. Critically, these newly formed aggregates of sequence detectors now respond in a discrete generalizing fashion when members of specific substitution classes of string elements are combined with each other. The discrete combinatorial neuronal assemblies (DCNAs) even respond in the same way to learned strings and to word sequences that never appeared in the input but conform to a rule. We also show how combinatorial information interacts with information about functional and anatomical properties of the brain in the emergence of discrete neuronal circuits that may implement rules and discuss the model in the wider context of brain mechanism for syntax and grammar. Implications for the evolution of human language are discussed in closing.

© 2009 Elsevier Ltd. All rights reserved.

The $>$10,000 words of a language can be combined in abundant ways to yield a virtually unlimited number of possible strings and a still gigantic number of sentences that conform to the grammar of the language. Considering only sequences made up of up to six words, the number of possible strings is $>$ one septillion (10^{24}) and, assuming that only one out of 1000 (or even a million) of these possible strings is in fact grammatical, a still extraordinary number $> 10^{21}$ (10^{18}) of correct sentences results. As the average human has only about 2.5×10^9 s to live, it is clear that only a small fraction of the grammatical strings can be learned item by item. Still, a random sample of uncommon sentences presented to competent speakers will inevitably lead to very similar judgments about their grammaticality. Sentences and their underlying combinatorial principles must therefore be deduced, or generalized, from the limited input by each speaker in a similar fashion. Linguists (Chomsky, 1957; Harris, 1951; Pinker, 1984; Steedman, 2000) have argued that this combinatorial system operates on groups of discrete lexical elements (word stems and affixes) and relates them to discrete higher-order classes of string segments, the syntactic categories, which can, in turn, be the substrate of higher-order rules. The beauty of this approach lies in the fact that the single rule

$$c \rightarrow ab$$

(to be read as: "symbol c is rewritten as *a followed by b*") covers large classes of string segments, so that the number of combined strings c increases exponentially with the number of string parts a, b. Large numbers of grammatical strings can therefore be described by a small set of abstract rules for combining discrete string segments. These descriptions form the common ground of linguistic theories, although rules and the principles underlying them have been formulated in different ways by different syntacticians (Chomsky, 1957; Steedman, 2000; Tesnière, 1953). An important linguistic proposal therefore is that surface elements of a sentence are linked by way of abstract representations operating in a discrete fashion. Here, we ask how such abstract discrete combinatorial representations may emerge.

The idea of a language as a discrete combinatorial system has been questioned in the neural network literature. It is

[*] Corresponding address: MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 2EF, UK. Tel.: +44 1223 355294x770; fax: +44 1223 359062.
E-mail addresses: friedemann.pulvermuller@mrc-cbu.cam.ac.uk (F. Pulvermüller), Andreas.Knoblauch@honda-ri.de (A. Knoblauch).

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.01.009

Author's personal copy

well known that grammatical knowledge can be extracted from the statistical properties of sentences, their patterns and probabilities of co-occurrence and substitution of lexical elements and syntactic phrases (Brent, 1993; Elman, 1990; Hanson & Negishi, 2002). Now, it has been argued that neural networks can extract statistical properties of strings but do not include representations or processing components that can be likened to linguistic rules operating on defined classes of discrete lexical and syntactic units (Elman et al., 1996; Rumelhart & McClelland, 1987; Seidenberg & Elman, 1999). In contrast to algorithmic rule systems, the generalization capacities documented in networks are, according to established views, not related to rule formation but rather to a non-algorithmic probabilistic process arising from superposition of patterns (Elman et al., 1996; Seidenberg & Elman, 1999). Even if complex symbol strings with similar structure lead to the emergence of *similar* activation landscapes in hidden units (Elman, 1990; Hanson & Negishi, 2002), this does not imply that there is a neuronal entity in these networks that uniquely processes all strings of a certain syntactic type, a network equivalent of a discrete combinatorial rule. The gradual adjustment of weights in probabilistic, interactive and domain-general systems yielding generalization behavior of networks is still compatible with the statement "No rules operate in the processing of language" (McClelland & Patterson, 2002).

One critique of the linguistic rule-based approach to combinatorial processes has been that it is not grounded in brain mechanisms. Recent proposals (Pulvermüller, 2003a; Schnelle, 1996a, 1996b; van der Velde & de Kamps, 2006) have tried to close this gap by postulating neuronal entities for discrete grammatical representations, but have not yet successfully addressed the question of how linguistic representations may emerge and be bound to words in the learning human brain. Linguistic structure can be represented explicitly in a network, for example by way of neuronal representations of symbolic grammatical principles.(grammatical constraints, (Smolensky, 1990, 1999)). Emergence of discrete grammatical representations may, however, also be driven, at least in part, by associative learning of combinatorial information immanent to word strings. This present neurocomputational study asks whether combinatorial information can, in principle, lead to the emergence of discrete neuronal representations carrying the representational and processing role of linguistic rules.

Neural network approaches to serial order problems have suffered, in a manner similar to abstract linguistic work, from using network architectures that are not in very good agreement with known features of the central nervous system. In the neural network literature, it has frequently been argued that incorporating neuroanatomical and neurofunctional principles into artificial networks can be beneficial, both for theoretical and practical purposes (Garagnani, Wennekers, & Pulvermüller, 2007; O'Reilly, 2001; Palm, 1982; Sommer & Wennekers, 2003; Wennekers, Garagnani, & Pulvermüller, 2006; Wermter et al., 2004). One of the important features of the cortical network is its auto-associative character. Neurons that are close to each other, in the same hypercolumn, area or region, have a high probability of being connected by excitatory synapses (Braitenberg & Schüz, 1998; Young, Scannell, & Burns, 1995) and can therefore strongly link with each other if they become frequently active at the same time. However, in one type of layered network frequently used to simulate serial order processing, such links between adjacent neural elements of one compartment are usually indirect, through intervening layers (Elman et al., 1996), therefore making it impossible to *directly* connect neural elements when building higher-order representations that could implement rules. A major basis of learning in the neocortex is unsupervised synaptic modification driven by coincident or correlated neuronal activation, but most neural network simulations still use supervised error-driven learning, which is more difficult to relate to neocortical mechanisms (O'Reilly, 2001). In this present work, we demonstrate that brain-inspired networks of artificial neurons with strong auto-associative links can learn, by Hebbian learning, discrete neuronal representations that can function as a basis of syntactic rule application and generalization.

As rules and the problem of rule generalization are defined in different ways in the cognitive and linguistic literature, we here formulate the problem addressed by this work:

Rule generalization: Given that a, b are lexical categories and A_i, B_j lexical atoms

$$a = \{A_1, A_2, \ldots, A_i, \ldots, A_m\}$$
$$b = \{B_1, B_2, \ldots, B_j, \ldots, B_n\},$$

the rule that a sequence ab is acceptable can be generalized from a set of l encountered strings A_iB_j even if the input is sparse, i.e. $l \ll n \times m$. A critical question is whether the combinatorial information in the input leads to the emergence of abstract representations that (i) bind lexical categories and (ii) are functionally discrete and anatomically distinct from lexical representations. Note that the general question can be asked at different levels (choosing lexical items, phrases or whole sentences as constituent elements) and the argument about the development of higher-order representations stays the same.

This study employs networks, which, like the cortex, include auto- and hetero-associative connections and mechanisms for regulating excitation. (Knoblauch & Palm, 2001; Palm, 1980; Willshaw, Buneman, & Longuet-Higgins, 1969). These networks are pre-structured insofar as they have built-in neuronal devices for sequence detection. As the networks map coincident neuronal activation (Gutig, Aharonov, Rotter, & Sompolinsky, 2003; Hebb, 1949; Tsumoto, 1992) driven by the co-occurrence and substitution patterns of string segments in sentences, they "grow" putative network equivalents of discrete rules, which we will call *discrete combinatorial neuronal assemblies* (DCNAs) here. We illustrate the learning processes, especially the interaction between structural network properties and the combinatorial information about string part substitutions that give rise to putative rule representations, and give examples of the specificity of the networks' generalization behavior.

1. General network structure

We started with a set of n string segments, or "words", each implemented as a discrete neural unit here called an input unit. In fact, neuroscience evidence indicates that not single neurons, but, instead, large overlapping neuronal assemblies – or "word webs" – are the biological counterpart of words (Demonet, Thierry, & Cardebat, 2005; Garagnani, Wennekers, & Pulvermüller, 2008; Pulvermüller, 1999). We simulated words by single input units to keep the complexity of the simulation at a manageable level, assuming that any processes invoked by combinatorial information of single unit activation would also become manifest in simulations replacing single units by neuronal groups. Neuroscience support for the existence of discrete neuronal units for words and morphemes comes from neurophysiological research revealing qualitatively different brain responses to linguistic and nonlinguistic stimuli (Pulvermüller, 2001; Shtyrov, Pihko, & Pulvermüller, 2005). A word-related input unit was fully activated by the appearance of its corresponding word in the input and lost activity exponentially thereafter (for parameters, see Methods).

In addition to the word-related input units, which together formed the network's "lexicon", there was an array of n^2 neuronal units each responding maximally to one specific sequence of input unit activations. Sequence detectors specifically responding to input patterns have been found in a range of animals, including

Author's personal copy

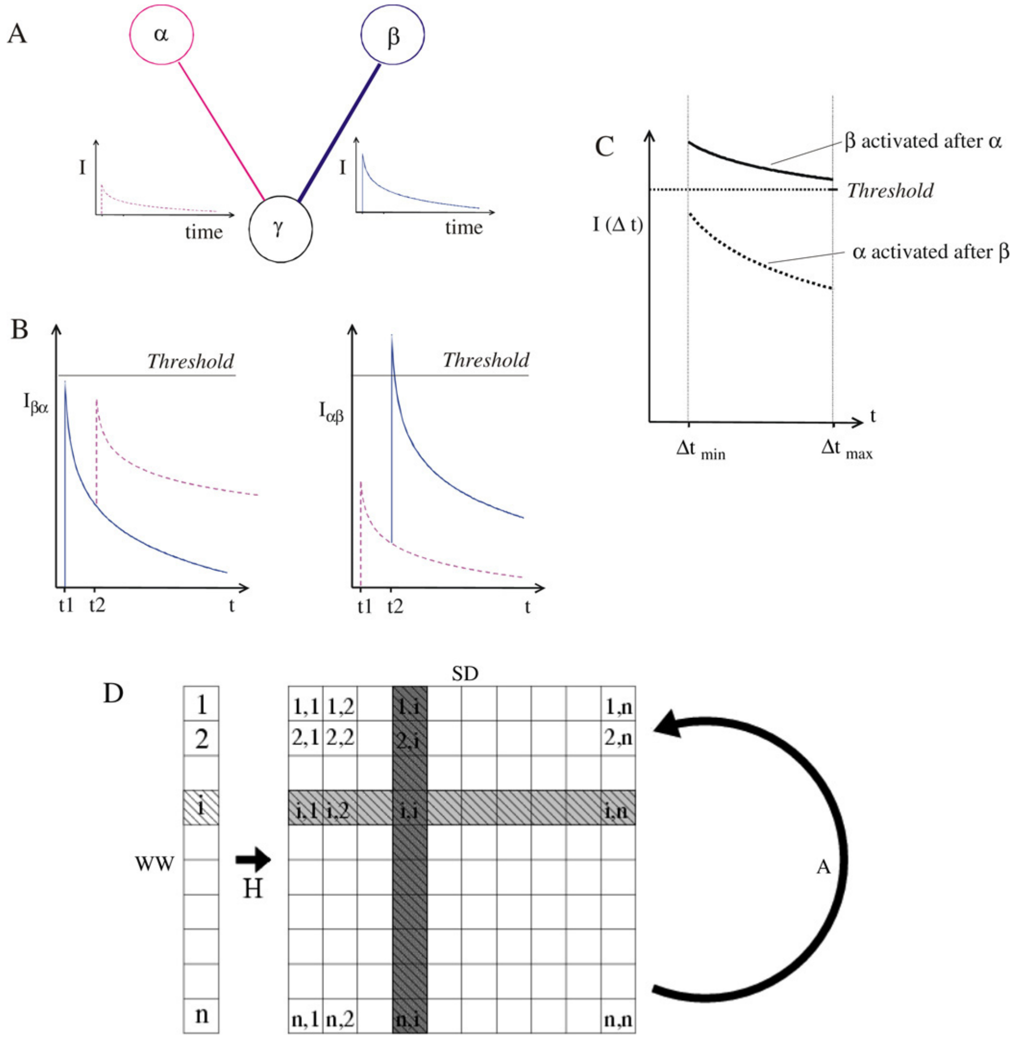


Fig. 1. Illustration of the mechanism for sequence detection and structure of the network model used in Simulation I and II. Parts A–C: Illustration of the mechanism of a sequence detector specialized for the string *AB*, which activates neuronal units $\alpha\beta$ in sequence. Figure part A: Additive inputs to the sequence detector through a weak (from α) and strong connection (from β). Figure part B: Stimulation with the critical sequence *AB* and the inverse sequence *BA* (upper diagram): Note the order selectivity of the SD unit, whose activity level only crosses the threshold for input *AB*. Figure part C: Maximal activation values of the sequence detector are shown as a function of the time delay between first and second word occurrence for both critical (*AB*) and inverse (*BA*) input sequence. Figure part D: Network model. The word-related input unit area of the network (labeled WW, as it includes "word webs") projects onto the sequence detector (SD) area via hetero-associative synaptic connections (H). Initially, the in-column connections from WW to area SDs are "strong" (dark gray), the in-row connections are "weak" (light gray), and all remaining connections, including the auto-associative connections (A), are "very weak" (see Materials and Methods). Thus, sequence detector SD_{ij} is selective for word i followed by word j. In addition, there are "very weak" auto-associative recurrent connections (A) within area SD.

fly, frog and monkey (Barlow, Hill, & Levick, 1964; Hubel, 1995; Reichardt & Varju, 1959), and it is therefore likely that sequence detectors exist in the human brain as well. The human perisylvian language cortex houses anatomically pre-structured neuronal wiring sufficient to form $>$100 sequence detectors for each word pair sequence possible in a language with $\sim 10^4$ lexical entries (Pulvermüller, 2003b). Sequence specificity was implemented by a pair of connections of each sequence detector, which provided it with input from two input units via a "weak" and a "strong" connection. Only if the strong input arrived at the sequence detector within a time interval Δt after the weak input, a non-linear interaction of input effects led to a powerful input to the sequence detector, which exceeded its threshold. Whereas the input sequence AB elicited a full sequence detector activation, the inverse sequence BA failed to activate this sequence detector (Figure 1 Knoblauch & Pulvermüller, 2005; Figure 1 Pulvermüller, 2003b). There was one specific elementary sequence detector for each of the n^2 possible sequential combinations of the n word-related input units in the networks (Fig. 1(D), see Methods for further details).[1] Because each section of cortex is known to have both rich hetero-associative connectivity with other cortical sites and even richer within-area auto-associative local connections (Braitenberg & Schüz, 1998; Young, Scannell, Burns, & Blakemore, 1994), we implemented global hetero-associative connections between input units and sequence detectors and auto-associative connections between each pair of sequence detectors. These connections all had "minimal" weights.

[1] Note that, for a vocabulary of 10^4 words, this implies that 10^8 sequence detectors are available for learning in the network at the onset of learning (Pulvermüller, 2003b). Neuroimaging evidence indicates that there are cortical areas, most likely in the periphery of the perisylvian cortex and certainly including part of Broca's area (Dapretto & Bookheimer, 1999), which are relevant for processing information about serial order and, potentially, grammatical rules. It appears that, within the perisylvian cortex, there are rich reciprocal connections between adjacent and even distant areas, thus providing a basis for the multiple weak links required for the large number of sequence detectors postulated (Catani, Jones, & Ffytche, 2005; Pandya & Yeterian, 1985; Young et al., 1995).

Author's personal copy

2. Methods

2.1. Neuron model

Each neuron was modeled as a simple leaky integrator unit. The membrane potential $x(t)$ follows the differential equation

$$\tau \cdot \mathrm{d}x_i(t)/\mathrm{d}t = -x_i(t) + \Sigma_j w_{ji} y_j(t),$$

where t is time, τ is the leak time constant, y_j is the output of a synaptically coupled neuron j, and w_{ji} is the (excitatory or inhibitory) strength of the synaptic coupling from neuron j to neuron i. The output y_i is a linear function of x_i with saturation, i.e., $y_i = x_i$ and $0 \le i \le 1$. For synaptic learning we used a Hebbian coincidence rule with constant decay rate,

$$\mathrm{d}w_{ij}(t)/\mathrm{d}t = -D + rf_{\mathrm{pre}}(y_i)f_{\mathrm{post}}(y_j),$$

where the synaptic weight $w_{ij}(t)$ was restricted to an interval $[0; w_{\max}]$. Here D is the decay rate, r is the learning rate, and f_{pre} and f_{post} are sigmoid functions of pre- and post-synaptic neural activity, respectively. For the simulations we used the Fermi function

$$f_{\mathrm{pre/post}}(y) = 1/(1 + \exp(-\beta(y - \Theta)))$$

with, typically, $\Theta = 0.8$, $\beta = 1000$, $D = 0.00001$, and $r = 1$. The model was implemented using the Felix++ simulation software (Knoblauch, 2003). Differential equations were numerically solved using the fourth-order Runge–Kutta method with a step size of 0.01 time units.

2.2. Network model

The network model consists of two connected arrays of neuronal elements or "areas". The first array included word-related input units (or "word webs", WW) and was considered the network analogue of the brain-internal neuronal assemblies processing lexical elements in the perisylvian cortex of the left hemisphere. This "lexicon area" comprised 20 simple leaky integrate units, one per word. The occurrence of a word in the input was simulated by an instantaneous increase of activity of its corresponding input unit followed by an exponential decay (leak time constant $\tau_{WW} = 10$) to mimic sustained excitation in active memory (Fuster, 1997, 2003).

The second array consisted of neuronal elements connected to pairs of word representations in the first (lexicon) area. Hetero-associative connections between areas was such that the second (grammar) array included 20 × 20 = 400 leaky integrate units (leak time constant $\tau_{SD} = 1$), one for each possible sequence of the 20 words represented in the lexicon area. The unit in the ith row and jth column of the array (Figs. 3 and 4) represents the sequence of word i followed by word j. At the start of the simulations, the neuronal units of the second array received a "weak" input from input unit i and a "strong" input from unit j (w parameters were set to 0.5 for weak and 1.0 for strong connections). These asymmetric connections make the neuronal element in the second array respond most strongly to the sequence ij of input unit activations (Figure 1c Knoblauch & Pulvermüller, 2005; Figure 1c Pulvermüller, 2003b). The neuronal elements in the second array therefore extract information about the serial order of words and can be considered elementary sequence detectors (SDs). Furthermore, in addition to the weak and strong connections of an SD to a pair of input units, there were initially "very weak" hetero-associative connections from the input units to the SD array ($w = 0.1$) and, importantly, auto-associative connections in the SD area of the network ($w = 0.1$). The SD_{ii} units on the diagonal (Figs. 3 and 4) received "very strong" input from the input units and can be thought of as functionally equivalent to word representations in the grammar area.

Anatomical realism also includes implementation of inhibitory circuits (Markram et al., 2004; O'Reilly, 2001). Following the previous work (Knoblauch & Palm, 2001; Palm, 1982), area SD included two inhibitory neuron populations, in addition to the excitatory SD cells. One population received only local input from the SDs, whereas neurons from the second inhibitory population, similarly to their excitatory neighbors, received also external inputs from the WW cells. The purpose of this architecture was to increase network capacity and stabilize overlapping cell assemblies (Aviel, Horn, & Abeles, 2005).

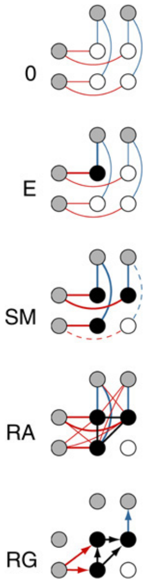


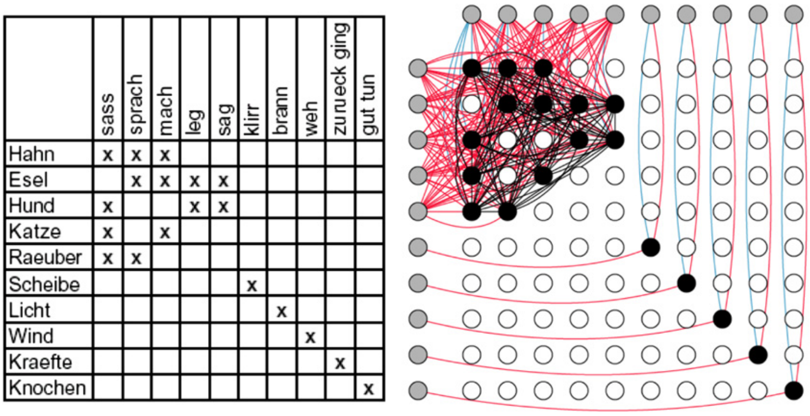
Fig. 2. Four overlapping stages of syntactic learning are illustrated schematically (from top to bottom): Starting stage (0), enabling of sequence detector (E), substitution mapping (SM), rule assemblage (RA), and rule generalization (RG). For explanation, see Results. Gray nodes correspond to word-related input units, segregated for first word elements (left) and second word elements (top). Central white or black nodes correspond to sequence detectors.

2.3. Learning and testing procedures

There were two training or learning phases: In the first training phase, we presented the word pairs from the training set while the effect of the auto-associative – or recurrent – connections in the sequence detector area remained weak (e.g., by balancing excitatory and inhibitory feedback). The training set of to-be-learned word strings corresponded to the non-zero table-entries of Figs. 3 and 4 (left parts). When a sequence was "presented", the first word stimulated its corresponding input unit for a single time step, followed by a delay of $\Delta t = 7$ time units, after which the subsequent word in the string stimulated its respective neuronal unit etc. In-between string "presentations", there was a break of 100 time units to allow network activity to go back to a resting level. In this regime, the SD units were largely unaffected by feedback and their activity and strengthening of their synaptic links was determined by bottom up, WW to SD, activation (as in Pulvermüller (2003b)). As indicated in Fig. 2, the functional result was a strengthening of the synaptic connections between sequence detectors and their respective pairs of input units i and j (enabling, E, and sequence mapping, SM). In addition, links to the units on the diagonal (SD_{ii}, SD_{jj}) were strengthened.

In the second learning phase, or "replay phase", the network was in a feedback dominated attractor regime, with auto-associative connections being fully effective. (Knoblauch & Palm, 2001), while the word sequences from the training set were presented again several times ($\Delta t = 0.1$). In this case, word sequences in the input activated a larger set of neuronal units in the SD area. This yielded the strengthening of (initially "very weak") hetero-associative connections from the activated input units to the set of activated SD units, and, importantly, to the strengthening

Author's personal copy



	sass	sprach	mach	leg	sag	klirr	brann	weh	zurueck ging	gut tun
Hahn	x	x	x							
Esel		x	x	x	x					
Hund	x			x	x					
Katze	x		x							
Raeuber	x	x								
Scheibe						x				
Licht							x			
Wind								x		
Kraefte									x	
Knochen										x

Fig. 3. Combinatorial information and network result in Simulation I. Diagram on the left: Matrix of substitutions and co-occurrences of nouns and verbs in a fairy tale (left). Data are given in a binary fashion. Diagram on the right: Network of word-related input units and sequence detectors that formed during learning of the set of strings. Grey circles in the periphery represent word-related input units corresponding to the words in the diagram on the left. The central matrix shows sequence detectors corresponding to word pair sequences. Filled black circles indicate enabled sequence detectors. Red and black lines display strengthened hetero- and auto-associative connections, respectively. At the upper left, an aggregate of enabled sequence detectors has formed, which represents a putative network correlate of a discrete combinatorial rule.

Table 1
Outcome of Simulation I using the network depicted on the right of Fig. 3 to determine verb (second string element) activation as a consequence of noun (first string element) input. *Top:* It can be seen that, after stimulation of first string segments (listed in column on the left), activation levels of second string segments (indicated at the top) were close to each other for second segments of learned substitution strings and for new, not previously learned, "substitution neighbors" (green array). Second elements of learned "unsubstitution strings", where first and second string elements always learned in only one sequence, were also activated when their corresponding first segments occurred, but no activation and therefore generalization to new strings could be observed. Bottom: Activation of learned and unlearned second string segments after first segment presentation to the network. Mean activation values, standard errors and minimum and maximum values for learned substitution and unsubstitution strings are listed alongside with those for unlearned substitution, unsubstitution, row and column neighbors. Note the similar results for learned and not-learned strings involving the substitution strings and the clear difference between learned sequences and not-learned neighbor sequences in the case of lack of substitutions.

	sass	sprach	mach	leg	sag	klirr	brann	weh	zurueck ging	gut tun
Hann	372	372	372	372	372	166	166	166	166	166
Esel	384	384	384	384	384	166	166	166	166	166
Hund	380	380	380	380	380	166	166	166	166	166
Katze	357	357	357	357	357	166	166	166	166	166
Raeuber	353	353	353	353	353	166	166	166	166	166
Scheibe	166	166	166	166	166	334	166	166	166	166
Licht	166	166	166	166	166	166	337	166	166	166
Wind	166	166	166	166	166	166	166	341	166	166
Kraefte	166	166	166	166	166	166	166	166	345	166
Knochen	166	166	166	166	166	166	166	166	166	349

		Learned		Not learned	
		V1 (subst)	V2 (not subst)	V1 (subst)	V2 (not subst)
N1 (subst)	mean (sde,n)	372.2 (3.4, $n = 14$)	$n = 0$	365.2 (4.0, $n = 11$)	166.2 (0.0, $n = 25$)
	min–max	352.8–383.9	–	352.8–383.9	166.2–166.2
N2 (not subst)	mean (sde,n)	$n = 0$	341.3 (3.1, $n = 25$)	166.2 (0.0, $n = 25$)	166.2 (0.0, $n = 20$)
	min–max	–	333.5–349.0	166.2–166.2	166.2–166.2

of auto-associative connections between the co-activated SD units *within* the sequence detector area (rule assemblage, RA, s. Fig. 2).

Finally, in the testing phase (where rules generalization, RG, was observed, *s.* Fig. 2) the network was again in the feedforward regime. Due to the learned hetero-associative connections, the network now was capable of generalizing word sequences that never occurred before. We tested this by stimulating individual input units $i = 1, \dots, 10$, corresponding to the first item of the learned word sequences, and examining the maximal activity state of the possible second word-related input units $j = 11, \dots, 20$ brought about by connections via the sequence detectors SD_{ij}. The results shown in Tables 1 and 2 therefore indicate the syntactic priming (Pickering & Branigan, 1999; Pulvermüller & Shtyrov, 2003) of word-related input unit *j* by word-related input unit *i*.

3. Results

3.1. General learning processes

Word sequences in the input activated their corresponding word-related input units. Because word-related input units retain activity for some time, this resulted in an overlapping pattern of activation and excitation of critical sequence detectors, which, in turn, brought about the modification of synaptic weights between the co-activated neural elements. A sequence detector was considered to be active when passing the activation threshold Θ of a bounded linear activation function. All hetero- and auto-associative connections were modified depending on their co-occurring pre- and post-synaptic excitation level, with weight change being proportional to the product of pre- and post-synaptic excitation level. Below, we illustrate activation and learning processes, which were observed in partly overlapping intervals, by referring to the paradigmatic example network shown in Fig. 2.

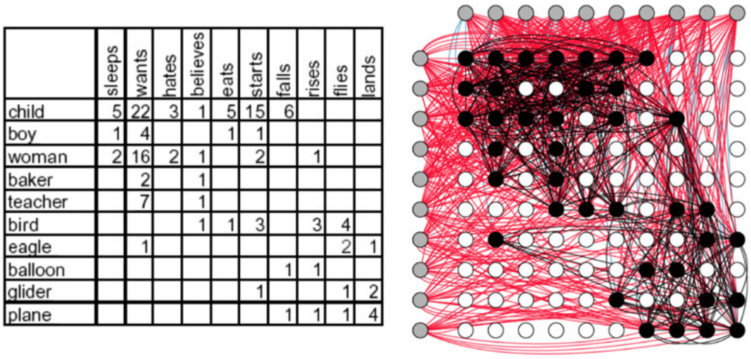
1. *Storage of word sequences:* Word-related input units were activated in a sequential manner, for example the word-related input units for string elements A_1 followed by B_1. This provides the optimal stimulus for the sequence detector A_1B_1 and stimulates competing sequence detectors less. Therefore, connections between this critical sequence detector and the stimulated word-related input units were strengthened

Author's personal copy

Table 2
Outcome of Simulation II using the network depicted on the right of Fig. 4 to determine verb (second string element) activation as a consequence of noun (first string element) input. The green squares at the upper left and the lower right indicate the set of first and second string elements frequently exchanged with each other, where discrete rule representations emerged. Rule generalization was specific as indicated by the, in average, high activation of not previously learned string representations when they were within a neighborhood of substituted strings. Activation of first elements spreads to second segments of learned substitution strings and unlearned substitution neighbors for each discrete assembly specifically, with very little cross-talk, indicating that two distinct discrete combinatorial neuronal assemblies were established. False positive deviations from the rule pattern (in orange) are due to the infrequent occurrence of atypical sequences ("eagle wants", "woman rises") or to cumulative effects of exceeding learning in both rows and columns of an elementary sequence detector. Misses indicate that some lexical items are not completely bound into the rule pattern. The table at the bottom shows average values, standard errors, minima, maxima, and number n of second string elements for which calculations were done: The learned strings were reproduced with similar activity values regardless of whether they were part of a rule pattern or not (left half of table). However, generalization of the rule pattern to new strings was specific to the neighborhood of the two rule patterns, respectively (left half of the table).

	sleeps	wants	hates	believes	eats	starts	falls	rises	flies	lands
Child	1161	2297	774	1929	1161	1925	1130	187	187	187
Boy	1150	2301	954	954	1150	1904	556	187	187	187
Woman	1159	1927	772	1926	1346	1545	187	1892	187	187
Baker	1264	2277	567	1900	1284	1664	187	187	187	187
Teacher	1324	2303	955	1919	940	1324	187	571	187	187
Bird	1334	1334	952	1902	1156	1922	187	1927	1534	425
Eagle	187	2312	187	1291	187	959	570	1730	1543	1157
Balloon	557	557	187	1315	936	1315	1137	1896	1324	557
Glider	187	563	187	187	1347	1933	563	1733	1546	1160
Plane	187	187	187	440	440	956	1146	1923	1538	1154

		Learned		Not learned	
		V1	V2	V1	V2
N1	mean (sde,n)	1657.2 (109.8)	1620.8 (214.8)	1156.8 (76.9)	236.6 (28.4)
	min–max	772.3–2303.0	1130.3–1927.0	567.0–1664.3	187.1–570.9
N2	mean (sde,n)	1790.0 (414.3)	1536.9 (93.6)	490.4 (99.3)	938.2 (167.4)
	min–max	1156.2–2311.9	1137.3–1932.9	187.1–1346.8	187.2–1733.3



	sleeps	wants	hates	believes	eats	starts	falls	rises	flies	lands
child	5	22	3	1	5	15	6			
boy	1	4			1	1				
woman	2	16	2	1		2		1		
baker		2		1						
teacher		7		1						
bird				1	1	3		3	4	
eagle		1							2	1
balloon							1	1		
glider						1			1	2
plane							1	1	1	4

Fig. 4. Combinatorial information and network result in Simulation II. Diagram on the left: Matrix of co-occurrences and substitutions of nouns and verbs obtained from the British National Corpus. Data are given in number of word pair occurrences in the 100-million-word corpus. Diagram on the right: Network of word-related input units and sequence detectors that formed during learning of the set of noun–verb strings (for explanation, see Fig. 3). At the top left and bottom right, two discrete combinatorial neuronal assemblies have formed.

specifically. As this happened repeatedly, the sequence detector is strongly bound to its word-related input units and is then called "enabled" (stage E in Fig. 2).

2. *Mapping of combinatorial information:* There is a new input sequence A_1B_2, which includes string elements already learned in one or more different sequences before. This input string therefore activates its critical sequence detector, and, in addition, partly activates the previously enabled sequence detectors connected to its word-related input units. If this applies to both first and second segments in a string, A_i and B_j, a set of sequence detectors, which represent the co-occurrence and mutual substitution of string segments in the past, will be attached to each word-related input unit by strengthening of hetero-associative links (stage SM in Fig. 2).
3. *Formation of discrete combinatorial representations:* Sequential activation of word-related input units that each have enabled sequence detectors attached already, leads to co-activation of these sequence detectors and therefore strengthening of their auto-associative connections in the sequence detector array and, thus, development of mutual functional links between them. (In contrast to the example in the figure, a more realistic scenario will lead to co-activation of multiple sequence detectors in each row and column.) In addition, there is further strengthening of hetero-associative connections between the activated elementary sequence detectors and their connected word-related input units. The enabled and co-activated sequence detectors, each of which represents one string whose segments have participated in substitutions with each other, will thus be bound into a discrete higher-order representation with strong specific functional links, a Discrete Combinatorial Neuronal Assembly (DCNA; stage RA in Fig. 2).
4. *Generalization to new strings:* Because of the strong internal links of the DCNA, future presentation of a learned sequence will not only activate its corresponding sequence detector, but the entirety of the DCNA. A new sequential activation of word-related input units that have never been active in sequence before but are both separately bound into the DCNA by way of their attached and enabled elementary sequence detectors is now possible: The DCNA produces sequential activation spreading between the neural counterparts of segments of a new string (illustration in part RG of Fig. 2).

Author's personal copy

In an auto-associative network consisting of sequence detectors, the learning of three strings A_1B_1, A_1B_2, A_2B_1 can therefore establish connections that provide a neural basis for the processing of a fourth new string A_2B_2, too (cf. Fig. 3). This basic form of generalization follows from associative learning in a pre-structured auto-associative network including sequence detectors. It remains to be shown that this type of learning can address aspects of syntactic rules and especially their specificity.

Below, we look at two data sets of string co-occurrences and substitutions and ask whether a brain-inspired neural network confronted with these data will build discrete combinatorial representations that could mechanistically explain aspects of rule generalization to new strings.

3.2. Simulation I: Emergence of a neuronal rule-equivalent

A network was confronted with a pattern of noun–verb combinations extracted from a German fairy tale. Small corpora such as fairy tales have proven useful in illustrating the working of neuronal networks, and previous research has indicated that, similar to mathematical approaches exploiting mutual information of words (Brent, 1993), neural networks can model the formation of lexical categories (Honkela, Pulkki, & Kohonen, 1995; Knoblauch & Pulvermüller, 2005). Fig. 2 displays the pattern of sequential co-occurrences of nouns and verbs within sentences of the text in matrix form. Nouns that substituted each other in the same verb context are represented by matrix columns with multiple filled circles in them. Verbs that replace each other in a given noun context are represented by matrix rows with multiple filled circles. The co-occurrence/substitution matrix shows a rectangle of dot accumulations at the upper left, indicating a high likelihood of substitutions of the animate nouns $N_a = \{$*Esel, Hund, Katze, Hahn, Räuber*$\}$ in the context of the action verbs $V_a = \{$*sass*, *sprach*, *machte*, *legte*, *sagte*$\}$ and, vice versa, a high probability of substitutions between the action verbs in the respective noun contexts. A rule $S \rightarrow N_aV_a$, which connects the head of the subject noun phrase with that of the verb phrase, implies all the observed co-occurrences of string segments along with additional ones that result from completion of the rule scheme and thus effective connections within the rectangle at the upper left. The lower right section of the matrix is sparsely populated with filled circles indicating absence of multiple substitutions, reflecting the fact that the remaining nouns and other heads of noun phrases were not multiply recombined with the remaining verbs.

Sequential activation of word-related input units according to the matrix established the corresponding sequence detectors in the network, so that each filled dot in the substitution matrix had a corresponding established sequence detector in the syntax network (cf. left and right diagram in Fig. 3). Whenever a string segment from an already learned string was replaced by a different string segment, the sequence detector specific to the new string was activated together with those of the previously learned string, thus leading to links between the sequence detectors of the related strings. Because there were multiple substitutions within both the N_a and V_a categories in the context of the respective other category, and substitutions always led to links between established sequence detectors (rule assemblage mechanism, RA), multiple links developed for the area of the network densely populated with established sequence detectors (upper left of diagrams in Fig. 3). These multiple links led to an increasingly stronger influence of the enabled sequence detectors for strings participating in substitutions of both of their segments and eventually to the formation of a functionally coherent neuronal assembly composed of elementary sequence detectors. Strings whose elements did not participate in multiple substitutions established their corresponding sequence detectors, but did not link them with others into a neuronal ensemble (lower right of diagrams in Fig. 3).

After learning, the network was tested for its processing of learned and novel ("unlearned") strings. Four types of strings were examined: Among the *learned strings*, there were those also participating in substitutions of both their first and second segments (*substitution strings*) and the rest, which did not participate in multiple substitutions (*unsubstitution strings*). Among the new or *unlearned strings*, some had their sequence detector fall into the high-exchange area of the substitution matrix, with established and assembly-bound sequence detectors in both the same row and column (*substitution neighbors*). Other unlearned strings had their sequence detector in a neighborhood of unsubstitution strings (*unsubstitution neighbors*).

Activation of first string elements participating in substitutions led to activation spreading to second segments of learned strings through the established sequence detectors of both types of previously learned sequences, substitution and unsubstitution strings (Table 1). It is noteworthy that the difference in activation between second elements of learned substitution strings and their not previously learned neighbors was only <2%. Also, the learned categories, substitution and unsubstitution strings, were primed equally well (9% difference). Critically, however, there was a profound difference between not previously learned neighbors of substitution strings and the neighbors of singular unsubstitution strings not participating in substitutions. Second elements of not previously learned neighbors of singular unsubstitution strings received only <50% of the activation of learned strings and, critically, also only <50% of the activation of not previously learned neighbors of substitution strings. This is clear evidence for discrete processing applying to both learned strings and new strings that would be covered by a rule operating on classes of lexical elements. The discrete neuronal representation built around the sequences participating in substitutions therefore binds substitution neighbors that have not been learned but would be covered by an abstract rule.

These results indicate that brain-inspired networks including sequence detectors, auto-associative connections and activation control learn discrete representations, DCNAs that store and represent combinatorial information. A DCNA equalizes learned and not previously learned strings that share string segments, thus leading to discrete behavior of the neural network. Learning generalizes from the learned material to new "substitution neighbor" strings whose composite parts had been involved in substitutions. Discrete and specific generalization required that items participate in both column- and row-wise substitutions.

3.3. Simulation II: Separating rules

The fairy tale example of simulation I only gave rise to one discrete representation in the neural network. To explore the specificity of algorithmic neuronal processes, it is important to store more than one combinatorial pattern in the same network. In this case, the task of the network is to store and separate representations of two combinatorial patterns. This task is relatively easy for combinatorial patterns operating on different sets of vocabularies; the task becomes most difficult if substitution patterns have overlapping vocabularies. As nouns and verbs can be subcategorised into fine-grained lexical subclasses that can overlap (e.g., nouns related to living and flying entities), we investigated whether a neural network could build distinct and discrete neuronal representations for the intersecting combinatorial patterns of realistic noun–verb substitutions.

A second simulation was therefore carried out to explore the simultaneous development of distinct combinatorial rules

Author's personal copy

in the same network type. Patterns of co-occurrence and substitution on the basis of the British National Corpus, a text database including 100-million-word tokens of English (see http://www.natcorp.ox.ac.uk, http://corpus.byu.edu/bnc). Semantic criteria were used for pre-selecting 10 nouns and 10 verbs of high frequencies: Nouns referred either to human subjects or to flying objects and verbs described actions and states, some of which semantically related to humans and others to flying.

The co-occurrence/substitution matrix listing nouns followed by verbs (Fig. 4, diagram on the left) showed two partly overlapping patterns of substitutions. Nouns referring to humans frequently occurred with action and internal state verbs, whereas animal and nonliving object names grouped with the verb set semantically related to flying. This dichotomy was by no means a strict one, as in a number of cases the flying object nouns co-occurred with internal state verbs ("eagle wants") or the human nouns with verbs otherwise mostly used in flying contexts ("woman rises"). However, usage in typical contexts predominated. We asked whether the network would build and separate neuronal representations that could be likened to syntactico-semantic subrules capturing the predominating combinatorial patterns.

The learning results (Fig. 4, diagram on the right) showed that frequently occurring word sequences established their corresponding elementary sequence detectors (black dots), thereby replicating the pairing of nouns and verbs observed in the corpus. In addition, the established sequence detectors were joined into two groups (black lines in upper left and lower right corner). When testing retrieval of second string elements, verbs, after activating the first string elements, nouns, categorial behavior of the network could again be demonstrated (Table 2, top): Second string segments of unlearned substitution neighbors in the territory of the first rule (green area in the upper left) were recruited when first elements covered by this rule occurred. This led to specific binding of nouns referring to humans with the group of action and internal state verbs. The same generalization processes were seen for the second combinatorial pattern: When first string elements belonging to the category of flying object nouns were presented, there was a general activation of word-related input units corresponding to verbs related to flying (green area in the lower right). In addition to this categorial behavior, stored exceptions could also be retrieved.

Similar to simulation I, these results demonstrate discrete combinatorial processes emerging in a brain-inspired network architecture. Furthermore, these processes distinguished semantic categories of strings on the basis of their substitution patterns. Established elementary sequence detectors were selectively joined together so that they selectively formed a neuronal assembly with those other sequence detectors with which they had frequently been co-active. These groups of elementary sequence detectors therefore became bound together selectively, which led to the formation of two DCNAs, each with numerous strong internal links (Fig. 4, diagram on the right). Strong connections from each of the assemblies to elementary sequence detectors outside were relatively sparse, as were strong connections between the two DCNAs. This demonstrates the specificity of combinatorial mechanisms developing in a network on the basis of string segment co-occurrence and substitution. The developing neuronal aggregates can differentiate and selectively generalize combinatorial patterns depending on the syntactic and semantic context of an incoming string segment.

4. General discussion

In networks incorporating neurophysiological and neuroanatomical features of the brain, the pattern of substitutions of segments of grammatical sentences led to the formation of neuronal aggregates that sequentially link specific classes of string segments to each other. Critically, these combinatorial neuronal ensembles also provide a binding link between constituents of unlearned substitution neighbor strings – precisely the novel string types to which a grammatical rule generalizes. The neuronal aggregates developing in the brain-inspired network are best described as *discrete* functional units that tend to act as a group, due to their strong internal connections, and are either activated by a specific input or not. They mediate the sequential relationship between abstract classes of string segments that are defined by their pattern of substitutions with other string segments. They are higher-order representations over and above the representations of the constituent elements they operate on. From a linguistic perspective, they therefore appear as possible neuronal equivalents of rules of grammar. Just like grammar rules, the discrete combinatorial neuronal assemblies are the basis of rule generalization from a limited set of input strings to a range of sentences, scaling exponentially with the number of lexical elements they are composed of. The discrete combinatorial neuronal assemblies, DCNAs, emerged, as a consequence of associative, Hebb-type learning of combinatorial information immanent to symbol strings and the networks' structural and functional features that mimic properties of brain connectivity and physiology. We will briefly discuss below 1/ the relationship between DNCAs and rules of syntax, 2/ network features critical for the emergence of higher-order discrete representations, 3/ the relationship between the present approach and previous symbolic and non-symbolic neural network models of syntax, 4/ relationships to statistical approaches to sentence structure and 5/ perspectives of the present approach.

4.1. DCNAs as a mechanistic basis of rules of grammar

A putative neuronal basis of rules of syntax and grammar is provided by aggregates of elementary sequence detectors, which are linked together due to the learning of strings and the substitutions of string segments between them (Simulation I). These neuronal aggregates can be specific enough to distinguish combinatorial patterns characteristic of fine-grained lexico-semantic sub-categories and rules operating on them (Simulation II).

The present simulations categorize lexical elements, nouns and verbs, into lexical categories and link these categories together. Although we took nouns and verbs as an example, links between other lexical categories (determiner–noun, adjective–noun, verb–preposition etc.) can be learned in the same way. The DNCAs can therefore be likened to syntactic rules of the form $c \rightarrow ab$, where a, b represent lexical categories and c the expansion of either a or b. This is an elementary problem in syntax and one may ask whether the emergent higher-order discrete representations are restricted to such linkage of lexical categories. In our view, the results demonstrate that combinatorial patterns of constituents can lead to the emergence of discrete higher-order representations binding, or merging, constituent categories. In this view, the general mechanism is applicable to constituent categories (a, b) of different kinds, single lexical elements (noun and verb, determiner and noun), but also larger constituents (phrases, sentences). Although this proposal needs to be worked out in more detail, the general mechanism of DNCA formation seems to explain the emergence of rules of syntax at higher levels in the very same way as it explains the emergence of syntactic category representations that bind lexical categories. Clearly, structural and functional properties of the brain-inspired networks along with associative learning are essential for these DCNAs to develop.

We simulate here the learning processes brought about by strings with multiple mutual substitutions between them and, critically, the emergence of putative neuronal correlates of binary rules as they form a basis of binding underlying syntactic tree

Author's personal copy

structures of many grammar theories (Steedman, 2000). As a range of symbolic approaches to the neural implementation of grammar are rooted in the rule concept (e.g., Chomsky (1965), Pinker (1994), Steedman (2000), Tesnière (1959) and van der Velde and de Kamps (2006)), the discrete neuronal binding mechanism may be useful for grounding syntax theories in neuronal circuits and synaptic learning. Critically, neurophysiological research has provided evidence that grammatical processing is reflected in early brain activation that appears to index the activation of discrete combinatorial representations in the brain (Friederici, Hahne, & Mecklinger, 1996; Friederici, Pfeifer, & Hahne, 1993; Hasting, Kotz, & Friederici, 2007; Neville, Nicol, Barss, Forster, & Garrett, 1991; Pulvermüller & Assadollahi, 2007; Shtyrov, Pulvermüller, Näätänen, & Ilmoniemi, 2003). The present research may therefore contribute to the integration of grammar theory with neuroscience research in support of the rule concept. More research is necessary to investigate the emergence and interaction of putative neuronal rule equivalents in larger networks operating on large vocabularies and corpora, along with the neurophysiological signs of rule processing.

In one approach to the neuronal basis of grammar (Knoblauch & Pulvermüller, 2005; Pulvermüller, 1993, 2002, 2003a), DCNAs (previously also called neuronal *sequence sets*) that operate on lexical categories are considered sufficient for representing and processing of simple sentences. For complex sentences, additional a priori mechanisms are postulated, including a neuronal pushdown store (Pulvermüller, 1993) and a mechanism for multiple activation of the same representation (Hayon, Abeles, & Lehmann, 2005; Pulvermüller, 2003a). This framework implies that syntactic rules covered by DNCAs directly operate on groups of lexical elements and that additional aspects of rule representations are linked to specific properties of human brain anatomy and function.

4.2. Network features critical for rule development

Why did the present network architecture yield discrete neuronal mechanisms functionally similar to discrete rules, whereas earlier search efforts for neural rule equivalents failed? The following features, which, as we have argued above, are all neurobiologically motivated, distinguish this present network model from the most common artificial neural networks used to address language questions:

1. rich reciprocal auto-associative connectivity,
2. built-in elementary sequence detectors specific to temporally ordered inputs,
3. unsupervised Hebbian learning,
4. sparse coding,
5. inhibitory circuits.

The beneficial effect of features 3–5 has been highlighted in earlier work (e.g., Markram et al. (2004), O'Reilly (2001), Palm and Sommer (1995), Wennekers et al. (2006) and Willshaw and Dayan (1990)). We wish to capitalize here on the importance of the first two features: Biological mechanisms for processing temporal order are prewired into the nervous systems of a range of animals, at different levels (Barlow et al., 1964; Hubel, 1995; Reichardt & Varju, 1959). There is therefore good reason to assume that the same type of mechanism is exploited in grammar processing. Rich auto-associative connectivity is evident especially for local cortical connections (Braitenberg & Schüz, 1998). Implementation of auto-associative connections in the grammar area of the network, which provides the mechanistic links between sequence detectors, is a precondition for yielded abstract discrete rule representations that form the basis of generalization.

The network structure mainly used in established neural models, for example the simple recurrent network, a three-layer perceptron with an additional memory layer (Elman, 1990), exhibit at least two features, which may hinder emergence of discrete representations: They avoid direct excitatory auto-associative connections within layers and "compress" the layer where the critical computations are performed. The "compressed" critical "hidden" layer includes fewer neural units than input or output layers, so that the coding immanent to it cannot be sparse. As we have argued, the cortex is essentially an auto-associative network structure with high connection probability between neighbors (Braitenberg & Schüz, 1998). Also, the primary cortices where input and output fibers originate include much less neurons than other cortical areas, arguing in favor of an expanded, rather than a compressed, hidden layer. We therefore suggest that the elementary neuroanatomical features incorporated in the present networks, especially the auto-associative connections and the relatively large number of prestructured neuronal units in the "grammar" area, are critical for what may be considered as neural rule formation.

4.3. Relationship to non-symbolic distributed network models

Having said this, it must be emphasized that research on distributed neural networks using versions of error-backpropagation learning have been extremely successful in addressing various variants of the serial order problem, including the processing of words in sentences. After Elman's seminal work, it is now well established that three-layered neural networks with an additional memory layer attached to the middle "hidden" layer, can learn sequences with syntactic structure and can generalize patterns to new symbol strings. A mechanism for this is the similarity of activation patterns in the hidden layer between symbols that appear regularly in similar contexts, as could be shown by Hierarchical Cluster Analysis and Linear Discriminant Analysis, LDA (Christiansen & Chater, 1999; Elman, 1990). Hanson and Negishi further showed, using Elman networks and LDA, that the states of a finite state grammar used to generate symbol strings are mapped onto similar activation patterns of the hidden layer (Hanson & Negishi, 2002). Critically, similar state-related activation patterns in the hidden layer were even achieved with different vocabularies, an important finding which the authors interpret as an index of neural rule formation. However, when new symbol strings generated by already learned finite state grammars were presented to the network, state-specific hidden layer activity was in-between previously encountered activity clusters, thus leaving it open whether the *same* rule or just a *similar*activity pattern was activated. These and similar studies have documented impressively

(i) that neural networks can learn behavioral patterns attributable to rules,
(ii) that they can generalize regularities to novel stings, and
(iii) that the networks reflect identical syntactic structure by similar functional states.

However, these results demonstrate the similarity of continuous neuronal activation patterns to structurally similar strings rather than rule equivalents at the mechanistic level of neuronal circuits. The similarities of hidden unit activity observed are still compatible with the statement that network performance is due to pattern overlay and similarity mapping of activation patterns, rather than to the formation of a qualitatively different, discrete neuronal entity processing variable symbols by the same mechanism (Elman et al., 1996). In other words, these results still allowed cognitive scientists to maintain, with reference to both networks and brains, that "No rules operate in the processing of language" (McClelland & Patterson, 2002).

The present work demonstrates a discrete mechanism – in terms of neuronal connections and network-anatomical changes

Author's personal copy

– applicable to at least some variants of rule formation and generalization in a specific kind of auto- and hetero-associative network employing Hebbian unsupervised learning, built-in prewired sequence detectors and auto-associative connectivity. This critical advance now makes it impossible to maintain a general statement about rulelessness of neural networks. The implication is that combinatorial rules of a discrete and abstract nature may form in the brain when symbol strings and substitutions between them are being encountered and produced. Thus, rules can, in a relevant sense, be learned. However, *tabular rasa* claims cannot be maintained either, as the networks were pre-structured. Neuronal rule acquisition in networks requires the exploitation of information built into the structure and function of the CNS.

4.4. *Relationship to symbolic network models*

Previous symbolic neural models showed that discrete neuronal representations can implement grammar mechanisms. The discrete neural blackboard architecture by van der Velde & de Kamps identifies linguistic entities with discrete neural representations and addresses challenging problems of the cognitive neuroscience of language, including the binding, multiple instantiation and structural representation problems (van der Velde & de Kamps, 2006). Similarly, the Neuronal Grammar framework used prewired discrete combinatorial neuronal sets and inhibitory feedback regulation mechanisms to simulate syntactic processing (Knoblauch & Pulvermüller, 2005; Pulvermüller, 2002, 2003a). These symbolic-neuronal approaches provide prewired circuitry that solves computational problems and generates predictions on neurodynamics during language processing and understanding. However, the principal question addressed by the present research, how the structural or discrete combinatorial representations may emerge, has not been answered by these models. Our present work demonstrates a mechanism for the formation of DCNAs and may therefore contribute to the foundation of symbolic network models of syntax and grammar.

An important contribution to the symbolic-neuronal modeling of grammar is Optimality Theory (Fodor & McLaughlin, 1990; Prince & Smolensky, 1997; Smolensky, 1999). Tensor networks (Hinton, 1990) represent the binding between vectors, each coding for an abstract entity by their vector product or tensor (Smolensky, 1990). Related approaches, vector symbolic architectures (Gayler, 2003, 2006), spatter codes (Kanerva, 1993) and holographic reduced representations (Plate, 2003), used modified tensor networks to develop symbolic models of binding between linguistic representations, including lexical items and roles, that connect to the level of distributed neuronal patterns. As the tensor product representation implies both excessive resource requirements and redundant representations, and multiple vector products can be assumed to represent the multiple links between symbols in a string, vector convolution, compression and other techniques were applied to obtain reduced representations. However, similar to other symbolic approaches to grammar, Optimality Theory and other tensor approaches to grammar do not provide a detailed mechanistic account of the learning processes underlying grammatical roles and rules. Rather, constraints related to principles of Universal Grammar were represented in a symbolic fashion by neural networks (Smolensky, 1999). Our present approach differs from variants of tensor networks although some parallels may be detected: Lexical items are represented by extremely sparse vectors of length m, n, and the link between them can be described in matrix form, by an m by n matrix of effective sequence detectors. In view of tensor networks, the formation of discrete neuronal rule representations in our simulations can be interpreted as one way to automatically reduce matrices of numerous sequence detector activations into the selection of one from a few DCNAs.

Hecht–Nielssen's confabulation theory presents another possibility to link symbolic representations of words and phrases to neuronal entities, neurons, modules and their interconnections (Hecht-Nielsen, 2005, 2007). He proposes architectures of heavily interconnected symbols represented locally in the cortex and thalamus. Different modules are active in parallel, whereas competition predominates within each local module. This architecture generates strings of symbols and words on the basis of previously learned "knowledge links". Similar to distributed non-symbolic processing approaches (see discussion above), the claim is that no rules exist in these architectures. Hecht–Nielsen emphasizes the importance of probabilistic relationships between non-adjacent words in a string and shows that his model makes use of them. We have previously emphasized that the well-known non-local relationships between syntactic objects are captured by sequence detectors and DCNAs (e.g., Pulvermüller (2002, 2003a)). Our simulations now show that DCNAs develop strong internal connections, which allows activity to reverberate and be maintained for longer periods of time (cf. Fuster (2003), Zipser, Kehoe, Littlewort, and Fuster (1993)), thus bridging the time-gap between constituents separate in time, but bound by syntactic links (Pulvermüller, 2003a). If DCNAs have formed on the basis of frequently recombined adjacent elements (for example, *birds fly*), these neuronal elements can, due to their prolonged reverberatory activity, also link linguistic elements when they are distant from each other in a sentence (*birds* with grey feathers grown over years *fly*).

4.5. *Syntactic-semantic categories, flexibility, and relationship to previous work in statistical language learning*

It is well known that statistical properties of word strings can be exploited to extract grammatical and syntactic information (Brent, 1993). Corpora tagged by Hidden Markov Models can be parsed automatically, thereby revealing information about grammatical features, e.g., sub-categorization features, of the lexical materials (Briscoe & Carroll, 1997). These procedures can classify nouns and verbs into fine-grained lexical classes, which are also characterized by specific semantic features (Lin, 1998). Similar results in lexico-semantic classification can be obtained from untagged text using neuronal network architectures, for example Elman networks (Elman, 1990) or Kohonen maps (Honkela et al., 1995), and mathematical techniques, for example independent component analysis (Honkela, Hyvärinen, & Väyrynen, 2005). It is therefore plausible that these methods exploit information immanent in the combination and recombination of string elements. In our simulations, we also observed the formation of sub-categories of lexical classes characterized by both syntactic and semantic features (e.g. nouns referring to living entities – N [+living], verbs referring to an action by a living being – V [+action], see Table 2). Syntactic-semantic categories emerged on the basis of unsupervised learning from a small text and a 10-million-word corpus. These emerged in a network structure inspired by the human language cortex, where lexical representations and processing devices for sequential information are side by side in a network with both rich auto- and hetero-associative connections. The fine-grained syntactic and semantic classes therefore seem to reflect combinatorial properties of the strings rather than properties of the algorithms or networks.

The lexical sub-categorization according to semantic criteria depends on the threshold at which DCNAs are operating. In Simulation II (Table 2), an activation threshold of 500 leads to separation of 2 rules (for living and flying objects), whereas at a threshold of 150, all nouns would provide the critical priming input for all verbs. This illustrates a potentially important point, the flexibility of category representations immanent to the present model. Whereas linguistic grammar theories usually define static categories, the auto-associative grammar area can flexibly merge

Author's personal copy

or separate word sub-categories depending on threshold. This feature could become relevant for explaining why rule selection sometimes depends on context, text form and conversation type.

Over and above lexical categorization, the present networks also provide a mechanism for syntactic binding between constituent classes, possibly including complementary higher-order syntactic categories. A two-step process of first tagging a corpus with lexical category labels and then building syntactic representations by binding together these categories could, therefore, be related to different network mechanisms. In the processing of strings, the lexical elements each first make contact with their DCNAs and, subsequently, the DCNAs fire and provide the syntactic binding between them (Pulvermüller, 2003a). Neurophysiological data indicate that both kinds of processes are extremely rapid, taking less than 200 ms (Pulvermüller & Shtyrov, 2003).

4.6. Rich local auto-associative connectivity: A critical feature in the evolution of language?

The necessity of strong auto-associative connectivity for building neuronal rules may be relevant in the context of theories of language evolution: A "grammar organ" in our brains may require a relatively high degree of auto-associative connectivity and connection probability of neurons within relevant cortical areas, especially left-perisylvian cortex. One way to implement this would be relatively large dendritic trees, or certain branches thereof, with particularly large numbers of synapses (Jacobs et al., 1993; Jacobs, Schall, & Scheibel, 1993). A different way to provide strong connections within left-perisylvian cortex is by way of long distance fibre bundles between left-frontal and left-temporal cortex (Catani et al., 2005; Saur et al., 2008; Rilling et al., 2008). The phylogenetic development towards optimising connectivity in left-perisylvian language cortex may therefore have been a key step in human language evolution also critical for setting up DCNA for syntactic processing.

4.7. Theoretical implications of this work

These results resolve a long-standing debate between neurocognitive modelers and linguists about the brain implementation of rules. By demonstrating the emergence of discrete neuronal aggregates in a brain-inspired network, we refute the claim that neural networks are, by necessity, rule-free, or implement "rules" of a type fundamentally different from the rules specified by linguistic algorithms (Elman et al., 1996; McClelland & Patterson, 2002). By defining and linking together complementary substitution classes of string segments, pre-structured associative networks form exactly the type of reduced algorithmic representation implied by linguistic rules. Neuronal correlates of at least one form of rules and rule generalization to new strings can result from associative learning, but this requires that information about specific structural and functional properties of the central nervous system be built into the network beforehand.

Acknowledgments

We thank two anonymous referees for their important comments on an earlier version and Günther Palm, Fermin Moscoso del Prado Martin, Yury Shtyrov, Mark Steedman and Thomas Wennekers for their help at different stages of this work. This work was supported by the Medical Research Council, UK (U1055.04.003.00001.01, 1055.04.003.00003.01) and by the European Community under the "Information Society Technologies Programme" (IST-2001-35282) and the "New and Emerging Science and Technology Programme" (NEST-2005-PATH-HUM contract 043374, NESTCOM). The authors do not have any conflict of interest related to this work.

References

Aviel, Y., Horn, D., & Abeles, M. (2005). Memory capacity of balanced networks. *Neural Computation*, *17*(3), 691–713.

Barlow, H. B., Hill, R. M., & Levick, W. R. (1964). Retinal ganglion cells responding selectively to direction and speed of image motion in the rabbit. *Journal of Physiology*, *173*, 377–407.

Braitenberg, V., & Schüz, A. (1998). *Cortex: Statistics and geometry of neuronal connectivity* (2 ed.). Berlin: Springer.

Brent, M. (1993). From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, *19*(3), 243–262.

Briscoe, E., & Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL conference on applied natural language processing* (pp. 1–9). *Vol. 5*.

Catani, M., Jones, D. K., & Ffytche, D. H. (2005). Perisylvian language networks of the human brain. *Annals of Neurology*, *57*(1), 8–16.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(2), 157–205.

Dapretto, M., & Bookheimer, S. Y. (1999). Form and content: Dissociating syntax and semantics in sentence comprehension. *Neuron*, *24*(2), 427–432.

Demonet, J. F., Thierry, G., & Cardebat, D. (2005). Renewal of the neurophysiology of language: Functional neuroimaging. *Physiological Reviews*, *85*(1), 49–95.

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

Elman, J. L., Bates, L., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Fodor, J., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, *35*(2), 183–204.

Friederici, A. D., Hahne, A., & Mecklinger, A. (1996). Temporal structure of syntactic parsing: Early and late event-related brain potential effects. *Journal of Experimental Psychology. Learning, Memory and Cognition*, *22*(5), 1219–1248.

Friederici, A. D., Pfeifer, E., & Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research*, *1*(3), 183–192.

Fuster, J. M. (1997). Network memory. *Trends in Neurosciences*, *20*(10), 451–459.

Fuster, J. M. (2003). *Cortex and mind: Unifying cognition*. Oxford: Oxford University Press.

Garagnani, M., Wennekers, T., & Pulvermüller, F. (2007). A neuronal model of the language cortex. *Neurocomputing*, *70*, 1914–1919.

Garagnani, M., Wennekers, T., & Pulvermüller, F. (2008). A neuroanatomically-grounded Hebbian learning model of attention-language interactions in the human brain. *European Journal of Neuroscience*, *27*(2), 492–513.

Gayler, R. W. (2003). Vector Symbolic Architectures answer Jackendoff's challenges for cognitive neuroscience. Paper presented at the Joint International Conference on Cognitive Science, University of New South Wales, Sydney, Australia.

Gayler, R. W. (2006). Vector symbolic architectures are a viable alternative for Jackendoff's challenges. *Behavioral and Brain Sciences*, *29*(1), 78-+.

Gutig, R., Aharonov, R., Rotter, S., & Sompolinsky, H. (2003). Learning input correlations through nonlinear temporally asymmetric Hebbian plasticity. *Journal of Neuroscience*, *23*(9), 3697–3714.

Hanson, S. J., & Negishi, M. (2002). On the emergence of rules in neural networks. *Neural Computation*, *14*(9), 2245–2268.

Harris, Z. S. (1951). *Structural linguistics*. Chicago: Chicago University Press.

Hasting, A. S., Kotz, S. A., & Friederici, A. D. (2007). Setting the stage for automatic syntax processing: The mismatch negativity as an indicator of syntactic priming. *Journal of Cognitive Neuroscience*, *19*(3), 386–400.

Hayon, G., Abeles, M., & Lehmann, D. (2005). A model for representing the dynamics of a system of synfire chains. *Journal of Computational Neuroscience*, *18*(1), 41–53.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: John Wiley.

Hecht-Nielsen, R. (2005). Cogent confabulation. *Neural Networks*, *18*(2), 111–115.

Hecht-Nielsen, R. (2007). *Confabulation theory*. New York: Springer-Verlag.

Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, *46*(1–2), 47–75.

Honkela, T., Hyvärinen, A., & Väyrynen, J. (2005). Emergence of Linguistic Features: Independent Component Analysis of Contexts. In A. Cangelosi (Ed.), *Proceedings of NCPW9, neural computation and psychology workshop* (in press).

Honkela, T., Pulkki, V., & Kohonen, T. (1995). Contextual relations of words in grimm tales analyzed by self-organizing map. proceedings of international conference on artificial neural networks. In F. Fogelman-Soulie' & P. Gallinari (Eds.), *Proceedings of the international conference on neural networks (ICANN)* (pp. 3–7).

Author's personal copy

Hubel, D. (1995). *Eye, brain, and vision* (2 ed.). New York: Scientific American Library.

Jacobs, B., Batal, H. A., Lynch, B., Ojemann, G., Ojemann, L. M., & Scheibel, A. B. (1993). Quantitative dendritic and spine analyses of speech cortices: A case study. *Brain and Language*, *44*, 239–253.

Jacobs, B., Schall, M., & Scheibel, A. B. (1993). A quantitative dendritic analysis of Wernicke's area in humans: II. gender, hemispheric, and environmental factors. *Journal of Comparative Neurology*, *327*, 97–111.

Kanerva, P. (1993). The spatter code for encoding concepts at many levels. In M. Marinaro, & P. G. Morasso (Eds.), *ICANN '94, Proceedings of the International Conference On Neural Networks* (pp. 226–229). New York: Springer Verlag.

Knoblauch, A. (2003). Synchronization and pattern separation in spiking associative memory and visual cortical areas, University of Ulm, Germany, Ulm.

Knoblauch, A., & Palm, G. (2001). Pattern separation and synchronization in spiking associative memories and visual areas. *Neural Networks*, *14*(6–7), 763–780.

Knoblauch, A., & Pulvermüller, F. (2005). Sequence detector networks and associative learning of grammatical categories. In S. Wermter, G. Palm, & M. Elshaw (Eds.), *Biomimetic neural learning for intelligent robots* (pp. 31–53). Berlin: Springer.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on computational linguistics and 36th annual meeting of the association for computational linguistics* (pp. 768–773) *Vol. 17*.

Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., & Wu, C. (2004). Interneurons of the neocortical inhibitory system. *Nature Reviews Neuroscience*, *5*(10), 793–807.

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Science*, *6*(11), 465–472.

Neville, H., Nicol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of Cognitive Neuroscience*, *3*, 151–165.

O'Reilly, R. C. (2001). Generalization in interactive networks: the benefits of inhibitory competition and Hebbian learning. *Neural Computation*, *13*(6), 1199–1241.

Palm, G. (1980). On associative memory. *Biological Cybernetics*, *36*(1), 19–31.

Palm, G. (1982). *Neural assemblies*. Berlin: Springer.

Palm, G., & Sommer, F. T. (1995). Associative data storage and retrieval in neural networks. In E. Domany, J. L. van Hemmen, & K. Schulten (Eds.), *Models of neural networks III* (pp. 79–118). New York: Springer Verlag.

Pandya, D. N., & Yeterian, E. H. (1985). Architecture and connections of cortical association areas. In A. Peters, & E. G. Jones (Eds.), *Cerebral cortex. Vol. 4. Association and auditory cortices* (pp. 3–61). London: Plenum Press.

Pickering, M. J., & Branigan, H. P. (1999). Syntactic priming in language production. *Trends in Cognitive Sciences*, *3*, 136–141.

Pinker, S. (1984). *Language, learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S. (1994). *The language instinct. How the mind creates language*. New York: Harper Collins Publishers.

Plate, T. A. (2003). *Holographic reduced representations: Distributed representations for cognitive structures*. Stanford, CA: CSLI Publications.

Prince, A., & Smolensky, P. (1997). Optimality: from neural networks to universal grammar. *Science*, *275*, 1604–1610.

Pulvermüller, F. (1993). On connecting syntax and the brain. In A. Aertsen (Ed.), *Brain theory—Spatio-temporal aspects of brain function* (pp. 131–145). New York: Elsevier.

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, *22*, 253–336.

Pulvermüller, F. (2001). Brain reflections of words and their meaning. *Trends in Cognitive Sciences*, *5*(12), 517–524.

Pulvermüller, F. (2002). A brain perspective on language mechanisms: From discrete neuronal ensembles to serial order. *Progress in Neurobiology*, *67*, 85–111.

Pulvermüller, F. (2003a). *The neuroscience of language*. Cambridge: Cambridge University Press.

Pulvermüller, F. (2003b). Sequence detectors as a basis of grammar in the brain. *Theory in Biosciences*, *122*, 87–103.

Pulvermüller, F., & Assadollahi, R. (2007). Grammar or serial order?: Discrete combinatorial brain mechanisms reflected by the syntactic mismatch negativity. *Journal of Cognitive Neuroscience*, *19*(6), 971–980.

Pulvermüller, F., & Shtyrov, Y. (2003). Automatic processing of grammar in the human brain as revealed by the mismatch negativity. *Neuroimage*, *20*, 159–172.

Reichardt, W., & Varju, D. (1959). Übertragungseigenschaften im Auswertesystem für das Bewegungssehen. *Zeitschrift für Naturforschung*, *14b*, 674–689.

Rilling, J. K., Glasser, M. F., Preuss, T. M., Ma, X., Zhao, T., Hu, X., et al. (2008). The evolution of the arcuate fasciculus revealed with comparative DTI. *Nature Neuroscience*, *11*(4), 426–428.

Rumelhart, D. E., & McClelland, J. L. (1987). Learning the past tense of English verbs: implicit rules or parallel distributed processing. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.

Saur, D., Kreher, B. W., Schnell, S., Kummerer, D., Kellmeyer, P., Vry, M. S., et al. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences USA*, *105*(46), 18035–18040.

Schnelle, H. (1996a). Approaches to computational brain theories of language—A review of recent proposals. *Theoretical Linguistics*, *22*, 49–104.

Schnelle, H. (1996b). *Die Natur der Sprache. Die Dynamik der Prozesse des Sprechens und Verstehens* (2 ed.). Berlin, New York: Walter de Gruyter.

Seidenberg, M. S., & Elman, J. L. (1999). Networks are not 'hidden rules'. *Trends in Cognitive Science*, *3*(8), 288–289.

Shtyrov, Y., Pihko, E., & Pulvermüller, F. (2005). Determinants of dominance: Is language laterality explained by physical or linguistic features of speech?. *Neuroimage*, *27*(1), 37–47.

Shtyrov, Y., Pulvermüller, F., Näätänen, R., & Ilmoniemi, R. J. (2003). Grammar processing outside the focus of attention: An MEG study. *Journal of Cognitive Neuroscience*, *15*(8), 1195–1206.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1–2), 159–216.

Smolensky, P. (1999). Grammar-based connectionist approaches to language. *Cognitive Science*, *23*(4), 589–613.

Sommer, F. T., & Wennekers, T. (2003). Models of distributed associative memory networks in the brain. *Theory in Biosciences*, *122*(1), 55–69.

Steedman, M. (2000). *The syntactic process*. Cambridge, MA: MIT Press.

Tesnière, L. (1953). *Esquisse d'une syntax structurale*. Paris: Klincksieck.

Tesnière, L. (1959). *Eléments de syntaxe structurale*. Paris: Klincksieck.

Tsumoto, T. (1992). Long-term potentiation and long-term depression in the neocortex. *Progress in Neurobiology*, *39*, 209–228.

van der Velde, F., & de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, *29*(1), 37–70. discussion 70-108.

Wennekers, T., Garagnani, M., & Pulvermüller, F. (2006). Language models based on Hebbian cell assemblies. *Journal de Physiologie (Paris)*, *100*, 16–30.

Wermter, S., Weber, C., Elshaw, M, Panchev, C., Erwin, H., & Pulvermüller, F. (2004). Towards multimodal neural network robot learning. *Robotics and Autonomous Systems*, *47*, 171–175.

Willshaw, D., & Dayan, P. (1990). Optimal plasticity from matrix memories: What goes up must come down. *Neural Computation*, *2*, 85–93.

Willshaw, D. J., Buneman, O. P., & Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature*, *222*(197), 960–962.

Young, M. P., Scannell, J. W., & Burns, G. (1995). *The analysis of cortical connectivity*. Heidelberg: Springer.

Young, M. P., Scannell, J. W., Burns, G., & Blakemore, C. (1994). Analysis of connectivity: neural systems in the cerebral cortex. *Review in Neuroscience*, *5*, 227–249.

Zipser, D., Kehoe, B., Littlewort, G., & Fuster, J. (1993). A spiking network model of short-term active memory. *Journal of Neuroscience*, *13*(8), 3406–3420.