Physica A 388 (2009) 732–746

Contents lists available at ScienceDirect

Physica A

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

A Tsallis' statistics based neural network model for novel word learning

Tarik Hadzibeganovic [a,b], Sergio A. Cannas [c,*,1]

[a] Cognitive Science Section, Department of Psychology, University of Graz, A-8010, Austria

[b] Cognitive Neuroscience Research Unit, Department of Psychiatry & Forensic Medicine, Faculty of Medicine, Hospital del Mar, Universitat Autònoma de Barcelona, 08003 Barcelona, Spain

[c] Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba, Ciudad Universitaria, 5000 Córdoba, Argentina

ARTICLE INFO

Article history:
Received 22 August 2008
Available online 5 November 2008

PACS:
87.19.lj
05.10.-a
87.19.lv
43.71.Hw

Keywords:
Novel word learning
Perceptron
Tsallis entropy
Nonextensive statistical mechanics

ABSTRACT

We invoke the Tsallis entropy formalism, a nonextensive entropy measure, to include some degree of non-locality in a neural network that is used for simulation of novel word learning in adults. A generalization of the gradient descent dynamics, realized via nonextensive cost functions, is used as a learning rule in a simple perceptron. The model is first investigated for general properties, and then tested against the empirical data, gathered from simple memorization experiments involving two populations of linguistically different subjects. Numerical solutions of the model equations corresponded to the measured performance states of human learners. In particular, we found that the memorization tasks were executed with rather small but population-specific amounts of nonextensivity, quantified by the entropic index q. Our findings raise the possibility of using entropic nonextensivity as a means of characterizing the degree of complexity of learning in both natural and artificial systems.

1. Introduction

As shown by Montemurro [1], Zipf–Mandelbrot law satisfies the first-order differential equation of the type $\frac{\mathrm{d}f}{\mathrm{d}s} = -\lambda f^q$, with its solutions asymptotically taking the form of pure power laws with decay exponent $1/(q-1)$. Further modification of the expression into $\frac{\mathrm{d}f}{\mathrm{d}s} = -\mu f^r - (\lambda - \mu) f^q$ (now with a new parameter and a new exponent), allows for the presence of two global regimes [2] characterized by the dominance of either exponent depending on the particular value of f. After this formalism was applied to experimental datasets on re-association in heme proteins [3], within the framework of non-extensive statistical mechanics [4,5], Tsallis suggested its potential usefulness in describing linguistic and neurocognitive phenomena (see e.g. Refs. [1,6]).

Ever since, there has been growing interest within a variety of fields [7–9], including biomedical engineering and computational neuroscience [10–15], in the non-extensive statistical mechanics based on Tsallis' generalized entropy $S_q = k\frac{1-\sum_i p_i^q}{q-1}$ ($\sum_i p_i = 1$; $q \in \mathcal{R}$), which in the limit of $q \to 1$ (and $k = k_b$) reduces to conventional Boltzmann–Gibbs entropy.

The parameter q that underpins the generalized entropy of Tsallis is linked to the underlying dynamics of the system and measures the amount of its non-extensivity. In statistical mechanics and thermodynamics, systems characterized by the property of nonextensivity are systems for which the entropy of the whole is different from the sum of the entropies of the respective parts. Such are usually the systems with interactions over long distances, with long memories of perturbations, and with very often fractal or multi-fractal structural properties. Since Tsallis' formalism is rooted on a non-extensive

[*] Corresponding author.
E-mail addresses: ta.hadzibeganovic@uni-graz.at (T. Hadzibeganovic), cannas@famaf.unc.edu.ar (S.A. Cannas).

[1] Member of CONICET, Argentina.

0378-4371/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2008.10.042

entropy, it appears to be a suitable candidate for describing systems with any kind of microscopic interactions (both short- and long-ranged). In other words, the generalized entropy of the whole is *greater* than the sum of the generalized entropies of the parts if $q < 1$ (superextensivity), whereas the generalized entropy of the system is *smaller* than the sum of the generalized entropies of the parts if $q > 1$ (subextensivity).

As noted by Hopfield [16] and then applied to attractor networks by Amit et al. [17], neural network models have direct analogies in statistical physics, where the investigated system consists of a large number of units each contributing individually to the overall, global dynamic behavior of the system. The characteristics of individual units represent the *microscopic* quantities that are usually not directly accessible to the observer. However, there are *macroscopic* quantities, defined by parameters that are fixed from the outside, such as the temperature $T = 1/\beta$ and the mean value of the total energy $\langle E\rangle$. The main aim of statistical physics is to provide a link between the microscopic and the macroscopic levels of an investigated system. An important development in this direction was Boltzmann's finding that the probability of occurrence for a given state $\{x\}$ depends on the energy $E(\{x\})$ of this state through the well-known Boltzmann–Gibbs distribution $P(\{x\}) = \frac{1}{Z}\exp[-\beta E(\{x\})]$, where Z is the normalization constant $Z = \sum_{\{x\}}\exp[-\beta E(\{x\})]$.

In the context of neural networks, statistical physics can be applied to study learning in the sense of a stochastic dynamical process of synaptic modification [18]. In this case, the dynamical variables $\{x\}$ represent synaptic couplings, while the error made by the network (with respect to the learning task for a given set of values of $\{x\}$) plays the role of the energy $E(\{x\})$. The usage of gradient descent dynamics as a synaptic modification procedure leads then to a stationary Boltzmann–Gibbs distribution for the synapses [18]. However, the gradient descent dynamics corresponds to a strictly *local* learning procedure, while non local learning dynamics may lead to a synaptic couplings distribution different from the Boltzmann–Gibbs one [19,20].

In the present study, we employ the *nonextensive* statistics theory of Tsallis to include some degree of non-locality in a two-level perceptron model. This q-generalized artificial neural network is further used to simulate the novel word learning process in two linguistically different populations of subjects. With respect to the computational simulations, our goal has been to investigate whether novel word learning occurs in an extensive or nonextensive manner. The core of the model is represented by a particular kind of *non-extensive cost function* that should induce a *non-local learning rule* in the neural network. In this sense, it is possible to think of non-extensivity as a particular form of globality or non-locality, at least in principle.

Alternatively, an implementation of a cost function that would induce a *local learning rule*, would cause the variation of the synapse between any two neurons at a given time to depend only on the instantaneous post-synaptic potentials (PSP) received by them, and *not* on the PSPs received by the rest of the neurons. It seems, therefore, more reasonable to assume that the full specification of a given neural representation depends on a non-local, distributed pattern of activity, emerging from the interaction of the constituents of whole neuronal ensembles, rather than from the activity in any particular, single neuron [21].

Representing linguistic knowledge by the distributed patterns of activity in neural networks has a long tradition in computational neuroscience [22–26]. More recently developed techniques for recording the simultaneous activity in populations of neuronal cells [27,28] provide substantial evidence for the non-local, distributed patterns hypothesis. Furthermore, there is growing evidence that neuronal populations distributed over distant cortical areas synchronize and work in synergy as *functional webs* during language processing [29].

Through the reciprocal links with the language areas, ventral visual stream, and the hippocampal formation, the anteroventral temporal cortex integrates a variety of aspects of letter-string information during processing such as visual, lexical, semantic and mnestic [30]. Novel word learning, which depends upon the structures in the medial temporal lobe, eventually becomes independent of these structures, relying more on other neocortical areas, such as those in temporal and temporo-parietal regions (see e.g. Ref. [31] for a review). Thus, the representation of the lexical information does not remain strictly limited to a particular area, but instead, it becomes distributed across different brain regions relevant for storing different aspects of information such as word meanings (temporal lobe) and word sounds (temporoparietal regions). For such reasons, it is necessary to investigate the effects of introducing non-local learning rules in neural network models for language learning, especially where they outperform purely local neural dynamics and better fit psychological and neuroscientific phenomenology.

A full understanding of the neural bases of learning also requires an accurate characterization of the learning processes as they occur in behavioral experiments. Learning is generally believed to include a gradual restructuring and strengthening of underlying connections between neural cells [32,33], which is behaviorally manifested in the gradual decrease of error after a series of repeated learning trials. Such asymptotic behavior is usually measured by using the learning curve, which is a plot of the magnitude or frequency of the response accuracy (or error) as a function of the number of learning trials. The agreement that is often found between the investigations of group-averaged learning behavior and the widely accepted neurobiological theories of individual animal learning, has caused many neuroscientists to use population-averaged learning curves for comparing the asymptotic learning behavior between differently treated groups of subjects (e.g. Refs. [34,35]). For a brief review, and an opposite viewpoint on this issue, see Ref. [36].

In the present study, two simple memorization tasks were carried out in two groups of learners with orthographically different native languages [37–39]. Subjects monitored 5 × 5 and 7 × 6 nonbinary letter matrices for a fixed number of seconds. Letter sequences in the matrix rows formed novel word items with very low summated type bigram frequencies (STBFs) and sparse orthographic neighborhoods (ONs). Learning was measured following each of the 10 stimulus exposures.

The sequences of letters were learned to a criterion of two consecutive perfect recalls. By plotting the individual error as a function of the number of successive experimental trials and then averaging the individual data over the examined populations, we obtained two distinct average learning curves — one for *shallow* (regular) and one for *deep* (irregular) orthography language speakers.

After obtaining empirical learning curves with human participants, we study both human and artificial novel word learning dynamics. To this end, we employ numerical simulations of a Langevin equation based two-level neural network model, with a non-extensive cost function. We show that the resulting learning algorithm with a non-local q-generalized learning rule can replicate the population-specific learning behavior to a high degree. The model further allows for the analysis of the population-specific learning efficiency, given the number of bits of random information an agent consumes as it proceeds in a learning task.

This paper is organized as follows. In Section 2, a neural network learning dynamics with a non-extensive cost function is introduced as the core of our language learning model. Model assumptions, parameters, and the general properties of the resulting dynamics are presented and then analyzed in Section 3. To test the model, we collected data from two simple memorization experiments conducted cross-linguistically with human participants (Section 4). The results from both experiments are presented in Section 5. Numerical simulations of the experimentally obtained human learning curves are given in Section 6. Finally, the discussion of the human and artificial neural network performance is presented in Section 7, followed by the conclusions and further research perspectives in Section 8.

2. A neural network model for novel word learning

A simple model of non local learning can be derived from an artificial neural network structure known as perceptron [40]. It consists of an input layer of N *binary* neurons $S_i = \pm 1$, and an output layer of N *analog* neurons (real variables) $\sigma_i \epsilon [-1, 1]$. Information is allowed to flow only from the input layer to the output layer, without backwards or lateral connections. The activation law for the output neurons is given by:

$$\sigma_i \left[\left\{ S_j \right\} \right] = \tanh \left[\frac{g}{\sqrt{N}} \sum_{j=1}^{N} J_{ij} S_j \right] \tag{1}$$

where the *gain* $g > 0$ is an arbitrary real number and $\left\{ J_{ij} \right\}$ are real valued *synaptic couplings*, whose values are restricted by the normalization [41]:

$$\sum_{i=1}^{N} J_{ij}^2 = N. \tag{2}$$

Let us consider the general case of learning (memorizing) a set of p binary patterns $\left\{ \xi_j^{\mu} \right\}$, with $j = 1, 2, \ldots, N$ and $\mu = 1, \ldots, p$, where $\xi_j^{\mu} = \pm 1$ are independent random variables with zero mean. A learning rule then consists of an algorithm which allows for an iterative modification of the synaptic couplings, such that these couplings evolve to a final configuration in which the network stores the input pattern associatively. In other words, starting from a random initial configuration of the synaptic couplings (subject to the constraint (2)), the algorithm leads to a set of final values that map every pattern $\left\{ \xi_j^{\mu} \right\}$, as well as any other state close enough, into an analog pattern $\left\{ \sigma_j \approx \xi_j^{\mu} \right\}$, $j = 1, 2, \ldots, N$, which is as similar as possible to $\{ \xi_j^{\mu} \}$ within the present constraints.

Every iteration step can be thought of as a discretized time step of a certain continuous-time synaptic modification process. In that case, the learning process can be described as a continuous-time *dynamics* for the synaptic couplings and modelled by a set of differential equations. Assuming that the learning process can be affected by a random environment, this leads to a *stochastic dynamics*, which is described by a set of stochastic differential equations. The above learning task can be carried out by different dynamics, the most widely used being the *gradient descent* method [40], ruled by the following set of Langevin equations:

$$\frac{\mathrm{d}J_{ij}}{\mathrm{d}t} = -\frac{\partial V}{\partial J_{ij}} + \eta_{ij}(t) \tag{3}$$

where $\eta_{ij}(t)$ are Gaussian uncorrelated random variables or white noise stochastic process, that is, they satisfy $\langle \eta_{ij}(t) \rangle = 0$ and $\langle \eta_{ij}(t) \eta_{i'j'}(t') \rangle = 2\ T\ \delta_{ii'} \delta_{jj'} \delta(t - t')$ [42]; here $\langle \ldots \rangle$ stands for an average over different realizations of the stochastic process and T is a parameter that gives the amplitude of the noise. The cost function V is some measure of the deviation of the network's output $\sigma_j(\{ \xi_j^{\mu} \})$ from the desired output $\left\{ \xi_j^{\mu} \right\}$, $\mu = 1, \ldots, p$. The cost function should be minimal whenever all the input-output pairs of patterns agree. In this way the learning process is mapped into an optimization problem. The dynamics of Eq. (3) lead, for long times, to an equilibrium Boltzmann–Gibbs probability distribution for the synaptic couplings [42]

$$p(\{J_{ij}\}) = \frac{1}{Z} \mathrm{e}^{-\beta V} \tag{4}$$

with

$$Z = \int \mathrm{d}\mu(\{J_{ij}\})\mathrm{e}^{-\beta V}; \tag{5}$$

$\beta \equiv 1/T$ and $\mathrm{d}\mu(\{J_{ij}\})$ is a normalized measure in the coupling space that takes into account the constraint (2). In physical contexts, T is the temperature of a thermal bath to which the system is exposed; we will adopt the same nomenclature here. At low temperatures $T \to 0$, the distribution (4) is strongly peaked at the minima of V and, in this way, the dynamics perform the desired optimization task.

The usual choices of V are *extensive* functions of the type $V = \sum_j V_j(J_{ij})$ where the sum runs over the output neurons, and V_j depends only on the synapses associated with the output neuron j. These kind of dynamics generate a *local* learning rule, *i.e.*, the updating of the coupling J_{ij} depends only on the local field at the output neuron j:

$$\frac{\mathrm{d}J_{ij}}{\mathrm{d}t} = -\frac{\partial V_j}{\partial J_{ij}} + \eta_{ij}(t). \tag{6}$$

In other words, the learning process at every output neuron is completely independent of the learning process at the rest of neurons, as if it were isolated, in a context-independent way.

Non-local learning dynamics can be obtained by a generalization of this method [20], in which the cost function V is replaced in Eq. (3) by a *non-extensive* function $\overline{V}$ defined by the map [19]

$$\overline{V} = \frac{1}{\beta(q-1)} \ln\left[1 + \beta(q-1)V\right] \tag{7}$$

where the index q is an arbitrary real number such that $q \geq 1$; $\overline{V}$ is a monotonically increasing function of V, and therefore it preserves its minima structure. Eq. (3) is then replaced by:

$$\frac{\mathrm{d}J_{ij}}{\mathrm{d}t} = -\frac{1}{1+\beta(q-1)V}\frac{\partial V}{\partial J_{ij}} + \eta_{ij}(t). \tag{8}$$

These dynamics induce *non-local* learning rules (compare Eqs. (6) and (8)). Consequently, one has to consider the full set of output neurons in the updating of every coupling, as can be inferred from the non-linear structure of Eq. (8). The parameter $q - 1$ measures the degree of non-locality of the learning process. These dynamics lead, for long times, to a generalized equilibrium probability distribution for the couplings J_{ij} of the form [19]:

$$p(\{J_{ij}\}) = \frac{\left[1 - \beta(1-q)V\right]^{1/(1-q)}}{Z_q} \tag{9}$$

with

$$Z_q = \int \mathrm{d}\mu(\{J_{ij}\})\left[1 - \beta(1-q)V\right]^{1/(1-q)}. \tag{10}$$

The probability distribution (9) can be derived by optimizing the Tsallis entropy [4]:

$$S_q\left[p(\{J_{ij}\})\right] = \frac{1}{q-1}\left\{1 - \int \mathrm{d}\mu(\{J_{ij}\})\left[p(\{J_{ij}\})\right]^q\right\} \tag{11}$$

with the constraint [5]:

$$\langle V\rangle_q \equiv \int \mathrm{d}\mu(\{J_{ij}\})\left[p(\{J_{ij}\})\right]^q V(\{J_{ij}\}) = \text{constant}. \tag{12}$$

Probability distributions derived from this entropy have recently been applied to a variety of statistical problems in cognitive neuroscience and related areas [6,10–15,20,43,44].

In the limit $q = 1$ the standard gradient descent Eq. (3) and the Boltzmann–Gibbs probability distribution Eq. (4) are recovered.

In order to simplify the analysis, let us first consider the case of learning a single pattern, i.e., $p = 1$. Let us denote $\xi_j \equiv \xi_j^1$, $j = 1, \ldots, N$. The performance of the learning dynamics can be quantified by the *quadratic error function*:

$$\varepsilon \equiv \frac{1}{4N}\sum_{j=1}^{N}\left(\sigma_j(\vec{\xi}) - \xi_j\right)^2 \tag{13}$$

with

$$\vec{J}_j \equiv \begin{pmatrix} J_{1j} \\ J_{2j} \\ \vdots \\ J_{Nj} \end{pmatrix} \qquad \vec{\xi} \equiv \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_N \end{pmatrix}. \tag{14}$$

In the limit $g \to \infty$ Eq. (13) gives the Hamming distance between the input and the output patterns, $\vec{\xi}$ and $\vec{\sigma}(\vec{\xi})$ respectively [20]. The relevant quantity to be compared with the experimental results is the time evolution of $\langle\langle\varepsilon\rangle\rangle$, where $\langle\langle\ldots\rangle\rangle$ denotes a double average over the initial conditions and the realizations of the noise.

One choice for the cost function V is given by:

$$V = \sum_j \left(1-\lambda_j\right)^2 \Theta\left(1-\lambda_j\right), \tag{15}$$

where the *stability parameters* λ_j are defined as $\lambda_j \equiv \xi_j \vec{J}_j.\vec{\xi}/\sqrt{N}$ [45] and $\Theta(x)$ is the Heaviside step function, i.e., $\Theta(x) = 1$ if $x \geq 0$ and $\Theta(x) = 0$ otherwise. Other choices of V give similar results, but this one has been shown to better reproduce the basic features of the learning curves obtained in the kind of experiments addressed in this work [20,44].

The error function (13) can be expressed in terms of the stability parameters as

$$\varepsilon = \frac{1}{4N}\sum_{j=1}^{N}\left[1+\tanh^2\left(g\lambda_j\right) - 2\,\tanh\left(g\lambda_j\right)\right], \tag{16}$$

where time evolution of the stability parameters is obtained from Eq. (8) as follows:

$$\frac{\mathrm{d}\lambda_j}{\mathrm{d}t} = -\frac{1}{1+\beta(q-1)V}\frac{\partial V_j}{\partial \lambda_j} + \sqrt{T}\eta'_j(t) \tag{17}$$

where

$$\eta'_j(t) \equiv \frac{1}{\sqrt{TN}}\sum_i \xi_i \eta_{ij}(t)$$

is also a white noise with $\langle\eta'_j(t)\rangle = 0$ and $\langle\eta'_j(t)\eta'_{j'}(t')\rangle = 2\delta_{jj'}\delta(t-t')$ and $V_j = \left(1-\lambda_j\right)^2 \Theta\left(1-\lambda_j\right)$. Starting from different initial random configurations for λ_j, $\langle\langle\varepsilon\rangle\rangle$ can be calculated as a function of time by solving Eq. (17) numerically.

The initial probability distribution for the synaptic couplings determines the value of $\langle\langle\varepsilon(0)\rangle\rangle$. If no *a priori* knowledge can be assumed, then the initial values of the $\vec{J}_j$ will be uniformly distributed in the N-dimensional hypersphere of radius $\sqrt{N}$; in that case it is easy to see that the initial values of the λ_j's are Gaussian distributed, with mean value zero and variance one. This, in turn, implies an initial error that is $1/4 \leq \langle\langle\varepsilon(0)\rangle\rangle \leq 1/2$ depending on the gain parameter g, where the lower bound $\langle\langle\varepsilon(0)\rangle\rangle = 1/4$ is obtained in the $g \to \infty$ case. However, in previous experiments [44], it was observed that many individuals memorized more than 50% of the information content of the displayed lettered matrix already in the first experimental trial, which considerably reduced the average initial error with respect to the expected one for an unbiased, randomly generated matrix of letters. In the present model, this corresponds to an initial error *smaller* than 1/4. Such "*a priori*" knowledge can be introduced in the model as a positive bias in the initial probability distribution of the stability parameters (or equivalently in the synaptic couplings) $\langle\lambda_j\rangle = a > 0$. This new parameter a measures the degree of deviation in the average previous knowledge with respect to a completely random pattern in the displayed stimulus. Both a and g parameters fix the initial error. In order to reduce the number of free parameters, we will set $g = 0.999$ in all the analysis (which accounts for a sharp distinction between $+1$ and -1 values in the binary outputs) and use only the parameter a to fix the initial error.

In the case of learning $p > 1$ patterns, for any pattern μ one must introduce a set of stability parameters $\{\lambda_j^\mu\}$, with $j = 1, \ldots, N$ and $\mu = 1, \ldots, p+1$. Instead of the set of N equations given by (17) we will have a set of $(p+1)\,N$ coupled Langevin equations. However, the learning process can be implemented one pattern per time (on-line learning). In other words, one can consider the problem of learning one single pattern, once the network has already learnt p previous ones. In the case of uncorrelated patterns, it can be shown that for $N \gg p$ these equations decouple and the stability parameter for each pattern evolves independently of the others [20]. So, in such a limit we recover Eq. (17).

3. General properties of the model

We now present general properties of the numerical solutions of Eq. (17), which can be solved by standard methods [46]. In Fig. 1 we show the typical behavior of $\langle\langle\varepsilon(t)\rangle\rangle$ (learning curves), for different values of q and typical values of N, a and T. We observe that learning is slower when q increases above unity, i.e., for non-local learning, as expected.

In Fig. 2 we show the influence of increasing N on the learning curves. We see that learning also becomes slower when the number of neurons increases, as expected. Moreover, after comparing Figs. 1 and 2, it seems that the effects of q and N are very similar. We can quantify the effect of both q and N on the learning process, by defining characteristic learning time τ, such that for $t > \tau$ the average learning error $\langle\langle\varepsilon(t)\rangle\rangle$ becomes smaller than 10% of the initial error (the value of 10% is arbitrary and the results do not depend on this choice).

In Fig. 3, we show $\tau(N, q)$ vs. the product $(q-1)N$ in a log–log plot for different values of q. We see that the learning time scales as $\tau(N, q) \sim f((q-1)N)$, where the function $f(x)$ behaves asymptotically as $f(x) \sim x^\alpha$ for large values of x. The linear fitting of Fig. 3 for large values of $(q-1)N$ gives an exponent $\alpha = 1.15 \pm 0.05$.

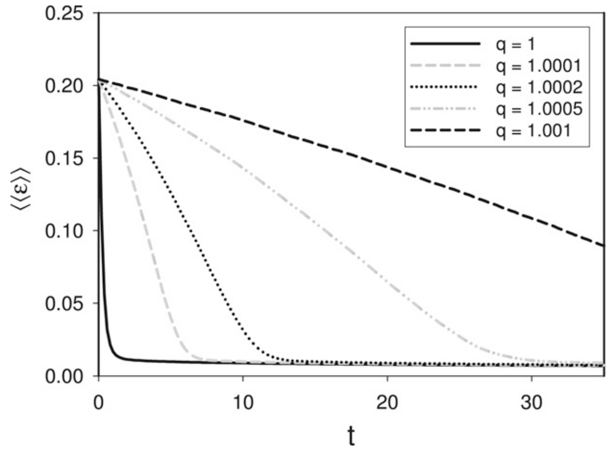

Fig. 1. Average error versus time (arbitrary units) for $N = 200$, $T = 0.001$, $a = 0.5$ and different values of q.

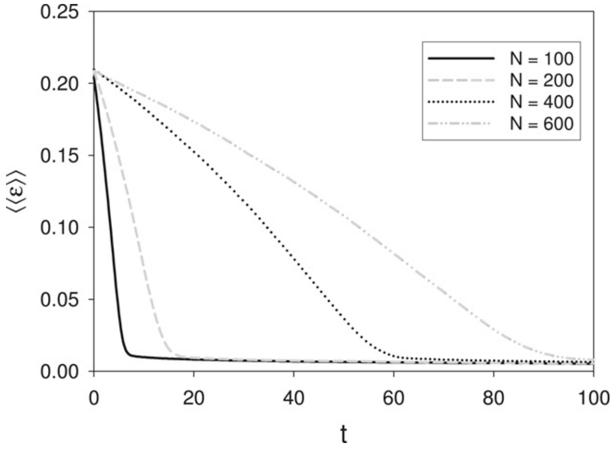

Fig. 2. Average error versus time (arbitrary units) for $q = 1.0050$, $T = 0.001$, $a = 0.5$ and different values of qN.

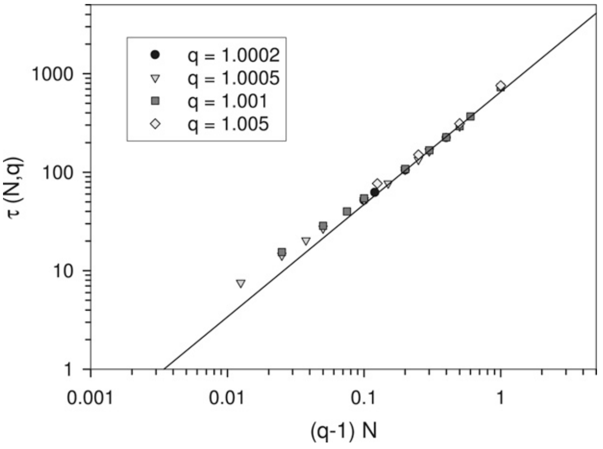

Fig. 3. Learning time τ versus $(q-1)N$ for different values of q ($T = 0.001$ and $a = 0.5$).

4. Method (Experiments 1 & 2)

To test the model against the empirical data, we conducted two simple memorization tasks in two orthographically different language populations. By monitoring and simulating the evolution of learning states in linguistically different subjects, a set of different q values may be assigned to the corresponding population-averaged learning curves, which may be of valuable practical importance when classifying the varying degrees of complexity of learning found among linguistically different populations.

4.1. Basic assumptions

The ability to repeat a nonword after only a single exposure has been shown to be a very good predictor of language learning ability in children [47,48] and adults [49]. However, early investigations of the effects of orthographic depth on language processing (e.g. Ref. [37]) showed that the letter-string stimuli are differently processed in orthographically different language populations, with particularly pronounced differences between *shallow* and *deep* orthography. When asked to read words and non-words, deep orthography language speakers took longer to begin reading each stimulus, and were much slower when generating pronunciations for novel letter-strings, compared to shallow orthography language speakers who were much faster in both word and nonword reading tasks [50]. The frequently found better recognition performance in a shallow orthography is usually attributed to the simple and isomorphic grapheme-to-phoneme connections, as opposed to ambiguous, many-to-one grapheme-to-phoneme relations in orthographically complex languages such as English.

Similarly to previous studies (e.g. Refs. [37,39]), we assume here that behaviorally, different accuracies and learning rates are expected for orthographically different adult linguistic populations, when novel word learning is considered, with speakers of a shallow orthography language performing significantly better than deep orthography language speakers. We expect a significant group-difference in the initiation of novel word representations in memory (due to different initial biases). Furthermore, the final mastering (i.e. memorization) of novel words should generally occur later in a deep orthography.

We attribute the expected differences between orthographies to their different abilities in establishing a system of mappings between the letters or graphemes of written novel words and the corresponding realizable phonemes. We assume that these mappings are faster established when the underlying orthographic system is more regular and consistent, ideally, when letters and respective sounds are in a one-to-one or bijective mapping-like relationship [38,51]. Thus, the ease of generating the pronunciation for novel words due to the faster letter-to-sound mapping ability might considerably enhance the learnability of novel word forms.

Previous studies showed that younger learners adopt different strategies when dealing with a deep alphabetic orthography, compared to the strategies that children use when learning a more transparent orthography [52]. Similarly, we ask whether the cross-linguistic differences in orthographic depth also affect the choice of strategies employed during novel word learning in adults. More importantly, we were interested in knowing to what extent the observable learning strategies can inform us about the more latent aspects of learning (e.g., about the nature of native language-based memorization skills), but also how the degree of efficiency of a given learning strategy can be quantified in a meaningful way. After increasing the information amount in the stimulus, as in Exp. 2, we analyze to what extent speakers of different language orthographies are able to reduce the degree of nonlocality of their strategic learning behavior in order to maintain constant, and thus, faster learning rates.

We note in passing that the aim of the present experiments was not to investigate the effects of orthographic depth by using a conventional ANOVA design. Various cross-linguistic aspects of the roles of orthography in language processing have already been widely investigated within such contexts (see e.g. Refs. [37–39]). Previous research assessed the effects of orthographic depth within a variety of alphabetic and non-alphabetic languages, but mainly in the domain of reading acquisition (see Ref. [53] and references therein). However, little is known about the influence of this variable on novel word learning in adults. We therefore aim to test for differences in rates of memorization of novel word stimuli among adult subjects speaking languages which vary significantly in orthographic complexity. Moreover, we aim to provide a novel, neural network based quantification and simulation of the observed cross-linguistic differences by using the model presented in Section 2.

With exception of only a few proto-attempts [54,55], this orthographically-based variation has not yet been submitted to a detailed computational investigation (see Ref. [39] for details). The present study makes further experimental comparison with an elaborated computational analysis in the domain of adult nonword learning, but now between the two writing systems placed more extremely on the dimension of orthographic depth, namely, a highly transparent, regular script (Croatian) and a deeply opaque, irregular English script.

4.2. Participants

A total of 103 students at the University of Graz participated in the experiments (7 participants were excluded due to their failure in following the exact procedural requirements). In the remaining sample, $N = 48$ were native speakers of English (as a deep orthography language), and the rest of $N = 48$ participants were native speakers of Croatian (as a shallow orthography language). All observers reported having normal or corrected-to-normal vision and never suffered from any type of a language-specific impairment according to self-report.

Table 1
Nonwords used in Exp. 1.

Stimuli	Neighbors	STBFs
GLNVS	0	962 (GL LN NV VS)
SLGNG	2	4271 (SL LG GN NG)
CRMST	2	5824 (CR RM MS ST)
BLMNK	2	2265 (BL LM MN NK)
STRLT	2	6871 (ST TR RL LT)
SLANG	3	10713 (SL LA AN NG)
START	4	11704 (ST TA AR RT)

Table 2
Nonwords used in Exp. 2.

Stimuli	Neighbors	STBFs
KVISRT	0	5728 (KV VI IS SR RT)
LPTRIM	0	6716 (LP PT TR RI IM)
SMNVKI	0	1405 (SM MN NV VK KI)
RDSPOB	0	3965 (RD DS SP PO OB)
KGHABI	0	4565 (KG GH HA AB BI)
OBGRAB	0	6237 (OB BG GR RA AB)
VKNABS	0	3396 (VK KN NA AB BS)

4.3. Apparatus & stimuli

Visual stimuli were presented on a 17'' CRT gamma-corrected color monitor (SRR 85 Hz) and controlled by a MATLAB program based on PSYCHTOOLBOX [56] using a PC (AMD Athlon 1.6 GHz) as the host computer. All letters in the display were black on a white background and were viewed binocularly in a darkened room at a normal viewing distance. For the 5 × 5 matrix type used in Exp. 1, the stimulus set consisted of uppercase consonant letters chosen randomly and independently from a total of 19 letters. The resulting information amount in the matrix was 106.2 bits. Vowels were omitted in order to minimize the possibility of observers interpreting the letter-strings as words. To illustrate the difference between the stimuli used in Exp. 1 and the real, existing English words and their properties, Table 1 shows nonword stimuli used in Exp. 1 and in addition, it lists two orthographic neighbors of the experimental stimuli SLGNG and STRLT, namely, SLANG and START. These words are characterized by the clearly larger neighborhood size and higher summated type bigram frequencies when compared to the nonword stimuli used in the experiment.

In the 7 × 6 matrix (Exp. 2), the number of alternative letters was 22 (vowels A, I, and O were added), so that the corresponding information amount was $42 * \log_2(22) = 187.3$ bits. The three added vowels were used to enhance the pronounceability of now longer strings of letters (6 per row). Some letters were not only excluded for reasons of information amount reduction, but this exclusion had additional psychological plausibility — all participants were told that O, when it occurred, was the letter "O" and was not considered a number (zero); Q was excluded from all trials because it is confusable with O, just as X is confusable with K or Y. Similarly, the letters F and W were excluded due to their feature overlaps with the letters E and V. On the other hand, X and Y do not exist as letters in some shallow orthography languages (for instance in Croatian).

WordGen 1.0 software [57] was used for stimulus generation and analysis. Tables 1 and 2 display (sub)lexical properties of the nonword stimuli used in Exps. 1 and 2 (the number of neighbors and the summated type bigram frequency, STBF, in English). The absolute STBF of the stimuli across languages is of course different, as expected by their different orthographic and phonological structures.

4.4. Procedure

All participants were tested individually. Before engaging in the tests, they were given a brief introduction on purposes of the study. They were familiarized with experimental items and the procedure until they felt that the protocol had been mastered. In Exp. 1, a 5 × 5 matrix of letters was presented during a fixed number of seconds (8 s). Then the screen became uniformly grey, to avoid masking effects that might have impaired the processing of the stimuli in the display. The participants were then required to reproduce the letters in an empty grid but were not allowed to exceed a total of 60 s in Exp.1 and 80 s in Exp. 2 for their reproductions. After a rest interval of 10 s following the written reproduction of items, the 5 × 5 matrix was shown again with the same stimulus parameters. The same procedure was repeated until all stimuli were reproduced correctly in two successive exhibitions. However, the matrix was never shown more than 10 times. In Exp. 2, the same basic experimental procedure was used as in Exp. 1, but now with 7 × 6 letter matrices and longer stimulus exposure times (10 s). Finally, following the last stimulus presentation, participants were asked to briefly describe how they proceeded to memorize and reproduce the presented stimuli.

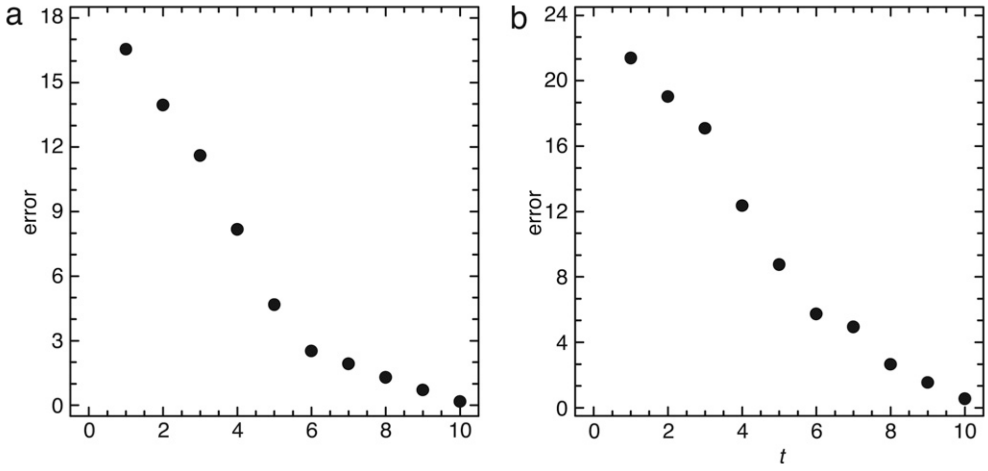

Fig. 4. Average error versus time (N of trials) for shallow (a) and deep (b) orthography observers associated with the learning of the 5 × 5 matrix (Exp. 1).

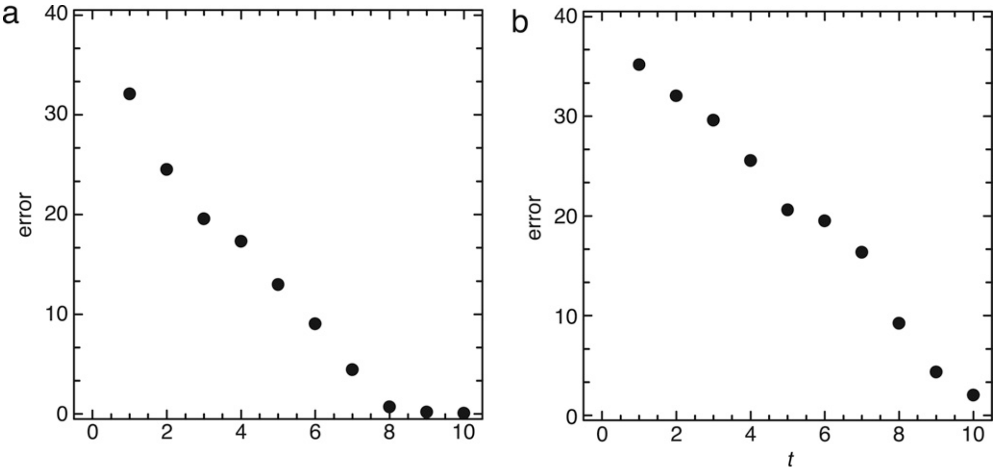

Fig. 5. Average error versus time (N of trials) for shallow (a) and deep (b) orthography observers associated with the learning of the 7 × 6 matrix (Exp. 2).

5. Results (Exps. 1 & 2)

Fig. 4 shows the average error associated with the learning of the 5 × 5 matrix of letters (Exp. 1), as a function of number of successive trials for both shallow (a) and deep orthography (b) observers.

In Exp. 1, the item recall improved with each successive stimulus presentation from 33.8% at 1st trial, to 53.6% at 3rd trial, and to 89.9% at 6th trial (shallow orthography), and from 14.5% at 1st trial, to 31.7% at 3rd trial, and to 77% at 6th trial (deep orthography).

On average, items were less well recalled in a deep orthography, than were in a shallow orthography group. A matched pairs t-test showed that the 4.83 letters (SD $=$ 2.42) mean group-difference in the initial trial was significant, $t(47) = 13.80$; $p < .001$, yielding the better average performance of shallow orthography language speakers. Thus, the learning curve associated with the performance of shallow orthography observers is shifted with respect to the other group (deep orthography) in the initial trial by 20.5 bits in the 48 participants data of Fig. 4. This difference between orthographies remained significant in all subsequent trials for the 5 × 5 matrix of Exp. 1. Detailed trial-by-trial comparisons further indicated that the learning generally progressed slower in a deep, relative to a shallow, orthography.

Fig. 5 shows the average error associated with learning the 7 × 6 matrix of letters (Exp. 2), as a function of number of trials, again for both orthographies.

In Exp. 2 (7 × 6 matrix of letters), the initial error was much larger (relative to Exp. 1) for both groups, as expected. Another matched pairs t-test revealed that the average initial error group-difference of 3.06 letters (SD $=$ 3.39) was significant, $t(31) = 5.09$; $p < .001$, with shallow orthography observers again performing significantly better than deep orthography observers. The learning curve associated with the performance of shallow orthography observers is shifted with respect to the other group (deep orthography) for the initial trial by 13.6 bits in the 32 participants data of Fig. 5. The average difference in performance between the two groups considerably increased after the first trial and remained significant in

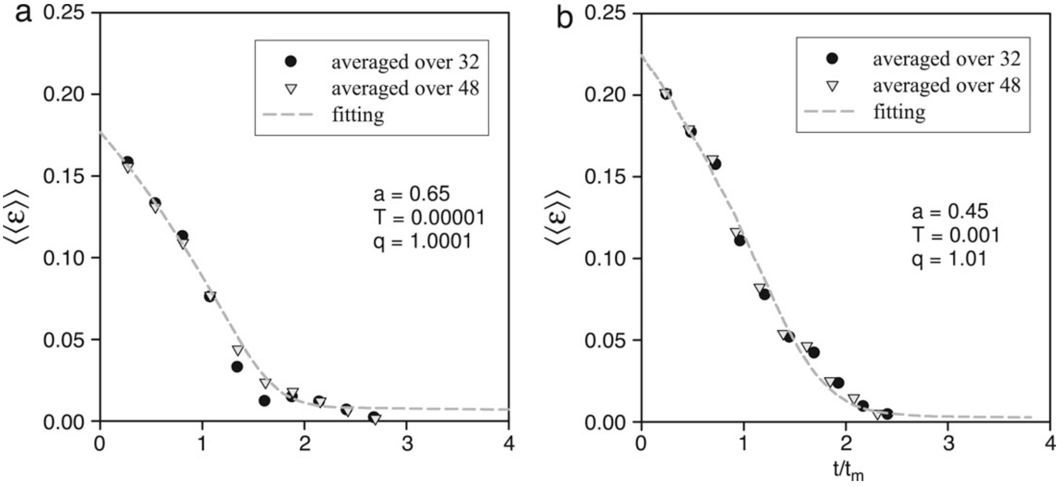

Fig. 6. Mean error versus rescaled time t_m. Shallow (a) and deep (b) orthography data fittings, Exp. 1.

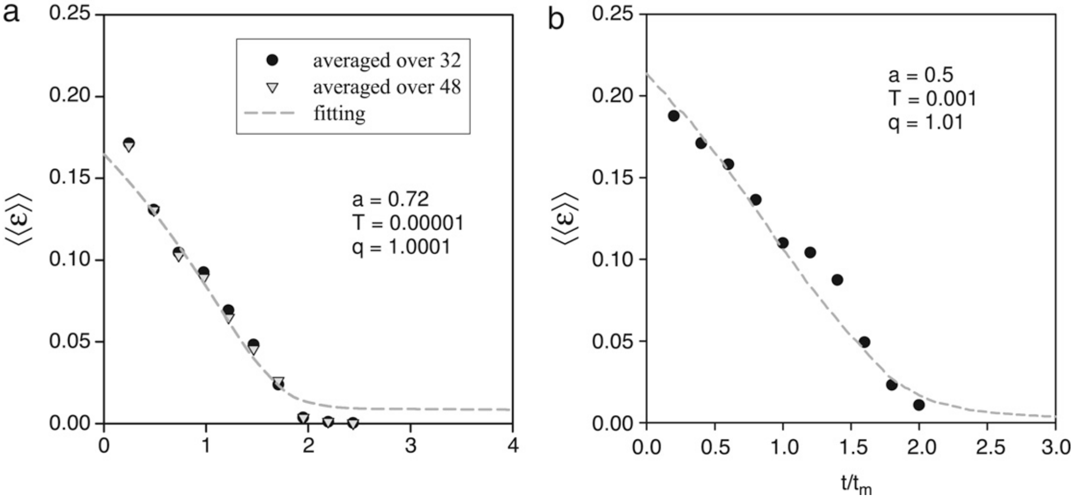

Fig. 7. Mean error versus rescaled time t_m. Shallow (a) and deep (b) orthography data fittings, Exp. 2.

the subsequent trials. The highest mean error group-difference of 11.90 letters (SD $= 6.19$) was observed in the seventh trial, $t(31) = 10.86$; $p < .001$. Learning was again found to be generally slower in the deep orthography group.

It was further observed that, as the information amount increased in Exp. 2, the measured average error after the initial trial decreased almost linearly with the learning time, i.e., with an increasing number of stimulus presentations, and this was found in both groups.

6. Numerical Simulations

Figs. 6 and 7 compare the rescaled empirical findings (from Exps. 1 & 2 for both groups of observers) with the theoretical learning curves $\langle\langle\varepsilon\rangle\rangle$ versus t/t_m for different values of a, T, and q, and for the cost function $V = \sum_j \left(1-\lambda_j\right)^2 \Theta\left(1-\lambda_j\right)$. Dots and triangles in the plots correspond to the experimental data, i.e., averaging over 32 and 48 individuals. Since the microscopic time scale of the experiment is not accessible, we needed to rescale the time appropriately for both the experimental and the neural network data, in order to make them comparable. We therefore defined, for every learning curve, a characteristic time t_m, as the time for which the mean error decays e.g. to half of its maximum value. This is expressed as $\langle\langle\varepsilon\rangle\rangle(t_m) = \frac{1}{2}\langle\langle\varepsilon\rangle\rangle(0)$ and we use t_m as the time unit.

Both a and g parameters fix the initial error. In order to reduce the number of free parameters, we set $g = 0.999$ throughout the analysis (which accounts for a sharp distinction between $+1$ and -1 values in the binary outputs) and use only the parameter a to fix the initial error. The values of T can be bounded by noting that the learning curves decay monotonically with t. Hence, the minimum value of the experimental curve can be taken as an upper bound for the

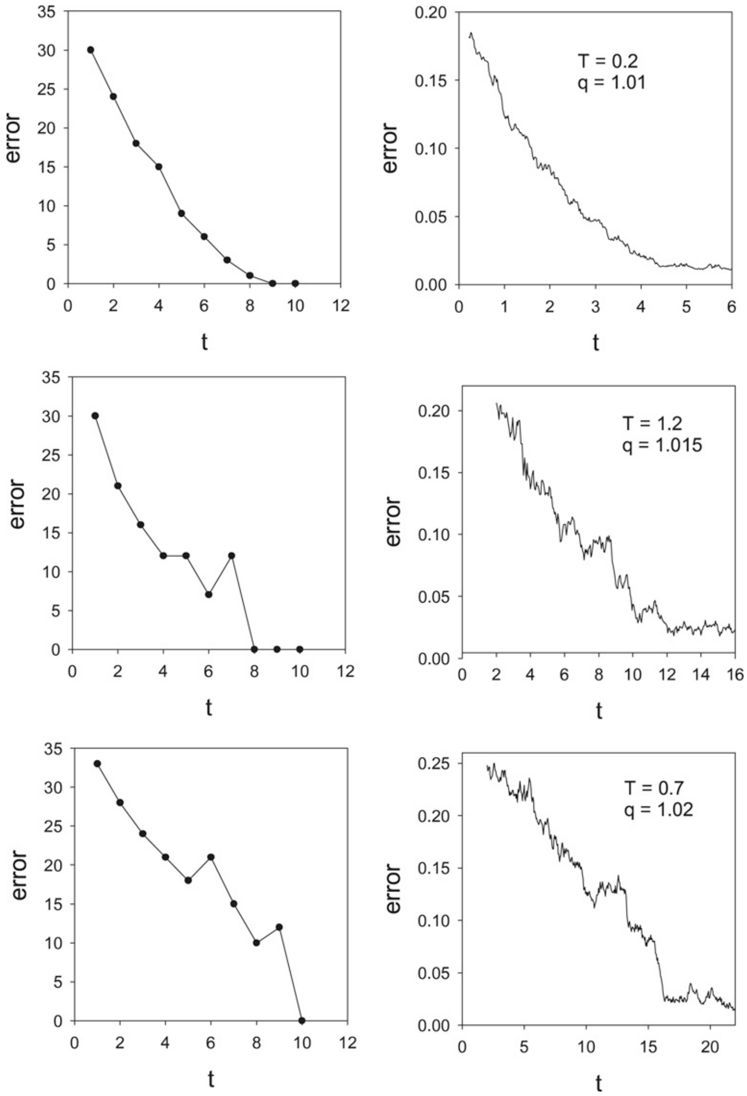

Fig. 8. Comparison between individual performances (left column; the abscissa is the number of learning trials) and perceptron performances (right column; the abscissa is the number of iterations) for different initial conditions and sequences of the random noise. The parameters of the perceptron were chosen as those that closely match the particular individual performance.

asymptotic value of $\langle\langle\varepsilon\rangle\rangle$ at $t \rightarrow \infty$. This value can be calculated numerically as a function of T, from the equilibrium distribution (9).

The best fitting learning curve for the shallow orthography observers' data (Exp. 1) was obtained with the gain parameter $g = 0.999$, $a = 0.65$, $T = 0.00001$, and $q = 1.0001$. The performance of deep orthography observers in Exp. 1 was best fitted by the model with the following parameter values: $g = 0.999$, $a = 0.45$, $T = 0.001$, and $q = 1.01$.

The empirical learning curve obtained from Exp. 2 with shallow orthography observers was best fitted by the model with the parameter values: $g = 0.999$, $a = 0.72$, $T = 0.00001$, and $q = 1.0001$. The corresponding learning curve for deep orthography observers' performance in Exp. 2 was best characterized by the following parameter values in the model: $g = 0.999$, $a = 0.5$, $T = 0.001$, and $q = 1.01$.

The smaller value of the a parameter (initial bias) in the English observers' case accounts for a larger initial error (due to smaller *pre-knowledge*); in mathematical terms, that means a smaller bias in the initial distribution of the synapses (more randomness). The larger value of T accounts for a larger relative error at the end of the tail, and the larger value of q accounts for a larger degree of non-locality (less efficient).

Finally, in Fig. 8 we show that not only the experimental average learning error, but also individual performances are qualitatively well reproduced by the perceptron model.

7. Discussion

Learning a list of n items usually takes a time $t(n)$, which increases more than proportionately with n. This was one of the first scientific observations made in studies of human memory. Consistently with that, our simulation results provide some insight into how learning strategies can contribute to modify the effective learning time. That is, our model predicts that the learning time scales as $\tau(N, q) \sim f((q-1)N)$, where the function $f(x)$ behaves asymptotically as $f(x) \sim x^{\alpha}$ for large values of x. With the linear fitting presented in Fig. 3 for large values of $(q-1)N$ we obtained an exponent $\alpha = 1.15 \pm 0.05$. This dependency of the learning time on q and N through the product $(q-1)N$ can be easily interpreted: to maintain the same learning rate with an increasing information amount (or number of bits) N in the learning task, the degree of non-locality of the learning behavior must be reduced by the same proportion.

Furthermore, similar effects of q and N on learning were observed. The learning curves exhibited, for all values of q, an exponential decay for long times. For short times, the qualitative behavior changed drastically when q departed from unity, showing a slow decay for $q > 1$, even for values near to one. Moreover, for $q = 1$, the learning curve was a convex function (positive curvature) for all t while, for $q > 1$, the curves were concave at short t, changing their curvature at intermediate times.

Thus, we found that learning was slower when q increased above unity, i.e., for non-local learning, as expected. These effects can easily be understood by looking at Eq. (8). For short times, the mean value of the cost function V is relatively high, and the non-local factor $[1 + \beta(q-1)V]^{-1}$ diminishes (for $q \neq 1$) the driven effect of the gradient term. As the system evolves $V \to 0$ and $1 + \beta(q-1)V \to 1$; therefore, for long times, the dynamics becomes the gradient descent one and $\langle\langle \varepsilon \rangle\rangle$ presents the $q = 1$ exponential decay.

To test the model introduced in Section 2, we employed two simple memorization tasks, showing that briefly displayed, novel letter-strings, are differently well recalled by English native speakers compared with students who learned Croatian as a first language. The results of Exps. 1 and 2 supported the assumed difference in the letter-string memorization ability between the investigated groups. Learning was thus found to progress considerably slower in a deep relative to a shallow orthography, and this slower learning dynamic was additionally characterized (quantified) in the modeling part of our study, by the higher values of the entropic index q, that were necessary for fitting the English learners' data.

Our results showed that, at short times, the learning dynamics induced by an extensive cost function (local learning rules) is very different from that induced by a nonextensive cost function (non-local learning rules). The excellent agreement with the experimental data is quite remarkable, especially if we consider the fact that the theoretical learning curves were reproduced with an extremely simplified model (a simple, two-level perceptron with a nonextensive cost function), far removed from the complexity of a multi-level non-linear dynamic system such as a human brain.

However, we showed that for the same range of tasks, humans performed similarly to non-extensive two-level perceptrons. It seems therefore that some aspects of the learning dynamics in biological systems could be independent of the detailed microscopic structure of the neural network, depending more on some overall properties. In this sense, some kind of universality probably exists in these processes, similarly to that appearing in critical phenomena, where the asymptotic behavior of most relevant variables can essentially be determined by a few macroscopic parameters [58]. Our results indicate that the non-locality of learning rules might represent one such universal property necessary for understanding learning processes in biological systems.

In the experiments with human subjects, the process of novel letter-string memorization was found to be of a different order of duration for orthographically dissimilar languages, and carried out with different degrees of nonextensivity. We attributed these expected differences in processing between orthographies to their different abilities in establishing a system of mappings between the letters of written novel words and the corresponding, realizable phoneme sequences. According to self-reports after both experiments, shallow orthography speakers could easily generate pronunciations for most letter-strings, quite differently from deep orthography learners who experienced difficulties when trying to pronounce the displayed stimuli, and to memorize them as whole word-like units.

In both populations, parallel to letter-string memorization, most individuals also gradually updated their strategic behavior. Thus, subjects learned both the strategy and the stimulus items within a given strategy. In fact, it is rather difficult, if not impossible, to differentiate between the processes of letter-string memorization and strategy acquisition. Indeed, on several occasions, it was observed that the error in later trials was larger than the error in the earlier trials, thereby contradicting the asymptotic behavior of the learning curve. As we could infer from individual self-reports, such rather unexpected behavior occurred whenever subjects tried to dedicate a considerable amount of time to the selection of a proper memorization strategy, rather than to the stimulus memorization process itself.

Reducing the number of bits of random information a learner consumes in the learning process while meaningfully extracting (classifying and/or clustering) the input information is the first sign of a strategic learning behavior. For instance, by means of chunking, many subjects decreased the total number of items held in memory by increasing the size of each single item. However, relying solely on one type of strategy, especially when information amount in the stimulus increases, is usually inefficient. The application of non-local learning strategies (which are generally less efficient) might then become necessary. In such cases, larger letter matrices need to be decomposed into several different subpart units on the basis of more global and non-linguistic factors.

As we learned from individual self-reports following the experiments, a variety of elementary shapes that could be extracted from the visual stimulus were considered for a non-local visual strategy development (e.g., extraction of letters in different locations forming a *T* or *L*, same letters aligned, columns forming visually better global shapes than rows, etc.). Thus, subjects had to look more for global patterns in the stimulus, because a local strategy did not pay off as the stimulus size increased.

We showed that less efficient and more global strategic behavior, as inferred from English learner's data, was characterized by considerably higher values of the non-extensive entropic index q. As q departed only slightly above unity, as in the case of Croatian shallow orthography learners, it corresponded to somewhat slower learning dynamics but also higher ability for devising efficient, particularly language-based, memorization strategies. Such slightly global learning behavior, characterized by only small departures of q from unity, seems to allow for introducing errors in some neurons in the network in the hope of decreasing the overall error. On the other hand, greater than just 'slight' departures of q from unity (as in the case of English learners), correlated with much slower learning dynamics but also with lower ability for devising efficient memorization strategies.

Our modeling results related to the Exp. 2 data in both orthographies, showed that the average errors corresponding to the first experimental trial performance were not successfully fitted by the model. We also noticed that this deviation of the simulated curve from the empirical one for the first trial was more pronounced in the deep orthography data than in the data obtained with shallow orthography observers. This difference between the first (not fitted) and the second (and subsequent) trial performance (well fitted) in Exp. 2, can be viewed as a state transition from a more local to a more global learning, which becomes necessary as the information amount in the stimulus increases, as was the case in Exp. 2. When facing larger letter matrices, individuals were able to remember only a few items in the first trial (the large initial error is systematic), which was most likely caused by memorizing individual letters in a local, letter-by-letter fashion. After the second trial, participants enter into a different regime, with a background of memorized letters and an initiation of a more global learning strategy.

It seems that this transition is characterized by an intellectual effort of a learner to analyze the lettered matrix in detail, find out what is peculiar to it, detect privileged letter sequences/rows, determine the relative positioning of the different parts or whole rows of letters in the display, etc. This state transition is also the period when the resources of long-term memory are mobilized, when particular criteria are provided with which the randomly generated matrix of letters can be characterized. With the successive increase of trials, the number of readily available criteria increases and the problem becomes one of selecting the strategy among a number of different available learning strategies.

Our findings support the orthographic depth hypothesis [37,39,53], and additionally address all of the three critical components of a theory of language acquisition [59]: We have analyzed the relevance of the initial state of knowledge (i.e., initial linguistic bias), the character of the strategic mechanisms responsible for the dynamics of novel word memorization, and the role of the novel input in the initiation of the learning process. We have seen that different initial biases lead to clearly different initial representations. Most importantly, we found that the adult novel word learning was highly influenced by prior linguistic experience. It seems that adult, linguistically well experienced learners, tend to detect and extract from the input stimulus what the structure of their native language tells them is the most relevant aspect of the randomly generated, novel linguistic pattern. In this sense, our results fit well with the Neural Commitment Theory [59–61] according to which any future language learning remains affected and directed by the initially acquired native-language information. As a consequence, previously established and frequently activated neural representations start to work as *perceptual magnets* [62] or *attractors* [63]. Thus, *a priori* linguistic knowledge alters a learner's perception of language [64], and we showed that this particular bias can enhance the learnability of novel, randomly generated letter-strings after only a few exposures.

8. Conclusions

In the present study, we proposed a Tsallis' statistics-based generalization of the gradient descent dynamics as a learning rule in a simple perceptron. The resulting model's equations were solved numerically for different values of an index q and for a particular cost function. We have demonstrated that even such a simple artificial device can learn unfamiliar novel words on the basis of a particular *nonextensive* dynamical equation, and that this device can mimic human and, moreover, population-specific learning behavior to a high degree.

When applied to experiments in which shallow and deep orthography language speaking adults were asked to memorize unfamiliar letter-strings, the model was able to account for many aspects of the experimental results, including the time-course and outcome of the learning, but also how it varies as a function of language orthography, and moreover, as a function of the initial linguistic knowledge of tested subjects.

The differences in strategic development found in our study can be understood in terms of structural differences in the relationship between orthographic and phonological information between the two analyzed languages. For instance, the relationships between graphemes and phonemes in a shallow orthography are more or less isomorphic, so that one grapheme is always connected to one phoneme, and vice versa. On the other hand, in a deep orthography like English, one grapheme may be linked to several phonemic alternatives and a cluster of phonemes may be connected to several orthographic clusters. Consequentially, when a pronunciation needs to be generated for a given novel word in English, several phonemic alternatives may compete for activation causing processing delays that might result in generally slower learning dynamics. Thus, we argue that the interlanguage differences we found in the two memorization tasks are most likely attributable to their different efficiencies in establishing a system of grapheme-to-phoneme mappings, differences that derive from their orthographic depth.

The work presented here, in its conception, was inspired by the study of Tsallis et al. [6], but now conducted with more complex, nonbinary stimuli and in another, linguistic domain, with different presentation times, including two linguistically different populations of tested subjects and a more sensitive way of measuring and evaluating the amount of acquired information.

We showed that the optimality of learning across languages indeed varies as a function of language regularity/consistency [65]. Thus, the more regular a given language, the faster the acquisition of novel, randomly generated letter-string items. This phenomenology could be highly important for further studies of e.g., second, third or bilingual and multilingual language development [66]. We further demonstrated that for a specific novel word learning task, human learners tend to perform similarly to very simple, but slightly 'non-extensive' neural networks. In other words, the experimental learning curves were best fitted by the non-extensive $q > 1$ model instead of the more efficient $q = 1$ extensive one (see Figs. 6 and 7).

The experiments presented here are still open to further interpretations. One can still propose many possible hypotheses about what is happening to cause the shape of the learning curve in the process of novel word learning and ask whether nonextensivity generally plays a universal role in language learning. However, our study is a first step towards a more complete work on this topic. The simulations with the nonextensive neural network model demonstrate that a considerable plurality of language learning processes can result from a very few and simple interactive mechanisms depending upon only several macroscopic parameters.

With regard to the future modeling, it would be interesting to try studying the much more difficult problem of a multilayer perceptron with a noisy local dynamics and observe whether it can lead to similar learning curves. Moreover, it remains a challenge for future research to simulate the novel word learning processes with our model in other shallow and deep orthography languages. It is also important to compare the results from the linguistic domain with those due to other effects and from other learning domains, especially when non-visual or non-auditory sensory modalities are involved during the learning process. Besides the declarative learning of novel words, we also need to check whether our model generalizes to other types of language learning such as implicit, procedural learning of the mental grammar.

Although we restricted the present study to the investigation of two specific languages, our neural network model is not intended to be dedicated to any particular artificial or natural language to which we believe our conclusions are most relevant. On the contrary, we wanted to demonstrate that the non-locality in learning arises naturally from the internal dynamics of a learning system and its interaction with the external stimulus, but independently of the particularities of its neural architecture.

Taken together, we have seen that our q-generalized neural network shows high sensitivity to different and, moreover, population-specific degrees of globality of learning. Such findings raise the possibility of using entropic nonextensivity as a means of characterizing the degree of complexity of learning in both natural and artificial systems [67–69]. As a good indicator of learning efficiency, specified by the entropic index q, the model may be useful in developing diagnostic monitoring tools that could be applied in a variety of learning domains.

Finally, we argue that the results and advantages of the presented model span beyond learning applications, and may help in studying other problems in neuroscience such as neurological impairments. More specifically, our q-generalized neural network could be used for diagnosing the degree of language impairment in dyslexic patients, with greater sensitivity and accuracy than with many currently available methods. When compared with healthy readers, dyslexics are typically slower and more inaccurate at translating letters into the sounds they stand for. This, however, becomes much more difficult to measure and compare between the speakers of differently transparent language orthographies who show an equal degree of neural impairment, but varying letter-to-sound conversion abilities [70,71]. Thus, it is a challenge for the present model to provide a method of easily and accurately diagnosing and classifying the cross-linguistically different degrees of dyslexia found under the same or similar brain impairment conditions.

Acknowledgments

This work was partially supported by grants from CONICET (Argentina) and SeCyT, Universidad Nacional de Córdoba (Argentina). The preparation of some theoretical parts of this paper was in part supported by the Austrian Research Association (ÖFG). We thank B. Turan and A. Kocan for their assistance in Exps. 1 & 2.

References

[1] M.A. Montemurro, Physica A 300 (2001) 567.
[2] R. Ferrer i Cancho, R.V. Solé, J. Quant. Ling. 8 (2001) 165.
[3] C. Tsallis, G. Bemski, R.S. Mendes, Phys. Lett. A 257 (1999) 93.
[4] C. Tsallis, J. Statist. Phys. 52 (1988) 479.
[5] E.M.F. Curado, C. Tsallis, J. Phys. A 24 (1991) L69. 24 (1991) 3187 (corrigenda); 25 (1992) 1019.
[6] A. Tsallis, C. Tsallis, A.C.N. de Magalhaes, F.A. Tamarit, Complexus 1 (2004) 181.
[7] M.A. Montemurro, D. Zanette, Glottometrics 4 (2002) 87.
[8] O. Sotolongo-Costa, A.H. Rodríguez, G.J. Rodgers, Entropy 2 (2000) 172.

[9] C. Beck, Physica A 306 (2002) 189.
[10] S. Tong, A. Bezerianos, Y. Zhu, R. Geocadin, D. Hanley, N.V. Thakor, Eng. Med. Biol. Soc. (2001); Proc. 23rd Ann. Intl. Conf. IEEE, vol. 2, 2001, p. 1926.
[11] S. Tong, A. Bezerianos, J. Paul, Y. Zhu, N.V. Thakor, Physica A 305 (2002) 619.
[12] N.V. Thakor, J. Paul, S. Tong, Y. Zhu, A. Bezerianos, Statistical signal processing, in: Proc. 11th IEEE Signal Processing Workshop, 2001, pp. 261–264.
[13] A. Capurro, L. Diambra, D. Lorenzo, O. Macadar, M.T. Martin, C. Mostaccio, et al., Physica A 265 (1999) 235.
[14] O.A. Rosso, M.T. Martin, A. Plastino, Physica A 320 (2002) 497.
[15] M. Mazza, W. Tedeschi, M. de Pinho, U.P.C. Neves, Neurocomputing 44 (2002) 915.
[16] J.J. Hopfield, Proc. Natl. Acad. Sci. USA 79 (1982) 2554.
[17] D.J. Amit, H. Gutfreund, H. Sompolinsky, Phys. Rev. Lett. 55 (1985) 1530.
[18] T.L.H. Watkin, A. Rau, M. Biehl, Rev. Modern Phys. 65 (1993) 499.
[19] D.A. Stariolo, Phys. Lett. A 185 (1994) 262.
[20] S.A. Cannas, D. Stariolo, F.A. Tamarit, Network: Comput. Neural Sci. 7 (1996) 141.
[21] F.T. Sommer, P. Kanerva, Behav. Brain Sci. 29 (2006) 86.
[22] T.J. Sejnowski, in: Proc. 24th Ann. Meeting Assoc. Computational Ling., 1986, p. 184.
[23] D.E. Rumelhart, J.L. McClelland, The PDP Research Group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. I, MIT Press, Cambridge, MA, 1986.
[24] T.J. Sejnowski, C.R. Rosenberg, Complex Syst. 1 (1987) 145.
[25] J.L. McClelland, M.St. John, R. Taraban, Lang. Cogn. Processes 4 (1989) (SI) 287.
[26] J.L. McClelland, T.T. Rogers, Nat. Rev. Neurosci. 4 (2003) 310.
[27] Y. Ikegaya, G. Aaron, R. Cossart, D. Aronov, I. Lampl, D. Ferster, R. Yuste, Science 304 (2004) 559.
[28] G. Buzsáki, Nat. Neurosci. 7 (2004) 446.
[29] F. Pulvermüller, Behav. Brain Sci. 22 (1999) 253;
F. Pulvermüller, Trends Cogn. Sci. 5 (2001) 517;
F. Pulvermüller, The Neuroscience of Language: On Brain Circuits of Words and Serial Order, Cambridge University Press, Cambridge, 2003.
[30] E. Halgren, C. Wang, D.L. Schomer, S. Knake, K. Marinkovic, J. Wu, I. Ulbert, NeuroImage 30 (2006) 1401.
[31] M.T. Ullman, Nat. Rev. Neurosci. 2 (2001) 717.
[32] B. Milner, L.R. Squire, E.R. Kandel, Neuron 20 (1998) 445.
[33] Y. Dudai, Ann. Rev. Psychol. 55 (2004) 51.
[34] M.E. Bach, R.D. Hawkins, M. Osman, E.R. Kandel, M. Mayford, Cell 81 (1995) 905.
[35] Y. Kishimoto, R. Fujimichi, K. Araishi, S. Kawahara, M. Kano, A. Aiba, Y. Kirino, Eur. J. Neurosci. 16 (2002) 2416.
[36] E.B. Papachristos, C.R. Gallistel, J. Exp. Anal. Behav. 85 (2006) 293.
[37] R. Frost, L. Katz, Memory Cogn. 17 (1989) 302.
[38] L. Katz, R. Frost, in: R. Frost, L. Katz (Eds.), Orthography, Phonology, Morphology and Meaning, Elsevier Science Publishers, Amsterdam, 1992, pp. 67–84.
[39] P.H.K. Seymour, M. Aro, J.M. Erskine, British J. Psychol. 94 (2003) 143.
[40] B. Müller, J. Reinhardt, Neural Networks: An Introduction, Springer-Verlag, Berlin, 1991.
[41] H.S. Seung, H. Sompolinsky, N. Tishby, Phys. Rev. A 45 (1992) 6056.
[42] N.G. Van Kampen, Stochastic Processes in Physics and Chemistry, Elsevier Sience V.B, 1992.
[43] A. Plastino, O.A. Rosso, Europhys. News 6 (2005) 224.
[44] T. Hadzibeganovic, S.A. Cannas, Abstracts of the Intl. Sci. Conf. on Modern Methods in Linguistics (Math. and Computer Ling.), University of Saints Cyril and Methodius, Slovakia, October 23–24, 2006, pp. 3–5;
T. Hadzibeganovic, S.A. Cannas, Proc. 3rd Intl Conf. on Discourse & Cogn. Ling, Hankuk Publishing Co, Seoul, 2007, pp. 96–110.
[45] H. Horner, Z. Phys. B 86 (1992) 291.
[46] H. Risken, The Fokker–Planck Equation, Springer-Verlag, Berlin, 1996.
[47] S.E. Gathercole, C. Willis, H. Emslie, A.D. Baddeley, Develop. Psychol. 28 (1992) 887.
[48] E. Service, V. Kohonen, Appl. Psycholing. 16 (1995) 155.
[49] P. Gupta, Quart. J. Exp. Psychol. 56A (2003) 1213.
[50] E. Paulesu, E. McCrory, F. Fazio, L. Menoncello, N. Brunswick, S.F. Cappa, et al., Nat. Neurosci. 3 (2000) 91.
[51] S. Kandel, S. Valdois, Lang. Cogn. Processes 21 (2006) 531.
[52] H. Wimmer, P. Hummer, Appl. Psycholing. 11 (1990) 349.
[53] N.C. Ellis, M. Natsume, K. Stavropoulou, L. Hoxhallari, V.H.P. van Daal, N. Polyzoe, et al., Read. Res. Quart. 39 (2004) 438.
[54] J.C. Ziegler, A.M. Jacobs, G.O. Stone, Behav. Res. Meth. Instr. Comput. 28 (1996) 504.
[55] R. Peereman, A. Content, Behav. Res. Meth. Instr. Comput. 31 (1999) 376.
[56] D.H. Brainard, Spat. Vision 10 (1997) 433.
[57] W. Duyck, T. Desmet, L.P.C. Verbeke, M. Brysbaert, Beh. Res. Meth. Instr. Comput. 36 (2004) 488.
[58] J.J. Binney, N.J. Dowrick, A.J. Fisher, M.E.J. Newman, The Theory of Critical Phenomena, Oxford Science, Oxford, 1993.
[59] P.K. Kuhl, Proc. Natl. Acad. Sci. USA 97 (2000) 11850.
[60] P.K. Kuhl, Nat. Rev. Neurosci. 5 (2004) 831.
[61] Y. Zhang, P.K. Kuhl, T. Imada, M. Kotani, Y. Tohkura, NeuroImage 26 (2005) 703.
[62] P.K. Kuhl, Percept. Psychophy. 50 (1991) 93.
[63] J.E. Flege, in: W. Strange (Ed.), Speech Perception and Linguistic Experience, York Press, Timonium, MD, 1995, pp. 233–277.
[64] G.K. Vallabha, J.L. McClelland, Cogn. Affect. Behav. Neurosci. 7 (2007) 53.
[65] U. Goswami, Educ. Psychol. Practise 21 (2005) 273.
[66] G. Meschyan, A.E. Hernandez, NeuroImage 29 (2005) 1135.
[67] C. Tsallis, Chaos Solitons Fractals 13 (2002) 371.
[68] M. Rajkovic, Physica A 340 (2004) 327.
[69] W. Bialek, I. Nemenman, N. Tishby, Physica A 302 (2001) 89.
[70] E. Paulesu, J.-F. Démonet, F. Fazio, E. McCrory, V. Chanoine, N. Brunswick, et al., Science 291 (2001) 2165.
[71] G. Silani, U. Frith, J.-F. Demonet, F. Fazio, D. Perani, C. Price, et al., Brain 128 (2005) 2453.