2012-09

Critical behavior in a cross-situational lexicon learning scenario

EPL, MULHOUSE, v. 99, n. 6, pp. 188-192, SEP, 2012
http://www.producao.usp.br/handle/BDPI/33440

Critical behavior in a cross-situational lexicon learning scenario

A Letters Journal Exploring the Frontiers of Physics

September 2012

EPL, **99** (2012) 60001 www.epljournal.org
doi: 10.1209/0295-5075/99/60001

Critical behavior in a cross-situational lexicon learning scenario

P. F. C. Tilles and J. F. Fontanari

Instituto de Física de São Carlos, Universidade de São Paulo - Caixa Postal 369,
13560-970 São Carlos, São Paulo, Brazil

received 12 June 2012; accepted in final form 4 September 2012
published online 5 October 2012

PACS 05.40.-a – Fluctuation phenomena, random processes, noise, and Brownian motion
PACS 89.75.Fb – Structures and organization in complex systems
PACS 05.70.Jk – Critical point phenomena

Abstract – The associationist account for early word learning is based on the co-occurrence between referents and words. Here we introduce a noisy cross-situational learning scenario in which the referent of the uttered word is eliminated from the context with probability γ, thus modeling the noise produced by out-of-context words. We examine the performance of a simple associative learning algorithm and find a critical value of the noise parameter γ_c above which learning is impossible. We use finite-size scaling to show that the sharpness of the transition persists across a region of order $\tau^{-1/2}$ about γ_c, where τ is the number of learning trials, as well as to obtain the learning error (scaling function) in the critical region. In addition, we show that the distribution of durations of periods when the learning error is zero is a power law with exponent $-3/2$ at the critical point.

Introduction. – The problem of early word learning has been subject of philosophical controversy for centuries [1]. The always visionary Augustine argued that the child makes the connections between words and their referents by understanding the referential intentions of others, thus anticipating the modern theory of mind in about fifteen centuries [2]. In the 17th century, Locke's empiricism supported the associationist viewpoint, which contends that the mechanism of word learning is sensitivity to covariation, *i.e.*, if two events occur at the same time, they become associated.

Here we examine a radical offshoot of the associationist approach to lexicon acquisition termed cross-situational or observational learning [3], which asserts that the meaning of a word can be determined by looking for something in common across all observed uses of that word [4]. In other words, learning takes place through the statistical sampling of the contexts in which a word appears.

A scenario to describe the lexicon acquisition process should take into account the inherent ambiguity of the learning task (*i.e.*, many distinct objects may be associated to the same word) as well as the noise effect of out-of-context words (*i.e.*, the uttered word may not refer to any object in the context). Whereas the noiseless scenario has been explored in great detail in the literature [5–7], where it was shown that the learning error decreases exponentially with the number of learning trials, a systematic study of the effect of noise is lacking.

To remedy this deficiency, we modify the minimal model of noiseless cross-situational learning [5–7] so as to include the effect of noise produced by out-of-context words. Using Monte Carlo simulations and finite-size scaling we identify and characterize a critical phenomenon that separates the asymptotic regime where the lexicon can be acquired without errors from the regime where learning is impossible. At the critical noise level, we find that the duration of the periods with zero error is distributed by a power-law distribution.

Cross-situational learning scenario. – We assume that there are N objects, N words and a one-to-one mapping between words and objects. At each learning event, C objects are chosen at random without replacement from the fixed list of N objects and one of these objects is named according to the word-object mapping. The C objects form the context which determines the interpretation of the uttered word and the learner's task is to guess which of the C objects that word refers to. This is then an ambiguous word learning scenario in which there are multiple object candidates for any word. The parameter C is a measure of the ambiguity of the learning task. In particular, in the case $C = N$ the word-object mapping is not learnable within a cross-situational scenario.

A learning episode comprises a context and a single target word. In an uncorrupted learning episode, the context must exhibit the correct object (*i.e.*, the object named by the target word according to the object-word mapping) plus $C-1$ distinct mismatching objects. Noise is added to the learning scenario by removing the correct object from the context, which will then exhibit C mismatching objects. Such corrupted and misguiding learning episodes occur with probability $\gamma \in [0,1]$. This type of noise is an integrant part of any realistic learning situation, arising usually from the unwarranted narrowing of the context by the learner.

To represent the one-to-one object-word mapping we use the index $i=1,\ldots,N$ to label the distinct objects and $h=1,\ldots,N$ to label the distinct words. Then, without lack of generality, the correct mapping is defined by assigning object $i=1$ to word $h=1$, object $i=2$ to word $h=2$ and so on. The problem faced by the learner is to determine the correct mapping given a sequence of learning episodes. Next we will describe a simple (perhaps, the simplest) procedure to accomplish this learning task.

Associative learning model. – We assume that learning is a change in the confidence with which the learner associates the target word h to a given object i and represent this confidence by a non-negative integer p_{ih}. Our associative accumulator learning procedure is described as follows. Before learning all confidences are set to zero, *i.e.*, $p_{ih}=0$ for $i,h=1,\ldots,N$, and whenever object i^* appears in a context with target word h^* the confidence $p_{i^*h^*}$ increases by one unit [8]. Hence, exactly C confidence values are updated at each learning trial.

To determine which object corresponds to word h the learner simply chooses the object index i for which p_{ih} is maximum. In the case of ties, the learner selects one object at random among those that maximize the confidence. From the definition of the correct word-object mapping, our learning algorithm achieves a perfect performance when $p_{hh} > p_{ih}$ for all h and $i \neq h$.

A critical feature of the accumulator model is that words are learned independently. This fact alone allows us to split the analysis of the vocabulary learning task in two parts. The first and most important part is the problem of learning the meaning (or the referent) of a *single* word. Once this is done, we can easily solve the problem of learning the N words given their sampling frequencies [7]. Hence, in this work we will focus on the single-word learning problem only.

Single-word learning. – Accordingly, we consider the learning of a single word, say word h, which is then uttered at all learning trials τ. We define the single-word learning error $\epsilon(\tau)$ for $\tau>0$ as follows. If $p_{hh} < p_{ih}$ for any $i \neq h$ then $\epsilon = 1$, otherwise if $p_{hh} = p_{ih}$ for n values of $i \neq h$ then $\epsilon = n/(n+1)$ with $n=0,\ldots,N-1$. At $\tau=0$ all confidences are set to zero and so $\epsilon = (N-1)/N$.

In the noiseless case ($\gamma=0$) we have $p_{hh} \geqslant p_{ih}$ for all $i \neq h$ since object $i=h$ is always part of the context. So errors are due to ties $p_{hh}=p_{ih}, i \neq h$ only. In fact, it can be shown analytically that in this case the average learning error vanishes like $[(C-1)/(N-1)]^\tau$ for large τ [5–7].

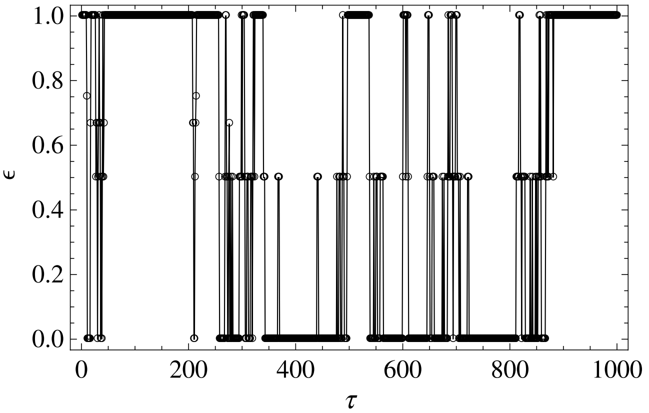

Fig. 1: Learning error *vs.* the number of learning trials τ for a single sample of the learning process using the accumulator learning model. The parameters are $N=20$, $C=6$ and $\gamma = \gamma_c = 0.7$. The lines are guides to the eye.

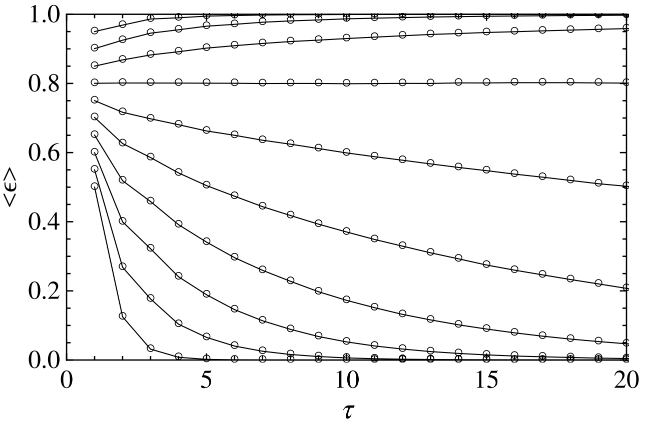

Fig. 2: Average learning error $\langle \epsilon \rangle$ as function of the number of learning trials for $N=5$, $C=2$ and (bottom to top) $\gamma = 0, 0.1, 0.2, \ldots, 0.9$. The critical value of the noise parameter is $\gamma_c = 0.6$ at which $\langle \epsilon_c \rangle = 0.8$. The symbols are the simulation results and the lines are guides to the eyes.

In the case the contexts are corrupted by noise with a probability γ an analytical approach is not possible and we have to resort to simulations to study the stochastic learning process. Figure 1 shows a typical evolution of the learning error at the critical noise level. Although this figure reveals a rich stochastic dynamics, it is rather uninformative from the learning perspective. In that sense, the behavior of the average learning error $\langle \epsilon \rangle$, shown in fig. 2, is more relevant. For a fixed τ, this average is calculated using typically 10^6 to 10^7 realizations of the learning process.

Figure 2 reveals that learning is possible provided that the noise parameter does not exceed a certain threshold γ_c. More pointedly, in the asymptotic regime $\tau \to \infty$ we find that $\langle \epsilon \rangle \to 0$ for $\gamma < \gamma_c$ and that $\langle \epsilon \rangle \to 1$ for $\gamma > \gamma_c$. The surprising finding is that at $\gamma = \gamma_c$, the average learning error becomes independent of $\tau \geqslant 0$.

There is a simple argument to determine γ_c as well as the error $\langle \epsilon_c \rangle$. We begin by noting that the borderline between learning and non-learning occurs when all N objects are equally likely of being selected to compose the contexts. In fact, since we assume that learning is based on the perception of differences in the co-occurrence of objects and target words, in the case all N objects have the same probability of being selected to form the contexts, such a purely observational learning is clearly unattainable. Hence in this case we have

$$\langle \epsilon_c \rangle = \epsilon\left(\tau = 0\right) = \frac{N-1}{N}, \tag{1}$$

which corresponds to the learning error prior to learning, as expected. The probability of selecting the correct object in a learning episode is given by the probability of generating a noise-free context, *i.e.*, $1-\gamma$, since the correct object is certain to be chosen in this case. A given confounding object can be selected in two ways. First, in a noise-free episode with probability $\left(1-\gamma\right)\left[\left(C-1\right)/\left(N-1\right)\right]$. Second, in a noisy episode with probability $\gamma\left[C/\left(N-1\right)\right]$. Note that in both ways the correct object is taken out from the list of eligible objects —in the first because it was already chosen and in the second because it cannot be chosen. Accordingly, γ_c is determined by equating the probability of selecting the correct object with the probability of selecting any given confounding object to compose the context in a learning episode,

$$1-\gamma_c = \left(1-\gamma_c\right)\frac{C-1}{N-1} + \gamma_c \frac{C}{N-1}, \tag{2}$$

from which we get

$$\gamma_c = 1 - \frac{C}{N}. \tag{3}$$

These expressions for $\langle \epsilon_c \rangle$ and γ_c proved correct for a vast selection of values of N and C. In addition, we can perform a simple consistency check on these expressions as follows. The average learning error at the first trial is given by

$$\langle \epsilon\left(\tau = 1\right) \rangle = \left(1-\gamma\right)\frac{C-1}{C} + \gamma \tag{4}$$

and by setting $\gamma = \gamma_c$ we recover eq. (1) as it should be since $\langle \epsilon_c \rangle$ is independent of τ (see fig. 2).

Analytical solution for $N = 2$. – For $N = 2$ (and so $C = 1$), the average learning error at an even number of trials τ is simply

$$\langle \epsilon \rangle = \frac{1}{2}\binom{\tau}{\tau/2}\left(1-\gamma\right)^{\tau/2}\gamma^{\tau/2} + \sum_{n=0}^{\tau/2-1}\binom{\tau}{n}\left(1-\gamma\right)^{n}\gamma^{\tau-n}, \tag{5}$$

where the first term corresponds to the case that both objects appear $\tau/2$ times and the second term to the case that the confounder appears more frequently than the correct object. (If the number of trial τ is odd, then the first term of eq. (5) must be discarded.) For $\tau \gg 1$ we find

$$\begin{aligned}\langle \epsilon \rangle \sim &\left(\frac{1}{2\pi\tau}\right)^{1/2}\left[4\gamma\left(1-\gamma\right)\right]^{\tau/2} \\ &-\frac{1}{2}\operatorname{erfc}\left[\tau^{1/2}\left(\frac{1-\gamma}{2\gamma}\right)^{1/2}\right] \\ &+\frac{1}{2}\operatorname{erfc}\left[\frac{\tau^{1/2}\left(1/2-\gamma\right)}{\left[2\gamma\left(1-\gamma\right)\right]^{1/2}}\right],\end{aligned} \tag{6}$$

where erfc is the complementary error function. The outcome of the limit $\tau \to \infty$ is determined by the last term of this expression, which goes to 0 (and so $\langle \epsilon \rangle \to 0$) for $\gamma < 1/2$ and to 1 (and so $\langle \epsilon \rangle \to 1$) for $\gamma > 1/2$, which proves that $\gamma_c = 1/2$ in this case. We note that setting $\gamma = 1/2$ in eq. (5) and using the identity

$$2\sum_{n=0}^{\tau/2-1}\binom{\tau}{n} + \binom{\tau}{\tau/2} = 2^{\tau} \tag{7}$$

yields $\langle \epsilon_c \rangle = 1/2$ regardless of the value of τ.

The behavior of the average learning error in the vicinity of the critical point is obtained by taking the limits $\gamma \to \gamma_c = 1/2$ and $\tau \to \infty$ in eq. (6), yielding

$$\langle \epsilon \rangle \sim \frac{1}{2}\operatorname{erfc}\left[\frac{\tau^{1/2}\left(\gamma_c-\gamma\right)}{\left[2\gamma_c\left(1-\gamma_c\right)\right]^{1/2}}\right]. \tag{8}$$

Finite-size scaling analysis. – Considering the "size" of the system as the number of learning trials τ, we proceed now to examine the sharpness of the phase transition at γ_c using finite-size scaling [9]. This threshold phenomenon is best appreciated in fig. 3, which exhibits the dependence of the average learning error on the distance to the critical parameter for different values of τ. As the number of trials τ increases, the difference between the regimes $\gamma < \gamma_c$ and $\gamma > \gamma_c$ becomes evident. All curves intersect at $\gamma = \gamma_c$ for which the average error is a constant given by eq. (1).

The key insight is obtained when one considers the average learning error as a function of the reduced variable $\left(\gamma_c - \gamma\right)\tau^{1/2}$. Use of this reduced variable produces the collapse of the data for different τ into a single scaling function as shown in the inset of fig. 3. We can improve the accuracy of the estimate of the scaling functions by fixing τ to a large value, say $\tau = 10^5$, and then varying γ in the vicinity of a fixed γ_c. The result of this procedure is illustrated in fig. 4 where we show the scaling functions for four different values of C and fixed N.

As illustrated in fig. 4, the data is fitted very well by the functional form

$$\langle \epsilon \rangle = \frac{1}{2}\operatorname{erfc}\left[a\left(N\right) + b\left(N, C\right)\left(\gamma_c - \gamma\right)\tau^{1/2}\right], \tag{9}$$

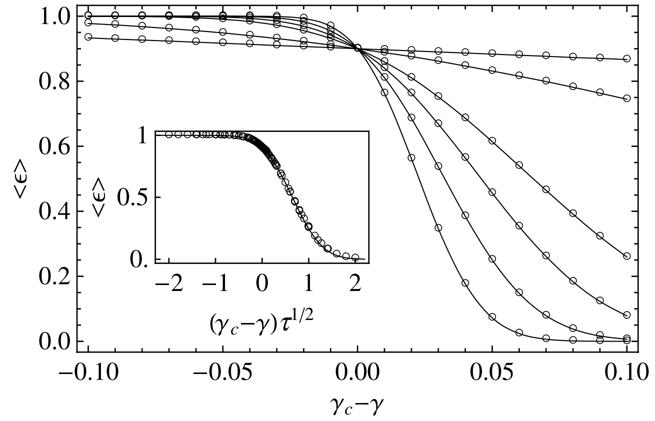

Fig. 3: Average learning error as function of the distance to the critical noise parameter for $N = 10$ and $C = 2$. The symbols are the simulation results for (top to bottom in the positive ordinate region) $\tau = 1, 10, 100, 200, 400$ and 800. The inset shows the data collapse when they are plotted *vs.* the reduced variable $(\gamma_c - \gamma)\,\tau^{1/2}$. The lines are guides to the eyes.

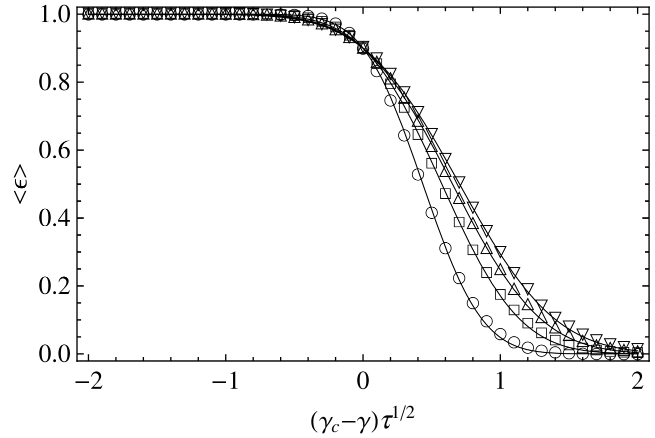

Fig. 4: Average learning error as function of the reduced variable $(\gamma_c - \gamma)\tau^{1/2}$ for $N = 10$ and $C = 1$ (◯), 2 (□), 3 (△) and 5 (▽). The symbols are the simulation results and the lines are given by the scaling function (9) with the parameter b obtained from the fitting of the data.

which has a single fitting parameter, $b\,(N, C)$. The parameter $a\,(N)$ is obtained by setting $\gamma = \gamma_c$ and then using the expression of $\langle \epsilon_c \rangle$, given by eq. (1). The final result is

$$a\,(N) = \mathrm{erfc}^{-1}\left[\frac{2\,(N-1)}{N}\right], \qquad (10)$$

where $\mathrm{erfc}^{-1}(x)$ stands for the inverse complementary error function. We note that $a\,(2) = 0$ and $a\,(N) < 0$ for $N > 2$.

In addition, we assume that $b\,(N, C) = b\,(\gamma_c)$ and plot this fitting parameter in fig. 5 for a large selection of values of N and C. More pointedly, for each value of N (represented by different symbols in the figure) we vary C from 1 to $N - 1$ to obtain scaling functions as those shown in fig. 4. Then these functions are fitted using eq. (9) in order to determine the fitting parameter b. For $N > 4$ the data is fitted very well by the function

$$b\,(\gamma_c) = \frac{b'}{[\gamma_c\,(1-\gamma_c)]^{1/2}} \qquad (11)$$

with $b' = 0.65$. Note that for $N = 2$, eq. (8) yields $b' = 1/\sqrt{2} \approx 0.71$.

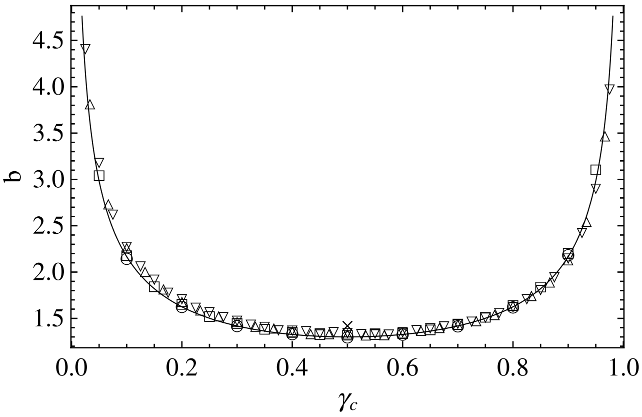

Fig. 5: Dependence of the fitting parameter b on the ratio γ_c for $N = 2$ (×), $N = 10$ (◯), $N = 20$ (□), $N = 30$ (△) and $N = 40$ (▽). The solid line is given by eq. (11).

Figure 5 reveals a most interesting symmetry: for fixed N the average learning error when plotted *vs.* the reduced variable $\kappa \equiv (\gamma_c - \gamma)\,\tau^{1/2}$ is invariant to the change $C \to N - C$ which implies $\gamma_c \to 1 - \gamma_c$. In particular, in fig. 4 for which $N = 10$, the results for $C = 9$ are identical to those displayed for $C = 1$, the results for $C = 8$ to those for $C = 2$ and so on. More pointedly, choosing $\kappa/\tau^{1/2} = 0.01$ this symmetry means that the associative algorithm performs identically for the parameter set $C = 1$ and $\gamma = 0.89$ and for the set $C = 9$ and $\gamma = 0.09$, so there is an exact balance between the performance deterioration due to an increase of the ambiguity of the task and the performance increment due to a decrease of the out-of-context-words noise. It should be emphasized that this symmetry, which is a direct consequence of eq. (11), is exact only in the limits $\tau \to \infty$ and $\gamma \to \gamma_c$.

A word is in order about our choices of the fitting functions eqs. (9) and (11) as well as of the scaling $\tau^{1/2}$. All these elements appear in the analytical study of the special case $N = 2$ that serves as guide for the empirical analysis of the general case. For arbitrary N and C, the probability that object 1 appears m_1 times, object 2 appears m_2 times, etc. in τ learning episodes is given by the multinomial distribution

$$\frac{\tau!}{m_1! \ldots m_N!}\,(1-\gamma)^{m_1}\left[(1-\gamma)\,\frac{C-1}{N-1} + \gamma\frac{C}{N-1}\right]^{\tau - m_1} \qquad (12)$$

with $\sum_i m_i = \tau$ and we have assumed that label 1 refers to the correct object. The learning error is obtained by summing over the m_i's while giving the appropriate weights to ties of distinct multiplicity and weight 1 to configurations such that $m_i > m_1$ for some $i > 1$. For large τ we move to a continuous variable approximation in which the sums are replaced by integrals and the multinomial is approximated by a multivariate Gaussian distribution where the means $\langle m_i \rangle$, variances $\mathrm{Var}(m_i)$ and covariances $\mathrm{Cov}(m_i, m_j)$ all scale linearly with τ. Since the relevant limits of the sums scale with τ also (see, *e.g.*, eq. (5)), the elimination of the dependence on τ in the integrand leads to the multidimensional Gaussian integrals whose limits scale with $\tau^{1/2}$ in our approximation scheme. This argument also explains our choice of the complementary error function in eq. (9). For $\gamma \to \gamma_c$ all event probabilities in the multinomial (12) become identical and equal to $1-\gamma_c$ (see eq. (2)). This implies that $\mathrm{Var}(m_i) = \tau\gamma_c\,(1-\gamma_c)$ for all i which then explains our choice of the functional form for the parameter $b\,(N,C) = b\,(\gamma_c)$ (see eq. (11)) that controls the sharpness of the learning error. Moreover, eq. (11), which stems from the statistical equivalence between the objects at the critical point, is responsible for the symmetry of the learning error discussed before.

To obtain the average learning error for an infinitely large lexicon, $N \to \infty$, we note first that $a(N) \sim -\ln^{1/2} N$ in this limit. To proceed further we have to consider two cases. First, if the context size C grows linearly with N (*i.e.*, $0 < \gamma_c < 1$) then $b(\gamma_c)$ is finite and the only diverging term in the argument of eq. (9) is $a(N) \to -\infty$ which leads to $\langle \epsilon \rangle \to 1$. Second, if C remains finite when $N \to \infty$ then $1-\gamma_c = 1/N \to 0$ and so $b(\gamma_c) \sim N^{1/2}$. Since the divergence of b to $+\infty$ is faster than the divergence of a to $-\infty$ we find $\langle \epsilon \rangle \to 0$ in this case.

Statistics of stasis. – A distinctive feature of the learning process revealed by fig. 1 is the existence of long periods when the learning error stands at zero value, *i.e.*, $p_{hh} > p_{ih}$ for all objects $i \neq h$. These periods or stases are characterized by repeated additions of credence units to the confidence values p_{ih} and they end when one (or more) of the $N-1$ confidences $p_{ih}, i \neq h$, equals p_{hh}.

We begin the analysis of the distribution $P_c(\Delta\tau)$ of the durations $\Delta\tau$ of the stases at the critical parameter γ_c by showing in fig. 6 how the total number of learning trials τ_0 (basically a cutoff time) affects this distribution. The rescaling $\tau_0^{3/2} P_c(\Delta\tau/\tau_0)$ makes the results essentially independent of the cutoff parameter τ_0 provided $\Delta\tau/\tau_0$ is not too small (data not shown). The curves exhibit a clear power-law behavior with exponent $-3/2$, which is the mean-field exponent for the size of avalanches in self-organized critical models [10].

A rough analogy between our learning model and a sand-pile–like model goes as follows. Let us interpret the confidence value associated to an object as the amount of sand at a given point in space, and assume that the sole trigger of the avalanches is the amount of sand at the position in the pile corresponding to the correct object. At each learning episode, C grains of sand are distributed at C distinct points in the pile according to the probabilities given by the noise and the sampling process. Whenever the amount of sand at the triggering site is greater than of any other site (here is the mean-field assumption: all sites are neighbors of each other) an avalanche takes place. Hence the periods of zero learning error correspond to an ongoing avalanche. The avalanche stops when the amount of sand of one or more neighbors equal that of the triggering site, thus flattening the slope of the pile in the direction that connects those sites.

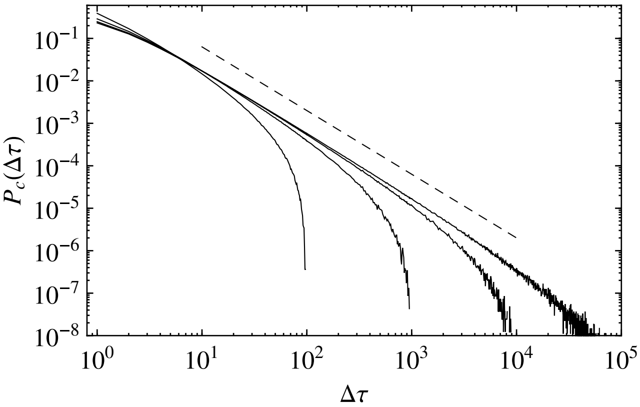

Fig. 6: Distribution of stases for $N = 20$, $C = 6$, $\gamma = \gamma_c = 0.7$, and (bottom to top) $\tau_0 = 10^3, 10^4$ and 10^5. The slope of the straight line is $-3/2$.

In addition, we find that away from the critical point the distribution $P(\Delta\tau)$ is exponential and that the average duration of the stases diverges like $\langle \Delta\tau \rangle \sim |\gamma_c - \gamma|^{-1}$ as $\gamma \to \gamma_c$.

As expected, these mean-field critical exponents are robust to changes in the model parameters N and C. In fact, for $N = 2$ and $C = 1$ the distribution $P(\Delta\tau)$ can be easily calculated analytically for any value of γ since this is the classical ruin problem in which a gambler with initial capital $z = 1$ plays against an infinitely rich adversary. The results for the duration of the game $\Delta\tau$ are simply $P_c(\Delta\tau) \approx (2/\pi)^{1/2}\,(\Delta\tau)^{-3/2}$ and $\langle \Delta\tau \rangle = (1/2)\,|\,\gamma_c - \gamma\,|^{-1}$ (see Chapt. XIV of [11]).

Changes in the number of objects N have no significant influence on $P_c(\Delta\tau)$ whereas changes in the context size C produce a shift on the distribution, without affecting the power-law exponent, as illustrated in fig. 7. In fact, an increase of C increases the frequency of short stases and, consequently, reduces the frequency of long ones. This is expected since the larger the context size, the greater the number of mismatching objects that have their confidences updated, and so the greater the odds of occurrence of the jump condition $p_{ih} \geqslant p_{hh}$ for some object $i \neq h$.

Finally, we note that although we have focused on the periods of the learning process when the error learning is 0, the very same conclusions hold for the periods when the learning error is 1.

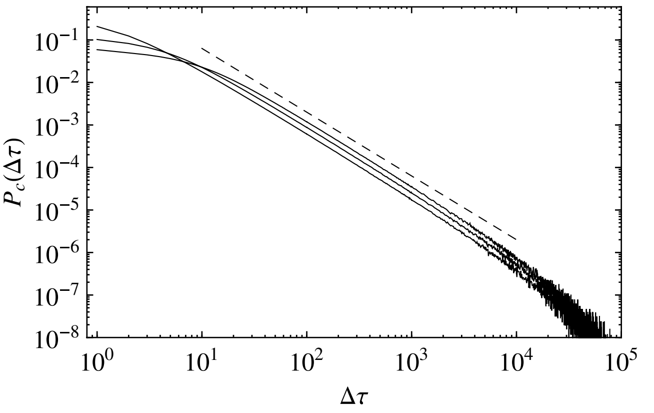

Fig. 7: Distribution of stases for $N = 20$, $\tau_0 = 10^5$ and (bottom to top at $\Delta\tau = 1$) $C = 1, 2, 5$. The slope of the straight line is $-3/2$.

Conclusion. – The view of language as a collective phenomenon arising out of local social interactions has prompted its modeling and investigation through statistical physics concepts and tools [12]. Words have been likened to genes and their evolution studied within a population genetics framework [13,14], whereas the competition between whole languages has been considered using population dynamics models [15–17]. The study of the bootstrap of a common lexicon among a large population of individuals has revealed a sharp phase transition towards shared conventions [18] as well as an unexpected connection with random occupancy problems in the case only two individuals interact but the lexicon size is very large [19].

The problem of acquiring, rather than bootstrapping, a fixed lexicon from observational learning is relevant to developmental psychology since it allows a quantitative appraisal of the associationist hypothesis on early word learning [1]. In particular, we show that the utterance of out-of-context words may result in severe limitations to learning, depending on the ratio C/N between the number of objects presented to the learner at a learning trial and the total number of objects. If this ratio is small (*i.e.*, γ_c is close to 1) then this noise effect is largely irrelevant and the lexicon can quickly be learned to perfection. However, for large values of this ratio (*i.e.*, γ_c is close to 0) learning becomes impossible regardless of the number of trials τ. Finite-size scaling shows that the threshold phenomenon persists across a region of size $\tau^{-1/2}$ around γ_c and offers the explicit functional form of the learning error in this region resorting to the single fitting parameter b' introduced in eq. (11).

The simplicity of our associative learning algorithm allowed us to consider the learning of the distinct words as independent stochastic processes. Interactions between words, such as the mutual exclusivity constraint that instructs children to associate novel words to unnamed objects [1], are well established in developmental psychology and it would be interesting to see whether and how they alter the characteristics of the critical phenomenon reported here.

* * *

This research was supported by The Southern Office of Aerospace Research and Development (SOARD), Grant No. FA9550-10-1-0006, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). PFCT was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

REFERENCES

[1] Bloom P., *How Children Learn the Meaning of Words* (MIT Press, Cambridge, Mass.) 2000.

[2] Adolphs R., *Nat. Rev. Neurosci.*, **4** (2003) 165.

[3] Pinker S., *Language Learnability and Language Development* (Harvard University Press, Cambridge, Mass.) 1984.

[4] Yu C. and Smith L. B., *Psychol. Rev.*, **119** (2012) 21.

[5] Smith K., Smith A. D. M, Blythe R. A. and Vogt P., *Lect. Notes Comput. Sci.*, **4211** (2006) 31.

[6] Blythe R. A., Smith K. and Smith A. D. M., *Cognit. Sci.*, **34** (2010) 620.

[7] Tilles P. F. C. and Fontanari J. F., arXiv:1204.1564v3 (2012).

[8] Bush R. R. and Mosteller F., *Stochastic Models for Learning* (Wiley, New York) 1955.

[9] Privman V., *Finite-Size Scaling and Numerical Simulations of Statistical Systems* (World Scientific, Singapore) 1990.

[10] Muñoz M. A., Dickman R., Vespignani A. and Zapperi S., *Phys. Rev. E*, **59** (1999) 6175.

[11] Feller W., *Introduction to Probability Theory and its Applications*, Vol. **1**, third edition (John Wiley & Sons, New York) 1968.

[12] Loreto V. and Steels L., *Nat. Phys.*, **3** (2007) 758.

[13] Fontanari J. F. and Perlovsky L. I., *Phys. Rev. E*, **70** (2004) 042901.

[14] Baxter G. J., Blythe R. A., Croft W. and McKane A. J., *Phys. Rev. E*, **73** (2006) 046118.

[15] Abrams D. M. and Strogatz S. H., *Nature*, **424** (2003) 900.

[16] Mira J. and Paredes Á., *Europhys. Lett.*, **69** (2005) 1031.

[17] Schulze C., Stauffer D. and Wichmann S., *Commun. Comput. Phys.*, **3** (2008) 271.

[18] Baronchelli A., Felici M., Loreto V., Caglioli E. and Steels L., *J. Stat. Mech.* (2006) P06014.

[19] Fontanari J. F. and Cangelosi A., *Interact. Stud.*, **12** (2011) 119.