Lexicon acquisition algorithms and random occupancy problems

José F. Fontanari*
Instituto de Física de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560-970 São Carlos, São Paulo, Brazil

Angelo Cangelosi†
Adaptive Behaviour & Cognition Research Group, University of Plymouth, Plymouth PL4 8AA, United Kingdom

Lexicon acquisition algorithms involve the repeated interaction between at least two agents who must reach a consensus on how to name N objects using H words. Here we consider minimal models of two types of learning algorithms: cross-situational learning in which the learner determines the meaning of a word by looking for something in common across all observed uses of that word, and operant conditioning learning in which there is strong feedback between speaker and hearer about the intended meaning of the words. Despite the stark differences between these learning schemes we show that they yield the same communication accuracy in the limits of large N and H, which coincides with result of the classical occupancy problem of randomly assigning N objects to H words.

PACS numbers: 89.75.Da, 89.75.Fb, 02.50.Ey, 02.50.Le

How a coherent lexicon can emerge in a group of interacting agents is a major open issue in the language evolution/acquisition research areas [1, 2]. In addition, it is the topic to which mathematical modeling can contribute the most, as the emergence of a lexicon from scratch implies some type of self-organization and, possibly, threshold phenomenon which can only be fully understood within a statistical mechanics framework [3–5].

There are basically two competing schemes for lexicon acquisition [6]. The first scheme, termed cross-situational or observational learning, is based on the intuitive idea that one way that a learner can determine the meaning of a word is to find something in common across all observed uses of that word [7]. Hence learning takes place through the statistical sampling of the contexts in which a word appears. Since the learner receives no feedback about its inferences, we refer to this scheme as unsupervised learning. The second scheme, known generally as operant conditioning, involves the active participation of the agents in the learning process, with intense exchange of non-linguistic cues to provide feedback on the hearer inferences. This supervised learning scheme has been applied to the design of a system for communication by autonomous robots – the so-called Talking Heads experiments [8, 9].

Many different computational implementations and variants of these learning schemes have been proposed in the literature (see, e.g., [10–12] for the unsupervised and [13, 14] for the supervised scheme). Except for the extensive statistical analysis of a variant of the supervised learning algorithm which reduces the problem to that of naming a single object [4, 5], the investigation of the effects of the parameters of those models have been usually limited to the display of the time evolution of some measure of the communication accuracy of the population.

Here we study minimal models of the supervised and unsupervised learning schemes which preserve the main ingredients of these classical language acquisition paradigms. In particular, we consider only two agents (a common assumption in language acquisition models, such as the popular iterated learning model [15]) who play in turns the roles of speaker and hearer. The agents live in a fixed environment composed of N objects and have H words available to name these objects. As we are interested in the limit where N and H are very large with the ratio $\alpha \equiv H/N$ finite, we do not need to account for the possibility of creation of new words as done in some variants of the supervised learning scheme.

We assume that each agent is characterized by a $N \times H$ verbalization matrix P, the entries of which $p_{nh} \in [0, 1]$ with $\sum_h^H p_{nh} = 1,\ \forall n$ yield the probability that object n is associated with word h. This assumption rules out the existence of objects without names, but it allows for words which are never used to name objects. To describe the communicative behavior of the agents through the verbalization matrix (i.e., the associations between objects and words for use both in production and interpretation) we need to specify how the speaker chooses a word for any given object as well as how the hearer infers the object the speaker intended to name by that word. To name an object, say object n, the speaker simply chooses the word h^* associated to the largest entry of row n of the matrix P, i.e. $h^* = \max_h \{p_{nh}, h = 1, \ldots, H\}$. To guess which object the speaker named by word h the hearer selects the object that corresponds to the largest of the N entries p_{nh}, $n = 1, \ldots, N$. In other words, the hearer chooses the object that it itself would be most likely to associate with word h [10, 11] (see [16] for the original version of this inference scheme).

Effective communication takes place when the two

agents reach a consensus on which word must be assigned to each object. To achieve this we must provide a prescription to modify their initially random verbalization matrices. Here we will consider two learning procedures that differ basically on whether the agents receive feedback (supervised learning) or not (unsupervised learning) about the success of a communication episode. However, before doing this we need to set the language game scenario where the agents interact.

From the list of N objects, the agent who plays the speaker role chooses randomly C objects without replacement. This set of C objects forms the context. Then the speaker chooses randomly one object in the context and produces the word associated to that object, according to the procedure sketched before. The hearer has access to that word as well as to the C objects that comprise the context. Its task is to guess which object in the context is named by that word. Once the verbalization matrices are updated the two agents interchange the roles and a new context is generated following the same procedure.

To control the convergence properties of the learning algorithms described next we discretize the entries p_{nh} so that they can take on the values $0, 1/M, 2/M, \ldots, 1 - 1/M, 1$. In addition, as there are two agents who alternate in the roles of speaker and hearer, henceforth we will add the superscripts I or J to the verbalization matrix in order to identify the agent it corresponds to. At the beginning of the language game each agent has a different, randomly generated verbalization matrix. More pointedly, to generate the row n of P^I we distribute with equal probability M balls among H slots and set the value of entry p^I_{nh} as the ratio between the number of balls in slot h and the total number of balls M. An analogous procedure is used to set the initial value of P^J.

Unsupervised learning

In this scheme, the list of objects in the context $n_1, \ldots, n_c$, and the accompanying word h^* is the only information fed to the learning algorithm. Hence in the unsupervised scheme only the hearer's verbalization matrix is updated. For concreteness, let us assume that agent I is the speaker and so agent J is the hearer. As pointed out before, the idea here is to model the cross-situational learning scenario [7] in which the agents infer the meaning of a given word by monitoring its occurrence in a variety of contexts. Accordingly, the learning procedure increases the entries $p^J_{n_1h^*}, \ldots, p^J_{n_ch^*}$ by the quantity $1/M$. In addition, for each object in the context, say n_1, a word, say h, is chosen randomly and the entry $p^J_{n_1h}$ is decreased by the quantity $1/M$, thus keeping the correct normalization of the rows of the verbalization matrix. (The possibility that $h = h^*$ is not ruled out.) This procedure which is inspired by Moran's model of population genetics [17] guarantees a minimum disturbance in the verbalization matrix. We note that during this learning stage the hearer does not need to guess which object in the context is named by word h^*. An extra rule is needed to keep the entries p^J_{nh} within the unit interval $[0, 1]$: we assume that once an entry reaches the values $p^J_{nh} = 1$ or $p^J_{nh} = 0$ it becomes fixed, so the extremes of the unit interval act as absorbing barriers for the stochastic dynamics of the learning algorithm.

Supervised learning

The setting is identical to that described before except that now the hearer must guess which object in the context the speaker named by h^* and then communicate its choice to the speaker (using some nonlinguistic means, such as pointing to the chosen object). In turn, the speaker must provide another nonlinguistic hint to indicate which object in the context it named by word h^*. Let us assume that the speaker associates word h^* to object n_1. If the hearer's guess happens to be the correct one then both entries $p^I_{n_1h^*}$ and $p^J_{n_1h^*}$ are incremented by the factor $1/M$. Furthermore, two words h_s and h_h are chosen randomly and the entries $p^I_{n_1h_s}$ and $p^J_{n_1h_h}$ are decreased by $1/M$ so the normalization of row n_1 is preserved in both verbalization matrices. Suppose now the hearer's guess is wrong, say, object n_2 instead of n_1. Then both entries $p^I_{n_1h^*}$ and $p^J_{n_2h^*}$ are decreased by the factor $1/M$ and, as before, two words h_s and h_h are chosen randomly and the entries $p^I_{n_1h_s}$ and $p^J_{n_2h_h}$ increased by $1/M$. As in the unsupervised case, the extremes $p^{I,J}_{nh} = 1$ and $p^{I,J}_{nh} = 0$ are absorbing barriers.

Our simulations of these learning algorithms show, not surprisingly, that after a transient the two agents become identical, in the sense that they are described by the same verbalization matrix. In addition, in the case of unsupervised learning the stochastic dynamics always leads to binary verbalization matrices, i.e., matrices whose entries p_{nh} can take on the values 1 or 0 only. Of course, once the dynamics produces a binary matrix it becomes frozen. This same outcome characterizes the supervised case as well, except in the cases that the lexicon size H is on the same order of the context size C. However, as we focus on the regime where C is finite and N and H are large we can guarantee that the stochastic dynamics leads to binary verbalization matrices regardless of the learning procedure.

Once the dynamics becomes frozen (and so the learning stage is over) we measure the average communication error ϵ as follows. The speaker chooses object n in the list of N objects and emits the corresponding word (there is a unique word assigned to any given object, i.e., there is a single entry 1 in any row of the verbalization matrix). The hearer must then infer which object is named by that word. Since the same word can name many objects (i.e., there may be many entries 1 in a given column), the probability ϕ_n that the hearer's guess is correct is simply the reciprocal of the number of objects named by that word. This probability is the communication accuracy regarding object n. The procedure is repeated for the N

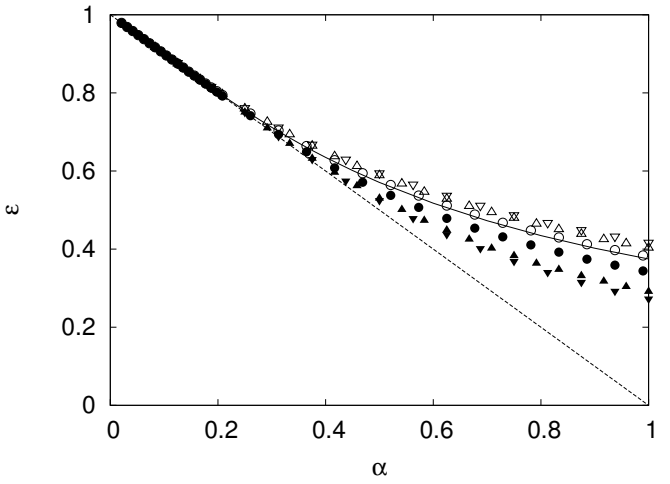

FIG. 1: Communication error ϵ as function of the ratio $\alpha = H/N$ between the lexicon size H and number of objects N for $N = 16(\bigtriangledown), 24(\bigtriangleup)$ and $96(\bigcirc)$. The open (filled) symbols represent the data for the unsupervised (supervised) algorithm. The error bars are smaller than the symbol sizes. The solid line is the result of the extrapolation for $N \to \infty$ (see Fig. 2) whereas the dashed line represents the optimal performance $1 - \alpha$. The parameters are $C = 2$ and $M = 10^4$.

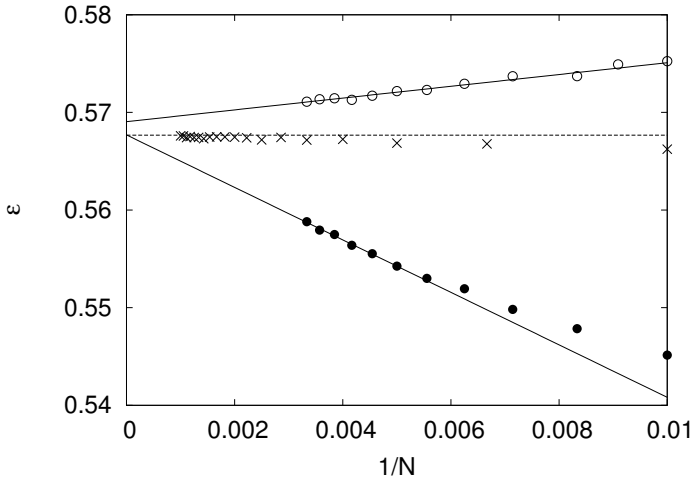

FIG. 2: Dependence of the communication error ϵ on the reciprocal of the number of objects $1/N$ for $\alpha = 0.5$ for the unsupervised ($\circ$) and supervised ($\bullet$) learning algorithms. The error bars are smaller than the symbol sizes. The linear fittings (solid straight lines) yield $\epsilon = 0.5690 \pm 0.0003$ (unsupervised) and $\epsilon = 0.5677 \pm 0.0004$ (supervised) for $N \to \infty$. The Monte Carlo estimate of the error for the random assignment of objects to words is given by the symbols $\times$ and the dashed horizontal line corresponds to the estimate of Eq. (3), $\epsilon_r = 0.5677$. The parameters are $C = 2$ and $M = 3\ 10^4$.

objects, so the average communication error is defined as $\epsilon = 1 - \phi$ where $\phi = \sum_n \phi_n/N$ is the average communication accuracy of the algorithm. We note that in the definition of these communication measures the context plays no role.

For $H \leq N$ the optimal (minimum) communication error ϵ_m is obtained by making a one-to-one assignment between $H-1$ words and $H-1$ objects, and then assigning the single remaining word to the remaining $N-H+1$ objects. This procedure yields $\epsilon_m = 1 - H/N = 1 - \alpha$. For $H > N$ we can obtain $\epsilon_m = 0$ simply by discarding $H - N$ words and making a one-to-one word-object assignment with the other N words. Figure 1 shows the comparison between this optimal result and the actual performances of the two learning algorithms as function of the ratio α. In this, as well as in the other figures of this paper, each symbol stands for the average over 10^4 independent samples or language games. The performance of the supervised algorithm deteriorates as the number of objects N increases, in contrast to that of the unsupervised algorithm which actually shows a slight improvement in this case. For $N \to \infty$ both algorithms produce the same communication error ϵ_r (see Fig. 2), which is shown by the solid line in Fig. 1. (A preliminary comparative analysis of these algorithms for $N = 8$ led to an incorrect claim about the general superiority of the supervised learning scheme [18].) For small α the performances of the two learning algorithms are practically indistinguishable from the optimal performance, but as we will show below the algorithms actually never achieve that performance, except for $\alpha = 0$.

A surprising finding was that for both supervised and unsupervised algorithms the average communication accuracy ϕ coincided with the ratio between the actual number of words used ($H_e \leq H$) and the number of objects N. This is what we expect when the objects are assigned randomly to the words, which is a classical occupancy problem discussed at length in Feller's book [19, Ch. IV.2]. In this occupancy problem, the probability that the number of words m not used in the assignment of the N objects to the H words (i.e., $m = H - H_e$) is

$$P_m(N,H) = \binom{H}{m} \sum_{\nu=0}^{H-m} \binom{H-m}{\nu} (-1)^\nu \left(1 - \frac{m+\nu}{H}\right)^N, \tag{1}$$

which in the limits $N \to \infty$ and $H \to \infty$ reduces to the Poisson distribution

$$p(m;\lambda) = \mathrm{e}^{-\lambda} \frac{\lambda^m}{m!} \tag{2}$$

where $\lambda = H \exp(-N/H)$ remains bounded [19, Ch. IV.2]. Hence the average communication accuracy resulting from the random assignment of objects to words is simply $(H - \langle m \rangle)/N$, which yields the communication error

$$\epsilon_r = 1 - \alpha + \alpha \mathrm{e}^{-1/\alpha}. \tag{3}$$

This equation describes perfectly the communication error of the two learning algorithms in the limit $N \to \infty$

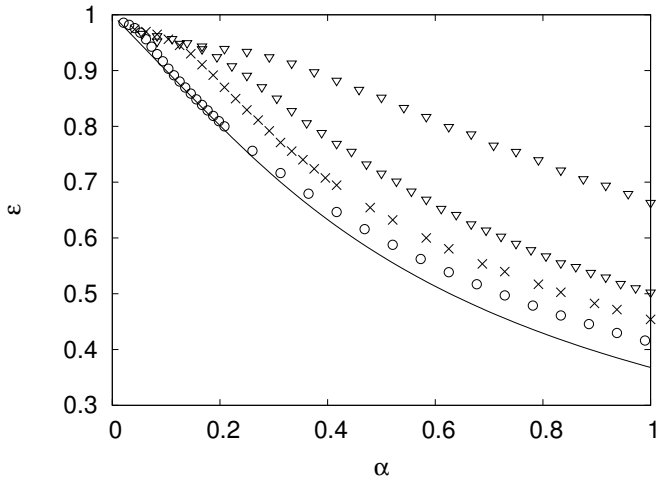

FIG. 3: Communication error ϵ of the unsupervised lexicon acquisition algorithm for context size $C = 4$ and $N = 24(\bigtriangledown), 36(\bigtriangleup), 48(\times)$, and $96(\bigcirc)$. The error bars are smaller than the symbol sizes. The learning rate is $1/M = 10^{-4}$ and the solid line is the result of Eq. (3).

(solid line in Fig. 1). We note that the (small) discrepancy observed in Fig. 2 for the extrapolated data of the unsupervised algorithm and the analytical prediction can be reduced to zero by decreasing the learning rate $1/M$. Equation (3) explains also why the performances of the algorithms are practically indistinguishable from the optimal performance for small α, since the difference between them vanishes as $\exp(-1/\alpha)$. In addition, Eq. (3) shows that in the limit of large α, the communication error vanishes as $1/\alpha$.

A word is in order about the effect of the context size C on the performance of the two learning algorithms, since Figs. 1 and 2 exhibit the results for $C = 2$ only. Simulations for larger values of C show that this parameter is completely irrelevant for the performance of the supervised algorithm. Of course, this is expected since regardless of the context size, at most two rows (object labels) of the verbalization matrices are updated. But the situation is far from obvious for the unsupervised algorithm since C determines the number of rows to be updated in each round of the game. However, the results summarized in Fig. 3 for $C = 4$ indicate that, despite strong finite-size effects particularly for small α, the communication error ultimately tends to ϵ_r in the limit of large N.

Therefore in the more realistic situation in which the number of objects N as well as the lexicon size H are very large, both supervised and unsupervised lexicon acquisition schemes yield the same communication accuracy, namely, the accuracy obtained by a random occupancy problem in which N objects are assigned randomly to H words. This surprising finding calls for a complete reappraisal of the current lexicon acquisition paradigms of Cognitive Science. It would be most interesting to devise sensible lexicon acquisition algorithms that reproduce the optimal communication performance or, at least, that exhibit an communication error that decays faster than the random occupancy result, $1/\alpha$.

The research at São Carlos was supported in part by CNPq and FAPESP, Project No. 04/06156-3. J.F.F. thanks the hospitality of the Adaptive Behaviour & Cognition Research Group, University of Plymouth, where this research was initiated. The visit was supported by euCognition.org travel grant NA-097-6.

* Electronic address: fontanari@ifsc.usp.br
† Electronic address: a.cangelosi@plymouth.ac.uk

[1] M.A. Nowak and D.C. Krakauer, Proc. Natl. Acad. Sci. USA **96**, 8028 (1999).
[2] L. Steels, in *Simulating the Evolution of Language*, edited by A. Cangelosi and D. Parisi (Springer-Verlag, London, 2002), pp. 211–226.
[3] V. Loreto and L. Steels, Nature Physics **3**, 758 (2007).
[4] A. Baronchelli, M. Felici, V. Loreto, E. Caglioli, and L. Steels, J. Stat. Mech. P06014 (2006).
[5] A. Baronchelli, L. Dall'Asta, A. Barrat, and V. Loreto, Phys. Rev. E **73**, 015102 (2006).
[6] T. Rosenthal and B. Zimmerman, *Social Learning and Cognition* (Academic Press, New York, 1978).
[7] J.M. Siskind, Cognition **61**, 39 (1996).
[8] L. Steels and J-C. Baillie, Rob. Aut. Syst. **43**, 163 (2003).
[9] L. Steels, Trends Cogn. Sci. **7**, 308 (2003).
[10] A.D.M. Smith, Lecture Notes in Artificial Intelligence **2801**, 499 (2003).
[11] A.D.M. Smith, Artificial Life **9**, 557 (2003).
[12] J. de Beule, B. de Vylder and T. Belpaeme, Proceedings of the Xth Conference on Artificial Life (MIT Press, Cambridge, 2006), pp. 466–472.
[13] L. Steels and F. Kaplan, Proceedings of IJCAI-99 (Morgan Kauffman, Los Angeles, 1998) pp. 862–867.
[14] T. Lenaerts, B. Jansen, K. Tuyls, and B. de Vylder, J. Theor. Biol. **235**, 566 (2005).
[15] H. Brighton, K. Smith and S. Kirby, Phys. Life Rev. **2**, 177 (2005).
[16] M. Oliphant and J. Batali, Center for Research on Language Newsletter, **11** (1) (1997).
[17] W.J. Ewens, *Mathematical Population Genetics* (Springer-Verlag New York, 2004).
[18] J.F. Fontanari and L.I. Perlovsky, Proceedings of the IEEE International Joint Conference on Neural Networks IJCNN06 (IEEE Press, Piscataway, 2006), pp. 2892–2897.
[19] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol I, 3rd Edition (Wiley, New York, 1968).